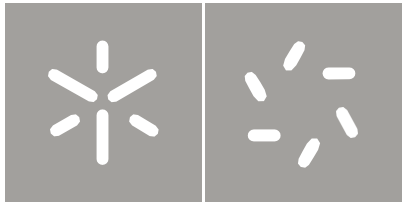
Universidade do Minho
Escola de Ciências

Carla Sofia Garcia Soares Ferreira

Estudo da Prevalência das doenças lisossomais em Portugal

outubro de 2018

Estudo da Prevalência das doenças lisossomais em Portugal
Carla Sofia Garcia Soares Ferreira
UMinho | 2018



Universidade do Minho
Escola de Ciências

Carla Sofia Garcia Soares Ferreira

Estudo da Prevalência das doenças lisossomais em Portugal

Tese de Mestrado
Mestrado em Estatística

Trabalho efetuado sob a orientação de
Professor Doutor Luis Filipe Meira Machado
Doutora Lúcia Maria Wanzeller Guedes de Lacerda

outubro de 2018

Agradecimentos

Nasci numa família que me ensinou que todas as experiências são uma aprendizagem. Começo por agradecer aos meus pais, Alice e Domingos, o amor, a determinação e a persistência que sempre me incutiram e transmitiram. São o meu reflexo de exemplo desde criança.

Gostaria também de agradecer a todos que, da sua forma, contribuíram neste percurso. Em especial:

Ao professor Luís Filipe Meira Machado e à Doutora Lúcia Lacerda pelo apoio, disponibilidade e sobretudo pela extrema paciência e compreensão sempre demonstrada ao longo destes meses. Deram-me sempre os melhores conselhos e foram uma constante nesta etapa de crescimento académico, profissional e pessoal. Desta forma deixo o meu mais sincero obrigada.

Ao meu irmão, Pedro e as minhas avós, em especial, a minha avó Joana. A todos os meus tios, em especial, Fátima e Paulo, e aos meus primos. Obrigada por acreditarem sempre em mim. Ao meu primo Hélio, obrigada por seres como és!

Aos meus amigos, a família que escolhi para partilhar todos os sucessos e insucessos, carinho e amizade. À Joana, Paulo, Sara, Ariana, Cátia, Mariana, Joana, Andreia, Sónia, Daniela, Mara e Sara. Não poderia ter melhores companheiros para partilharem todos os desafios da minha vida.

À Susana e à Rita que durante estes dois anos foram uma presença constante e um apoio essencial.

Por fim, a toda a equipa que integra o Centro de Genética Médica Jacinto Magalhães por me receberem de braços abertos.

Foram anos repletos de alegrias, tristezas e onde todos os desafios que me surgiram permitiram chegar ao término desta fase com um sentimento enorme de gratidão.

Resumo

As doenças lisossomais de sobrecarga (DLSs) são um grupo de doenças hereditárias do metabolismo que resultam de acumulação de substratos no lisossoma. Realizaram-se estudos de prevalência para os casos estudados, num dos laboratórios do Centro de Genética Médica Jacinto Magalhães, CGMJM, e a estimação de intervalos de referência para *probandos*.

Pretendeu-se estabelecer estimativas de prevalências de um grupo de patologias, para o intervalo de atividade do laboratório e para o intervalo relativo ao período de nascimento dos casos diagnosticados. Adicionalmente, estabeleceu-se intervalos de referência para um parâmetro analítico sujeito a critérios de seleção como, sexo e qualidade da amostra. Esta estimação possui por base a necessidade de atualização dos valores do laboratório, de forma a que o diagnóstico possa ir de encontro à realidade da sua população.

Foram estabelecidas comparações entre proporção binomial e taxa, entre vários métodos existentes de estimação intervalar como, Clopper-Pearson, Wald, Agresti-Coull, Wilson, mid-P, Jeffreys e para uma taxa, com base na distribuição Binomal e Poisson. Para estimação dos intervalos de referência, foram seguidas as recomendações IFCC (1984a,1984b) e CLSI (2000). Na identificação de *outliers* recorreu-se ao método de Tukey com análise gráfica do *boxplot* e ao método de Horn. No final, foi estimado pelo método robusto com 95% de confiança, o intervalo de referência para os *probandos* doentes, e pelo método não paramétrico com 95% de confiança, o intervalo de referência para os *probandos* normais.

O método de estimação intervalar determinado como adequado para este estudo é o método de Wilson. A decisão teve por base investigações, tais como Agresti & Coull (1998), Brown *et al.* (2001), Brown *et al.* (2002), Cepeda-Cuervo *et al.* (2008) entre outras, que comparavam todos os métodos referidos anteriormente, de forma exaustiva. Na estimação dos intervalos de referência para um parâmetro analítico, os *outliers* foram identificados com base o método de Tukey, e pode-se afirmar que ambos os intervalos estimados são representativos dos indivíduos do laboratório.

Palavras-chave: doenças lisossomais de sobrecarga, prevalência, intervalos de referência, estimação intervalar, proporção, taxa, Binomial, Poisson, CLSI, IFCC, Tukey, Horn, método não paramétrico, método robusto.

Abstract

Lysosomal storage disorders (LSDs) are a group of inherited metabolic diseases that result from accumulation of substrates in the lysosome. Prevalence studies were performed in one of the laboratories of the Centro de Genética Médica Jacinto Magalhães, CGMJM, and the estimation of reference intervals for probands.

It was intended to establish prevalences estimates of a group of pathologies, for the laboratory activity interval and for the interval relative to the period of birth of the diagnosed cases. In addition, reference intervals were established for an analytical parameter subject to selection criteria such as sex and sample quality. This estimation is based on the need to update the laboratory values, so that the diagnosis can meet the reality of its population.

Comparisons between binomial proportion and rate were established between several existing interval estimation methods, such as Clopper-Pearson, Wald, Agresti-Coull, Wilson, mid-P, Jeffreys, and for a rate, based on Binomal and Poisson distributions. For the estimation of reference intervals were followed the IFCC (1984a, 1984b) and CLSI (2000) recommendations. For outlier's identification were used the Tukey method with graphical analysis of the boxplot and the Horn method. In the end, it was estimated by the robust method with 95% confidence, the reference interval for probands with disease, and by the non-parametric method with 95% confidence, the reference interval for probands without disease.

The method for interval estimation determined as appropriate for this study is the Wilson method. The decision was based on investigations such as Agresti & Coull (1998), Brown *et al.* (2001), Brown *et al.* (2002), Cepeda-Cuervo *et al.* (2008), among others, that compared all the above methods exhaustively. In reference intervals estimation for an analytical parameter, outliers were identified based on the Tukey method, and both estimated intervals are representative of the laboratory individuals.

Key-words: lysosomal storage disorders, prevalence, reference intervals, interval estimation, proportion, rate, Binomial, Poisson, CLSI, IFCC, Tukey, Horn, non-parametric method, robust method.

Índice de Conteúdos

Índice de Figuras

Índice de Tabelas

Glossário

Probando – Em Genética Médica, trata-se de um individuo na qual se deteta uma doença ou caraterística genética inicialmente, servindo como ponto de partida para um estudo familiar.

Lista de Abreviaturas

CGMJM – Centro de Genética Médica Jacinto Magalhães

UM – Universidade do Minho

DLSs – Doenças lisossomais de sobrecarga

SNSs – Serviços Nacionais de Saúde

$\boldsymbol{IC}$ – Intervalo de Confiança

DBS – *Dry Blood Spotting*

VP – Valores do parâmetro

1. Introdução

1.1. Entidade acolhedora

O Centro de Genética Médica Doutor Jacinto Magalhães (CGMJM) foi criado em 1980. Encontra-se integrado no Centro Hospitalar e Universitário do Porto desde 2013. É composto por uma estrutura de trabalho multidisciplinar, clínica e laboratorial, tendo como principal objetivo a prestação de cuidados de saúde, a investigação e formação na área da genética médica.

Integrada nas instalações do CGMJM está a Central de aquisição, armazenamento e distribuição dos produtos dietéticos hipoproteicos, sendo a consulta de nutrição, desde 1990, responsável pela sua gestão a nível nacional. Estes produtos dietéticos são comparticipados pelo Ministério da Saúde, para todas as Doenças Hereditárias do Metabolismo que deles necessitem, conforme o Despacho nº14 319, 2ª série de 29 de junho de 2005, publicado no Diário da República.

A área clínica – Unidade de Genética Médica – é responsável pelas consultas de Genética, Pré-Natal, Nutrição e Psicologia. A área laboratorial encontra-se repartida nas seguintes unidades: Unidade de Bioquímica Genética, Citogenética e Genética Molecular, Centro Hospitalar do Porto (2016).

O CGMJM é um centro de referência para diagnóstico de doenças genéticas, dedicado particularmente a patologias raras. É o único no país a efetuar o diagnóstico de algumas doenças, ao nível do gene, cromossoma e via metabólica. Possui uma oferta analítica a mais de 400 patologias, desde doenças hereditárias do metabolismo, neuromusculares, atraso mental, cromossomopatias, entre outras. A entidade acolhedora preza por adquirir um acompanhamento continuo relativamente as patologias sobre as quais providencia diagnóstico e, adquirir métodos e materiais que permitam uma evolução continua como entidade certificada para esse fim.

1.2. Objetivos

O presente trabalho foi elaborado com o objetivo de providenciar uma consulta detalhada de ferramentas estatísticas, relativamente a dois temas muito presentes nas áreas da Medicina e Bioquímica Genética. O primeiro tema toma como foco o estudo de prevalências de doenças lisossomais na população Portuguesa, enquanto que, no segundo tema, o desafio foi estabelecer novos intervalos de referência para um parâmetro bioquímico no diagnóstico de uma patologia.

As doenças lisossomais de sobrecarga, designadas de DLSs, são um grupo de doenças metabólicas hereditárias caraterizadas pela acumulação de substrato dentro do lisossoma, devido a mutações na codificação dos genes responsáveis pela produção da enzima necessária para que ocorra o processo de digestão celular. A maioria das doenças lisossomais de sobrecarga tem uma transmissão autossómica recessiva, excetuando as que se encontram ligadas ao cromossoma sexual X. A Unidade de Bioquímica Genética do Centro de Genética Médica Doutor Jacinto Magalhães (CGMJM), é o laboratório que em Portugal realiza o diagnóstico pré e pós-natal das DLSs há 30 anos.

Ainda que consideradas doenças raras, os estudos de prevalência são uma poderosa ferramenta para validar a importância do diagnóstico e abordagens terapêuticas. O cálculo da prevalência é ainda determinante para permitir estimar o impacto de uma doença na população e nos custos que o seu tratamento tem para os Serviços Nacionais de Saúde, SNSs.

Neste estudo utilizaram-se a distribuição Binomial e a distribuição de Poisson para obtenção destes valores, e a partir dos quais, foram aplicados métodos para a estimação de intervalos de confiança.

O primeiro tema foi, inicialmente, o grande foco desta colaboração entre a UM e o CGMJM. A evolução do trabalho permitiu a conclusão do primeiro tema e a realização de um segundo tema, relativo ao estabelecimento de valores de referência para um parâmetro analítico.

Estabelecer intervalos de referência num laboratório é muito importante, mas extremamente desafiante. Para a construção destes intervalos de referência, é primordial a avaliação da qualidade da colheita, através da análise da enzima de referência, para que os dados sejam o mais realistas possíveis, e consequentemente os resultados derivados destes sejam fidedignos.

O segundo tema toma uma abordagem mais prática, indicando qual o processo a realizar de modo a estabelecer os valores do intervalo pretendido. É realizada a análise exploratória clássica aos dados, descritos diferentes métodos de identificação de *outliers,* verificada a normalidade dos dados e, por fim, a estimação do intervalo de referência.

Este trabalho foi realizado com dados reais anonimizados. Fica, por razões de confidencialidade acordadas por ambas as partes desta cooperação, omitida a qualificação dos grupos de doenças lisossomais e, inclusive, a respetiva designação das patologias, sendo desta forma, atribuída a denominação de "Grupo" e "Patologia". Relativamente ao segundo tema, a designação atribuída ao parâmetro para a qual está a ser estabelecido um novo intervalo de referência é "parâmetro".

Como aluna do Mestrado em Estatística da Escola de Ciências da Universidade do Minho, e sendo Licenciada em Matemática com formação complementar em Biologia pela Faculdade de Ciência da Universidade do Porto, após o contacto com colegas da área da Biologia que se ressentiam no que corresponde ao domínio de ferramentas estatísticas, o meu objetivo pessoal com este trabalho, para obtenção do grau de Mestre, era elaborar um guia cuidado e de forma a orientar esses mesmos colegas quando deparados com desafios como os que me foram colocados.

1.3. Projeto

O presente estágio tem como local de ação o Centro de Genética Médica Doutor Jacinto Magalhães, CGMJM, particularmente na Unidade de Bioquímica Genética, coordenada pela Doutora Lúcia Lacerda, após proposta de vários temas possíveis de estudo. Como representante do Mestrado em Estatística da Escola de Ciências da Universidade do Minho na entidade acolhedora, a grande meta estabelecida por mim era conciliar as duas grandes áreas que me preenchem a nível pessoal e obter um forte enriquecimento académico, e poder providenciar à instituição os melhores resultados possíveis enquanto estagiária.

O estágio teve como data de partida janeiro de 2018, onde numa abordagem inicial se pretendia assentar a viabilidade da interação das áreas, Estatística e Bioquímica Genética. Era notória a intenção dos profissionais da área nesta colaboração tal como, a disponibilidade imediata da Doutora Lúcia Lacerda, responsável da Unidade de Bioquímica Genética a dar início a este projeto, pois iria potencializar a publicação de artigos, apresentação de poster para exposição no 14º Simpósio Internacional da Sociedade Portuguesa de Doenças Metabólicas (SPDM) e, a atualização do valor de referência para um parâmetro enzimático e estimativa do respetivo intervalo de referência para *probandos* normais e doentes. Garantidas todas as condições e estabelecidos todos os contactos entre a Escola de Ciências da Universidade do Minho e o Departamento de Ensino, Formação e Investigação do Centro Hospitalar do Porto, DEFI, o estágio teve como data de início 22 de janeiro de 2018 e data de finalização a 31 de outubro de 2018.

O início do trabalho teve lugar após apresentação da instituição e da Unidade que acolheu este estágio, Unidade de Bioquímica Genética do Centro de Genética Médica Doutor Jacinto Magalhães, CGMJM.

1.4. Estrutura

O presente trabalho é composto por dois temas: Prevalência e Intervalos de Referência. No primeiro tema pretende-se, através de estimativas de prevalências, entender de que forma estas patologias raras possuem impacto na população. Realizadas essas mesmas estimativas, o desafio deteve-se na determinação do método de estimação intervalar mais adequado a este estudo. No segundo tema, o foco era estimar um novo intervalo de referência para *probandos* do laboratório. O objetivo era que estes valores fossem atualizados de modo a que sejam representativos dos indivíduos do laboratório.

No Capítulo 2 foram descritas as metodologias que permitem a obtenção dos resultados finais. É descrita a distribuição Binomial e métodos de estimação intervalar associados, assim como, a distribuição de Poisson e respetivo intervalo de confiança para uma taxa. São também descritas as metodologias que estiveram na base da obtenção do intervalo de referência, desde a seleção de indivíduos de referência e dimensão da amostra, análise de *outliers*, normalidade dos dados e métodos de estimação do intervalo.

No Capítulo 3 foram introduzidos os diferentes grupos de doenças lisossomais, as variáveis da base de dados e respetiva análise exploratória. É descrito o cálculo para a estimação da prevalência e resultados associados. Nos resultados, são efetuadas comparações e determinado o método de estimação intervalar que melhor se aplica ao estudo.

No Capítulo 4 foram obtidos intervalos de referência para um parâmetro analítico do laboratório. É descrito, de forma anónima, o parâmetro analítico do laboratório sobre o qual foi realizado o estudo, as variáveis que compõe a base de dados e respetiva análise exploratória da base de dados dos *probandos* doentes e *probandos* normais. Nos resultados, encontram-se os intervalos estimados e, relativamente aos resultados obtidos para os *probandos* normais, uma comparação ao intervalo atualmente praticado pela entidade acolhedora.

Por fim, no Capítulo 5 foram expostas as conclusões relativamente aos resultados dos estudos, limitações existentes e possíveis abordagens futuras a realizar.

1.5. *Software* utilizado

Na realização do presente trabalho foram utilizados o *software* R versão x64 3.4.2. e Excel.

O *software* R é um conjunto integrado de recursos de *software* para armazenamento e manipulação de dados, cálculo, modelação e exibição gráfica, R Project (2018). A linguagem do *software* R foi utilizada para computação e estabelecimento de cálculos mais complexos. Os gráficos apresentados neste trabalho foram construídos no *software* R e Excel.

Os pacotes utilizados neste trabalho para obtenção da estimativa de prevalência e respetivos intervalo de confiança são: ***binom***, ***DescTools***, ***epiR***, ***propsCIs***, ***stats, boot*** e ***bootstrap***. No Apêndice B encontram-se os métodos de estimação intervalar e respetivos comandos a utilizar.

Os pacotes utilizados neste trabalho para estimação dos intervalos de referência são: ***nortest*** e ***reference intervals***. No Apêndice B encontram-se os comandos a utilizar.

2. Metodologia

Neste capítulo são abordados os conceitos estatísticos, com base na distribuição Binomial e Poisson, para estimação intervalar com suporte em trabalhos que comparam as performances de diferentes métodos. Pretende-se concluir qual o método de estimação intervalar, com base nessas publicações, que para um estudo de prevalência melhor se aplica. São também descritas as metodologias referentes à estimação de um intervalo de referência para um parâmetro analítico do laboratório.

2.1. Estimação

2.1.1. Estimação pontual e intervalar

A estimação de parâmetros é o processo que consiste no uso de dados de uma amostra para estimar valores de parâmetros populacionais desconhecidos. A estimação de parâmetros pode ser realizada de duas formas: estimação pontual e/ou estimação intervalar. Na estimação pontual, considere-se uma dada amostra $(X_1, X_2, \dots, X_n)$ de uma população com função de densidade de probabilidade dada por $f(x|\theta)$, com $\theta \in \Theta$. As estatísticas utilizadas com o propósito de estimar θ designam-se por $T(X_1, X_2, \dots, X_n)$. Desta forma, é necessário distinguir estimador de estimativa, sendo que um estimador é uma função da amostra que representamos por $T(X_1, X_2, \dots, X_n)$ e, estimativa um valor concreto assumido pelo estimador para uma amostra em particular e que representamos por $T(x_1, x_2, \dots, x_n)$, Murteira *et al.* (2015, p. 401).

Definição 1: Um estimador pontual é uma estatística $\widehat{\Theta}$ cujos valores particulares $\hat{\theta}$ constituem estimativas do parâmetro populacional θ.

Avaliar a precisão das estimativas pontuais é uma dificuldade que se tenta contornar, quando complementado com a estimativa intervalar associada. Um intervalo de confiança é um intervalo real que contém o parâmetro em estudo com uma determinada confiança.

Definição 2 (Murteira *et al.* (2015, p. 428-429)): Seja $(X_1, X_2, \dots, X_n)$ uma amostra aleatória da população com função densidade de probabilidade dada por $f(x|\theta)$, com $\theta \in \Theta$. Considere-se duas estatísticas $\mathrm{T}_1 = \mathrm{T}_1(X_1, X_2, \dots, X_n)$ e $\mathrm{T}_2 = \mathrm{T}_2(X_1, X_2, \dots, X_n)$, tais que $P(\mathrm{T}_1 < \theta < \mathrm{T}_2) = 1 - \alpha$, $\forall\theta \in \Theta$ $(0 < \alpha < 1)$ onde α não depende de θ. Um intervalo aleatório para θ é um intervalo $(\mathrm{T}_1, \mathrm{T}_2)$, nas condições anteriores.

Quando se dispõe de uma amostra particular, sejam $t_1 = \mathrm{T}_1(x_1, x_2, \dots, x_n)$ e $t_2 = \mathrm{T}_2(x_1, x_2, \dots, x_n)$ os valores assumidos pelas estatísticas T_1 e T_2, respetivamente. A qualquer intervalo $(\mathrm{t}_1, \mathrm{t}_2)$, que seja uma concretização do intervalo aleatório $(\mathrm{T}_1, \mathrm{T}_2)$, chama-se intervalo de confiança, IC, a $(1 - \alpha) \times 100\%$ para θ.

O valor $1 - \alpha$ designa o grau de confiança que se tem em que uma amostra particular dê origem a um desses intervalos $(\mathrm{t}_1, \mathrm{t}_2)$. Por este motivo, conclui-se que não é correto dizer que $1 - \alpha$ é a probabilidade de θ pertencer ao intervalo $(\mathrm{t}_1, \mathrm{t}_2)$, uma vez que as respetivas extremidades não são aleatórias. O valor de α representa o nível de significância.

Aplicada ao caso discreto, a definição anterior sofre uma alteração e, a equação $P(\mathrm{T}_1 < \theta < \mathrm{T}_2) = 1 - \alpha$ é substituída por, $P(\mathrm{T}_1 < \theta < \mathrm{T}_2) \geq 1 - \alpha$, $\forall\theta \in \Theta$ $(0 < \alpha < 1)$, sendo o grau de confiança dado pelo mínimo das probabilidades $P(\mathrm{T}_1 < \theta < \mathrm{T}_2)$, a verificar a desigualdade anterior, quando θ percorre o espaço-parâmetro Θ.

2.2. Distribuição Binomial

A distribuição de Bernoulli aparece associada à experiência aleatória, designada como prova de Bernoulli, em que se observa a realização ou não realização de um determinado evento A com probabilidade $P(A) = p$. A realização de A diz-se um "sucesso", e a sua não realização diz-se "insucesso", com probabilidade $P(\bar{A}) = 1 - p$. A variável X que segue a distribuição de Bernoulli, possui como parâmetro p, e é representada pela seguinte notação:

$$X \sim Ber(p).$$

A função de probabilidade é dada por,

$$f_X(x|p) = P[X = x] = p^x(1-p)^{1-x} \ ,$$

onde:

- $p \in \mathbb{R}, \ 0 < p < 1$;
- $x = 0,1$.

Considere-se uma sucessão de n provas de Bernoulli independentes, isto é, uma sucessão de experiências aleatórias independentes em que em cada uma delas se observa a realização ou não realização do evento A, com probabilidade $P(A) = p$ constante de experiência para experiência, Murteira *et al.* (2015, p. 243-247). Associado a esta sucessão de n provas de Bernoulli independentes, enuncia-se o seguinte Corolário retirado de Murteira (1990):

Corolário: Se as variáveis $X_i, i = 1,2, \dots, n$ são independentes e identicamente distribuídas (e possuem distribuição de Bernoulli),

$$X_i \sim Ber(\ p) \Longrightarrow \sum_{i=1}^{n} X_i \sim Bin(n, p).$$

A distribuição de Bernoulli é um caso particular da distribuição Binomial de parâmetros 1 e p, podendo ser representada como $Bin(1, p)$, isto é, $n = 1$.

Definição 3: Seja X uma variável aleatória discreta. A variável X, representativa do número de vezes que o evento A ocorreu nas n repetições da experiência E é denominada de variável aleatória que segue uma distribuição Binomial. A experiência E apresenta dois resultados possíveis, sucesso ou insucesso.

Esta distribuição tem os seguintes pressupostos:

I. n provas independentes e sempre sobre as mesmas circunstâncias;
II. a probabilidade de sucesso é p, e a probabilidade de insucesso é $q = 1 - p$.

Note-se que no total de n provas, e assumindo x sucessos, existem então, $n - x$ insucessos no espaço amostral.

A variável X que segue a distribuição Binomial, possui os parâmetros n e p, e é representada pela seguinte notação:

$$X \sim Bin(n, p).$$

A probabilidade de sucesso em cada prova, isto é, a função de probabilidade é dada por,

$$f_X(x|p) = P[X = x] = \binom{n}{x} p^x (1-p)^{n-x} \quad ,$$

onde:

- $n \in \mathbb{N}, \ n > 0$;
- $p \in \mathbb{R}, \ 0 < p < 1$;
- $x = 0,1,2, \dots, n$.

A função geradora de momentos é dada por,

$$M(s) = E[e^{sX}] = \{(1-p) + pe^s\}^n.$$

Calculando as sucessivas derivadas da função geradora de momentos no ponto $s = 0$, obtêm-se os momentos da variável X. O valor esperado da variável X, designado de momento de ordem 1, é $E[X] = np$ e a variância, designada de momento de ordem 2, é dada por $Var[X] = np(1-p)$.

2.2.1. Proporção binomial

A proporção é uma fração que permite avaliar o número de indivíduos que numa população, P, estão abrangidos pela caraterística pretendida. É considerada uma estimativa pontual. Define-se proporção da seguinte forma:

$$P = \frac{X}{N},$$

onde:

- X, corresponde ao número de indivíduos com a caraterística pretendida na população;
- N, corresponde ao número de indivíduos da população.

2.2.2. Intervalos de Confiança

A obtenção dos intervalos de confiança para a proporção binomial foi efetuada recorrendo a diferentes metodologias e, calculados com base numa hipótese nula bilateral. Nos subcapítulos seguintes encontram-se os métodos utilizados para o cálculo dos intervalos de confiança.

2.2.2.1. Intervalo de confiança de Clopper-Pearson

O intervalo de confiança de Clopper-Pearson, ou, intervalo de confiança exato, Pearson e Clopper (1934), é baseado na inversão[1] da hipótese nula, $H_0: p = p_0$ do teste de hipóteses bilateral binomial contra a hipótese alternativa $H_1: p \neq p_0$, Agresti e Coull (1998).

Para valores $0 < x < n$, o limite inferior e superior do IC é dado por, Agresti e Coull (1998),

$$IC = \left(\sum_{k=x}^{n} \binom{n}{k} {p_0}^k (1-p_0)^{n-k} = \alpha/2 \,; \sum_{k=0}^{x} \binom{n}{k} {p_0}^k (1-p_0)^{n-k} = \alpha/2 \right),$$

onde:

- α representa o nível de significância.

Agresti e Coull (1998) indicam ainda que os limites do intervalo podem ser também estimados recorrendo à distribuição Beta, onde o IC é dado por,

$$IC = \left(B_{\alpha/2,x,n-x+1}, B_{1-\alpha/2,x+1,n-x} \right),$$

onde:

- o limite inferior é o quantil $\frac{\alpha}{2}$ da distribuição Beta(x,n-x+1);
- o limite superior é o quantil $1 - \frac{\alpha}{2}$ da distribuição Beta(x+1,n-x).

[1] "Intervalos de inversão de teste funcionam sob a definição de que um intervalo de confiança, sobre uma estatística observada, inclui uma variação de parâmetros que, quando testados, não rejeitariam essa estatística observada." Influential Points (2018)

2.2.2.2. Intervalo de confiança de Wald

O intervalo de confiança de Wald, Wald e Wolfowitz (1939), designado também como uma *aproximação teórica Normal* de um IC para uma proporção resulta da inversão do teste de Wald para uma proporção p e é definido por, Agresti e Coull (1998),

$$IC = \left(\hat{p} - z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right),$$

onde:

- $z_{1-\alpha/2}$ é o quantil $1 - \frac{\alpha}{2}$ da distribuição Normal (0,1).

É conhecido o facto deste intervalo possuir uma baixa performance, Brown *et al.* (2001), à exceção de quando n é elevado, Agresti e Coull (1998). Newcombe e Merino (2006) afirmam que quando $\hat{p}$ é próximo de 0 ou 1, este pode originar resultados questionáveis.

2.2.2.3. Intervalo de confiança de Agresti-Coull

O intervalo de confiança de Agresti-Coull, Agresti e Coull (1998), obtém-se a partir do estimador de Wald com uma ténue modificação, podendo ter a designação de intervalo de confiança de Wald ajustado. Esta modificação foi feita adicionando ao estimador dois sucessos e dois insucessos.

O estimador pontual é,

$$\hat{p}_w = \frac{x+2}{n+4}.$$

A variância para a proporção estimada é dada por,

$$\hat{s}_w^2 = \frac{\hat{p}_w(1-\hat{p}_w)}{n+4}.$$

Os limites do IC são definidos por, Agresti e Coull (1998),

$$IC = \left(\hat{p}_w - z_{1-\alpha/2} \sqrt{\frac{\hat{p}_w}{n+4}(1-\hat{p}_w)}, \hat{p}_w + z_{1-\alpha/2} \sqrt{\frac{\hat{p}_w}{n+4}(1-\hat{p}_w)} \right),$$

onde:

- $z_{1-\alpha/2}$ é o quantil $1-\frac{\alpha}{2}$ da distribuição Normal (0,1).

2.2.2.4. Intervalo de confiança de Wilson

O intervalo de confiança de Wilson, Wilson (1927), designado também como, intervalo de confiança *Score*, é apelidado desta forma pois deriva da inversão do que é denominado de teste *score* para p.

O limite inferior e superior para o IC são definidos por, Agresti e Coull (1998),

$$IC = \left(\frac{\hat{p} + \frac{z^2_{1-\alpha/2}}{2n} - z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2_{1-\alpha/2}}{4n^2}}}{1 + \frac{z^2_{1-\alpha/2}}{n}}, \frac{\hat{p} + \frac{z^2_{1-\alpha/2}}{2n} + z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2_{1-\alpha/2}}{4n^2}}}{1 + \frac{z^2_{1-\alpha/2}}{n}} \right),$$

onde:

- $z_{1-\alpha/2}$ é o quantil $1-\frac{\alpha}{2}$ da distribuição Normal (0,1).

Como Agresti e Coull (1998) referem, o intervalo de Clopper-Pearson, ou intervalo exato, mantem-se ainda bastante conservador mesmo que a amostra seja consideravelmente grande, quando $\hat{p}$ tende a estar próximo de 0 e 1. Brown *et al.* (2002) reforça que o intervalo de Clopper-Pearson é conservador e sugere que um método que tem uma melhor performance, associado a este, resulta na utilização do intervalo de confiança mid-P baseado no teste Binomial exato.

2.2.2.5. Intervalo de confiança mid-P exato

Ao intervalo de confiança mid-P exato são atribuídas boas propriedades de localização e cobertura de 95% em torno do valor médio. Este intervalo é obtido por inversão de teste, isto é, pela inversão da adaptação do teste exato que utiliza o valor mid-P, Agresti e Coull (1998).

Vollset (1993), Agresti e Coull (1998), Newcombe (1998) e Brown *et al.* (2001) sugerem que para fins práticos, este método adequa-se muito bem quando se pretende obter a estimação de intervalos de confiança para uma proporção.

O respetivo intervalo de confiança mid-P exato é o intervalo mid-P adaptado do intervalo de Clopper-Pearson.

Os limites do IC são definidos por, Rothman e Boice (1979),

$$IC = \begin{pmatrix} \frac{\alpha}{2} = \frac{1}{2}\binom{n}{x} P_{LB}^{x}(1-P_{LB})^{n-x} + \sum_{k=x+1}^{n} \binom{n}{k} P_{LB}^{k}(1-P_{LB})^{n-k}, \\ \frac{\alpha}{2} = \frac{1}{2}\binom{n}{x} P_{UB}^{x}(1-P_{UB})^{n-x} + \sum_{k=0}^{x-1} \binom{n}{k} P_{UB}^{k}(1-P_{UB})^{n-k} \end{pmatrix},$$

onde:

- x, corresponde ao número de indivíduos com a caraterística pretendida na amostra;
- n, corresponde ao número de indivíduos da amostra;
- α, nível de significância;
- P_{LB}: proporção do limite inferior;
- P_{UB}: proporção do limite superior.

2.2.2.6. Intervalo de credibilidade de Jeffreys

A abordagem bayesiana, para estimação de um intervalo, define-se pela avaliação da distribuição à posteriori do parâmetro.

Definição 4 (Cepeda-Cuervo *et al.* (2008)): Se $\theta \in \Theta$ é uma quantidade desconhecida, $C \subset \Theta$ é uma região de $100(1-\alpha)\%$ de credibilidade para θ se $P(\theta \in C|x) \geq 1 - \alpha$.

Para este caso, $1 - \alpha$ é denominado de nível de credibilidade. Se θ é um escalar, a região C é dada, usualmente, pelo intervalo $[c_1, c_2]$.

O intervalo de Jeffreys, como referido por vários autores, é um intervalo que considera uma distribuição à *priori* não informativa. É calculado partindo do pressuposto que p segue a distribuição Beta$(1/2\,, 1/2)$ à *priori*; portanto, p segue a distribuição Beta$(X + 1/2\,, n - X + 1/2)$ à *posteriori*. É referenciado por possuir boas propriedades frequentistas, referido em Brown *et al.* (2001) e Brown *et al.* (2002), com base em Wasserman (1991).

O limite inferior e superior para o IC é dado por, Brown *et al.* (2001),

$$IC = (B_{\alpha/2,x+1/2,n-x+1/2},\, B_{1-\alpha/2,x+1/2,n-x+1/2}),$$

onde:

- o limite inferior é o quantil $\frac{\alpha}{2}$ da distribuição Beta(x+$\frac{1}{2}$,n-x+$\frac{1}{2}$);
- o limite superior é o quantil $1 - \frac{\alpha}{2}$ da distribuição Beta(x+$\frac{1}{2}$,n-x+$\frac{1}{2}$).

Tratando-se de patologias raras, e as proporções resultarem em valores próximos de 0, entre as metodologias existentes, é necessário existir rigor na escolha dos métodos entre os quais se pretende estabelecer uma comparação. Analisando os trabalhos de He e Wu (2009) e referências anteriormente citadas, estes são os métodos em que existe concordância para sejam estabelecidas comparações, entre todos os métodos de estimação intervalar para proporções binomiais.

2.3. Distribuição Poisson

Uma experiência que possa ser denominada como rara apresenta uma forte relação com a distribuição de Poisson. Esta distribuição é utilizada quando se está perante o acontecimento de eventos raros num grande número de indivíduos e $p \approx 0$ e, somente nesses casos, esta pode ser equiparável à distribuição Binomial, UMass (2007). Nestas condições, o parâmetro da distribuição a utilizar nesta aproximação é dado por $\lambda = np$.

Definição 5: Seja X uma variável aleatória discreta. A variável X representa o número de eventos A que ocorreram num determinado período, sendo esta denominada como uma variável aleatória de Poisson. Uma experiência de Poisson segue as seguintes propriedades, Interative Mathematics (2018):

I. possui dois resultados possíveis, sucesso ou insucesso;
II. as condições da experiência permanecem constantes no decorrer do tempo;
III. o número de sucessos registados nos intervalos da partição são independentes entre si;
IV. a probabilidade de se registar um sucesso num intervalo, é praticamente proporcional à dimensão do intervalo.

A variável X que segue a distribuição de Poisson, com parâmetro λ, é representada pela seguinte notação:

$$X \sim Poisson(\lambda).$$

A probabilidade de ocorrência de um evento corresponde à função de probabilidade e é dada por,

$$f_X(x|\lambda) = P[X = x] = \frac{e^{-\lambda}\lambda^x}{x!},$$

onde:

- $x \in N_0$,
- $\lambda > 0$.

A função geradora de momentos é dada por,

$$M(s) = E[e^{sX}] = e^{\lambda[e^s - 1]}.$$

Calculando as sucessivas derivadas da função geradora de momentos no ponto $s = 0$, obtêm-se os momentos da variável X. O valor esperado da variável X, designado de momento de ordem 1, é $E[X] = \lambda$ e a variância, designada de momento de ordem 2, é dada por $Var[X] = \lambda$. Na distribuição de Poisson os primeiros dois momentos são idênticos, isto é, $E[X] = Var[X]$.

2.3.1. Taxa

A razão representa a divisão de uma quantidade pela outra, em que não existe necessariamente relação entre o numerador e denominador. A taxa é uma particularização de razão, onde temos obrigatoriamente implícita no denominador, uma medida de tempo. A taxa de ocorrência é definida pela seguinte fórmula:

$$\lambda = \frac{X}{N},$$

onde:

- X, corresponde ao número de indivíduos com a caraterística pretendida na população;
- N, corresponde ao número de indivíduos da população.

2.3.2. Intervalo de confiança

Sendo X uma variável que segue uma distribuição de Poisson e λ, a taxa de ocorrência, pretende-se obter o intervalo de confiança do número de ocorrências do evento A.

O limite inferior e superior do IC é dado por, Ulm (1990),

$$IC = \left(\frac{\chi^2_{\frac{\alpha}{2},2n}}{2T}, \frac{\chi^2_{1-\frac{\alpha}{2},2(n+1)}}{2T} \right),$$

onde:

- $\chi^2_{v1,v2}$ representa o quantil v1 da distribuição Qui-Quadrado, com v2 graus de liberdade;
- n, número de eventos verificados;
- T, número de indivíduos em risco naquele período de tempo.

Se n atingir valores muito elevados, pelo Teorema do Limite Central,

$$IC = \left(\hat{\lambda} - z_{1-\alpha/2} \times \sqrt{\frac{\hat{\lambda}}{n}}, \hat{\lambda} + z_{1-\alpha/2} \times \sqrt{\frac{\hat{\lambda}}{n}} \right),$$

onde:

- $z_{1-\alpha/2}$ é o respetivo quantil da distribuição Normal (0,1).

Existem ainda, outros métodos de estimação intervalar. Neste trabalho apenas foram considerados para a obtenção de resultados e respetiva comparação os métodos enunciados nos subcapítulos anteriores, mas a título elucidativo, é referido um outro método que surge como alternativa a esses, o método *Bootstrap*. Introduzido por Efron e Tibshirani (1983), é uma técnica estatística e computacional de reamostragem. O princípio deste método é reamostrar a partir de um conjunto de dados, diretamente ou através de um modelo ajustado a estes, de forma a gerar réplicas sem recorrer a cálculos analíticos. A forma de obtenção das réplicas *bootstrap* é a mesma para o caso paramétrico e não paramétrico. Deste modo, pode-se classificar o método *bootstrap* em paramétrico e não paramétrico.

O método possui inúmeras vantagens como soluções de problemas complexos, e é prático pois não necessita de muitas suposições para estimar parâmetros de interesse das distribuições. É uma técnica robusta na construção de um intervalo de confiança, possuindo diferentes métodos na obtenção dos intervalos de confiança *bootstrap*, sendo eles intervalo *bootstrap* padrão, intervalo *bootstrap*-t, intervalo *bootstrap* percentil, entre outros (Alves E.J., 2013).

Neste trabalho não serão apresentados os resultados obtidos por este método, mas no Apêndice B, encontra-se os comandos do *software* R que permitem a estimação do intervalo de confiança pelo método.

2.4. Intervalo de referência

A linha cronológica da metodologia utilizada para estimar intervalos de referência é traçada de forma idêntica de autor para autor. A metodologia descrita neste trabalho tem como base as recomendações da *Internacional Federation of Clinical Chemistry* (IFCC), publicadas em 1984, e *Clinical and Laboratory Standards Institute* (CLSI), publicados em 2000. Ceriotti F. (2007, p.115) decreve que um intervalo de referência "é um intervalo que quando aplicado à população atendida por um laboratório, inclui corretamente a maioria dos sujeitos com características semelhantes ao grupo de referência e exclui os outros.". Com base nesta definição, cumprindo todas as regras inerentes à metodologia, pretende-se estimar um intervalo de referência que responda corretamente às necessidades de diagnóstico do laboratório.

As recomendações dadas pela IFCC e CLSI possuem diferentes etapas. Muitos autores seguem estas mesmas recomendações, reajustando alguns passos de acordo com os dados que possuem e o objetivo final. Os procedimentos a serem estabelecidos seguem a seguinte linha cronológica:

2.4.1. Definição da população e seleção dos indivíduos de referência
2.4.2. Dimensão da amostra de referência
2.4.3. Identificação de *outlier*s e respetiva exclusão
2.4.4. Verificação da Normalidade dos dados
2.4.5. Estimação do intervalo de referência

Nos subcapítulos seguintes são descritas, de forma pormenorizada, todas as etapas a realizar.

2.4.1. Definição da população e seleção dos indivíduos de referência

A definição da população de referência é a primeira etapa a concretizar. A população de referência abrange todos os indivíduos que possuem as caracteristicas necessárias para serem integrados no estudo. A estimação de intervalos de referência reflete o comportamento dos indivíduos em estudo existindo duas categorias para estes, quando avaliada a sua

condição de saúde: não possui a doença ou possui a doença. Podem ser estimados intervalos de referência para cada uma delas, desde que esteja identificada a condição de saúde e o objetivo final. A definição de indivíduo saudável varia consoante o parâmetro analítico que se pretende estudar e, envolve o histórico do indivíduo, assim como uma diversidade de exames realizados.

Selecionados os indivíduos de referência para o estudo, ao parâmetro analítico pode ser necessário aplicar critérios que induzam à exclusão de indivíduos da amostra ou a necessidade de estes serem divididos em partições. É fundamental realizar uma seleção mediante certos critérios específicos do parâmetro, na sua maioria critérios biológicos e socio-económicos, como idade, sexo, peso, genética, qualidade da amostra, etnia, etc. Devido à existência de partições pode haver a necessidade de existirem vários intervalos para o mesmo parâmetro. Por este motivo, todos os valores coletados devem ser completamente descritos e padronizados, de acordo com o uso pretendido, IFCC (1984a).

Os indivíduos de referência são os indivíduos resultantes da população de referência estabelecida.

2.4.2. Dimensão da amostra de referência

Na maioria das situações, o número de indivíduos é limitado, e em consequência, o intervalo estimado é excessivamente largo, IFCC (1984b).

A dimensão da amostra de referência está diretamente relacionada com a distribuição dos dados. Se os dados seguirem uma distribuição Normal utiliza-se o método paramétrico para estimação do intervalo, caso contrário, se os dados não seguirem uma distribuição Normal, utiliza-se o método não paramétrico ou o método robusto.

Assumindo que as observações, originais ou após alguma transformação matemática, seguem uma distribuição Normal e se pretende aplicar o método paramétrico, Ferreira e Andriolo (2008) indicam com base na literatura, que a amostra deve conter mais de 30 observações para que seja estimado o intervalo e, Friedrichs *et al.* (2012) indicam que 20 observações, inclusive, são suficientes. No entanto, se os dados não seguirem uma distribuição Normal, IFCC (1984b) e CLSI (2000) indicam que se deve recorrer, no mínimo,

a 120 observações caso se pretenda aplicar o método não paramétrico. O método robusto pode ser aplicado em amostras que variam a sua dimensão entre as 20 e 120 observações.

Katayev *et al.* (2010) indicam que para esta problemática, a confiabilidade num intervalo de referência está relacionada com o número de observações totais usadas. Ainda indicam que a nível estatístico, torna-se mais robusto analisar um vasto conjunto de observações podendo incluir indivíduos não saudáveis, do que apenas 120 observações pertencentes a indivíduos que se assumem sem doença.

A amostra de referência não deve conter observações identificadas como *outliers*. Ainda assim, quando essas são detetadas, a sua exclusão deve ser avaliada de forma cuidada.

Os métodos de estimação para intervalos de referência são descritos posteriormente.

2.4.3. Identificação de *outlier*s e respetiva exclusão

Estabelecida a amostra de referência é necessário analisar as observações identificadas como *outliers*, verificando a sua natureza e, se justificável, proceder à sua exclusão. Existem diferentes formas de definir *outliers*, a definição aplicada é enunciada da seguinte forma:

Definição 6 (Grubbs (1969))**:** Uma observação *outlier* é uma observação que se parece desviar acentuadamente dos restantes valores do conjunto de dados.

Os *outliers* detetados numa amostra ou população, assumem diferentes naturezas. Estas observações surgem como resultado de desvio de um procedimento experimental, erros de cálculo ou registo de valores numéricos, Gubbs (1969). A existência de *outliers* pode induzir a variações de erro elevadas, distorção nas estimativas de parâmetros e nas estatísticas de testes paramétricos ou não paramétricos, Osborne e Overbay (2004). Estes mesmo autores, Osborne e Overbay (2004), sugerem que uma investigação para averiguar a sua origem na base de dados deve ser tomada como primeira instância.

Quase todos os métodos para deteção de *outliers* baseiam-se no pressuposto dos dados seguirem uma distribuição Normal. Quando os dados não assumem a Normalidade, as probabilidades associadas a esses testes serão diferentes. Enquanto se encontram em desenvolvimento métodos que não sejam sensíveis ao pressuposto da Normalidade, compete ao investigador realizar uma interpretação correta dos resultados obtidos, Gubbs (1969).

Horn e Pesce (2003) sugerem que como primeiro passo na análise dos dados, seja realizada a observação do histograma e *boxplot.* Estes permitem por métodos gráficos, observar o comportamento e enviesamento da variável. Mediante a observação gráfica, se justificável, aplicar um dos métodos de deteção *de outliers*. A maioria dos testes existentes para *outliers* assumem como pressuposto a Normalidade dos dados, CLSI (2000). No caso de os dados não seguirem uma distribuição Normal, serão aplicados os métodos de Tukey e Horn.

O método de Tukey (1977) é uma alternativa clássica para detetar *outliers*. O método de Tukey identifica a existência de *outliers* através de distâncias calculadas a partir dos valores do 1° e 3° quantil. Quando aplicado este método no *software* R é possível, recorrendo ao *boxplot*, a visualização dessas mesmas observações.

Assumido um intervalo inter-quantil, $IQR = Q3 - Q1$, são calculadas 4 distâncias,

$$q1 = Q3 - 1{,}5 \times IQR \qquad Q1 = Q3 - 3 \times IQR$$

$$q3 = Q3 + 1{,}5 \times IQR \qquad Q3 = Q3 + 3 \times IQR$$

Este método reconhece *outliers* quando estes se localizam entre $q1$ e $Q1$ ou $q3$ e $Q3$, e também, quando a sua distância ao 1° e 3° quantil é superior a $Q1$ ou $Q3$.

O método de Horn, Horn *et al.* (2001), assume uma transformação Box-Cox aos dados, sendo necessário determinar o valor de λ (Ver apêndice B).

$$y = \begin{cases} (x^{\lambda} - 1)/\lambda & se\ \lambda \neq 0 \\ \ln(x + c) & se\ \lambda = 0 \end{cases} .$$

De seguida, recorre-se a 50% do meio da amostra e retira-se as observações que sejam identificadas pelo método de Tukey, Horn *et al.* (2001).

Os *outliers* detetados devem ser avaliados e não automaticamente excluídos. Estes podem indicar informação importante sobre os dados, e devem por isso, ser devidamente analisados para identificar possíveis causas do seu aparecimento. Ferreira e Andriolo (2008) referem que até 10% da amostra pode ser desprezada devido a observações *outlier*.

2.4.4. Verificação da Normalidade dos dados

A verificação da Normalidade dos dados é um procedimento essencial para a determinação do método a aplicar na estimação do intervalo de referência. Inicialmente, procede-se a uma análise gráfica dos dados, histograma e *boxplot*, que permitirá ao estatístico ter a perceção inicial de como os dados se encontram distribuídos. De seguida, aplica-se um teste de hipóteses, onde a hipótese nula e a hipótese alternativa são:

$$H_0: Os\ dados\ seguem\ uma\ distribuição\ Normal.$$
$$vs$$
$$H_1: Os\ dados\ não\ seguem\ uma\ distribuição\ Normal.$$

Estes testes permitem verificar se a hipótese nula é valida, isto é, não se rejeita H_0 se o valor de prova (*p-value*) for maior que α. Os principais testes para a validação da Normalidade, Ghasemi e Zahediasl (2012), são:

I. Teste de Shapiro-Wilk (amostras menores a 30 observações);
II. Teste de Kolmogorov-Smirnov (com correção de Lilliefors);
III. Teste de Andersen-Darling.

No caso dos testes de Kolmogorov-Smirnov e Anderson-Darling, estes são utilizados para testar a qualidade de ajuste a uma determinada distribuição, e, tem-se que a hipótese nula e a hipótese alternativa são:

$$H_0: Os\ dados\ seguem\ uma\ determinada\ distribuição.$$
$$vs$$
$$H_1: Os\ dados\ não\ seguem\ uma\ determinada\ distribuição.$$

O teste de Anderson-Darling é uma modificação do teste de Kolmogorov-Smirnov, sendo que este é mais poderoso e atribui maior relevância ao peso das caudas das distribuições, Engineering Statistics Handbook (2018).

O resultado do teste de hipóteses e a análise gráfica aos dados permite verificar se estes seguem, ou não, uma distribuição Normal.

2.4.5. Estimação do intervalo de referência

De acordo com o Capítulo 2.4.4., a verificação da Normalidade dos dados determina qual o método de estimação intervalar a aplicar. Se os dados seguirem uma distribuição Normal, aplica-se o método paramétrico; caso contrário, aplica-se o método não paramétrico ou o método robusto.

O método paramétrico estima o intervalo de referência de forma equiparada à estimação de um intervalo de confiança para a média com $(1-\alpha)\%$ de confiança. Mesmo quando se procede a uma transformação dos dados, desde que estes verifiquem a Normalidade, a fórmula utilizada é, IFCC (1984b),

$$(\bar{x} - z_{1-\alpha/2}\frac{s_x}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{s_x}{\sqrt{n}}),$$

onde:

- $\bar{x}$, média amostral;
- s_x, desvio-padrão amostral;
- n, tamanho da amostra;
- $z_{1-\alpha/2}$ é o respetivo quantil da distribuição Normal (0,1).

O método não paramétrico recorre à atribuição de *rankings*, IFCC (1984b). O processo inicia-se organizando as observações desde a mais pequena, com valor 1, à mais elevada, com valor n. Realizada a ordenação, calcula-se o quantil $0{,}025$ da forma $0{,}025(n+1)$ e o quantil $0{,}0975$ realizando o cálculo, $0{,}0975(n+1)$. Assume-se que o limite inferior do intervalo de referência é correspondente ao valor do *rank* que assume a posição $0{,}025(n+1)$, e o limite superior, ao valor do *rank* que assume a posição $0{,}0975(n+1)$. Se o valor obtido não for inteiro, os limites do intervalo de referência devem ser obtidos realizando uma interpolação entre os valores de referência inferior e superior aos obtidos para cada limite, CLSI (2000). Caso os valores sejam muito próximos de um inteiro, efetua-se uma simples aproximação. Neste método, como se realiza a atribuição de *ranks* aos valores de referência torna-se evidente que poderão surgir valores repetidos, e uma forma de combater essa problemática é aumentando o número de casas decimais.

Quando se pretende estimar um intervalo de referência e se possui uma amostra mínima de 120 indivíduos, CLSI (2000) recomenda o uso do método não paramétrico devido à sua simplicidade de execução e confiabilidade, pois este não necessita de uma suposição acerca da distribuição dos valores.

Após a descrição detalhada das etapas a seguir, em algumas situações, os dados reais não se adequam a uma decisão final que atenda os pressupostos a cumprir para utilização do método paramétrico ou não paramétrico, nomeadamente distribuição dos dados e dimensão da amostra. O método proposto por Horn (1988) e, abordado num novo estudo do autor juntamente com outros, Horn *et al.* (1998), é denominado como método robusto e substitui as usuais três estimativas $\bar{x}, s^2$e s^2/n por estimativas robustas de localização e dispersão. O intervalo com $(1-\alpha)100\%$ de confiança é dados por:

$$\begin{pmatrix} T_{bi}(c_1) - t_{n-1}(1-\alpha/2)[s_T^2(c_1) + s_{bi}^2(c_2)]^{1/2}; \\ T_{bi}(c_1) + t_{n-1}(1-\alpha/2)[s_T^2(c_1) + s_{bi}^2(c_2)]^{1/2} \end{pmatrix},$$

onde:

- T_{bi}, estimador de localização *biweigth*[2] da constante de ajuste c_1;
- $s_T^2(c_1)$, estimador *biweigth*[2] da variabilidade de $T_{bi}(c_1)$;
- $s_{bi}^2(c_2)$, estimador *biweigth*[2] de espalhamento da constante de ajuste c_2.

Como nota importante a reter, dependendo do parâmetro para o qual está a ser estimado o intervalo de referência, esse pode exigir que sejam estimados diferentes intervalos, mediante o critério estabelecido. Quando se verifica esta necessidade, realiza-se todas as etapas anteriormente descritas mediante o critério aplicado, sendo que é válido verificar se existem diferenças significativas entre os grupos de cada partição. Existem vários testes de hipóteses que permitem validar se essas diferenças se verificam, sob o pressuposto: as amostras são independentes ou emparelhadas entre si. São aplicados diferentes testes de hipóteses para este fim, sendo que estes são escolhidos tendo por base a distribuição dos dados.

[2] O termo *biweigth* encontra-se referenciado em Horn *et al.* (1998).

3. Prevalência

Foram elaborados estudos de prevalência de nascimento das DLSs em Poorthius *et al.* (1999) e Meikle *et al.* (1999), na Holanda e na Austrália, respetivamente. Pinto *et al.* (2004) estimaram a prevalência em Portugal e Poupětová *et al.* (2010) estudaram o impacto destas patologias na República Checa. A mais recente abordagem à estimação da prevalência foi elaborada por Elmonem *et al.* (2016), em crianças egípcias.

Com base na casuística, foi estimado que a prevalência em 2004 destas patologias na população Portuguesa era de 25/100,000 nados vivos.

3.1. Doenças Lisossomais

As doenças lisossomais de sobrecarga, DLSs, são um grupo de doenças hereditárias do metabolismo caracterizadas pela acumulação de substratos dentro do lisossoma, devido a mutações em genes que codificam enzimas necessárias para a digestão intracelular. As deficiências resultam em disfunção metabólica que pode induzir à morte das células.

Na sua grande maioria, estas patologias são de transmissão autossómicas recessivas, isto é, associadas aos autossomas, cromossomas não responsáveis pela determinação do sexo do indivíduo. Quando a patologia se encontra associada aos autossomas, atinge em igual proporção indivíduos do sexo masculino e do sexo feminino, e é transmitida pelos dois progenitores, portadores ou afetados pela doença, Instituto Unidos pela Vida (2010).

As patologias adquiridas pelo cromossoma sexual X, atingem indivíduos dos diferentes sexos em diferentes proporções, onde indivíduos do sexo masculino são sempre afetados. No caso dos indivíduos do sexo feminino estes são portadores, mas não são afetados.

Estas patologias são consideradas raras pois apenas um número limitado de indivíduos na população total "menos de um em cada 2000", APN (2015), é afetado pela doença. As doenças raras ocorrem com pouca frequência na população, mas possuem um elevado número de consequências tanto a nível clínico como social. A nível clínico, pois sendo o diagnóstico tardio, pode já não existir forma de suavizar os danos, colocando em risco a vida

do indivíduo. A nível social, atualmente, esta tem sido uma área de crescimento a nível científico e investimento no desenvolvimento de abordagens terapêuticas.

Para algumas das doenças lisossomais existem abordagens terapêuticas que permitem melhorar a qualidade de vida, entre estas, substituição enzimática, redução de substrato e terapia de chaperones. Mais recentemente, a terapia génica é a nova abordagem em implementação, encontra-se ainda em fase de ensaios clínicos. Os custos associados ao tratamento destas patologias são muito elevados.

Estas doenças possuem muitas formas de serem agrupadas, não existindo apenas uma forma de classificação, pois o critério pode variar desde a natureza bioquímica do material primário acumulado ao grupo de características de apresentação clínica.

Um dos critérios divide as DLSs nos Grupos I, II, III, IV V e VI. Nos subcapítulos seguintes é feita uma descrição destes subgrupos de DLSs.

3.1.1. Grupo I

O Grupo I é representativo do maior grupo de doenças lisossomais de sobrecarga (DLSs), assinaladas por serem doenças multissistêmicas, isto é, afetam diferentes órgãos, levando a grave disfunção multiorgânica. À exceção de apenas uma patologia deste grupo, são na sua maioria autossómicas recessivas.

São patologias caracterizadas pela falta da produção de uma enzima para digestão de açúcares. Existem diferentes tipos de doenças dentro deste grupo, onde a sua tipologia se encontra diretamente ligada à enzima que o organismo não consegue produzir. Atualmente o diagnóstico é mais eficaz, comparado com tempos anteriores, pois alguns dos sintomas são comuns a outras doenças.

O tratamento mais requisitado para as doenças relacionadas com este grupo é a substituição enzimática, que previne a acumulação do substrato primário caracterizante deste grupo no organismo. Os cientistas encontram-se a desenvolver estudos, acreditando que um dia serão encontrados novos tratamentos e planos de prevenção para estes indivíduos, tornando-se, por muitos fatores, necessário ao longo da vida destes pacientes, um acompanhamento especializado, NINDS (2015).

3.1.2. Grupo II

O Grupo II constitui um grupo de DLSs caraterizadas pela deficiência no catabolismo de glicoproteínas. A parte glicosídica das glicoproteínas é degradada por enzimas hidrolíticas.

Este grupo apresenta formas clínicas leves a graves. Alguns sintomas como infeções recorrentes, deficiência auditiva, ataxia e défice de acuidade visual. Não existe nenhum tratamento específico, sendo que as principais opções de tratamento se direcionam para melhorar os sintomas apresentados. Para a maioria destas doenças, a abordagem terapêutica encontra-se em investigação.

3.1.3. Grupo III

O Grupo III é um grupo caracterizado pela acumulação de lípidos em várias células e tecidos do corpo. Os lípidos são um componente importante na constituição das membranas e da bainha mielina que cobre e protege os nervos, que devido à acumulação progressiva induz a danos permanentes nas células e tecidos, particularmente, no cérebro, sistema nervoso periféricos, fígado, baço e medula óssea.

Os sintomas aparecem no nascimento, adolescência ou, até mesmo, na vida adulta.

3.1.4. Grupo IV

O Grupo IV é um grupo caracterizado pela acumulação de lipofuscina no citoplasma de células que não se dividem como os neurônios e fibras do tecido muscular.

Clinicamente, os pacientes apresentam sintomas como atraso mental, cegueira progressiva, epilepsia, atraso no desenvolvimento de capacidades motoras e leva, na maioria dos casos, a morte prematura. Até ao presente, existe apenas tratamento para uma patologia dentro deste grupo.

3.1.5. Grupo V

O Grupo V inclui as patologias lisossomais de deficiência enzimática que não se englobam em nenhum dos restantes grupos.

3.1.6. Grupo VI

O Grupo VI é caracterizado por uma anormal acumulação de carbohidratos e lípidos nas células. Da acumulação de compostos não degradados resultam, entre outras, grave deformação do esqueleto.

Estas patologias podem ser detetadas no nascimento, ou nas primeiras fases de desenvolvimento. Não existe cura ou terapias específicas para este grupo, apenas tratamentos que permitem apaziguar os sintomas e concedem ao paciente a possibilidade de uma melhora, ainda que reduzida, de qualidade de vida. Estas terapias encontram-se ainda em estudo.

3.2. Base de dados

A base de dados inicial era constituída por 1587 indivíduos. Esta possuía todos os indivíduos e todas as patologias diagnosticadas na Unidade de Bioquímica Genética entre 1982 a 2017. Sendo o objeto de estudo, as doenças lisossomais de sobrecarga (DLSs), aplicou-se um filtro aos 1587 indivíduos, dando origem a uma base final constituída por 970 indivíduos. A base de dados final inclui diagnósticos pré e pós-natais de todas as patologias que possuem doentes. Estabelecendo uma relação entre a base de dados e as distribuições Binomiais e Poisson, a base de dados apenas possui sucessos/eventos.

Incluído no grupo das DLSs, a Unidade de Bioquímica Genética oferece diagnóstico a mais patologias, mas que efetivamente, ainda não possuem indivíduos diagnosticados. Por esse motivo, essas não aparecem na base de dados.

3.2.1. Variáveis da base de dados

A base de dados final possui 9 variáveis. Estas encontram-se enumeradas e descritas na Tabela 1.

Tabela 1: *Descrição das variáveis na base de dados.*

Variável	Classe	Níveis	Descrição
Número do caso	Numérica		Número de identificação do indivíduo.
Data de nascimento			Data de nascimento do paciente, no formato dd/mm/aaaa.
Sexo	Qualitativa Nominal	2	Sexo do individuo, classificado como, "M" = Masculino "F" = Feminino.
Número de família	Qualitativa Nominal	682	Número da família a que pertence cada indivíduo. Como se trata de doenças genéticas, sendo diagnosticado um individuo, a partir deste pode ser realizado um estudo à família e identificar novos indivíduos doentes. O número de família é muito importante a nível clínico.
Patologia	Qualitativa Nominal	46	Doença lisossomal.
Parentesco	Qualitativa Nominal		Grau de parentesco, em relação ao caso índex na família. Caso índex = caso inicial em estudo, na família.
Estatuto	Qualitativa Nominal	4	Cada estatuto, corresponde à condição do indivíduo, relativamente à patologia.

			"0" – Doente com patologia de transmissão autossómica recessiva. "4" – Doente do sexo masculino com patologia de transmissão ligada ao cromossoma X (hemizigoto). "6" – Heterozigoto. "7" – Heterozigotia (Portador).
Ano de Estudo			Ano em que o indivíduo começou a ser estudado no CGMJM.
Fetos	Qualitativa Nominal	2	"0" – Não feto "1" – Feto.

A variável "Fetos" foi criada posteriormente, para ser possível uma identificação imediata dos casos pré-natais e pós-natais, respetivamente.

3.2.2. Análise Exploratória

A elaboração de uma análise estatística permite retirar conclusões a partir dos dados, de forma a enriquecer o conhecimento do estatístico acerca da realidade da mesma.

Não existe apenas uma forma de explorar a natureza dos dados recolhidos tendo, por esse motivo, uma enorme repercussão nas restantes etapas a utilização das ferramentas estatísticas corretas. Elaborada uma correta análise exploratória, esta permite retirar algumas conclusões iniciais, e futuramente comparar com os resultados obtidos.

As duas variáveis de interesse a serem analisadas são, a variável que indica a patologia que o indivíduo possui e, a que indica se este é um feto ou não. A análise exploratória para este estudo, em particular, baseia-se na construção de diagramas de barras, Murteira (1993). Este é utilizado para realizar comparações entre as categorias de uma variável qualitativa ou

quantitativa discreta. Na base de dados descrita no Capítulo 3.2.1 verifica-se a existência de variáveis qualitativas, apenas.

Na Tabela 2, encontra-se a tabela das frequências absolutas de casos na base de dados, pré e pós-natal, para cada patologia.

Tabela 2: *Frequência absoluta de casos pré-natal e pós-natal, por patologia.*

	Número de Diagnósticos			Número de Diagnósticos	
Patologia	**Pós-natal**	**Pré-natal**	**Patologia**	**Pós-natal**	**Pré-natal**
Patologia I	34	4	Patologia XXVI	2	0
Patologia II (♂)	32	3	Patologia XXVII	1	0
Patologia III	3	0	Patologia XXVIII	56	0
Patologia IV	24	1	Patologia XXIX (♂)	1	0
Patologia V	7	0	Patologia XXX	4	1
Patologia VI	15	2	Patologia XXXI	142	0
Patologia VII	2	0	Patologia XXXII	4	1
Patologia VIII	22	1	Patologia XXXIII	17	0
Patologia IX	6	1	Patologia XXXIV	39	4
Patologia X	8	0	Patologia XXXV (♂)	125	0
Patologia XI	1	0	Patologia XXXV (♀)	138	0
Patologia XII	1	0	Patologia XXXVI	25	2
Patologia XIII	3	0	Patologia XXXVII	14	0
Patologia XIV	4	1	Patologia XVIII	24	3
Patologia XV	5	0	Patologia XXXIX	1	0
Patologia XVI	13	2	Patologia XL	1	0

	Número de Diagnósticos			Número de Diagnósticos	
Patologia	**Pós-natal**	**Pré-natal**	**Patologia**	**Pós-natal**	**Pré-natal**
Patologia XVII	4	0	Patologia XLI	3	0
Patologia XVIII	1	0	Patologia XLII	27	8
Patologia XIX	2	0	Patologia XLIII	1	0
Patologia XX	8	0	Patologia XLIV	21	0
Patologia XXI	19	1	Patologia XLV	48	1
Patologia XXII	1	0	Patologia XLVI	4	1
Patologia XXIII	12	1			
Patologia XXIV	2	0			
Patologia XXV	1	0			

Na Figura 2, encontra-se o diagrama de barras para a frequência absoluta de casos na base de dados, pré e pós-natal, para cada patologia.

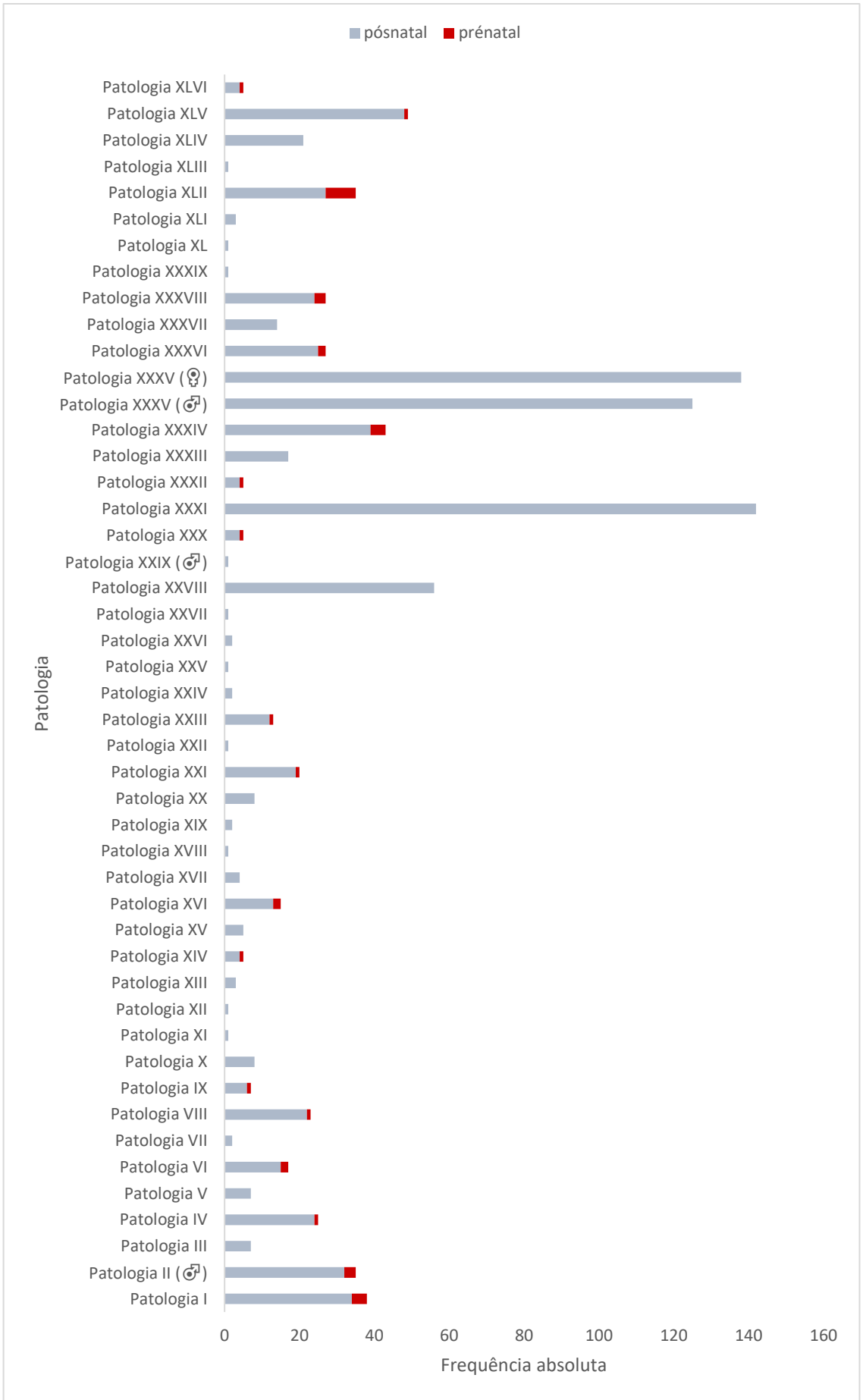


Figura 1: *Diagrama de barras de casos, pré-natal e pós-natal, por patologia.*

Pela análise da Tabela 2 e da Figura 1, pode-se concluir que as patologias com mais individuos diagnosticados são a Patologia XXXV, XXXI, XXVIII e XLV. Os individuos diagnósticados com a Patologia II, XXIX e XXXV encontram-se divididos por sexo, pois esta é uma patologia genética ligada ao cromossoma X. Verifica-se que a patologia que possui mais fetos diagnosticados é a Patologia XLII.

Na Figura 2, encontra-se o diagrama de barras para a frequência absoluta de indivíduos e fetos, por sexo.

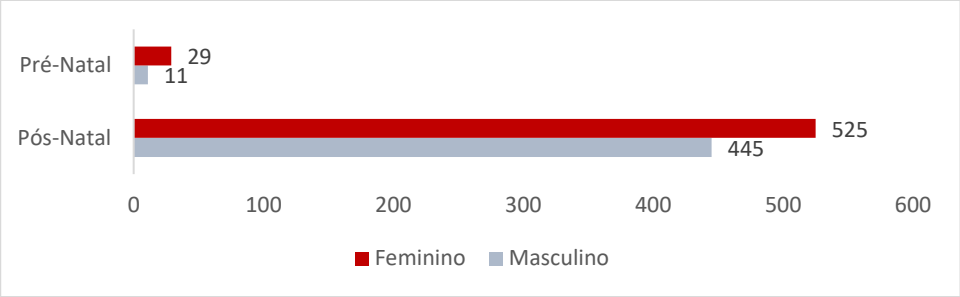


Figura 2: *Diagrama de barras do número de casos, por sexo.*

Analisando a Figura 2, verifica-se que existem 525 indivíduos do sexo feminino, em que 29 são fetos. O número de indivíduos do sexo masculino presentes na base de dados são 445 indivíduos do sexo masculino, entre os quais 11 são fetos.

3.3. Estimativa da prevalência

Como revisto em Wagner MB (1998), a forma mais simples de determinar a frequência de uma doença é através da contagem dos indivíduos afetados. Dada essa informação torna-se relevante estudar medidas que relacionem o número de casos diagnosticados com o tamanho da população. Desta forma, é possível compreender o impacto que a patologia assume na população.

O objetivo deste estudo é medir a frequência ou magnitude que estas doenças genéticas têm na população Portuguesa, desde o início do registo do diagnóstico até à atualidade. Sendo o objeto de estudo as DLSs, estabeleceu-se a ponte de ligação entre o conceito matemático e o conceito epidemiológico, sendo que a medida que foi calculada será designada posteriormente de prevalência.

Definição 7 (Wagner MB (1998)): A prevalência é uma medida que pretende medir a proporção de indivíduos que, num determinado momento, se encontram diagnosticados com uma determinada doença numa população.

É considerada uma medida estática, relacionada com um momento no tempo, ainda que a recolha de dados ocorra em dias, meses ou anos. Resulta de uma fração, onde no numerador se encontra o número de indivíduos diagnosticados até ao momento e no denominador estão presentes os casos diagnosticados e não diagnosticados. O cálculo da prevalência é o cálculo de uma proporção binomial. Pode também ser retratada como uma taxa.

A prevalência pode ser medida pontualmente ou ao longo de um período. A prevalência pontual é efetuada num determinado ponto específico do tempo, enquanto que a prevalência de período abrange todos os casos existentes nesse período, IPTSP (2018).

A prevalência retratada nos resultados deste estudo é referente à prevalência de período. Efetuou-se a estimativa das seguintes prevalências:

I. Prevalência de período, relativa ao período de atividade desta unidade laboratorial, entre 1982 a 2017.

II. Prevalência de nascimento, ou *birth prevalence*, relativa ao período entre o ano de nascimento do indivíduo mais velho e o ano de nascimento do indivíduo mais novo, para cada patologia.

O denominador utilizado no cálculo da prevalência corresponde a todos os indivíduos em risco de possuírem a doença, isto é, todos os nados vivos naquele período. Esses valores foram obtidos, somando o número de nados vivos nascidos em cada ano. O número de nados vivos até 2016, utilizado para a estimativa das prevalências, foi recolhido a partir do documento disponibilizado anualmente pelo INE, denominado de Estatísticas Demográficas, INE (2017). O número de nados vivos referente a 2017 foi recolhido a partir de uma tabela disponibilizada pelo INE, INE (2018).

A fórmula para a prevalência de nascimento é, numa primeira instância, referida em Poorthuis *et al.* (1999) e, recentemente em Poupětová *et al.* (2010). Foi com base nesta bibliografia que se decidiu aplicar a mesma definição, para futuramente, serem comparados valores de prevalência com base no mesmo registo. Nas patologias em que se possui apenas um caso diagnosticado, o denominador utilizado é o mesmo que na prevalência de período, Poupětová *et al.* (2010).

A prevalência foi estimada por 100,000 nados vivos. Tomou-se a iniciativa de estimar, também, a prevalência de período de atividade da unidade laboratorial, por uma questão de curiosidade na diferença de valores.

Nas patologias ligadas ao cromossoma X, o denominador corresponde ao número de nados vivos num período de tempo relativo ao sexo dos indivíduos para o qual se está a calcular a prevalência. Numa patologia ligada ao cromossoma X, considerando apenas indivíduos do sexo masculino, o denominador é constituído pelo número de nados vivos masculinos nesse período de tempo; no caso contrário, o mesmo se sucede para indivíduos do sexo feminino. À exceção dessas patologias, a estimativa da prevalência de período de atividade do CGMJM assume sempre o mesmo valor no denominador. Relativamente à prevalência de período, neste trabalho apenas se apresenta o número de nados vivos que engloba indivíduos do sexo masculino e feminino, no período considerado.

Sendo x, o número de casos pré e pós-natal diagnosticados, o cálculo é realizado aplicando a seguinte fórmula:

$$P_a = \frac{x}{3{,}991{,}824} \times 100{,}000.$$

Na estimação da prevalência de nascimento, quando se trata de patologias ligadas ao cromossoma X, sucede-se o mesmo que fora descrito acima. Na prevalência de nascimento o valor do denominador varia sempre perante o período que é estabelecido. Sendo x, o número de casos pré e pós-natal diagnosticados, o cálculo é realizado aplicando a seguinte fórmula:

$$P_b = \frac{x}{nados\ vivos\ do\ periodo\ de\ tempo\ estabelecido} \times 100{,}000.$$

Foram, posteriormente, obtidos os intervalos de confiança de todas as metodologias apresentadas anteriormente, para a prevalência de nascimento.

3.4. Resultados

De modo a exibir os resultados obtidos neste estudo, foram elaboradas tabelas onde constam apenas 6 das 46 patologias da base de dados. No Apêndice A encontram-se as estimativas das prevalências de período e de nascimento para todas as patologias de grupo.

Na Tabela 3 encontram-se os valores de prevalência estimados.

Tabela 3: *Estimativas da prevalência de nascimento e período.*

	Número de casos					
Patologia	**Pós-natal**	**Pré-natal**	**Prevalência de período/ Taxa**	**Período de nascimento**	**Número de nados vivos**	**Prevalência de nascimento / Taxa**
Patologia II (♂)	32	3	1,7018[a]	1967 - 2012	3,262,991[a]	1,0726[a]
Patologia X	8	0	0,2004	1973 - 2008	4,710,206	0,1698
Patologia XVI	1	0	0,0251	1982 - 2017	3,991,824	0,0251
Patologia XXIV	2	0	0,0501	1999 - 2001	348,784	0,5734
Patologia XXVIII	56	0	1,4029	1935 – 2016	12,865,119	0,4353
Patologia XXXI	142	0	3,5573	1917 - 2014	16,262,226	0,8732

[a] inclui apenas nados vivos masculinos

Na Tabela 3 torna-se evidente a diferença do valor da estimativa de prevalência, quando aplicados dois períodos de tempo diferentes. Analisando os valores da estimativa da prevalência de período, a patologia que apresenta maior prevalência é a Patologia XXXI, enquanto que a patologia com menor prevalência é a Patologia XVI. Sendo o denominador constante, as patologias com maior prevalência resultam das que possuem maior número de casos diagnosticados e, o contrário também se verifica. Observando os valores da estimativa da prevalência de nascimento, a patologia que apresenta maior prevalência é a Patologia II, enquanto que a Patologia XVI, em concordância com a prevalência de período, é a que apresenta menor prevalência.

Os valores da prevalência de nascimento alteram-se em grande escala comparativamente às prevalências de período pois, o denominador varia atendendo ao período de nascimento dos casos diagnosticados. As patologias em que se verifica essa abrupta diferença são a Patologia II XXIV, XXVIII e XXXI.

É visível que a diferença de valores provém do período de tempo que se assume no cálculo da prevalência. É de extrema importância ser reportada sempre no esclarecimento do método, qual o número de nados vivos utlizado e em que período de tempo este se baseou, de forma a que os resultados possam ser replicados.

Neste trabalho são apresentados apenas os resultados de estimação intervalar para a prevalência de nascimento, pois é aquela que está na base da redação de um artigo científico. Na Tabela 4 apresentam-se as respetivas estimativas intervalares para a prevalência de nascimento, para cada um dos métodos referidos no Capítulo 2, a 95% de confiança e 95% de credibilidade.

Tabela 4: *Intervalos de confiança de Clopper-Pearson, Wald, Agresti-Coull, Wilson e mid-p exato e intervalo de Poisson a 95% de confiança; intervalo de credibilidade de Jeffreys a 95% de credibilidade.*

Patologia	Prevalência de nascimento/ Taxa	Método	Intervalo	Comprimento
Patologia II (♂)	1,0726[a]	Clopper-Pearson	0,7471 – 1,4918	0,7447
		Wald	0,7173 – 1,4280	0,7101
		Agresti-Coull	0,7665 – 1,4965	0,7300
		Wilson	0,7713 – 1,4917	0,7204
		Mid-p exato	———	———
		Jeffreys	0,7599 – 1,4739	0,7140
		Poisson	0,7471 – 1,4918	0,7447
Patologia X	0,1698	Clopper-Pearson	0,0733 – 0,3347	0,2614
		Wald	0,0522 – 0,2875	0,2353
		Agresti-Coull	0,0796 – 0,3417	0,2621

Patologia	Prevalência de nascimento/ Taxa	Método	Intervalo	Comprimento
		Wilson	0,0861 – 0,3352	0,2491
		Mid-p exato	———	———
		Jeffreys	0,0803 – 0,3205	0,2402
		Poisson	0,0733 – 0,3347	0,2614
Patologia XVI	0,0251	Clopper-Pearson	0,0006 – 0,1396	0,1390
		Wald	0,0000 – 0,0742	0,0742
		Agresti-Coull	0,0000 – 0,1571	0,1571
		Wilson	0,0044 – 0,1419	0,1375
		Mid-p exato	———	———
		Jeffreys	0,0027 – 0,1171	0,1144
		Poisson	0,0006 – 0,1396	0,1390
Patologia XXIV	0,5734	Clopper-Pearson	0,0694 – 2,0714	2,0020
		Wald	0,0000 – 1,3681	1,3681
		Agresti-Coull	0,0114 – 2,2368	2,2254
		Wilson	0,1573 – 2,0910	1,9337
		Mid-p exato	———	———
		Jeffreys	0,1192 – 1,8396	1,7204
		Poisson	0,0694 – 2,0714	2,0020
Patologia XXVIII	0,4353	Clopper-Pearson	0,3288 – 0,5653	0,2365
		Wald	0,3213 – 0,5493	0,2280
		Agresti-Coull	0,3343 – 0,5662	0,2319
		Wilson	0,3352 – 0,5652	0,2168

Patologia	Prevalência de nascimento/ Taxa	Método	Intervalo	Comprimento
		Mid-p exato	———	———
		Jeffreys	0,3322 – 0,5609	0,2287
		Poisson	0,3288 – 0,5653	0,2365
Patologia XXXI	0,8732	Clopper-Pearson	0,7355 – 1,0292	0,2937
		Wald	0,7296 – 1,0168	0,2872
		Agresti-Coull	0,7404 – 1,0296	0,2892
		Wilson	0,7409 – 1,0291	0,2882
		Mid-p exato	———	———
		Jeffreys	0,7383 – 1,0259	0,2876
		Poisson	0,7355 – 1,0292	0,2937

Os intervalos de confiança para cada método, apresentados no Capítulo 2, encontram-se na Tabela 4. Antes de iniciar uma análise cuidada da tabela apresentada, denota-se que não foi possível obter estimativas para os intervalos de confiança para o método mid-P exato. No *software* R, os pacotes existentes possuem funções que estão desenhadas, maioritariamente, para aplicar o método em tabelas de contingência 2×2, e por tal motivo, a biblioteca utilizada para estimar o intervalo pelo método mid-p exato era a única que se aplicava ao estudo em questão, mas ainda assim, não respondeu de acordo com o que seria suposto.

Nota-se que o intervalo estimado pelo método de Poisson exato é idêntico ao intervalo estimado pelo método de Clopper-Pearson. No *software* R, a estimação deste intervalo quando n é elevado assume o teste binomial, que se encontra na base do método de Clopper-Pearson. Elandt-Johnson (1975) distinguiu no seu trabalho, se prevalência devia ser assumida como uma proporção ou uma taxa, afirmando que quando dividimos o número de casos num determinado momento pelo tamanho da população nesse respetivo momento se trata sempre de uma proporção. Não se dando por satisfeito, este assinala ainda que o termo "prevalence

rate" Elandt-Johnson (1975, p.271) é um conceito impraticável. Por este motivo, assumir prevalência como uma taxa é errado e foram apenas considerados, para uma tomada de decisão final, os métodos de estimação intervalar que recorrem à distribuição Binomial.

Relativamente aos métodos de estimação intervalar recorrendo á distribuição Binomial, o intervalo de confiança que apresenta menor amplitude é, na globalidade, o intervalo de confiança de Wald. Brown *et al.* (2001) mostraram que o desempenho deste intervalo é pobre e errático, salientado ainda que as apreciações atribuídas em redações influentes mostram que este possui bastantes falhas, aconselhando a sua não utilização. Cepeda-Cuervo *et al.* (2008) afirmam, com base nos estudos de Agresti e Coull (1998), Brown *et al.* (2002) e Newcombe e Merino (2006), que o uso do intervalo de Wald não é recomendado, pois este possui bastantes problemas, principalmente, quando o valor estimado da proporção é próximo de 0 ou de 1.

Comprova-se que os intervalos estimados com maior comprimento resultam dos métodos de Clopper-Pearson e Agresti-Coull. Brown *et al.* (2002) confirmam no seu estudo que o intervalo de Agresti-Coull é, em média, longo e bastante conservador quando o valor de p é próximo de 0 ou de 1. Este afirma ainda que os intervalos de Wilson e Jeffreys são comparáveis, e entre ambos, o de Jeffreys apresenta menor amplitude, como verificamos na Tabela 4, à exceção na Patologia XXVIII. Agresti e Coull (1998), Brown *et al.* (2001), Brown *et al.* (2002), Cepeda-Cuervo *et al.* (2008) e He e Wu (2009) concluem os seus estudos afirmando que os intervalos recomendados na estimação intervalar de proporções binomiais são os intervalos de confiança de Wilson e Jeffreys. A dúvida na escolha final do método que se aplica melhor a este estudo reside entre estes últimos dois métodos.

Bilder e Loughin (2015) apresentam um exemplo onde recorrem ao método de Wilson para estimar o intervalo de confiança para a prevalência de Hepatite C, após estabelecer uma comparação entre métodos.

Finalizando, o método para a obtenção da estimativa intervalar para proporções binomiais recomendado é o método de Wilson. No Apêndice A, encontram-se estimadas as prevalências de todas as patologias e o respetivo intervalo de confiança de Wilson da prevalência de nascimento.

4. Intervalos de Referência

A Medicina sempre foi uma área prioritária em questões de evolução científica para o Ser Humano. Todos os dias, profissionais da área se questionam sobre o estado de saúde dos seus pacientes, e que exames estes deverão requisitar para ter respostas às suas questões. A seleção de exames deve ser baseada na probabilidade de ser deduzido o diagnóstico, assente no histórico clínico, exame físico e prevalência da doença suspeita, razão pela qual a consulta e o exame clínico devem preceder a requisição dos exames.

Como referido em Ferreira e Andriolo (2008), existem várias normas que, quando examinadas de forma rigorosa, permitem assumir que o resultado de um exame laboratorial é fidedigno e útil no diagnóstico, monitorização ou predisposição de um indivíduo para uma dada doença. As condutas clínicas, desde a admissão ou não de pacientes nas unidades hospitalares, a alteração do esquema terapêutico e, por fim, a alta dada ao paciente são dependentes dos resultados dos exames laboratoriais e consequentemente, da interpretação dos intervalos de referência.

Quando analisados, os resultados das análises que possuem valores anómalos ditam se o paciente se encontra num estado de saúde normal ou num estado de saúde com possibilidade de diagnóstico positivo para a doença. O intervalo de referência é o que permite validar se esse indivíduo se encontra em risco de doença ou não.

Mantendo-se o anonimato do parâmetro, no subcapítulo seguinte é realizada a descrição e contextualização deste, em relação ao objetivo final que os resultados permitem obter.

4.1. Parâmetro

A degradação de macromoléculas, como os glicopeptídeos e glicolipídios, ocorre no lisossoma via enzimas catabólicas. Os defeitos que ocorrem nessas enzimas lisossomais causam doenças lisossomais de sobrecarga, DLSs.

As deficiências enzimáticas resultam de alterações nas sequências genómicas que podem ser mutações pontuais, defeitos na junção de união dos exões/intrões, e rearranjos parciais dos genes.

Para obter um diagnóstico clínico conclusivo e válido, é necessário determinar a deficiência enzimática em amostras biológicas como, por exemplo, plasma e os leucócitos de sangue periférico.

Existem várias técnicas clínicas para deteção de atividade deficiente do parâmetro, uma delas consiste na análise de gota de sangue seco, designadas por *dry blood spotting* (DBS), sendo esta a técnica utilizada nas amostras analisadas neste trabalho.

Para se tornar claro se a amostra recolhida tem qualidade no apoio a um diagnóstico, é realizada uma análise paralela a uma enzima de referência termolábil[3]. Se o valor desta estiver inserido no intervalo de referência dos resultados considerados normais, pode-se então prosseguir com a análise ao parâmetro; caso contrário, é feito um pedido de sangue total para efetuar a análise em outra amostra biológica.

Sendo a patologia de transmissão ligada ao cromossoma X, os *probandos* foram divididos por sexo, sendo que os intervalos de referência estimados para este parâmetro são referente apenas a *probandos* do sexo masculino, isto é, indivíduos do sexo masculino sobre os quais o médico tem suspeita de doença. Foi determinada a estimação de dois intervalos de referência, o primeiro referente a *probandos* diagnosticados com a doença, e outro referente a *probandos* sem diagnóstico.

4.2. Base de dados

A base de dados para este estudo possui todos os *probandos* recebidos com suspeita clínica de doença. Quando o médico suspeita da presença de doença nestes indivíduos, prescreve uma requisição analítica, de forma a confirmar a suspeita de diagnóstico.

[3] Substância sujeita a alterações, decomposição ou destruição por ação do calor.

Os valores de análises de sangue seco de todos os parâmetros analisados perfazem um volume total de 24,966 valores. Inicialmente, realizou-se a filtração do código da análise referente ao parâmetro em estudo, ao sexo dos indivíduos selecionando apenas os indivíduos de sexo masculino e, o filtro relativo à enzima de referência associada. Os valores do parâmetro e da enzima de referência foram agrupados pelo número da requisição, que é única. Verificou-se que alguns indivíduos possuíam na base mais do que uma requisição realizada e, nesses casos foi selecionada a requisição mais recente para constar na base de dados final.

A enzima de referência permite avaliar a qualidade da amostra. Caso não possua a qualidade necessária o laboratório pede ao médico que seja enviada nova colheita de sangue. A base de dados é, portanto, constituída por valores de análise referentes ao parâmetro pretendido, e em conjunto com o resultado associado à enzima de referência. Os indivíduos cujos resultados das análises à enzima de referência não se encontraram no intervalo estipulado para uma amostra de qualidade foram removidos da base de dados.

Como referido anteriormente, a base apenas contém *probandos* do sexo masculino, podendo-se afirmar que o critério aplicado na escolha da amostra de referência foi o sexo do indivíduo, em paralelo, com a qualidade da amostra. Os resultados existentes na base de dados são relativos a amostras de mancha de sangue seco em papel.

4.2.1. Variáveis da base de dados

Estabelecidos e aplicados os critérios de seleção para os indivíduos de referência, na Tabela 5 apresenta-se a descrição da base de dados, constituída por 958 indivíduos.

Tabela 5: *Descrição das variáveis da base de dados.*

Variável	Classe	Níveis	Descrição
Número do caso	Numérica		Número de identificação do indivíduo.
Número da requisição	Numérica		Número de requisição da análise.

Data de nascimento			Data de nascimento do paciente, no formato dd/mm/aaaa.
Sexo	Qualitativa Nominal	2	Sexo do individuo, classificado como, "M" = Masculino "F" = Feminino.
Número de família	Qualitativa Nominal	30	Número da família a que pertence cada indivíduo. Como se trata de doenças genéticas, sendo diagnosticado um individuo, efetua-se o estudo familiar para identificar novos indivíduos doentes.
Patologia	Qualitativa Nominal	3	Doença lisossomal.
Parentesco	Qualitativa Nominal	8	Grau de parentesco, em relação ao caso índex na família. Caso índex = caso inicial em estudo, na família.
Estatuto	Qualitativa Nominal	5	Cada estatuto, corresponde à condição do indivíduo, relativamente à patologia. No laboratório, quando um individuo da família é estudado torna-se necessário atribuir um estatuto, consoante o resultado final obtido. "0" – Doente com patologia de transmissão autossómica recessiva; "4" – Doente do sexo masculino com patologia de transmissão ligada ao cromossoma X (hemizigoto); "5" – Testado com resultado

			bioquímico normal; "29" – Testado com resultado molecular normal; "s/estatuto" – Testado normal sem histórico familiar.
Ano de Estudo	Qualitativa Nominal	9	Ano em que o indivíduo começou a ser estudado no laboratório.
Valores do parâmetro (VP)	Numérica		Resultados analíticos do parâmetro em estudo, por requisição e individuo, em $pmol/h/\text{punção}$ de mancha de sangue seco.
Valores da enzima de referência	Numérica		Resultados analíticos da enzima de referência, por requisição e individuo, em $pmol/h/\text{punção}$ de mancha de sangue seco.

Como descrito inicialmente, a base de dados possui *probandos* em risco de possuírem as doenças que a análise do parâmetro permite determinar. A variável Patologia apresenta 3 níveis que identificam as 3 patologias presentes na base de dados. O parâmetro em estudo valida o diagnóstico de 2 dessas patologias sendo que 1 dessas é uma variante da expressão genética da considerada patologia-mãe. Os indivíduos que possuem a outra patologia podem ser assumidos como doentes, mas não para as patologias que a análise ao parâmetro em questão permite diagnosticar. Estes indivíduos são abrangidos pelo estatuto "0". Estes e todos os restantes indivíduos sem nenhuma patologia associada entram no estudo como *probandos* sem diagnóstico para as patologias que o parâmetro em estudo diagnostica.

Como referido anteriormente, existem dois objetivos fulcrais recorrendo à presente base de dados. Nos subcapítulos seguintes encontra-se realizada a análise exploratória aos dados, que conduzem à tomada de decisão sobre o método a aplicar, para a estimação dos intervalos de referência.

4.2.2. Análise Exploratória

4.2.2.1. Base de dados: *Probandos* Doentes

A primeira etapa para estimação do intervalo consiste em selecionar os indivíduos de referência. Na base de dados, descrita na Tabela 5, os *probandos* doentes são aqueles que possuem o nível "4" da variável Estatuto. Como efeito desta seleção, resultam 43 indivíduos, uma amostra relativamente pequena em relação ao número de indivíduos da base original.

Na Tabela 6 encontra-se a análise descritiva da variável Valores do parâmetro, designada como VP.

Tabela 6: *Análise descritiva da variável VP para probandos doentes.*

	Variável VP
Mínimo	0,000
Média	1,448
Mediana	0,800
Máximo	7,010

Na Tabela 6 conclui-se que os valores do parâmetro variam entre o mínimo, 0,000 $pmol/h/punção$ e o máximo, 7,010 $pmol/h/punção$. O valor médio do parâmetro para *probandos* doentes é de 1,448 $pmol/h/punção$. Pretende-se analisar com se comportam os valores do parâmetro em estudo nesta amostra. Na Figura 3 encontra-se o respetivo histograma com o esboço da distribuição da variável VP a preto e o esboço da distribuição Normal a vermelho, e *boxplot* dos valores do parâmetro para estes indivíduos.

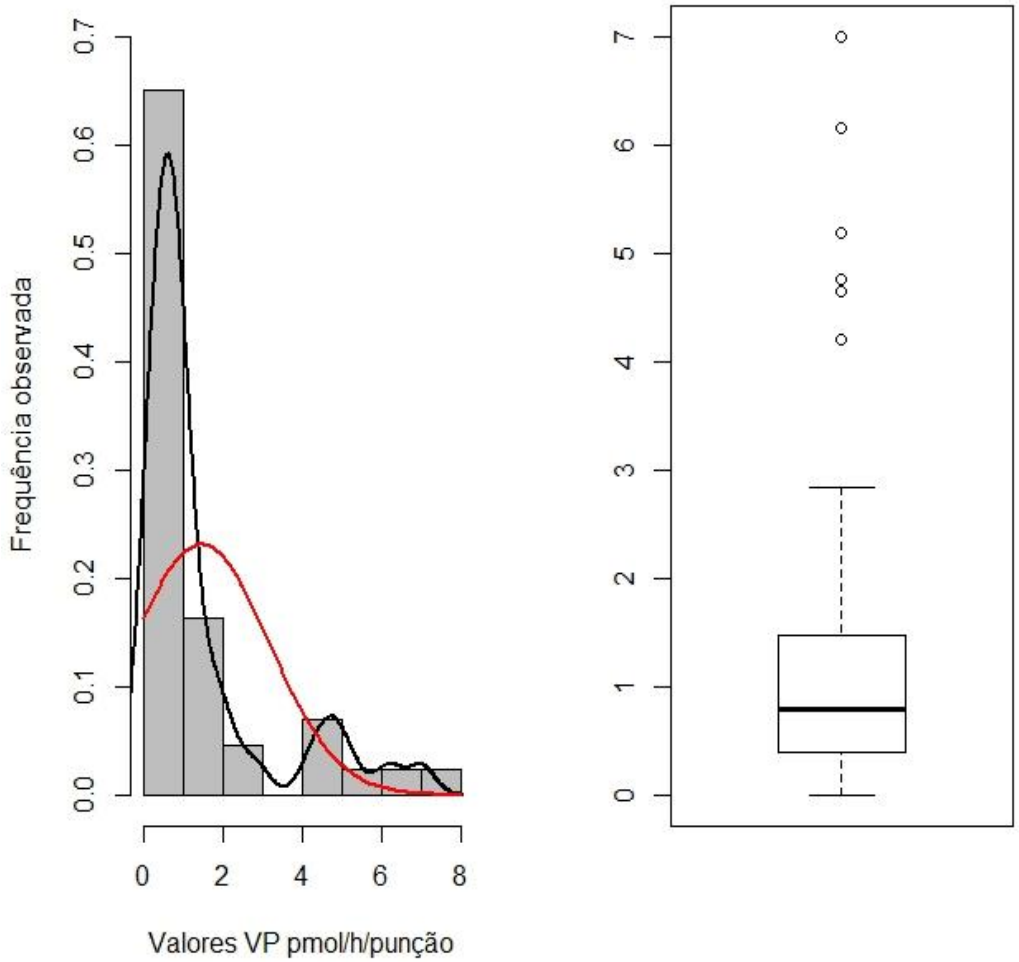


Figura 3: *Histograma e boxplot da base de dados de probandos doentes.*

A Tabela 6 mostra que o valor da média não é coincidente com o valor da mediana podendo-se assumir que a distribuição é enviesada à direita, pois o valor da média é superior ao valor da mediana. Pela observação da Figura 3, o histograma e *boxplot* sugerem que os dados não seguem uma distribuição Normal, evidenciando-se também a presença de observações *outliers*, por análise gráfica do *boxplot*.

Identificadas estas 6 observações, avaliou-se a natureza das mutações desses indivíduos, comprovando-se que são mutações que indicam a presença de doença, sendo uma variante que origina valores mais elevados em relação as restantes (pseudo-deficiência). Encontrado um motivo biológico válido para eliminação destes indivíduos, procedeu-se à eliminação dos mesmos da base de dados, e realizou-se de novo a construção dos mesmos gráficos para verificar se ainda existem observações que destabilizem a tendência dos dados. Foram identificadas ainda 4 observações *outliers*, em que realizada uma avaliação cuidada, apenas 1 observação foi retirada pois representava o valor mais extremo e possuía a mesma mutação identificada no grupo de *outliers* removidos anteriormente. As restantes 3 observações não foram excluídas pois não existia motivo biológico que o justificasse. A exclusão excessiva de observações *outliers* que não possuam motivo para tal, além de não ser correta podem ainda retirar variabilidade à amostra.

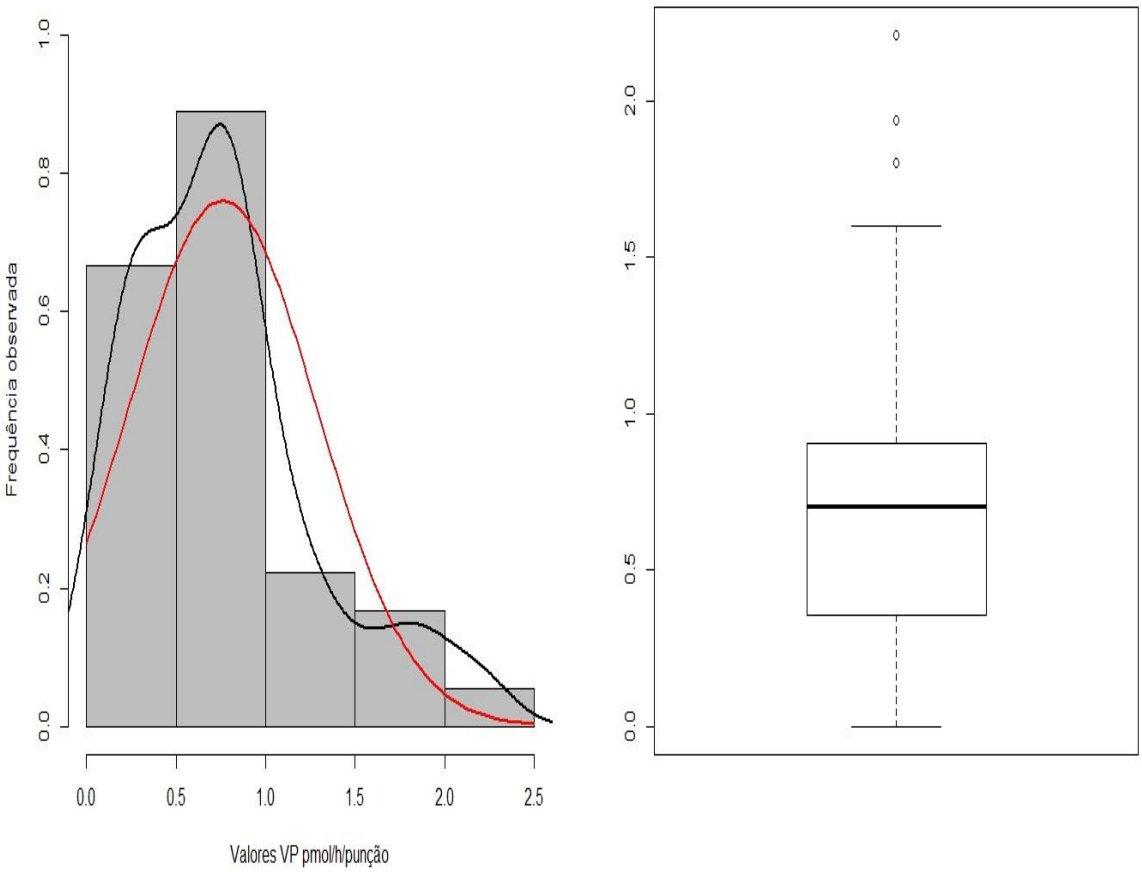


Figura 4: *Histograma e boxplot da base de dados de probandos doentes após exclusão de observações outliers.*

Como resultado final, após os ajustes necessários à base de dados, foram utilizados os valores de 36 observações. Na Figura 4 encontra-se o respetivo histograma com o esboço da distribuição da variável VP a preto e o esboço da distribuição Normal a vermelho, e *boxplot* dos valores do parâmetro, após a eliminação de observações *outliers*. Nesta base de dados recorreu-se apenas ao método de Tukey com análise gráfica do *boxplot* para identificação de observações *outliers* e respetiva exclusão.

Considerando um nível de significância de 5%, o teste de Anderson-Darling obteve um $p - valor = 0{,}0225$, isto é, rejeita-se H_0 a um nível de significância de 5%, podendo-se afirmar que os valores do parâmetro não seguem uma distribuição Normal.

Realizadas todas as etapas precedentes à determinação do método de estimação intervalar a aplicar, pode-se concluir que o método de estimação do intervalo de referência para a base de dados de *probandos* doentes é o método não paramétrico ou o método robusto.

Sendo o valor mínimo do parâmetro 0,00 $pmol/h/punção$, não se reconhece nenhuma transformação que se adeque aos dados e permita resolver a evidente assimetria. Como os dados não seguem uma distribuição Normal e não se possui o mínimo de 120 observações, não é possível aplicar o método não paramétrico. De acordo com o Capítulo 2.4.2., tendo em consideração a dimensão da amostra, o método robusto é o correto a ser aplicado. Horn *et al.* (1998) indicam que, após estabelecer comparações entre métodos para diferentes tamanhos de amostras, quando o tamanho da amostra é reduzido e a aplicação do método não paramétrico não é recomendada ou, se não for possível aplicar nenhuma transformação adequada aos dados de forma a alcançar a Normalidade, o método robusto proposto no Capítulo 2.4.5. deve ser o aplicado.

O método robusto proposto em Horn (1988) e sobre o qual estabeleceu novo estudo em 1998, foi o método escolhido para estimação do intervalo de referência para a base de dados de *probandos* doentes.

4.2.2.2. Base de dados: *Probandos* Normais

De acordo com o Capítulo 2.4.1., a primeira etapa a realizar é determinar os indivíduos de referência. Para este intervalo o objetivo final focasse na estimação de um intervalo de referência para indivíduos em risco, que após observação do valor do parâmetro são diagnosticados como normais, isto é, não apresentam doença. Esta seleção abrange indivíduos com os níveis "0", "5", "29" e "s/estatuto" da variável Estatuto.

Os indivíduos da base de dados que se encontram dentro dos critérios assentes no parágrafo anterior perfazem um total de 915 *probandos*, de um conjunto original de 958 *probandos*.

Na Tabela 7 encontra-se a análise descritiva da variável Valores do parâmetro, designada como VP.

Tabela 7: *Análise descritiva da variável VP para probandos normais.*

	Variável VP
Mínimo	1,710
Média	10,012
Mediana	9,430
Máximo	30,800

Na Tabela 7 conclui-se que os valores do parâmetro variam entre o mínimo, 1,710 $pmol/h/punção$, e o máximo, 30,800 $pmol/h/punção$. O valor médio do parâmetro para *probandos* normais é de 10,012 $pmol/h/punção$. Procede-se com a análise gráfica do histograma com o esboço da distribuição da variável VP a preto e o esboço da distribuição Normal a vermelho, e *boxplot* da base de dados dos *probandos* normais se encontra a curva na Figura 5.

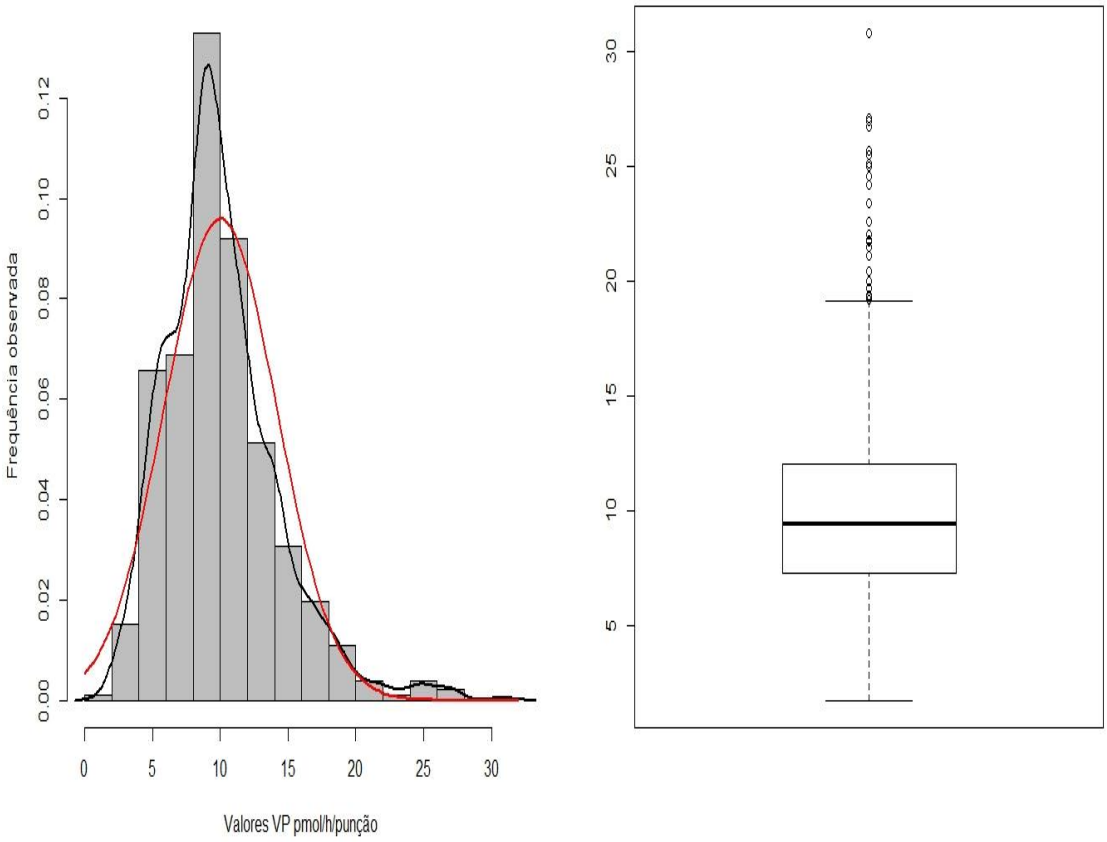


Figura 5: *Histograma e boxplot da base de dados dos probandos normais.*

A Tabela 7 mostra que o valor da média não é coincidente com o valor da mediana podendo-se assumir que a distribuição é enviesada à direita, pois o valor da média é superior ao valor da mediana. Pela observação da Figura 5, o histograma e *boxplot* sugerem que os dados não seguem uma distribuição Normal, evidenciando-se também a presença de observações *outliers* acima do 3ºquantil. Verifica-se assim a existência de valores elevados, para o parâmetro, no extremo superior da distribuição.

Pela análise do *boxplot*, conclui-se que este apenas detetava as observações *outliers* no extremo superior da distribuição. Dado que o valor mínimo das observações, Tabela 7, é de 1,710 $pmol/h/punção$, este valor seria considerado como doente e, abordada esta questão com a entidade acolhedora, tornou-se claro que o *boxplot* não estava a identificar as observações *outliers* no extremo inferior da distribuição. Foi constatado previamente que existiam indivíduos com observações muito baixas em relação ao que seria de esperar para

probandos normais e, alguns destes entrariam mesmo na gama de *probandos* diagnosticados com a doença. Explorou-se a causa, e foi reconhecido que apesar da enzima de referência indicar qualidade para a amostra, existiam novos pedidos de sangue total ao médico, para estes pacientes. A nova amostra de sangue origina um resultado que teria por base a realização de uma análise molecular de forma a detetar a mutação, caso existente.

De acordo com esta problemática, a identificação de *outliers* pelo *boxplot* não estava a responder como pretendido. Como a distribuição se mostra assimétrica, uma alternativa para a realização de análise estatística quando se evidencia a não Normalidade dos dados, é realizar transformações que estejam de acordo com a sua natureza, Osborne J.W. (2002).

Como a unidade de medida da variável VP se encontra em $pmol/h/punção$, a transformação mais comum a estes dados é a aplicação da transformação logarítmica. Qualquer número pode ser escrito como x^y, o que a transformação logarítmica permite é, mediante a base escolhida, determinar o y que originaria o valor inicial não transformado. O logaritmo de base e, 2 ou 10 são as transformações mais comuns que se aplicam aos dados. Como afirma Osborne J.W. (2002), quando são observados valores extremos na distribuição dos dados o aconselhado é aplicar o logaritmo de base 10.

Aplicado o logaritmo de base 10 aos dados, realiza-se de novo a análise gráfica do histograma com o esboço da distribuição da variável VP a preto e o esboço da distribuição Normal a vermelho, e *boxplot* da base de dados dos *probandos* normais na Figura 6.

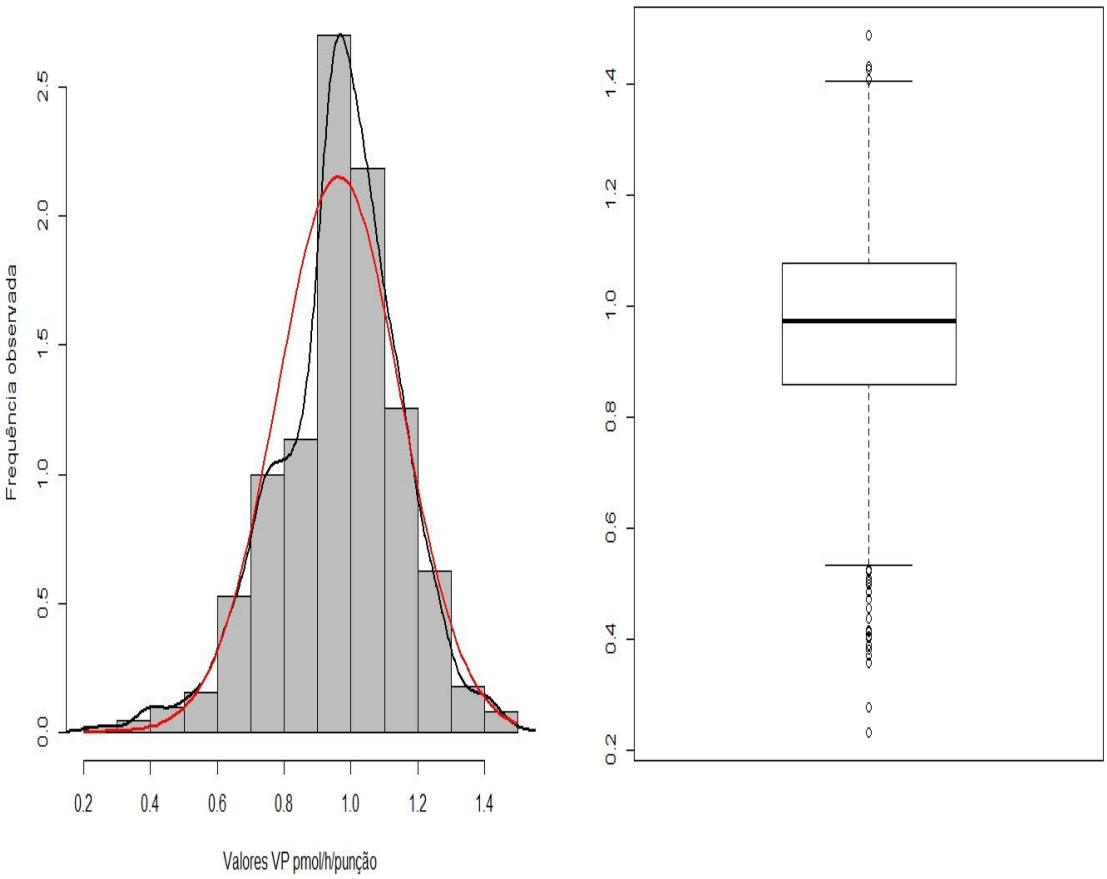


Figura 6: *Histograma e boxplot da base de dados de probandos normais após transformação logarítmica dos dados.*

Observando a Figura 6, pode-se concluir que a distribuição dos valores apresenta um ligeiro enviesamento à esquerda, e o *boxplot* identifica as observações *outliers* em ambos os extremos. Realizada a análise gráfica à variável VP, foram identificadas 40 observações *outliers*, o correspondente a 4% do total de indivíduos da base. Estas foram analisadas de forma a destacar as causas que validassem a sua presença e respetiva exclusão pelo método de Tukey.

Das 40 observações detetadas, os motivos que validaram a sua exclusão da base de dados são: registo de novo pedido de sangue pois o resultado do parâmetro não possibilitava um diagnóstico final conclusivo, amostra não recolhida no próprio laboratório, confirmação de amostras mal colhidas e, valores altos em que se confirma a realização, por parte de alguns indivíduos, de hemodiálise.

Obteve-se como resultado final uma base de dados composta por 875 indivíduos. Na Figura 7 encontra-se o histograma com o esboço da curva da distribuição da variável VP a preto e o esboço da curva da distribuição Normal a vermelho, e *boxplot* da variável VP após eliminação das observações *outliers*.

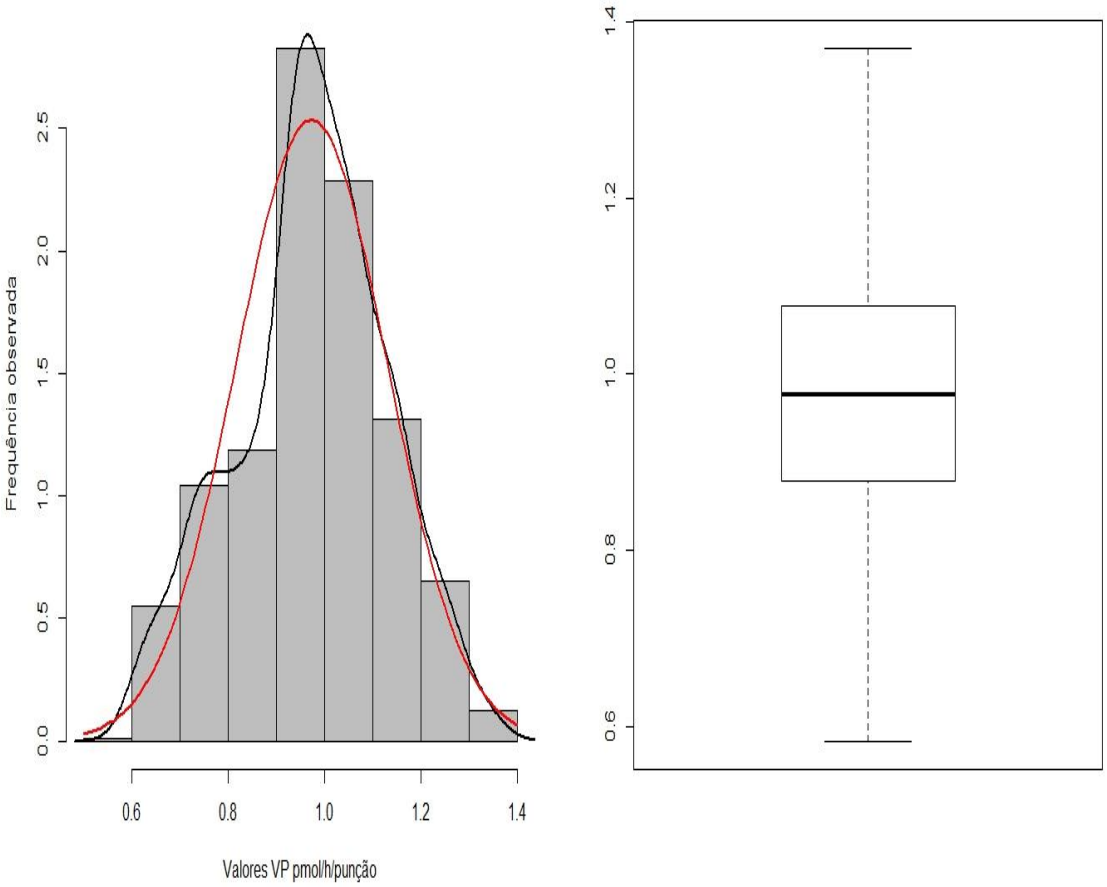


Figura 7: *Histograma e boxplot da base de dados de probandos normais após exclusão de observações outliers.*

Considerando um nível de significância de 5%, o teste de Kolmogorov-Smirnov com correção de Lilliefors obteve um $p - valor = 8{,}704 \times 10^{-8}$, isto é, rejeita-se H_0 a um nível de significância de 5%, podendo-se afirmar que os dados não seguem uma distribuição Normal.

Como descrito anteriormente, é o resultado dos testes de Normalidade aos dados que permitem concluir que método utilizar na estimação do intervalo de referência. Como os dados não seguem uma distribuição Normal e a dimensão da amostra é superior a 120 observações, aplica-se o método não paramétrico.

Quando aplicada a transformação logarítmica de base 10 aos dados e, após a estimação dos extremos do intervalo de referência, é necessário a cada extremo aplicar o inverso da transformação realizada, isto é, o intervalo de referência final é obtido da seguinte forma:

$$(10^{l_i}, 10^{l_s}).$$

A título de comparação, e de forma a estimar o intervalo que melhor descreve o comportamento dos *probandos* normais, foi também aplicado o método de Horn para identificação de observações *outliers*, com o objetivo de se verificar se o resultado é relativamente díspar em relação ao método de Tukey.

O método de Horn identificou como outliers 25 observações. Estas foram devidamente analisadas e conclui-se que compõem um subconjunto das observações detetadas pelo método de Tukey. Eliminadas essas observações, não foram detetadas mais observações discrepantes, resultando uma base de dados final composta por 890 indivíduos. Realizado o teste de Kolmogorov-Smirnov, verifica-se que os dados não seguem uma distribuição Normal e sendo a dimensão da amostra superior a 120 observações, foi aplicado o método não paramétrico na estimação do intervalo.

No Apêndice B encontram-se os comandos do *software* R que permitem obter os valores estimados para os extremos dos intervalos de referência.

4.3. Resultados

Realizadas as etapas necessárias que precedem a aplicação dos métodos descritos no Capítulo 2.4.5, concretizam-se as estimativas dos respetivos intervalos de referência.

Na Tabela 8 encontram-se os respetivos intervalos de referência para os *probandos* doentes pelo método robusto com identificação de *outliers* pelo método de Tukey, com 95% de confiança, e *probandos* normais pelo método não paramétrico com identificação de outliers pelo método de Tukey e de Horn.

Tabela 8: *Estimativas dos intervalos de referência para probandos doentes pelo método robusto com identificação de outliers pelo método de Tukey, com 95% de confiança, e probandos normais pelo método não paramétrico com identificação de outliers pelo método de Tukey e Horn, com 95% de confiança.*

Base de dados	Intervalo de Referência	Método
Probandos Doentes	0,000 – 1,753	Robusto com identificação de *outliers* pelo método de Tukey.
Probandos Normais	4,456 – 18,413	Não Paramétrico com identificação de *outliers* pelo método de Tukey e correspondente análise gráfica do boxplot.
	4,063 – 18,300	Não Paramétrico com identificação de *outliers* pelo método de Horn.

Como referido no Capítulo 4.2.2.2., o intervalo de referência para os *probandos* normais foi estimado pelo mesmo método, mas com diferentes abordagens para identificação dos *outliers* da base de dados. Pelo método de Tukey foram identificadas 40 observações *outliers*, e pelo método de Horn, 25 observações. O intervalo estimado para os *probandos* doentes teve por base a estimação pelo método robusto. Na Tabela 8 estão apresentados os resultados finais.

Em comparação aos intervalos de referência existentes para o parâmetro do laboratório, apenas é possível estabelecer uma comparação com o intervalo de referência para a base de

dados de *probandos* normais. A necessidade de estimação de um intervalo para *probandos* doentes surgiu com a necessidade de obter um intervalo que caracterizasse estes indivíduos. Analisado o intervalo de referência para *probandos* doentes, concluiu-se que este era representativo dos indivíduos testados e diagnosticados com a patologia no laboratório.

Existem indivíduos que possuem valores que se encontram fora dos limites do anterior intervalo para *probandos* normais e não apresentavam, a nível bioquímico e molecular, motivos de doença. Pretende-se saber em que categoria estes indivíduos se enquadraram e se, de facto, as conclusões retiradas pelo laboratório atendiam assertivamente a estes.

O intervalo estimado para *probandos* normais, aplicando o método de eliminação de *outliers* de Horn possui maior amplitude comparativamente ao intervalo estimado pelo método de Tukey. Torna-se necessário analisar se existem diferenças significativas entre ambos. Esta decisão final foi baseada nos *outliers* detetados na amostra.

Na Figura 8 encontra-se representado a cinzento os *probandos* normais e respetivos intervalos estimados pelo método de Tukey (linha preta) e Horn (linha azul), a vermelho os *probandos* doentes e respetivo intervalo estimado (linha vermelha), o antigo intervalo de referência para *probandos* normais (linha verde) e a laranja, os *probandos* normais com estatuto 5 e 29 sem doença.

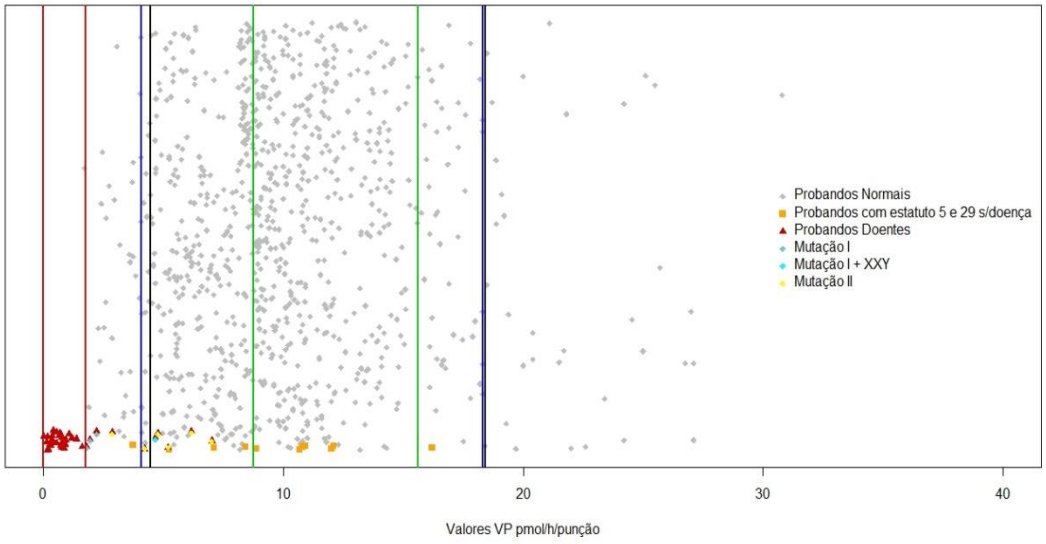


Figura 8: *Diagrama de dispersão dos probandos doentes, normais e com estatuto 5 e 29 sem doença e, representação das mutações detetadas nos probandos doentes.*

Na Figura 9 encontra-se representado a cinzento o histograma referente aos *probandos* normais e respetivos intervalos estimados pelo método de Tukey (linha preta) e Horn (linha azul), a vermelho o histograma referente aos *probandos* doentes e respetivo intervalo estimado (linha vermelha), o antigo intervalo de referência para *probandos* normais (linha verde) e a laranja, os *probandos* normais com estatuto 5 e 29 sem doença.

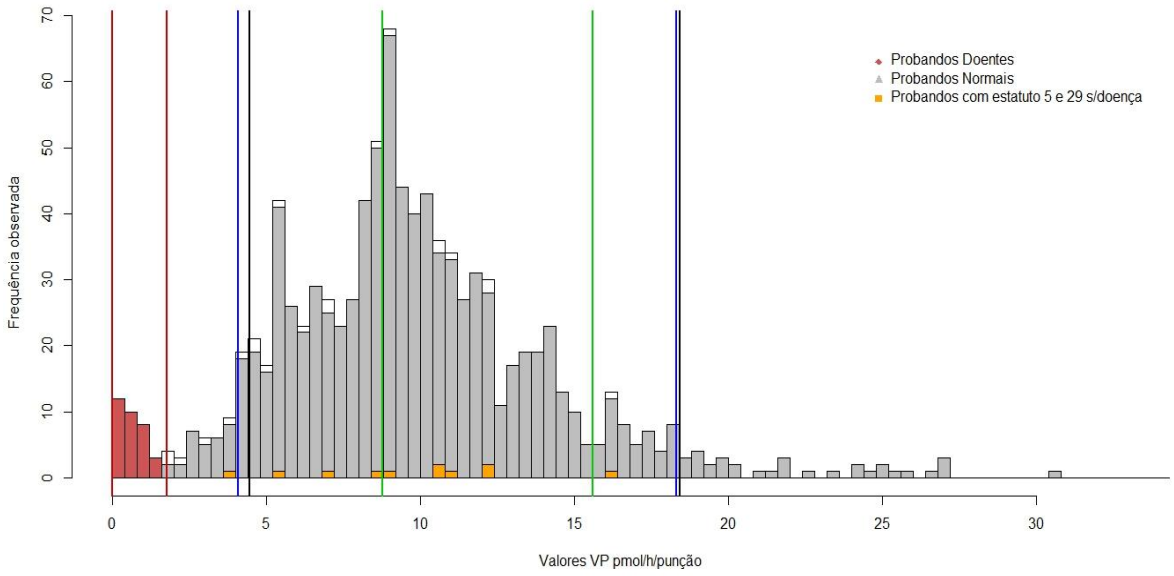


Figura 9: *Histograma dos probandos doentes, normais e com estatuto 5 e 29 sem doença.*

Pela Figura 8 e 9, pode-se concluir que o intervalo anteriormente adotado na entidade não é representativo dos indivíduos testados e não diagnosticados com a patologia, verificando-se a existência de uma grande quantidade de observações que extravasam os limites anteriormente utilizados. Num estudo familiar é necessário excluir a presença de mutação familiar. Ao serem estudados indivíduos de uma família, com doença excluída por via bioquímica e molecular, alguns destes também se apresentavam fora desses limites. Comprova-se assim, que o intervalo anteriormente adotado não refletia os indivíduos testados e diagnosticados como normais no laboratório, reforçando a necessidade de este ser atualizado e, efetivamente ser adotado um novo intervalo de referência para *probandos* normais.

Em análise ao intervalo estimado para *probandos* doentes, este é representativo dos indivíduos testados e diagnosticados com a patologia no laboratório. Na Figura 8, os pontos

que saem do limite superior deste intervalo representam os *probandos* doentes com uma mutação associada que origina valores para o parâmetro superiores aos restantes dessa mesma gama. Por motivos de confidencialidades estas não são descritas. Comparando os intervalos estimados para *probandos* normais, pelo método não paramétrico com identificação de *outliers* pelo método de Tukey e de Horn, conclui-se que não existiam diferenças consideráveis entre ambos, e em análise, foi determinado que o intervalo estimado a adotar para *probandos* normais corresponde ao que tem por base o método de Tukey para identificação de *outliers*. A decisão final pelo intervalo que tem por base o método de Tukey para identificação de *outliers* deteve como motivo o facto de este detetar um maior número de observações que causam variações no resultado final. Como os motivos se revelam bastante válidos e este permite, no limite inferior, reconhecer um maior número de *outliers*, este foi o intervalo de referência escolhido para ser adotado.

5. Considerações Finais

Em análise ao presente trabalho elaborado, podem ser retiradas algumas considerações e estipuladas metas futuras que permitam melhorar determinados resultados obtidos.

Como resultado final do estudo de prevalência realizado, conclui-se que a prevalência de nascimento das doenças lisossomais de sobrecarga, DLSs, em Portugal diminuiu face ao valor anteriormente publicado em Pinto *et al.* (2004). As abordagens terapêuticas aplicadas a indivíduos que sejam diagnosticados com uma patologia deste grupo de doenças lisossomais estão a ser eficazes e pelo acompanhamento continuo que estes indivíduos possuem pode-se afirmar que a sua qualidade de vida tende a melhorar.

Existem variados métodos que permitem obter as estimativas intervalares de proporções binomiais. Estes encontram-se em constante estudo, de forma a melhorar os existentes ou abraçar o desafio na criação de novos métodos com este fim. Atualmente existem processos muito eficazes quando se pretende estudar se estes têm os atributos necessários para que se assuma que perfazem um bom estimador intervalar.

Ainda que constando entre os métodos mais conhecidos, o método mid-p exato é um método menos convencional e existem ainda poucos pacotes que permitam estimar este intervalo, Influential Points (2018). Existiram algumas limitações na obtenção da fórmula e pacotes que o estimassem de acordo com o pretendido.

Futuramente, seria de grande importância a construção de uma biblioteca no *software* R que permitisse a estimação deste intervalo sem as limitações ocorridas.

Deste trabalho resultou a exposição de um poster no 14º Simpósio Internacional da Sociedade Portuguesa de Doenças Metabólicas (SPDM), de 15 a 17 de março de 2018, e o início da redação de um artigo para publicação em revista científica.

O estudo e validação de um intervalo de referência para um determinado parâmetro analítico de um laboratório é uma tarefa onde o estatístico necessita de reagir com visão crítica e apurada sensibilidade em todas as etapas e resultado final.

Os intervalos de referência estimados mostraram ser representativos de cada uma das categorias dos *probandos* do laboratório. Verificou-se também que o anterior intervalo de

referência para *probandos* normais já não representava corretamente esta categoria de indivíduos, validando desta forma a implementação do novo intervalo de referência estimado.

Entre todos os passos realizados para determinação do resultado final, o que mais potencializou dualidade fora a tomada decisão do método de exclusão de *outliers* a aplicar*,* dando origem à questão sobre qual seria o que se adequaria melhor à amostra e parâmetro em causa. A nível computacional, verificou-se a existência de variados *softwares* que atendem a esta temática e foi utilizado além do *software* R, um outro *software* estatístico que originou os mesmos resultados, aplicando o método de identificação de *outliers* de Tukey. (Ver Apêndice B).

Salienta-se que a metodologia, após uma vasta pesquisa, adotou-se corretamente aos dados em questão. Num trabalho futuro, pretende-se realizar outras abordagens que permitam estabelecer comparação com os resultados finais obtidos.

A investigação que relaciona a Estatística e as áreas da saúde está em constante atualização. O laboratório pretende acompanhar essa constante evolução, e adotar a nível interno estudos estatísticos que permitam a consciencialização aprofundada e apropriada da sua população. Após esta colaboração, foi de ambas as partes, demonstrada uma enorme vontade na continuidade de colaborações que permitam que este eleve o seu nome, no propósito deste prestar, cada vez melhor, diagnósticos à população.

Tratando-se de um parâmetro que permite identificar a presença de uma doença ligada ao cromossoma X, ficou acordado com a entidade acolhedora, fora o contexto deste estágio, a realização da estimação de um intervalo de referência para o mesmo parâmetro analítico, sendo a população alvo os *probandos* do sexo feminino do laboratório.

Bibliografia

Alves, E.J. (2013). *Métodos de bootstrap e aplicações em problemas biológicos* (Dissertação de mestrado, Universidade Estatual Paulista "Júlio de Mesquita Filho", Campus de Rio Claro, São Paulo, Brasil).

Associação Portuguesa de Neuromusculares (2015). *DOENÇA RARA?*. Acedido em: 19 de fevereiro de 2018, em: http://apn.pt/apn/doenca-rara/

Agresti, A. & Coull, B.A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. The American Statistician, 52:119-126.

Agresti, A. & Gottard, A. (2005). Comment: Randomized confidence intervals and mid-p approach. Statistical Science, 20:367-371.

AUSVET (2018). EpiTools epidemiological calculators – *Calculate confidence limits for a sample proportion.* Acedido em: 5 de fevereiro de 2018, em: http://epitools.ausvet.com.au/content.php?page=CIProportion&SampleSize=3976892&Positive=2000000&Conf=0.95&method=5&Digits=4

Bilder, C.R. & Loughin, T.M. (2015). *Analysis of Categoricl Data with R.* Chapman and Hall/CRC Press.

Brown, L.D., Cai, T.T., DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, 16:101–133.

Brown, L.D., Cai, T.T., DasGupta, A. (2002). Confidence Intervals for a binomial proportion and asymptotic expansions. *The Annals of Statistics*, 30:160-201.

Centro Hospitalar do Porto (2016). *Apresentação*. Acedido em: 18 de fevereiro de 2018, em: http://www.chporto.pt/cgmjm#apresentacao

Cepeda-Cuervo, E., Aguilar, W., Cervantes, V., Corrales, M., Díaz I.Rodríguez, D. (2008). Intervalos de confianza e intervalos de credibilidade para una proporción. *Revista Colombiana de Estadística*, 31(2):211-228

Cerotti, F. (2007). Prerequesites for use of common reference intervals. *Clinical Biochemistry Reviews,* 28:115-121.

Clinical and Laboratory Standards Institute, CLSI (2000) *How to define and determine reference intervals in the clinical laboratory: approved guideline*. (2ªedição). CLSI Documento C28-A2. Wayne: Clinical and Laboratory Standards Institute.

Clopper, C.J. & Pearson, E.S. (1934, dezembro). The Use of Confidence or Fiducial Limits Illustrated on the Case of the Binomial. *Oxford University Press*, pp. 404-413.

Efron, B. & Tibshirani, R. (1983). An Introduction to the bootstrap. Chapman and Hall, New York.

Elandt-Johnson, R.C. (1975). DEFINITIONS OF RATES: SOME REMARKS ON THEIR USE AND MISUSE. *American Journal of Epidemiology,* 102(4):267-271.

Elmonem, M.A., Mahmoud, I.G., Mehaney, D.A., Sharaf, S.A., Hassan, S.A, Orab,i A., Salem, F., Girgis, M.Y., El-Badawy, A., Abdelwahab, M., Salah, Z., Soliman, N.A., Hassan, F.A. Selim, L.A. (2016) Lysosomal storage disorders in Egyptian Children. *Indian J Pediatr,* 83(8):805-813.

Engineering Handbook Hand (2018). Acedido em: 27 de setembro de 2018, em: https://www.itl.nist.gov/div898/handbook/eda/section3/eda35e.htm

Epidemiology and Beyond (2012). Acedido em: 8 de abril de 2018, em: http://epid.blogspot.pt/2012/08/how-to-calculate-confidence-interval-of.html

Ferreira, C.E.S. & Andriolo, A. (2008). Intervalos de referência no laboratório clínico. *Jornal Brasileiro de Patologia e Medicina Laboratorial*, 1(44):11-16.

Friedrichs, K.R., Harr, K.E., Freeman, K.P., Szladovits, B., Walton, R.M., Bamhart, K.F., Blanco-Chavez, J. (2012). ASVCP reference intervals guidelines: determination of de novo reference intervals in veterinary species and other related topics. *Vet Clin Pathol*, 41(4): 441-453.

Ghasemi, A. & Zahediasl, S. (2012). Normality tests for statistical analysis: a guide for non-statisticians. *Int J Endocrinol Metab* 10(2):486-489.

Grubbs, F.E. (1969). Procedures for Detecting Outlying Observations in Samples. *Technometrics*, 11(1):1-21.

Kleinbaum D.G., Kupper L.L., Morgenstern H. (1982), *Epidemiologic Research: Principles and Quantitative Methods.* Belmont (Calif.): Lifetime learning publications.

He, X. & Wu, S.-J. (2009) Confidence Intervals for the Binomial Proportion with Zero Frequency. Paper SP-10.

Horn, P.S. (1988). A biweight prediction interval for random samples. *Journal of the American Statistical Association*, 83(401):249-256.

Horn, P.S., Pesce A.J., Copeland, B.E. (1998). A robust approach to reference interval estimation and evaluation. *Clinical Chemistry*, 44(3):622-631.

Horn, P.S., Feng, L., Li, Y., Pesce, A.J. (2001). Effect of outliers and nonhealthy individuals on reference interval estimation. *Clinical Chemistry*, 47(12):2137-2145.

Horn, P.S. & Pesce, A.J. (2003). Reference intervals: an update. *Clinical Chimica Acta,* 334(1-2): 5-23.

InfluentialPoints (2018), *Confidence intervals of proportions and rates*. Acedido 20 de fevereiro de 2018, em: http://influentialpoints.com/Training/confidence_intervals_of_proportions-principles-properties-assumptions.htm

Instituto Nacional de Estatística, INE (2017), *Estatísticas Demográficas – 2016*. Acedido em 1 de março de 2018, em: https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_publicacoes&PUBLICACOESpub_boui=277094583&PUBLICACOESmodo=2

Instituto Nacional de Estatística, INE (2018), *Nados vivos (Nº) por local de residência da mãe (NUTS-2013) e sexo; mensal*. Acedido em 1 de março de 2018, em: https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_indicadores&indOcorrCod=0008085&contexto=bd&selTab=tab2

IPTSP (2018), *Estudos de Prevalência*. Acedido em: 13 de março de 2018, em: https://posstrictosensu.iptsp.ufg.br/up/59/o/Modulo1-Estudosdeprevalencia.pdf

Instituto Unidos pela Vida (2010) - *O que é Genética, Hereditária, Autossómica e Recessiva?* Acedido em: 18 de fevereiro de 2018, em: http://unidospelavida.org.br/o-que-e-genetica-hereditaria-autossomica-e-recessiva/

Interative Mathematics (2018). *The Poisson Probability Distribution*. Acedido em: 9 de abril de 2018, em: https://www.intmath.com/counting-probability/13-poisson-probability-distribution.php

Jurecka, A., Lugowska, A., Golda, A., Czartoryska, B., Tylki-Szymañska, A. (2015). Prevalence rates of mucopolysaccharidoses in Poland. J Appl Genetics, 56:205-210.

Katayev, A., Balciza, C., Seccombe, D.W. (2010). Establishing reference intervals for clinical laboratory test results Is there a better way? *American Journal of Clinical Pathology*, 133:180-186.

Khan, S.A., Peracha, H., Ballhausen, D., Wiesbauer, A., Rohrbach, M., Gautschi, M., Mason, R.W., Giugliani, R., Suzuki, Y., Orii, K.E., Orii, T., Tomatsu, S. (2017). Epidemiology of mucopolysaccharidoses. Molecular Genetics and Metabolism, 121:227-240.

Krabbi, K., Joost, K., Zordania, R., Talvik, I., Rein, R., Huijmans, J.G.M., Verheijen, F.V., Õunap, K. (2012) The birth-prevalence of mucopolysaccharidoses in Estonia. Genetic Testin and molecular biomarkers, 16:846-849.

MEDCALC (2018). Acedido em: 30 de setembro de 2018, em: https://www.medcalc.org

Meikle, P.J., Hopwood, J.J., Clague, A.E., Carey, W.F. (1999). Prevalence of lysosomal storage disorders. JAMA, 54:281-249.

Metha, A. & Winchester, B. (2012), *LYSOSOMAL STORAGE DISORDERS A PRATICAL GUIDE* (1°edição). John Wiley & Sons, LDA

Murteira, B. (1990). *Probabilidades e Estatística* (2° edição). Lisboa: McGraw-Hill.

Murteira, B. (1993), *Análise Exploratória de dados – Estatística Descritiva*, Lisboa: McGraw-Hill.

Murteira, B., Silva, C., Andrade e Silva, J., Pimenta, C., Pimenta, F. (2015), *Introdução à Estatística* (3ªedição). Lisboa: Escolar Editora.

National Institute of Neurological Disorders and Stroke, NINDS (2015) - *Mucolipidoses Fact Sheet*. Acedido em: 19 de fevereiro de 2018, em: https://www.ninds.nih.gov/Disorders/Patient-Caregiver-Education/Fact-Sheets/Mucolipidoses-Fact-Sheet

Newcombe, R.G. (1998) Two-sided confidence intervals for the single proportion: Comparison of seven methods. Statistics in Medicine, 17:857-872.

Newcombe, R.G. & Merino, S.C. (2006). Intervalos de confianza para las estimaciones de proporciones y las diferencias entre ellas. *Interdisciplinaria*, *23*(2):141-154.

Osborne, J.W. (2002). Notes on the use of data transformation. *Pratical Assessment Research & Evaluation*, 8(6):1-7.

Osborne, J.W. & Overbay, A. (2004). The power of outliers (and why researchers should always check for them). Practical Assessment, Research and Evaluation, 9(6):1-8.

Pinto, R., Caseiro, C., Lemos, M., Lopes, L., Fontes, A., Ribeiro, H., Pinto, E., Silva, E., Rocha, S., Marcão, A., Ribeiro, I., Lacerda, L., Ribeiro, G., Amaral, O., Miranda, MC.S. (2004). Prevalence of lysosomal storage diseases in Portugal. Eur J Hum Genet, 12:87-92.

Poorthuis, B., Wevers, R.A., Kleijer, W.J., Groener, J.E.M., de Jong, J.G.N, van Weely, S., Niezen-Koning, K.E., Diggelen, O.P. (1999). The frequency of lysosomal storage diseases in The Netherlands. Hum Genet, 6:105-151.

Poupetová, H., Ledvinová, J., Berná, L., Dvorákova, L., Kozich, V., Elleder, M. (2010). The birth prevalence of lysosomal storage disorders in the Czech Republic: comparison with the data in different populations. J Inherit Metab Dis, 33:387-96.

Rothman, K. J. & Boice, J.D. (1979). *Epidemiologic analysis with a programmable calculator.* [Bethesda, Md.]: U.S. Dept. of Health, Education, and Welfare, Public Health Service, National Institutes of Health.

R Project (2018). *About R – What is R?* Acedido em 10 de abril de 2018, em: https://www.r-project.org/about.html

Silvestre, A.L. (2007), *Análise de Dados e Estatística Descritiva*, Lisboa: Escolar Editora.

Solberg, H. E. (1984a). IFCC recommendation: The theory of reference values. Part 2. Selection of individuals for the production of reference values. *The Journal of Automatic Chemistry*, *13*9, 205F-213F.

Solberg, H. E. (1984b). IFCC recommendation: The theory of reference values. Part 5. Statistical treatment of collected reference values. Determination of reference limits. *The Journal of Automatic Chemistry*, *137*, 97F-114F.

Sullivan, K.M. & Soe, M.M. (2006). Documention for Confidence Intervals for a proportion.

Tukey, J.W. (1977). *Exploratory data analysis*. Massachusetts: Addison-Wesley.

Ulm, K. (1990). SIMPLE METHOD TO CALCULATE THE CONFIDENCE INTERVAL OF A STANDARDIZED MORTALITY RATIO (SMR). *American Journal of Epidemiology*, 131(2): 373–375

University of Massachusetts Amherst, UMASS (2007). Statistics - *The Poisson Distribution*. Acedido em: 6 de abril de 2018, em: https://www.umass.edu/wsp/resources/poisson/

Vassarstats (2018). *The Confidence Interval of a Proportion.* Acedido em: 5 de fevereiro de 2018, em: http://vassarstats.net/prop1.html

Vollset, S.E. (1993). Confidence intervals for a binomial proportion. Statistics in Medicine, 12:809-824

Wagner, M.B. (1998). Medindo a ocorrência da doença: prevalência ou incidência? *Jornal de Pediatria*, 74:157-162

Wald, A. & Wolfowitz, J. (1939). Conidence Limits for Continuous Distribution Functions. The Annals Mathematical Statistics, 10:105—118

Wasserman, L. (1991). An inferential interpretation of default priors. Technical report, Dept. Statistics, Carnegie Mellon University.

Wilson, E.B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22:209–212.

Apêndices A

Tabela A 1: *Estimativa das prevalências e respetivo intervalo de confiança de Wilson, com 95% de confiança, do Grupo I.*

	Número de casos						
Patologia	**Pós-natal**	**Pré-natal**	**Prevalência de período**	**Período de nascimento**	**Número de nados vivos**	**Prevalência de nascimento**	**Intervalo**
Patologia I	34	4	0,9520	1955 - 2014	8,910,262	0,4265	0,3107 – 0,5853
Patologia II (♂)	32	3	1,7018[a]	1967 - 2012	3,262,991[a]	1,0726[a]	0,7713 – 1,4917
Patologia III	7	0	0,1754	1985 - 2011	3,039,968	0,2303	0,1115 – 0,4754
Patologia IV	24	1	0,6263	1966 - 2012	6,428,095	0,3889	0,2634 – 0,5742
Patologia V	7	0	0,1754	1988 - 2005	2,049,361	0,3416	0,1655 – 0,7051
Patologia VI	15	2	0,4259	1979 - 2008	3,651,012	0,4656	0,2907 – 0,7457
Patologia VII	2	0	0,0501	1984 - 2001	2,126,893	0,0940	0,0258 – 0,3429
Patologia VIII	22	1	0,5762	1962 - 2007	6,795,719	0,3384	0,2255 – 0,5079
Patologia IX	6	1	0,1754	1988 - 2015	3,000,119	0,2333	0,1130 – 0,4817

[a] inclui apenas nados vivos masculinos

Tabela A 2: *Estimativa das prevalências e respetivo intervalo de confiança de Wilson, com 95% de confiança, do Grupo II.*

	Número de casos						
Patologia	**Pós-natal**	**Pré-natal**	**Prevalência de período**	**Período de nascimento**	**Número de nados vivos**	**Prevalência de nascimento**	**Intervalo**
Patologia X	8	0	0,2004	1973 - 2008	4,710,206	0,1698	0,0861 – 0,3352
Patologia XI	1	0	0,0251	1982 –2017	3,991,824	0,0251	0,0044 – 0,1419
Patologia XII	1	0	0,0251	1982 – 2017	3,991,824	0,0251	0,0044 – 0,1419
Patologia XIII	3	0	0,0752	1982 - 1983	295,298	1,0159	0,3455 – 2,9872
Patologia XIV	4	1	0,1253	1954 - 2001	7,804,572	0,0641	0,0274 – 0,1500
Patologia XV	5	0	0,1253	1994 - 2001	901,686	0,5545	0,2369 – 1,2982

Tabela A 3: *Estimativa das prevalências e respetivo intervalo de confiança de Wilson, com 95% de confiança, do Grupo III.*

	Número de casos						
Patologia	**Pós-natal**	**Pré-natal**	**Prevalência de período**	**Período de nascimento**	**Número de nados vivos**	**Prevalência de nascimento**	**Intervalo**
Patologia XVI	13	2	0,3758	1983 – 2010	3,230,191	0,4644	0,2814 – 0,7662
Patologia XVII	4	0	0,1002	1984 - 1996	1,551,792	0,2578	0,1002 – 0,6628
Patologia XVIII	1	0	0,0251	1982 - 2017	3,991,824	0,0251	0,0044 – 0,1419

Tabela A 4: *Estimativa das prevalências e respetivo intervalo de confiança de Wilson, com 95% de confiança, do Grupo IV.*

	Número de casos						
Patologia	**Pós-natal**	**Pré-natal**	**Prevalência de período**	**Período de nascimento**	**Número de nados vivos**	**Prevalência de nascimento**	**Intervalo**
Patologia XIX	2	0	0,0501	1974 – 1994	2,987,293	0,0670	0,0184 – 0,2441
Patologia XX	8	0	0,2004	1967 - 2003	5,302,354	0,1509	0,0765 – 0,2977
Patologia XXI	19	1	0,5010	1973 - 2006	4,503,120	0,4441	0,2875 – 0,6861
Patologia XXII	1	0	0,0251	1982 – 2017	3,991,824	0,0251	0,0044 – 0,1419
Patologia XXIII	12	1	0,3257	1984 - 2007	2,780,429	0,4676	0,2733 – 0,8000
Patologia XXIV	2	0	0,0501	1999 - 2001	348,784	0,5734	0,1573 – 2,0910
Patologia XV	1	0	0,0251	1982 –2017	3,991,824	0,0251	0,0044 – 0,1419

Tabela A 5: *Estimativa das prevalências e respetivo intervalo de confiança de Wilson, com 95% de confiança, do Grupo V.*

	Número de casos						
Patologia	**Pós-natal**	**Pré-natal**	**Prevalência de período**	**Período de nascimento**	**Número de nados vivos**	**Prevalência de nascimento**	**Intervalo**
Patologia XXVI	2	0	0,0501	1980 - 2003	3,961,057	0,0505	0,0138 – 0,1841
Patologia XXVII	1	0	0,0251	1982 - 2017	3,991,824	0,0251	0,0044 – 0,1419
Patologia XXVIII	56	0	1,4029	1935 – 2016	12,865,119	0,4353	0,3352 – 0,5652
Patologia XXIX (♂)	1	0	0,0486[a]	1982 – 2017	2,056,544[a]	0,0486[a]	0,0086 – 0,2755
Patologia XXX	4	1	0,1253	1975 - 1988	2,126,100	0,2352	0,1005 – 0,5506

[a] inclui apenas nados vivos masculinos

Tabela A 6: Estimativa das prevalências e respetivo intervalo de confiança de Wilson, com 95% de confiança, do Grupo VI.

	Número de casos						
Patologia	Pós-natal	Pré-natal	Prevalência de período	Período de nascimento	Número de nados vivos	Prevalência de nascimento	Intervalo
Patologia XXXI	142	0	3,5573	1917 - 2014	16,262,226	0,8732	0,7409 – 1,0291
Patologia XXXII	4	1	0,1253	1983 - 2000	2,158,415	0,2317	0,0989 – 0,5423
Patologia XXXIII	17	0	0,4259	1948 – 2012	10,146,707	0,1675	0,1046 – 0,2683
Patologia XXXIV	39	4	1,0772	1968 - 2010	5,832,397	0,7373	0,5474 – 0,9930
Patologia XXXV (♂)	125	0	6,0782[a]	1932 - 2011	6,853,369[a]	1,8239[a]	1,5310 – 2,1729
Patologia XXXV (♀)	138	0	7,1308[b]	1917 - 2008	7,742,826[b]	1,7823[b]	1,5087 – 2,1055
Patologia XXXVI	25	2	0,6764	1958 – 2017	8,554,451	0,3156	0,2169 – 0,4592
Patologia XXXVII	14	0	0,3507	1955 - 1982	5,330,470	0,2626	0,1565 – 0,4409
Patologia XVIII	24	3	0,6764	1980 - 2014	4,043,424	0,6678	0,4589 – 0,9716
Patologia XXXIX	1	0	0,0251	1982 - 2017	3,991,824	0,0251	0,0044 – 0,1419
Patologia XL	1	0	0,0251	1982 - 2017	3,991,824	0,0251	0,0044 – 0,1419
Patologia XLI	3	0	0,0752	1985 - 1995	1,298,748	0,2310	0,0786 – 0,6792
Patologia XLII	27	8	0,8768	1985 - 2016	3,467,589	1,0093	0,7258 – 1,4037
Patologia XLIII	1	0	0,0251	1982 - 2017	3,991,824	0,0251	0,0044 – 0,1419
Patologia XLIV	21	0	0,5261	1953 – 2015	9,409,238	0,2232	0,1460 – 0,3412
Patologia XLV	48	1	1,2275	1959 – 2008	7,537,337	0,6501	0,4918 – 0,8594
Patologia XLVI	4	1	0,1253	1985 – 2011	3,039,968	0,1645	0,0703 – 0,3851

[a] inclui apenas nados vivos masculinos

[b] inclui apenas nados vivos femininos

Apêndices B

Tabela B 1: Pacotes e comandos para o software R.

Pacotes	Intervalos	Comandos
binom	Clopper–Pearson, Wald, Agresti–Coull, Wilson, Jeffreys conjugado com uma distribuição Beta	binom.confint(x =__ , n =__ , conf.level = __, methods = "__ ")
DescTools	Wald, Wilson, Agresti-Coull, Jeffreys, Wilson (com correção de continuidade), Clopper-Pearson	BinomCI(x=__ n=__ method="__")
epiR	Clopper–Pearson e Wilson	epi.prev(pos =__, tested = __, se =__, sp =__, method = " __", units =__, conf.level =__)
propCIs	Mid-p	midPci(x =__, n =__, conf.level = __)
stats	Poisson	poisson.test(x, T = __, r = __, alternative = __, conf.level = __)
boot	*Bootstrap* não paramétrico	boot.ci(boot(__), conf = __, type = __)
bootstrap	*Bootstrap* não paramétrico	bootstrap(x=__,n=__,theta=__,func=__)
MASS	*Transormação Box-Cox*	boxcox(x, lambda =__, eps = __)
nortest	*Teste de Anderson-Darling*	ad.test(x)
	Teste de Kolmogorov-Smirnov	ks.test(x, "distribuição a testar", mean=mean(x), sd=sd(x))
nortest	*Kolmogorov-Smirnov (com correção de Lilliefors)*	lillie.test(x)

reference intervals	*Intervalo de referência não paramétrico*	nonparRI(x=__, indices = 1:length(__), refConf = __)
reference intervals	*Intervalo de referência robusto*	robust(x=__, indices = 1:length(__), refConf = __)

Além de *softwares* estatísticos próprios, existem também ferramentas online que possibilitam o cálculo destes valores. Durante a pesquisa de pacotes próprias do *software* R para estimação da prevalência e intervalos de confiança, foram encontrados alguns websites que possibilitam obter os mesmos resultados, apenas se encontram extremamente limitados relativamente aos métodos existentes. As calculadoras online referidas encontram-se em AUSVET (AUSVET, 2018) e Vassarstats (Vassarstats, 2018).

Na técnica *Bootstrap*, em necessidade de aplicação, sugere-se ao leitor que explore todas as funcionalidades dos pacotes referidas na Tabela B1.

No estudo da estimação de intervalos de referência para *probandos* de um determinado parâmetro analítico foi identificado um *software* estatístico que possibilita a estimação e realização de toda a metodologia descrita, existindo apenas a restrição da necessidade de licença. Foi testado, com um período experimental de 15 dias e o resultado final foi coincidente, para ambos os intervalos que se pretendia estimar. O *software* referido é MEDCALC (MEDCALC, 2018).