UNIVERSIDADE DA BEIRA INTERIOR
Ciências

# Modelos Mistos Ortogonais com Amostras de Dimensão Aleatória

**Anacleto César Xavier Mário**

Tese para obtenção do Grau de Doutor em

**Matemática e Aplicações**

(3º ciclo de estudos)

Orientadora: Prof. Doutora Célia Maria Pinto Nunes
Coorientador: Prof. Doutor Dário Jorge da Conçeição Ferreira

**Covilhã, dezembro de 2020**

UNIVERSIDADE DA BEIRA INTERIOR
Ciências

**Modelos Mistos Ortogonais com Amostras de Dimensão Aleatória**

**Anacleto César Xavier Mário**
Orientadora: Prof. Doutora Célia Maria Pinto Nunes
Coorientador: Prof. Doutor Dário Jorge da Conçeição Ferreira

Tese para obtenção do Grau de Doutor em Matemática e Aplicações (3º ciclo de estudos).

Júri:
Presidente: Doutor Paulo Jorge da Silva de Almeida, Professor Catedrático e Presidente da Faculdade de Ciências da Universidade da Beira Interior

Vogais:

- Doutor Paulo Eduardo Aragão Aleixo Neves de Oliveira, Professor Catedrático da Faculdade de Ciências e Tecnologia da Universidade de Coimbra;
- Doutor Carlos Manuel Agra Coelho, Professor Catedrático da Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa;
- Doutora Maria Eduarda da Rocha Pinto Augusto da Silva, Professora Associada da Faculdade de Economia da Universidade do Porto;
- Doutora Teresa Paula Costa Azinheira Oliveira, Professora Associada da Universidade Aberta;
- Doutora Célia Maria Pinto Nunes, Professora Auxiliar da Universidade da Beira Interior.

**Covilhã, 26 de Novembro de 2020**

# Dedicatória

Dedico este trabalho primeiramente a Deus, por iluminar os meus caminhos pela força e coragem para suportar e superar as dificuldades. Aos meus amados pais, César Mário e Valentina Mário, e aos meus avós (in memória) que sempre foram exemplo de dignidade. Aos meus queridos Irmãos, Mano Ricardo, Mariano, Belchior, Cassôva, Fidel, Silvestre, Mayra e Sara, aos meus filhos, vocês são a motivação da realização, desenvolvimento pessoal e satisfação deste trabalho.

# Agradecimentos

Sinto imensa gratidão pelo apoio direto ou indireto de múltiplas pessoas e instituições na realização da presente tese. Não gostaria de correr risco de injustiça de não mencionar alguns dos contributos, expresso os meus agradecimentos:

- Professora Doutora Célia Maria Pinto Nunes, orientadora desta tese, fico sem palavras suficientes para agradecer, pela sua orientação, dedicação, paciência, conselhos e ensinamento durante todo o processo.

- Professor Doutor Dário Jorge da Conceição Ferreira, grato pelos valiosos ensinamentos e apoio para a elaboração desta tese.

- Professor Doutor João Tiago Mexia e Professora Doutora Sandra S. Ferreira, quero agradecer a vossa disponibilidade, incentivo e apoio pela investigação e conclusão desta tese.

- Aos professores do departamento de Matemática, os meus agradecimentos pela ajuda prestada durante a elaboração da presente tese.

- Filipa Raposo quero agradecer toda a colaboração que me foi prestada.

- Não queria deixar de agradecer ao professor doutor Alfredo Maria de Jesus Paulo, professor doutor Francisco Xavier Chitoma, aos Senhores Tomás Muelelivela, Custódio Kupesala e Fernando Mango, pela força e pelo caminho que sempre me prestaram desde a tomada de decisão para estudar em Portugal.

- Os meus agradecimentos são devidos a várias instituições, pelo apoio prestado e de forma particular à Direção da UBI e seus colaboradores.

- Não poderia esquecer as famílias Cassôco e Tavares pelo apoio, simpatia, e paciência que demonstraram desde o primeiro dia que conheci Portugal.

- Os meus agradecimentos também são extensivo aos meus amigos e colegas que de forma direta ou indireta contribuíram para o alcance dos meus objetivos, estando presente ou ausente de Angola, gostaria de destacar de forma carinhosa o Teófilo e Carlitos.

A todos o meu singelo e profundo agradecimento!

# Resumo

A aplicação da análise de variância (ANOVA) a situações em que as dimensões das amostras não são previamente conhecidas, pode ser uma situação bastante comum. Tal ocorre, por exemplo, quando a recolha das observações é realizada num período fixo de tempo ou quando podem ocorrer falhas de observações.

Nestes casos é mais correto considerar as dimensões das amostras como realizações de variáveis aleatórias independentes. São consideradas três distribuições distintas para as dimensões das amostras:

- a distribuição de Poisson, quando a ocorrência das observações corresponde a processos de contagem;

- a distribuição Binomial, caso exista um limite superior para a dimensão das amostras que nem sempre é atingido devido à ocorrência de falhas de observações;

- a distribuição Geométrica [Binomial Negativa], quando o número de observações corresponde ao número de ocorrências até ao primeiro sucesso [$s$-ésimo sucesso].

O objetivo do presente trabalho é estender a teoria dos modelos mistos ortogonais ao caso em que as dimensões das amostras são desconhecidas. A formulação do modelo é feita considerando situações de estabilidade, o que significa que as estatísticas de teste têm a mesma distribuição, quer para a parte de efeitos fixos, quer para a parte de efeitos aleatórios do modelo, quando a hipótese nula se verifica.

A aplicabilidade da abordagem proposta é ilustrada através de estudos referentes ao desemprego, em alguns países da União Europeia, com dados obtidos através da PORDATA - Base de dados de Portugal contemporâneos.

Os resultados obtidos sugerem que a utilização da nossa abordagem, pode evitar a ocorrência de falsas rejeições, o que é confirmado através da realização de alguns estudos com dados simulados.

# Palavras-chave

ANOVA, modelos mistos ortogonais, amostras de dimensão aleatória, situações de estabilidade, desemprego na União Europeia.

# Abstract

Applying analysis of variance (ANOVA) where the samples dimensions are not known in advance is a very common situation. This occurs, for example, when observations are collected within a fixed time period or when some observations failures may occur.

In these cases it is more appropriate to consider the sample sizes as realizations of independent random variables. Three different distributions are considered for the sample sizes:

- the Poisson distribution, when the occurrence of observations corresponds to counting processes;
- the Binomial distribution, when we have an upper bound for the sample sizes, which is not always achieved, since failures may occur;
- the Geometric [Negative Binomial] distribution, for samples constituted by the observations taken until a success [until s successes].

The aim of the present work is to extend the theory of the orthogonal mixed models to situations where the sample sizes are not known in advance. The model formulation is done considering stable statistics, which means that the test statistics have the same distribution whether referring to the fixed or the random effects part of the model, when the tested hypothesis holds.

The applicability of the proposed approach is illustrated through some studies on real data, considering the unemployement in the European Union, obtained from PORDATA - Base de dados de Portugal contemporâneos.

The obtained results suggest that false rejections may be avoided when applying our approach, which is confirmed by carrying out some simulation studies.



# Keywords

ANOVA, orthogonal mixed models, random sample sizes, stability situations, unemployment in the European Union.

# Índice

# Lista de Tabelas

# Lista de Acrónimos

# Capítulo 1

# Introdução

Um modelo linear misto é um modelo estatístico que contém efeitos fixos e efeitos aleatórios. Esses modelos são amplamente utilizados nas mais diversas áreas de pesquisa, nomeadamente em ciências biológicas, sociais e económicas, já que nestas áreas as bases de dados são muitas vezes complexas e confusas, podendo apresentar diferentes fatores, como por exemplo, espécies, locais, género, etc. Os tamanhos das amostras também podem deixar algo a desejar, especialmente se pretendemos ajustar um modelo com muitos parâmetros. Foi para lidar com esses dados ”confusos” que os modelos lineares mistos foram desenvolvidos, permitindo-nos utilizar todos os dados, mesmo quando não temos amostras de dimensão elevada.

Portanto, os modelos lineares mistos surgiram devido à necessidade de se conseguir aceder a quantidades de variação causada por certas fontes em modelos de efeitos fixos, por exemplo, a quantidades de variação que não são controladas pelo investigador ou aquelas cujos níveis são selecionados aleatoriamente de uma grande população de níveis, ver e.g. Khuri et al. (1998).

As variâncias destas fontes de variação são atualmente designadas como componentes de variância e têm sido amplamente investigadas nos últimos anos, veja-se por exemplo, Khuri and Sahai (1985), Searle et al. (1992) e Searle (1995). Entretanto, várias técnicas de estimação para as componentes de variâncias têm sido propostas, veja-se por exemplo Rao and Kleve (1988) e Searle (1995) ou mais recentemente Bailey et al. (2016), Ferreira et al (2013) e Nunes et al. (2008).

Entre estas técnicas, destacamos a Análise de Variância (ANOVA), ver e.g. Searle et al. (1992). A ANOVA foi introduzida por Ronald A. Fisher em 1918 quando estudava problemas na área da agricultura (Scheffé, 1959), e é atualmente um dos métodos estatísticos mais usados em aplicações práticas nas mais diversas áreas da ciência. O objectivo principal da ANOVA é a comparação de mais do que dois grupos no que respeita à localização.

Em muitas das situações práticas, onde se utiliza a ANOVA, pode não ser possível saber previamente as dimensões das amostras. Essas situações ocorrem quando há um período de tempo fixo para a recolha das observações. Um exemplo disso é a recolha de dados de pacientes com várias patologias que chegam às urgências de um hospital durante um período fixo de tempo. O número de pacientes para cada patologia não é conhecido antecipadamente e se resolvermos repetir o estudo durante um período diferente, com o mesmo comprimento, vamos obter certamente amostras com dimensão diferente, ver e.g. Moreira et al. (2013) e Nunes et al. (2014, 2019a, 2019c). Um outro bom exemplo foi apresentado em Nunes et al. (2012a), em que esta abordagem foi aplicada ao caso em que uma das patologias era rara.

Em situações como estas, em que se desconhecem as dimensões das amostras, e assumindo que se têm $m$ diferentes tratamentos, consideramos ser mais correto assumir essas dimensões como realizações, $n_1, \cdots, n_m$, de variáveis aleatórias independentes, $N_1, \cdots, N_m$.

As dimensões das amostras foram abordadas como aleatórias pela primeira vez em Bunge and Nagaraja (1991), Nguyen Bac Van (1988) and Singh (1980). Em Singh (1980), um número aleatório de observações foi usado para estimar a média, enquanto que em Nguyen Bac Van (1988) o número aleatório de observações foi modelado por um processo de Poisson. No trabalho de Bunge e Nagaraja (1991), os autores lidaram com as distribuições de certas estatísticas de registros a partir de um número aleatório de observações. Mais recentemente, este tópico foi abordado em Barsotti et al. (2016), onde foi proposta uma metodologia para testar hipóteses num ambiente Markoviano com observações em número aleatório. Um outro trabalho interessante é o artigo de Esquível et al. (2016), onde alguma inferência estatística clássica foi estendida ao caso em que se tem um número aleatório de observações.

A extensão da ANOVA ao caso em que as dimensões das amostras são consideradas como aleatórias tem vindo a ser aplicada a modelos de efeitos fixos, ver Capistrano (2015), Mexia et al. (2011), Moreira et al. (2013), Nunes et al. (2012a, 2013, 2014, 2015, 2019a). Em Capistrano et al (2015) e Nunes et al. (2012b) este tópico foi igualmente abordado considerando modelos de efeitos aleatórios.

O objetivo principal do presente trabalho é estender a ANOVA ao caso em que as dimensões das amostras não são conhecidas à partida, considerando modelos mistos ortogonais. Os modelos mistos com amostras de dimensão aleatória já foram considerados em Capistrano (2015) e Nunes et al. (2019c), onde a formulação do modelo foi feita através do uso de uma classe mais ampla de modelos, designadas por extensões $L$, ver Ferreira et al. (2009) e Moreira et al. (2009). Nestes dois trabalhos a inferência foi realizada apenas para a parte de efeitos aleatórios do modelo.

Com o presente trabalho pretendemos continuar o estudo da teoria dos modelos mistos, no contexto das amostras de dimensão aleatória, considerando situações de estatilidade, ver Nunes et al (2019b). Estamos perante situações de estatibilidade quando as estatísticas de teste têm a mesma distribuição, quer para a parte de efeitos fixos quer para a parte de efeitos aleatórios do modelo, quando a hipótese testada se verifica, Ferreira (2006).

Esta abordagem deverá basear-se na escolha adequada das distribuições das variáveis aleatórias independentes, $N_1, \ldots, N_m$. Iremos considerar três situações distintas:

- situações em que a ocorrência das observações corresponde a processo de contagem, levando-nos a assumir que as variáveis aleatórias, $N_1, \cdots, N_m$, seguem uma distribuição de Poisson, ver e.g. Mexia et al. (2011) e Nunes et al. (2014, 2019a, 2019c);

- situações em que existe um limite superior para as dimensões das amostras que nem sempre é atingindo uma vez que podem ocorrer falhas de observações, levando-nos a pressupor que as dimensões das amostras, $N_1, \cdots, N_m$, seguem uma distribuição Binomial, ver e.g. Nunes et al. (2015, 2019a, 2019b);

- situações em que o número de observações corresponde ao número de ocorrências até ao primeiro sucesso [$s$-ésimo sucesso], levando-nos a considerar que $N_1, ..., N_m$, seguem uma distribuição Geométrica [Binomial Negativa], ver Mário et al. (2019).

De seguida apresentamos a estrutura do presente trabalho. No Capítulo 2 é feita uma compilação de alguns conceitos e resultados importantes na área da Estatística. Primeiro apresentamos alguns resultados algébricos, onde são abordados conceitos sobre matrizes e álgebras de Jordan comutativas. De seguida apresentamos resultados referentes a algumas distribuições teóricas, discretas e contínuas, que serão utilizadas como distribuições das dimensões das amostras e distribuições das observações. Este capítulo termina com uma introdução aos modelos lineares mistos.

No Capítulo 3, apresentamos a extensão da ANOVA de efeitos fixos, ao caso em que as dimensões das amostras não são conhecidas. Assumimos como distribuição para as dimensões das amostras as três situações anteriormente referidas (distribuição de Poisson, distribuição Binomial e distribuição Geométrica). São obtidas as estatísticas de teste e as suas distribuições condicional (assumindo as dimensões das amostras como fixas) e não condicional (assumindo as dimensões das amostras como aleatórias). São ainda apresentadas duas aplicações com dados reais, referentes ao desemprego na União Europeia, por forma a mostrar a aplicabilidade desta abordagem. Terminamos o capítulo com a apresentação de um estudo prático com dados simulados, por forma a comparar a metodologia proposta com a ANOVA usual.

Quanto ao Capítulo 4, a abordagem proposta é aplicada aos modelos mistos ortogonais. A formulação do modelo é feita considerando situações de estabilidade (Ferreira, 2006). Mais uma vez consideramos as três situações descritas anteriormente para as distribuições das dimensões das amostras e as estatísticas de teste são obtidas assim como as suas distribuições condicional e não condicional. São apresentadas duas aplicações com dados reais e uma com dados simulados, por forma a comparar a metodologia proposta com a clássica. Finalizamos este capítulo com uma generalização ao cruzamento de $u$ fatores.

O presente trabalho termina com a apresentação das principais conclusões, obtidas no decorrer do mesmo, e com a referência a alguns desenvolvimentos que pretendemos realizar no futuro.

# Capítulo 2

# Resultados preliminares

Neste capítulo apresentamos uma compilação de alguns resultados importantes bem conhecidos na área da Estatística. Primeiro são apresentados alguns resultados algébricos onde serão abordados conceitos sobre matrizes inversas generalizadas e inversa de Moore-Penrose, Álgebras de Jordan, matrizes de projeção ortogonal e produto de Kronecker de matrizes. Posteriormente apresentamos resultados referentes a algumas distribuições teóricas que iremos utilizar nos capítulos seguintes. Finalmente é apresentada uma introdução aos modelos lineares mistos, onde se inserem os modelos mistos com estrutura ortogonal por blocos.

Estes conceitos servirão de apoio à obtenção dos resultados nos capítulos seguintes, podendo grande parte deles ser encontrados, por exemplo, em Schott (1997) ou em Graham (1981), Horn and Johnson (1985) e Rao (1973). A maioria das demonstrações não serão incluídas, uma vez que podem ser encontradas na bibliografia indicada.

## 2.1 Conceitos e resultados algébricos

Apresentamos de seguida algumas definições e proposições que, como já foi referido, serão necessárias para o desenvolvimento da metodologia apresentada nos capítulos seguintes.

**Definição 2.1** *(**Matriz simétrica**) Se $\boldsymbol{A}$ é uma matriz simétrica, então $\boldsymbol{A} = \boldsymbol{A}^{'}$, com $\boldsymbol{A}^{'}$ a transposta de $\boldsymbol{A}$.*

**Definição 2.2** *(**Matriz ortogonal**) Uma matriz $n \times n,$ é ortogonal se $\boldsymbol{A}\boldsymbol{A}^{'} = \boldsymbol{A}^{'}\boldsymbol{A} = \boldsymbol{I}_n,$*

*em que $\boldsymbol{I}_n$ representa a matriz identidade de ordem $n.$*

Vejamos algumas propriedades das matrizes ortogonais.

**Proposição 2.1** *Sejam $\boldsymbol{P}$ e $\boldsymbol{Q}$ matrizes ortogonais de ordem $m$ e $\boldsymbol{A}$ uma matriz qualquer também de ordem $m.$ Então, sendo $|.|$ o determinante de uma matriz, tem-se*

- $|\boldsymbol{P}| = \pm 1;$

- $|\boldsymbol{P}'\boldsymbol{A}\boldsymbol{P}| = |\boldsymbol{A}|$;

- $\boldsymbol{P}\boldsymbol{Q}$ *é uma matriz ortogonal.*

**Proposição 2.2** *Se $\boldsymbol{A}$ é uma matriz simétrica com valores próprios $\lambda_1, \cdots, \lambda_n$, e vetores próprios normalizados $\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_n$, tem-se*

$$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}(\lambda_1, \cdots, \lambda_n)\boldsymbol{P}^{'} = \sum_{i=1}^{n} \lambda_i \boldsymbol{\gamma}_i \boldsymbol{\gamma}_i^{'},$$

*onde $\boldsymbol{P}$ é uma matriz cujas colunas são os vetores $\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_n$, e $\boldsymbol{D}(\lambda_1, \cdots, \lambda_n)$ é uma matriz diagonal cujos elementos principais são $\lambda_1, \cdots, \lambda_n$. Ao somatório, $\sum_{i=1}^{n} \lambda_i \boldsymbol{\gamma}_i \boldsymbol{\gamma}_i^{'}$, chama-se decomposição espectral de $\boldsymbol{A}$.*

**Dem:** A matriz $\boldsymbol{P}$ é ortogonal uma vez que $\boldsymbol{P}$ tem como colunas os vetores normalizados $\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_n$. Assim,

$$\boldsymbol{P}\boldsymbol{P}^{'} = \boldsymbol{I}_n \Leftrightarrow \boldsymbol{A}\boldsymbol{P}\boldsymbol{P}' = \boldsymbol{A} \Leftrightarrow \boldsymbol{A}\left[\boldsymbol{\gamma}_1 \cdots \boldsymbol{\gamma}_n\right] \begin{bmatrix} \boldsymbol{\gamma}_1 \\ \vdots \\ \boldsymbol{\gamma}_n \end{bmatrix} = \boldsymbol{A} \Leftrightarrow$$

$$\Leftrightarrow \left[\boldsymbol{A}\boldsymbol{\gamma}_1 \cdots \boldsymbol{A}\boldsymbol{\gamma}_n\right] \begin{bmatrix} \boldsymbol{\gamma}_1 \\ \vdots \\ \boldsymbol{\gamma}_n \end{bmatrix} = \boldsymbol{A} \Leftrightarrow \left[\lambda_1\boldsymbol{\gamma}_1 \cdots \lambda_n\boldsymbol{\gamma}_n\right] \begin{bmatrix} \boldsymbol{\gamma}_1 \\ \vdots \\ \boldsymbol{\gamma}_n \end{bmatrix} = \boldsymbol{A} \Leftrightarrow$$

$$\Leftrightarrow \boldsymbol{P}\boldsymbol{D}\left(\lambda_1 \cdots \lambda_n\right)\boldsymbol{P}' = \boldsymbol{A} \Leftrightarrow \sum_{i=1}^{n} \lambda_i \boldsymbol{\gamma}_i \boldsymbol{\gamma}_i^{'} = \boldsymbol{A}.$$

■

**Corolário 2.2.1** *Se $\boldsymbol{A}$ é uma matriz simétrica, então existe uma matriz ortogonal $\boldsymbol{P}$ tal que*

$$\boldsymbol{P}^{'}\boldsymbol{A}\boldsymbol{P} = \boldsymbol{D}(\lambda_1, \cdots, \lambda_n),$$

*onde $(\lambda_1, \cdots, \lambda_n)$ são os valores próprios de $\boldsymbol{A}$.*

**Dem:** Com $\boldsymbol{P} = [\boldsymbol{v}_1, \cdots, \boldsymbol{v}_n]$, tem-se que

$$\begin{aligned}
\boldsymbol{PAP}' = [\boldsymbol{Av}_1 \cdots \boldsymbol{Av}_n] \begin{bmatrix} \boldsymbol{v}_1' \\ \vdots \\ \boldsymbol{v}_n' \end{bmatrix} &= [\lambda_1 \boldsymbol{v}_1 \cdots \lambda_n \boldsymbol{v}_n] \begin{bmatrix} \boldsymbol{v}_1' \\ \vdots \\ \boldsymbol{v}_n' \end{bmatrix} \\
&= \begin{bmatrix} \lambda_1 \boldsymbol{v}_1' \boldsymbol{v}_1 & \cdots & \lambda_n \boldsymbol{v}_n' \boldsymbol{v}_n \\ \vdots & \ddots & \vdots \\ \lambda_1 \boldsymbol{v}_n' \boldsymbol{v}_1 & \lambda_i \cdots & \lambda_n \boldsymbol{v}_n' \boldsymbol{v}_n \end{bmatrix} = D(\lambda_1, \cdots, \lambda_n),
\end{aligned}$$

■

uma vez que $\boldsymbol{P}$ é ortogonal.

**Definição 2.3** *(**Característica de uma matriz**) A Característica de uma matriz $\boldsymbol{A}$, car(**A**), é igual ao número máximo de linhas (ou colunas) linearmente independentes.*

**Proposição 2.3** *Se $\boldsymbol{A}$ é uma matriz $n \times n$ com valores próprios $\lambda_1, \cdots, \lambda_n$, então:*

$$|\boldsymbol{A}| = \prod_{i=1}^{n} \lambda_i;$$

*e*

$$tr(\boldsymbol{A}) = \sum_{i=1}^{n} \lambda_i,$$

onde $tr(\boldsymbol{A})$ representa o traço da matriz $\boldsymbol{A}$.

### 2.1.1 Matrizes inversas e inversa generalizadas

**Proposição 2.4** *Uma matriz $\boldsymbol{A}$ diz-se regular se existir a sua inversa, $\boldsymbol{A}^{-1}$. Caso contrário a matriz diz-se singular.*

**Definição 2.4** *(**Matriz inversa**) Uma matriz não singular $\boldsymbol{A}$, tem uma única inversa, $\boldsymbol{A}^{-1}$, em que*

$$\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}.$$

**Definição 2.5** *(**Matriz idempotente**) Uma matriz $\boldsymbol{A}$ diz-se idempotente se e só se*

$$\boldsymbol{A}\boldsymbol{A} = \boldsymbol{A}.$$

**Proposição 2.5** *Seja $\boldsymbol{A}$ uma matriz não singular e idempotente, então $\boldsymbol{A}^{-1}\boldsymbol{A}^2 = \boldsymbol{I}$.*

**Dem:** Se $\boldsymbol{A}$ é idempotente e não singular então $\boldsymbol{A}^2 = \boldsymbol{A}\boldsymbol{A} = \boldsymbol{A}$ e a sua inversa $\boldsymbol{A}^{-1}$ existe. Se multiplicarmos $\boldsymbol{A}^2$ por $\boldsymbol{A}^{-1}$, obtemos

$$\boldsymbol{A}^{-1}\boldsymbol{A}^2 = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}.$$

■

**Definição 2.6** ***(Matriz inversa generalizada)*** *A matriz inversa generalizada de uma matriz $\boldsymbol{A}$ do tipo $m \times n$, é uma matriz $\boldsymbol{A}^-$, que satisfaz a igualdade*

$$\boldsymbol{A}\boldsymbol{A}^-\boldsymbol{A} = \boldsymbol{A},$$

com $\boldsymbol{A}^-$ uma matriz de ordem $n \times m$.

Toda a matriz $\boldsymbol{A}$ tem uma inversa generalizada. De seguida iremos focar-nos na inversa generalizada de Moore-Penrose, por ser esta que irá ser utilizada nos desenvolvimento apresentados neste trabalho.

**Definição 2.7** ***(Matriz inversa generalizada de Moore-Penrose)*** *A inversa de Moore-Penrose de uma matriz $\boldsymbol{A}$ do tipo $m \times n$, é uma matriz $n \times m$, representada por $\boldsymbol{A}^+$, que satisfaz as seguintes condições:*

- $\boldsymbol{A}\boldsymbol{A}^+\boldsymbol{A} = \boldsymbol{A}$
- $\boldsymbol{A}^+\boldsymbol{A}\boldsymbol{A}^+ = \boldsymbol{A}^+$
- $(\boldsymbol{A}\boldsymbol{A}^+)' = \boldsymbol{A}\boldsymbol{A}^+$
- $(\boldsymbol{A}^+\boldsymbol{A})' = \boldsymbol{A}^+\boldsymbol{A}$.

A matriz inversa de Moore-Penrose, $\boldsymbol{A}^+$, é obtida pela decomposição dos valores próprios da matriz base, $\boldsymbol{A}$. Para toda a matriz $\boldsymbol{A}$, existe uma e uma só matriz $\boldsymbol{A}^+$.

Sempre que $\boldsymbol{A}$ é regular, então

$$\boldsymbol{A}^+ = \boldsymbol{A}^{-1}.$$

### 2.1.2 Matrizes de projeção ortogonal

Suponhamos que $S$ é um subespaço do espaço vetorial $E$.

**Definição 2.8** ***(Complemento ortogonal)*** *O complemento ortogonal de $S$, representado por $S^\perp$, é o conjunto de todos os vetores de $E$ que são ortogonais a cada vetor de $S$, ou seja,*

$$S^\perp \;=\; \{\boldsymbol{v} : \boldsymbol{v} \in E \text{ e } \boldsymbol{v}'\boldsymbol{s} = \boldsymbol{0},\ \forall \boldsymbol{s} \in S\}. \tag{2.1.1}$$

É óbvio que se $S$ é um subespaço vetorial de $E$ então o seu complemento ortogonal, $S^{\perp}$, também é um subespaço vetorial de $E$. Se $E$ for um espaço vetorial de dimensão $m$ e $S$ um subespaço vetorial de dimensão $r < m$, então o complemento ortogonal, $S^{\perp}$ é um subespaço vetorial de $E$ de dimensão $m - r$.

**Definição 2.9** *(**Base ortogonal**) Uma base $B = \{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_m\}$ do espaço vetorial $E$ diz-se ortogonal se o conjunto dos vetores que a constitui forem ortogonais.*

**Definição 2.10** *(**Base ortonormada**) Uma base $B = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_m\}$ de um espaço vetorial $E$ diz-se ortonormada, se $B$ é uma base ortogonal e todos os seus vetores são unitários ( $||\boldsymbol{v}_i|| = 1$, $i = 1, \ldots, m$, em que $\| \cdot \|$ representa a norma Euclidiana), o que significa que*

$$\boldsymbol{v}_i^{'}\boldsymbol{v}_j = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases} .$$

Suponhamos que os vetores coluna de uma matriz $\boldsymbol{Z}_1$ do tipo $m \times r$ formam uma base ortonormada para o subespaço vetorial $S$ (de dimensão $r$), que é um subespaço vetorial de $E$.

Suponha-se ainda que as colunas da matriz $\boldsymbol{Z}_2$ do tipo $m \times (m - r)$ formam uma base ortonormada para o subespaço vetorial $S^{\perp}$ (de dimensão $m - r$).

A matriz $\boldsymbol{Z}_1\boldsymbol{Z}_1^{'}$ é designada por matriz de projeção ortogonal, $MPO$, sobre o subespaço vetorial $S$. Similarmente, $\boldsymbol{Z}_2\boldsymbol{Z}_2^{'}$ será a $MPO$ sobre o subespaço vetorial $S^{\perp}$.

Assim, é de notar que

$$\boldsymbol{Z}\boldsymbol{Z}^{'} = [\boldsymbol{Z}_1\boldsymbol{Z}_2]\begin{bmatrix} \boldsymbol{Z}_1 \\ \boldsymbol{Z}_2 \end{bmatrix} = \boldsymbol{Z}_1\boldsymbol{Z}_1^{'} + \boldsymbol{Z}_2\boldsymbol{Z}_2^{'} = \boldsymbol{I}_m,$$

corresponde à matriz de projeção ortogonal sobre o espaço vetorial $E$.

Uma vez que

$$\boldsymbol{Z}\boldsymbol{Z}^{'} = \boldsymbol{Z}_1\boldsymbol{Z}_1^{'} + \boldsymbol{Z}_2\boldsymbol{Z}_2^{'}$$

e que $\boldsymbol{Z}\boldsymbol{Z}^{'} = \boldsymbol{I}_m$ vem

$$\boldsymbol{Z}_2\boldsymbol{Z}_2^{'} = \boldsymbol{I}_m - \boldsymbol{Z}_1\boldsymbol{Z}_1^{'}.$$

Ainda que um subespaço vetorial não tenha uma base ortonormada única, a $MPO$ formada a partir desta base ortonormada é única, ver por exemplo Schott (1997).

**Proposição 2.6** *É condição necessária e suficiente para que uma matriz quadrada $\boldsymbol{P}$, de ordem $m$, seja uma matriz de projeção ortogonal, $MPO$, que*

*a ) **P** seja idempotente ( $\boldsymbol{P}^2 = \boldsymbol{P}$);*

*b ) **P** seja simétrica ( $\boldsymbol{P}^{'} = \boldsymbol{P}$).*

**Definição 2.11** *(**Espaço imagem**) O espaço imagem de uma matriz $\boldsymbol{A}$ é dado por*

$$R(\boldsymbol{A}) = \{\boldsymbol{A}\boldsymbol{x} : \boldsymbol{x} \in E\}.$$

**Definição 2.12** *(**Espaço nulo**) O espaço nulo de uma matriz $\boldsymbol{A}$, é o espaço que contém as soluções do sistema de equações lineares homogéneo $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$, isto é,*

$$\mathcal{N}(\boldsymbol{A}) = \{\boldsymbol{x} : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}\}.$$

Sendo $\boldsymbol{P} = \boldsymbol{Z}_1\boldsymbol{Z}^{'}_1$, a $MPO$ sobre $S$, então $R(\boldsymbol{P}) = S$, visto que

$$R(\boldsymbol{P}) = \{\boldsymbol{P}\boldsymbol{x} : \boldsymbol{x} \in E\} = \{\boldsymbol{Z}_1\boldsymbol{Z}^{'}_1\boldsymbol{x} : \boldsymbol{x} \in E\} = S,$$

e $\mathcal{N}(\boldsymbol{P}) = S^{\perp}$, uma vez que

$$\mathcal{N}(\boldsymbol{P}) = \{\boldsymbol{x} : \boldsymbol{P}\boldsymbol{x} = \boldsymbol{0}\} = \{\boldsymbol{x} : \boldsymbol{Z}_1\boldsymbol{Z}^{'}_1\boldsymbol{x} = \boldsymbol{0}\} = S^{\perp}.$$

A soma direta ortogonal de subespaços vetoriais, representado por $\boxplus$, pode também ser representada pelo espaço imagem de matrizes de projeção ortogonal, como se pode ver de seguida.

Apresentam-se agora alguns resultados importantes sobre $MPO$.

**Proposição 2.7** *Sejam as $MPO$, $\boldsymbol{P}_1, \cdots, \boldsymbol{P}_k$, tais que $\boldsymbol{P}_i\boldsymbol{P}_j = \boldsymbol{0}$, com $i \neq j$; $i, j = 1, \cdots, k$. Então*

- $\boldsymbol{P} = \sum_{i=1}^{k} \boldsymbol{P}_i$ *é uma $MPO$,*
- $R(\boldsymbol{P}_i) \cap R(\boldsymbol{P}_j) = 0$, *com $i \neq j$,*
- $R(\boldsymbol{P}) = \boxplus_{i=1}^{k} R(\boldsymbol{P}_i)$.

**Proposição 2.8** *Seja $\boldsymbol{P}$ uma $MPO$ associada a um subespaço vetorial $S$. Suponhamos que $S = \boxplus_{i=1}^{k} S_i$, então existem $MPO$ únicas, $\boldsymbol{P}_1, ..., \boldsymbol{P}_k$ tais que*

$$\boldsymbol{P} = \sum_{i=1}^{k} \boldsymbol{P}_i$$

*e*

$$\boldsymbol{P}_i \boldsymbol{P}_j = \mathbf{0}, \ \ i \neq j.$$

**Proposição 2.9** *Seja $\boldsymbol{A}$ uma matriz do tipo $n \times k$. As matrizes de projeção ortogonal sobre $R(\boldsymbol{A})$ e $R(\boldsymbol{A}^{'})$ são dadas por, respetivamente.*

$$\boldsymbol{A}(\boldsymbol{A}^{'}\boldsymbol{A}^{+})\boldsymbol{A}^{'}$$

*e*

$$(\boldsymbol{A}^{'}\boldsymbol{A})^{+}(\boldsymbol{A}^{'}\boldsymbol{A}).$$

Da proposição 2.6 vem que, se $\boldsymbol{P}$ é $MPO$ então $\boldsymbol{P}^{+} = \boldsymbol{P}$.

**Proposição 2.10** *Se $\boldsymbol{P}$ é uma $MPO$, os seus valores próprios serão iguais a $0$ ou $1$.*

**Definição 2.13** *(**Matrizes de projeção ortogonal mutuamente ortogonais**) Duas matrizes de projeção ortogonal, $\boldsymbol{W}_1$ e $\boldsymbol{W}_2$, são mutuamente ortogonais, $MPOMO$, se*

$$\boldsymbol{W}_2 \boldsymbol{W}_1 = \mathbf{0}.$$

**Proposição 2.11** *Se $\boldsymbol{W}_1$ e $\boldsymbol{W}_2$ são $MPOMO$ então $\boldsymbol{W}_1 + \boldsymbol{W}_2$ é uma $MPO$.*

### 2.1.3 Produto de Kronecker de matrizes

Nesta secção apresentar-se-ão alguns resultados sobre o produto de Kronecker de matrizes que será denotado por $\otimes$.
O produto de Kronecker de matrizes, ao contrário do produto usual de matrizes, está definido para quaisquer tipos de matrizes. Esta operação entre matrizes foi amplamente estudada por exemplo por Graham (1981), Steeb (1991), e Steeb and Hardy (2011).

**Definição 2.14** *(**Produto de Kronocker**) Dada a matriz $\boldsymbol{A} = [a_{i,j}]$ de ordem $r \times s$ e uma matriz $\boldsymbol{B}$ de ordem $p \times q$, o produto de Kronocker entre estas duas matrizes será uma matriz do tipo $rp \times sq$ dada por*

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} a_{1,1}\boldsymbol{B} & \cdots & a_{1,s}\boldsymbol{B} \\ \vdots & \ddots & \vdots \\ a_{r,1}\boldsymbol{B} & \cdots & a_{r,s}\boldsymbol{B} \end{bmatrix}.$$

O produto de Kronecker não verifica a propriedade comutativa mas satisfaz a propriedade associativa.

De seguida são apresentados alguns resultados sobre o produto de Kronocker de matrizes que podem ser encontradas por exemplo em Schott (1997):

- consideremos as matrizes $\boldsymbol{A}$, $\boldsymbol{B}$ e $\boldsymbol{C}$ quaisquer, tem-se que

$$\boldsymbol{A}\otimes(\boldsymbol{B}\otimes\boldsymbol{C})=(\boldsymbol{A}\otimes\boldsymbol{B})\otimes\boldsymbol{C}=\boldsymbol{A}\otimes\boldsymbol{B}\otimes\boldsymbol{C};$$

- sejam $\boldsymbol{A}$ e $\boldsymbol{B}$ matrizes de ordem $r\times s$, e $\boldsymbol{C}$ e $\boldsymbol{D}$ matrizes de ordem $p\times q$, tem-se que

$$(\boldsymbol{A}+\boldsymbol{B})\otimes(\boldsymbol{C}+\boldsymbol{D})=\boldsymbol{A}\otimes\boldsymbol{C}+\boldsymbol{A}\otimes\boldsymbol{D}+\boldsymbol{B}\otimes\boldsymbol{C}+\boldsymbol{B}\otimes\boldsymbol{D};$$

- sejam $\boldsymbol{A}$ e $\boldsymbol{B}$ duas matrizes quaisquer e $\alpha$ um escalar, tem-se que

$$(\alpha\boldsymbol{A})\otimes\boldsymbol{B}=\alpha(\boldsymbol{A}\otimes\boldsymbol{B})=\boldsymbol{A}\otimes(\alpha\boldsymbol{B});$$

- sejam as matrizes $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ e $\boldsymbol{D}$ matrizes do tipo $m\times n$, $r\times s$, $n\times p$ e $s\times t$, respetivamente. Se o produto usual de matrizes está definido, então tem-se

$$(\boldsymbol{A}\otimes\boldsymbol{B})(\boldsymbol{C}\otimes\boldsymbol{D})=\boldsymbol{AC}\otimes\boldsymbol{BD};$$

- sejam $\boldsymbol{A}$ e $\boldsymbol{B}$ duas matrizes quaisquer, tem-se

$$(\boldsymbol{A}\otimes\boldsymbol{B})^{'}=\boldsymbol{A}^{'}\otimes\boldsymbol{B}^{'};$$

- consequentemente, tem-se ainda

$$(\boldsymbol{A}\otimes\boldsymbol{B})^{'}=\boldsymbol{A}\otimes\boldsymbol{B},$$

  com $\boldsymbol{A}$ e $\boldsymbol{B}$ duas matrizes simétricas;

- quaisquer que sejam as matrizes $\boldsymbol{A}$ e $\boldsymbol{B}$,

$$(\boldsymbol{A}\otimes\boldsymbol{B})^{+}=\boldsymbol{A}^{+}\otimes\boldsymbol{B}^{+}$$

- $car(\boldsymbol{A} \otimes \boldsymbol{B}) = car(\boldsymbol{A}) \otimes car(\boldsymbol{B})$.

**Proposição 2.12** *Sejam $\boldsymbol{A}$ e $\boldsymbol{B}$ duas matrizes idempotentes, então $\boldsymbol{A} \otimes \boldsymbol{B}$ é uma matriz idempotente.*

**Proposição 2.13** *Sejam $\boldsymbol{A}$ e $\boldsymbol{B}$ duas matrizes ortogonais, então $\boldsymbol{A} \otimes \boldsymbol{B}$ é uma matriz ortogonal.*

**Proposição 2.14** *Sejam $\boldsymbol{A}$ e $\boldsymbol{B}$ $MPO$, então $\boldsymbol{A} \otimes \boldsymbol{B}$ também será uma $MPO$.*

### 2.1.4 Álgebras de Jordan

A introdução das álgebras de Jordan ocorreu no ano 1934 e foi realizada por Pascual Jordan, ver Jordan et al. (1934). Mais tarde Seely designou estas estruturas por espaço vetoriais quadráticos e foram usados no desenvolvimento da inferência estatística, ver Seely (1970a, 1970b, 1971, 1977), e Seely and Zyskind (1971).

Depois, Michalski e Zmyślony em (1996) e (1999), usaram as álgebras de Jordan primeiro para testar hipóteses sobre componentes de variância e, posteriormente sobre funções lineares dos parâmetros em modelos mistos.

Atualmente as álgebras de Jordan continuam a ser muito utilizadas, veja-se por exemplo Vanleuwen et al. (1998, 1999), Fonseca et al. (2006, 2008, 2009), Jesus et al. (2007, 2009), Carvalho et al. (2015) e Ferreira at al. (2017).

De seguida serão apresentadas algumas definições e resultados importantes sobre álgebras de Jordan.

**Definição 2.15** *(**Álgebra**) uma álgebra $\mathcal{A}$ é um espaço linear provido de uma operação binária, $\circledast$, geralmente chamado produto, que satisfaz as seguintes condições: $\forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathcal{A}$ e $\alpha \in \mathbb{R}$*

*a)* $\boldsymbol{x} \circledast (\boldsymbol{y} + \boldsymbol{z}) = \boldsymbol{x} \circledast \boldsymbol{y} + \boldsymbol{x} \circledast \boldsymbol{z}$;

*b)* $(\boldsymbol{x} + \boldsymbol{y}) \circledast \boldsymbol{z} = \boldsymbol{x} \circledast \boldsymbol{z} + \boldsymbol{y} \circledast \boldsymbol{z}$;

*c)* $\alpha(\boldsymbol{x} \circledast \boldsymbol{y}) = (\alpha \boldsymbol{x}) \circledast \boldsymbol{y} = \boldsymbol{x} \circledast (\alpha \boldsymbol{y})$.

**Definição 2.16** *(**Álgebra comutativa**) Uma álgebra $\mathcal{A}$ é comutativa se e somente se, $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{A}$, se tem $\boldsymbol{x} \circledast \boldsymbol{y} = \boldsymbol{y} \circledast \boldsymbol{x}$.*

**Definição 2.17** *(**Álgebra associativa**) Uma álgebra $\mathcal{A}$ é associativa se e somente se, $\forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathcal{A}$, se tem $(\boldsymbol{x} \circledast \boldsymbol{y}) \circledast \boldsymbol{z} = \boldsymbol{x} \circledast (\boldsymbol{y} \circledast \boldsymbol{z})$.*

#### 2.1.4.1 Álgebra de Jordan comutativas

Começamos com a seguinte proposição apresentada em Schott (1997).

**Proposição 2.15** *Se $\boldsymbol{A}_1, \cdots, \boldsymbol{A}_m$ são matrizes simétricas, então existe uma matriz ortogonal $\boldsymbol{P}$ tal que $\boldsymbol{P}^{'}\boldsymbol{A}_i\boldsymbol{P} = \boldsymbol{D}_i$ é uma matriz diagonal, cujos valores próprios de $\boldsymbol{A}_i$ correspondem aos seus elementos principais, se e somente se $\boldsymbol{A}_i\boldsymbol{A}_j = \boldsymbol{A}_j\boldsymbol{A}_i$, $\forall(i,j);\ i,j = 1, \cdots, m.$*

Assim, $\boldsymbol{A}_1, \cdots, \boldsymbol{A}_m$ pertencem à família $\mathcal{D}(\boldsymbol{P})$ de matrizes diagonalizadas por $\boldsymbol{P}$, que é uma álgebra de Jordan comutativa, $AJC$.

Então podemos concluir que as matrizes de $M = \{\boldsymbol{M}_1, \cdots, \boldsymbol{M}_w\}$ pertencem a uma $AJC$ se e somente se elas comutam, tendo-se a seguinte definição.

**Definição 2.18** *(**Álgebra de Jordan comutativa**) Álgebras de Jordan comutativas, $AJC$, são espaços lineares constituídos por matrizes simétricas que comutam e contêm os quadrados das suas matrizes.*

Cada $AJC$, $\mathcal{A}$, tem uma única base, designada como base principal, $bp(\mathcal{A})$, cujos elementos são $MPOMO$, ver Seely (1971). )

Se $\boldsymbol{M}$ e $\boldsymbol{W}$ são matrizes pertencentes a uma $AJC$ com base principal $Q = \{\boldsymbol{Q}_1, ..., \boldsymbol{Q}_n\}$, então

$$\begin{cases} \boldsymbol{M} = \sum_{j=1}^{m} m_j\boldsymbol{Q}_j \in \mathcal{A} \\ \boldsymbol{W} = \sum_{j=1}^{w} w_j\boldsymbol{Q}_j \in \mathcal{A} \end{cases},$$

com $m \neq w$ e $m, w \neq 0$. Assim,

$$\begin{aligned} \boldsymbol{MW} = \left(\sum_{j=1}^{m} m_j\boldsymbol{Q}_j\right)\left(\sum_{j=1}^{w} w_j\boldsymbol{Q}_j\right) &= \sum_{j=1}^{m}\sum_{j=1}^{w} m_j w_j \boldsymbol{Q}_j\boldsymbol{Q}_j \\ &= \sum_{j=1}^{m} m_j w_j \boldsymbol{Q}_j \in \mathcal{A}, \end{aligned}$$

verificando-se assim que qualquer $AJC$ contém o produto das suas matrizes.

Se $\boldsymbol{Q}$ é uma $MPO$ pertencente à $AJC$ com base principal $Q = \{\boldsymbol{Q}_1, \cdots, \boldsymbol{Q}_m\}$ , então

$$\boldsymbol{Q} = \sum_{j=1}^{m} b_j\boldsymbol{Q}_j.$$

Sendo $\boldsymbol{Q}$ idempotente e $\boldsymbol{Q}_1, \cdots, \boldsymbol{Q}_m$ idempotentes e mutuamente ortogonais, então

$$\sum_{j=1}^{m} b_j \boldsymbol{Q}_j = \boldsymbol{Q} = \boldsymbol{Q}^2 = \sum_{j=1}^{m} b_j^2 \boldsymbol{Q}_j,$$

o que significa $b_j^2 = b_j$, e assim $b_j = 0$ ou $b_j = 1, j = 1, \cdots, m$. Então

$$\boldsymbol{Q} = \sum_{j \in \mathcal{G}} \boldsymbol{Q}_j,$$

com $\mathcal{G} = \{j : b_j \neq 0\}$, o que significa que uma $MPO$ de uma $AJC$ é uma soma de matrizes da sua base principal. Não obstante

$$R(\boldsymbol{Q}) = \boxplus_{j \in \mathcal{G}} R(\boldsymbol{Q}_j),$$

onde $R(\boldsymbol{Q})$ é o espaço imagem de $\boldsymbol{Q}$. Logo, também se tem

$$car(\boldsymbol{Q}) = \sum_{j \in \mathcal{G}} car(\boldsymbol{Q}_j).$$

O conjunto $\mathcal{G}$ poder ter apenas um elemento. Quando $\mathcal{G} = \{s\}$, teremos

$$\sum_{j \in \mathcal{G}} \boldsymbol{Q} = \boldsymbol{Q}_s.$$

Segue-se a seguinte proposição.

**Proposição 2.16** *Se $\boldsymbol{Q}$ é uma $MPO$ com caraterística 1, pertencente à $AJC$ $\mathcal{A}$, então $\boldsymbol{Q}$ pertence à base principal de $\mathcal{A}$.*

Seguem-se algumas definições importantes.

**Definição 2.19** *($AJC$ **regular**) Se $\boldsymbol{Q} = \frac{1}{n}\boldsymbol{J}_n \in \mathcal{A}$, com $\boldsymbol{J}_n = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'$, então $\mathcal{A}$ é uma $AJC$ regular.*

**Definição 2.20** *($AJC$ **completa**) Quando a $AJC$ $\mathcal{A}$ contém matrizes invertíveis, esta diz-se completa.*

Consideremos que as matrizes de $\mathcal{A}$ são do tipo $n \times n$. Então se $\mathcal{A}$ é uma $AJC$ completa tem-se

$$\sum_{j=1}^{w} \boldsymbol{Q}_j = \boldsymbol{I}_n,$$

uma vez que se tem $\sum_{j=1}^{w} car(\boldsymbol{Q}_j) = n$.

Com

$\boldsymbol{M} = \sum_{j=1}^{m} b_j \boldsymbol{Q}_j \in \mathcal{A}$, em que $b_j \neq 0, j = 1, \cdots, m$, correspondem aos valores próprios de $\boldsymbol{M}$ com multiplicidades $c_j, j = 1, \cdots, m$, então

$$|\boldsymbol{M}| = \prod_{j=1}^{m} b_j^{c_j}$$

$$\boldsymbol{M}^{-1} = \sum_{j=1}^{m} b_j^{-1} \boldsymbol{Q}_j.$$

**Proposição 2.17** *Sejam $\boldsymbol{\mathcal{A}}_1$ e $\boldsymbol{\mathcal{A}}_2$ duas $AJC$.*
*Se $\mathcal{A}_1 \subset \mathcal{A}_2$ e $\mathcal{A}_1$ é completa, então $\mathcal{A}_2$ também é completa.*

## 2.2 Modelos teóricos

Nesta secção abordar-se-ão as distribuições teóricas que utilizaremos nos capítulos seguintes. Como distribuições discretas iremos apresentar as principais propriedades da distribuição Binomial, Poisson, Geométrica e Binomial Negativa. No que respeita às distribuições contínuas faremos uma pequena abordagem às distribuições Qui-quadrado, $F$ e $\overline{F}$. As demonstrações destes resultados podem ser encontradas na mais diversificada bibliografia, veja-se por exemplo, Pestana e Velosa (2006) e Meyer (1970).

### 2.2.1 Modelos Discretos

#### Distribuição Binomial

Uma variável aleatória $X$ com distribuição Binomial representa o número de sucessos obtidos na realização de $n$ provas independentes de Bernoulli.

Seja $X$ uma variável aleatória Binomial com parâmetros $n$ e $p$, em que $n$ é o número de provas e $p$ a probabilidade de sucesso, colocamos $X \sim B(n,p)$ . A sua função de probabilidade será dada por, ver por exemplo Meyer (1970),

$$pr(X = x) = \begin{pmatrix} n \\ x \end{pmatrix} p^x (1-p)^{n-x}, \; x = 0, \cdots, n. \quad (2.2.2)$$

A função geradora de momentos de $X$, $\varphi_X(t)$, é dada por

$$\varphi_X(t) = E\left(e^{tX}\right) = \sum_{x=0}^{n} e^{tx} \begin{pmatrix} n \\ x \end{pmatrix} p^x (1-p)^{n-x} = (pe^t + 1 - p)^n.$$

Desta função facilmente se deduz

$$E(X) = \frac{d}{dt}\varphi_X(0) = n(p+1-p)^{n-1}p = np,$$

e

$$\begin{aligned} Var(X) &= E(X^2) - E^2(X) \\ &= \frac{d^2}{dt}\varphi_X(0) - \left(\frac{d}{dt}\varphi_X(0)\right)^2 \\ &= n(n-1)p^2 + pn - (np)^2 \\ &= np(1-p). \end{aligned}$$

A função geradora de probabilidade é dada por

$$\chi_X(t) = E(t^X) = \varphi_X(\ln t) = (pt + 1 - p)^n.$$

De seguida apresentamos um resultado importante sobre a soma de variáveis aleatórias independentes com distribuição Binomial.

**Proposição 2.18** *Considerando as variáveis aleatórias independentes* $X_i \sim B(n_i, p), i = 1, \cdots, m$ *então*

$$S = \sum_{i=1}^{m} X_i \sim B(\sum_{i=1}^{m} n_i, p).$$

### Distribuição de Poisson

Uma variável aleatória $X$ com distribuição de Poisson representa o número de sucessos num determinado intervalo de tempo.

Uma variável aleatória $X$ com distribuição de Poisson com parâmetro $\lambda$, $\lambda > 0$, tem como função de probabilidade

$$pr(X = x) \quad = \quad \frac{e^{-\lambda}\lambda^x}{x!}, \ x = 0, 1, \cdots . \tag{2.2.3}$$

Colocamos $X \sim \ P(\lambda)$.

A função geradora de momentos de $X$, é dada por

$$\begin{aligned} \varphi_X(t) &= E[e^{tX}] = \sum_{x=0}^{\infty} \frac{e^{tx}e^{-\lambda}\lambda^x}{x!} \\ &= e^{-\lambda}\sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!} = e^{\lambda(e^t-1)}. \end{aligned}$$

Através de $\varphi_X(t)$ facilmente se obtém

$$E(X) = \frac{d}{dt}\varphi_X(0) = \lambda$$

e

$$Var(X) = E(X^2) - (E(X))^2 = \frac{d^2}{dt}\varphi_X(0) - \left(\frac{d}{dt}\varphi_X(0)\right)^2 = \lambda(\lambda+1) - \lambda^2 = \lambda.$$

A função geradora de probabilidades é dada por

$$\chi_X(t) = E(t^X) \quad = \quad \varphi_X(\ln t) = e^{\lambda(t-1)}.$$

Apresentamos de seguida dois resultados importantes referentes à distribuição de Poisson.

**Proposição 2.19** *Considerando as variáveis aleatórias independentes* $X_i \sim P(\lambda_i), i = 1, \cdots, m$*, então*

$$S = \sum_{i=1}^{m} X_i \sim P\left(\sum_{i=1}^{m} \lambda_i\right).$$

**Proposição 2.20** *Seja* $X \sim B(n,p)$. *Se* $n \to \infty$ *e* $p \to 0$ *então*

$$\lim_{n\to\infty, p\to 0} pr(X=x) = \frac{e^{-\lambda}(\lambda)^x}{x!},$$

com $\lambda = np$.

**Dem:** Considerando $X \sim B(n,p)$ e $np = \lambda$, tem-se

$$\begin{aligned} pr(X=x) &= \binom{n}{x} p^x (1-p)^{n-x} \\ &= \frac{n!}{x!(n-x)!} p^x \frac{n^x}{n^x} \left(1 - \frac{np}{n}\right)^{n-x} \\ &= \frac{n!}{x!(n-x)!} \frac{(np)^x}{n^x} \left(1 - \frac{np}{n}\right)^{n-x} \\ &= \frac{n(n-1)\cdots(n-x+1)}{x!} \frac{\lambda^x}{n^x} \left(1 - \frac{\lambda}{n}\right)^{n-x} \\ &= \frac{1}{x!}\left(1-\frac{1}{n}\right)\ldots\left(1-\frac{x-1}{n}\right)\lambda^x\left(1-\frac{\lambda}{n}\right)^{n-x}. \end{aligned}$$

Quando $n \to \infty$ e $p \to 0$, enquanto $\lambda = np$ permanece fixo, e uma vez que $\lim_{n\to\infty}(1 - \frac{1}{n})\cdots(1 - \frac{x-1}{n}) = 1$, tem-se

$$pr(X=x) = \frac{1}{x!}\lambda^x(1-\frac{\lambda}{n})^{n-x} \to \frac{\lambda^x}{x!}e^{-\lambda}.$$

$\blacksquare$

## Distribuição Geométrica e Binomial Negativa

Nesta secção vamos apresentar as principais propriedades da distribuição geométrica e Binomial negativa.

### Distribuição Geométrica

Uma variável aleatória $X$ com distribuição geométrica representa o número de provas a realizar até se obter o primeiro sucesso. Seja a variável $X$ com distribuição Geométrica de parâmetro $q$ (probabilidade de ocorrência de um sucesso), colocamos $X \sim Geo(q)$.

A sua função de probabilidade é dada por

$$pr(X=x) = (1-q)^{x-1}q, \ x = 1, 2, \cdots.$$

A função geradora de momentos de $X$ é

$$\varphi_X(t) = E(t^X) = \frac{qe^t}{1-e^t(1-q)}.$$

Pode deduzir-se o valor esperado e variância, obtendo-se

$$E(X) = \frac{1}{q}$$

e

$$Var(X) = \frac{1-q}{q^2}.$$

**Distribuição Binomial Negativa**

A variável aleatória $X$ com distribuição Binomial Negativa representa o número de provas a realizar até se obterem $k$ sucessos. Seja $X$ uma variável com distribuição Binomial Negativa com parâmetros $k$ e $q$ (probabilidade de ocorrência de um sucesso). Colocamos $X \sim BN(k, q)$.
A sua função de probabilidade é dada por

$$pr(X = x) = \begin{pmatrix} x-1 \\ k-1 \end{pmatrix} q^k (1-q)^{x-k}, x = k, k+1, \cdots$$

A função geradora de momentos desta variável é

$$\varphi_X(t) = \left( \frac{qe^t}{1 - e^t(1-q)} \right)^k,$$

podendo facilmente deduzir-se o valor esperado e a variância, obtendo-se, respetivamente

$$E(X) = \frac{k}{q}$$

e

$$Var(X) = \frac{k(1-q)}{q^2}.$$

É fácil verificar que a distribuição Geométrica é um caso particular da distribuição Binomial Negativa quando $k = 1$.

Seguem-se alguns resultados importantes referentes a esta distribuição.

**Proposição 2.21** *Consideremos as variáveis aleatórias independentes* $X_i \sim Geo(q)$, $i = 1, \cdots, m$, *então*

$$S = \sum_{i=1}^{m} X_i \sim BN(m, q).$$

**Proposição 2.22** *Seja* $X \sim BN(k, q)$. *Se* $k \to \infty$ *e* $(1-q) \to 0$, *tem-se*

$$\lim_{k \to \infty, \, (1-q) \to 0} pr(X = x) = \frac{e^{-\lambda} \lambda^x}{x!},$$

*com* $\lambda = (1-q)k$.

**Dem:** Com $X \sim BN(k,q)$, tem-se

$$pr(X=x) = \left( \begin{array}{c} x+k-1 \\ k-1 \end{array} \right) q^k(1-q)^x, x=0,1,\cdots$$

vindo, com $\lambda = (1-q)k \Leftrightarrow q = 1-\frac{\lambda}{k}$ e visto que $\Gamma(k)=(k-1)!$ e $\Gamma(k+1)=k\Gamma(k)$,

$$\begin{aligned} pr(X=x) &= \frac{(x+k-1)!}{(k-1)!x!} q^k(1-q)^x \\ &= \frac{\Gamma(x+k)}{x!\Gamma(k)} \left(1-\frac{\lambda}{k}\right)^k (\frac{\lambda}{k})^x \\ &= \frac{\lambda^x}{x!}\frac{\Gamma(x+k)}{k^x\Gamma(k)} \left(1-\frac{\lambda}{k}\right)^k \\ &= \frac{\lambda^x}{x!} \left(1-\frac{\lambda}{k}\right)^k. \end{aligned}$$

Quando $k \to \infty$ e $(1-q) \to 0$, enquanto $\lambda = (1-q)k$ permanece fixo, tem-se

$$pr(X=x) = \frac{\lambda^x}{x!}(1-\frac{\lambda}{k})^k \to \frac{\lambda^x}{x!}e^{-\lambda}.$$

■

### 2.2.2 Modelos Contínuos

#### Distribuição normal

Seja $X$ uma variável aleatória que segue uma distribuição normal com parâmetros $\mu$ e $\sigma$, colocamos $X \sim \mathcal{N}(\mu,\sigma)$. A sua função densidade de probabilidade é dada por

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}(\frac{x-\mu}{\sigma})^2}, \ -\infty < x < +\infty, -\infty < \mu < +\infty, \sigma > 0.$$

Temos o caso particular da variável $Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$, com valor médio nulo e variância igual a 1, cuja função densidade é dada por

$$f_z(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}, \ -\infty < x < +\infty.$$

Seguem-se alguns resultados importantes referentes à distribuição normal.

**Proposição 2.23** *Consideremos as variáveis aleatórias independentes* $X_i \sim \mathcal{N}(\mu_i,\sigma_i),\ i = 1,2,\cdots,m$*, então*

$$S = \sum_{i=1}^{m} X_i \sim \mathcal{N}\left(\sum_{i=1}^{m} \mu_i ; \sqrt{\sum_{i=1}^{m} \sigma_i^2}\right).$$

**Corolário 2.23.1** *Sejam as variáveis aleatórias independentes* $X_i \sim \mathcal{N}(\mu, \sigma)$*,* $i = 1, ..., m$*, então*

$$S = \sum_{i=1}^{m} X_i \sim \mathcal{N}\left(m\mu, \sigma\sqrt{m}\right).$$

### Distribuição Qui-quadrado

Consideremos as variáveis aleatórias independentes

$$Z_i \sim \mathcal{N}(0,1),\ i = 1, \cdots, m.$$

A variável

$$X = \sum_{i=1}^{m} Z_i^2,$$

segue uma distribuição qui-quadrado central com $m$ graus de liberdade. Consideramos

$$X \sim \chi_m^2.$$

A sua função densidade é dada por

$$f(x) = \begin{cases} \frac{1}{2\Gamma(\frac{m}{2})}(\frac{x}{2})^{\frac{m}{2}-1} e^{\frac{-x}{2}}, & x > 0 \\ 0, & o.v. \end{cases},$$

sendo $\Gamma(m) = \int_0^\infty x^{m-1} e^{-x} dx,\ m > 0.$

A função geradora de momentos é dada por

$$\varphi_X(t) \quad = \quad E(e^{tX}) = \frac{1}{(1-2t)^{\frac{m}{2}}},\ t < \frac{1}{2},$$

e consequentemente obtém-se

$$E(X) = \frac{d}{dt}\varphi_X(0) = m$$

e

$$Var(X) = \frac{d^2}{dt^2}\varphi_X(0) - \left(\frac{d}{dt}\varphi_X(0)\right)^2 = m^2 + 2m - m^2 = 2m.\frac{k(1-p)}{q^2}.$$

Segue-se um resultado importante referente à distribuição qui-quadrado.

**Proposição 2.24** *Consideremos as variáveis aleatórias independentes* $X_i \sim \chi^2_{k_i},\ i = 1, \cdots, m$, *então*

$$S = \sum_{i=1}^{m} X_i \sim \chi^2_{\sum_{i=1}^{m} k_i}.$$

**Qui-quadrados não centrais**

Nesta subsecção consideramos os qui-quadrados não centrais, ver por exemplo Mexia (1995) e Nunes (2005).

Consideremos as variáveis aleatória independentes $X_i \sim N(\mu_i, 1), i = 1, \cdots, m$, então

$$X = \sum_{i=1}^{m} X_i^2$$

seguirá uma distribuição qui-quadrados não central com $m$ graus de liberdade e parâmetro de não centralidade $\delta = \sum_{i=1}^{m} \mu_i^2$. Colocamos $X \sim \chi^2_{m,\delta}$.

Neste caso tem-se

$$E(X) = m + \delta$$

e

$$Var(X) = 2m + 4\delta.$$

Seguem-se alguns resultados importantes.

**Proposição 2.25** *Considerando as variáveis aleatórias independentes* $X_1 \sim \chi^2_{m_1,\delta_1}$ *e* $X_2 \sim \chi^2_{m_2,\delta_2}$, *então*

$$S = X_1 + X_2 \sim \chi^2_{(m_1+m_2),(\delta_1+\delta_2)}.$$

Seja $g(x|m, \delta)$ a função densidade de um qui-quadrado não central com $m$ graus de liberdade e parâmetro de não centralidade $\delta, \chi^2_{m,\delta}$. Esta densidade é uma mistura, ver Robbins (1948) e Robbins and Pitman (1949), tendo-se

$$g(x|m, \delta) = e^{\frac{-\delta}{2}} \sum_{j=0}^{+\infty} \frac{(\frac{\delta}{2})^j}{j!} g(x|m + 2j),$$

onde $g(x|m + 2j)$ corresponde à função densidade de um $\chi^2_{m+2j}$. Podemos portanto concluir que a densidade de um qui-quadrado não central é uma contribuição linear de densidades de qui-quadrados centrais.

## Distribuições $F$ e $\overline{F}$

Vamos agora considerar as distribuições $F$ e $\overline{F}$, ver por exemplo Mexia (1995) e Nunes (2005).

Dados os qui-quadrados independentes $\chi^2_m$ e $\chi^2_n$,

$$\Im_{m,n} = \frac{n\chi^2_m}{m\chi^2_n},$$

seguirá uma distribuição $F$ central com $m$ e $n$ graus de liberdade. Colocamos $\Im_{m,n} \sim F(z|m,n)$.
A densidade de $\Im_{m,n}$ é definida por

$$f(z|m,n) = \frac{\Gamma\frac{(m+n)}{2}}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \frac{\frac{m}{n}(\frac{m}{n}z)^{\frac{m}{2}-1}}{(1+\frac{m}{n}z)^{\frac{m+n}{2}}}, z > 0, \tag{2.2.4}$$

e tem-se

$$E(\Im_{m,n}) = \frac{n}{n-2}, n > 2$$

e

$$Var(\Im_{m,n}) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}, n > 4.$$

A densidade da variável aleatória

$$\mathcal{T}_{m,n} = \frac{\chi^2_m}{\chi^2_n},$$

é facilmente obtida através de (2.2.4), vindo

$$\bar{f}(z|m,n) \quad = \quad \frac{\Gamma\frac{(m+n)}{2} z^{\frac{m}{2}-1}}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})(1+z)^{\frac{m+n}{2}}}, z > 0. \tag{2.2.5}$$

Consideramos que $\mathcal{T}_{m,n}$ segue uma distribuição $\overline{F}$ central com $m$ e $n$ graus de liberdade. Colocamos $\mathcal{T}_{m,n} \sim \overline{F}(z|m,n)$.

Ter-se-á ainda

$$E(\mathcal{T}_{m,n}) \quad = \quad E(\frac{m}{n}\Im_{m,n}) = \frac{m}{n}E(\Im_{m,n}) = \frac{m}{n} \cdot \frac{n}{n-2} = \frac{m}{n-2},\ n > 2,$$

### Distribuições $F$ e $\overline{F}$ não centrais

Consideremos agora

$$\mathcal{T}_{m,n,\delta} = \frac{\chi^2_{m,\delta}}{\chi^2_n}.$$

$\mathcal{T}_{m,n,\delta}$ seguirá uma distribuição $\overline{F}$ não central com $m$ e $n$ graus de liberdade e parâmetro de não centralidade $\delta$. Consideramos $\mathcal{T}_{m,n,\delta} \sim \overline{F}(z|m,n,\delta)$.

Como vimos anteriormente, a densidade de $\chi^2_{m,\delta}$ é uma mistura. Assim, e uma vez que

$$\mathcal{T}_{m,n,\delta} = \frac{\chi^2_{m,\delta}}{\chi^2_n}$$

corresponde ao quociente de dois qui-quadrados independentes, a densidade de $\mathcal{T}_{m,n,\delta}$ será também uma mistura com os mesmos coeficientes, ver Robbins (1948) e Robbins e Pitman (1949). Tem-se então

$$\bar{f}(z|m,n,\delta) \quad = \quad e^{\frac{-\delta}{2}} \sum_{i=0}^{+\infty} \frac{(\frac{\delta}{2})^i}{i!} \bar{f}(z|m+2i,n). \tag{2.2.6}$$

Portanto, a função distribuição de $\mathcal{T}_{m,n,\delta}$ será dada por, ver e.g. Nunes (2005),

$$\begin{aligned} \overline{F}(z|m,n,\delta) &= pr(\mathcal{T}_{m,n,\delta} \le z) \\ &= e^{\frac{-\delta}{2}} \sum_{i=0}^{+\infty} \frac{\frac{\delta}{2}^i}{i!} \overline{F}(z|m+2i,n). \end{aligned} \tag{2.2.7}$$

Derivando (2.2.7) em ordem a $\delta$, obtém-se, com $j = i-1$,

$$\begin{aligned} \frac{\partial \overline{F}(z|m,n,\delta)}{\partial \delta} &= -\frac{1}{2}\overline{F}(z|m,n,\delta) + e^{\frac{-\delta}{2}} \sum_{i=1}^{\infty} \frac{(\frac{\delta}{2})^{i-1}}{(i-1)!} \overline{F}(z|m+2i,n) \\ &= -\frac{1}{2}\overline{F}(z|m,n,\delta) + e^{\frac{-\delta}{2}} \sum_{j=0}^{\infty} \frac{(\frac{\delta}{2})^{j}}{j!} \overline{F}(z|m+2+2j,n) \\ &= \frac{\overline{F}(z|m+2,n,\delta) - \overline{F}(z|m,n,\delta)}{2}, \end{aligned}$$

e uma vez que $\overline{F}(z|m+2,n,\delta)$ < $\overline{F}(z|m,n,\delta)$, ver novamente Nunes (2005), tem-se

$$\frac{\partial \overline{F}(z|m,n,\delta)}{\partial \delta} < 0,$$

propriedade que será utilizada no Capítulo 4 do presente trabalho.

Por outro lado

$$\Im_{m,n,\delta} = \frac{n}{m}\frac{\chi^2_{m,\delta}}{\chi^2_n},$$

seguirá uma distribuição $F$ não central com $m$ e $n$ graus de liberdade e parâmetro de não centralidade $\delta$, $F(z|m,n,\delta)$. Facilmente se verifica que

$$F(z|m,n,\delta) = \overline{F}\left(\frac{m}{n}z|m,n,\delta\right).$$

A densidade de $\Im_{m,n,\delta}$ virá

$$f(z|m,n,\delta) = \left(\frac{m}{n}\right)\overline{f}\left(\frac{m}{n}z|m,n,\delta\right).$$

Nos próximos capítulos iremos trabalhar quer com a distribuição $F$ quer com a distribuição $\overline{F}$).

### 2.2.3 Propriedades de monotonia das distribuições $F$ e $\overline{F}$

Nesta subsecção apresentar-se-á uma propriedade de monotonia das distribuições $F$ e $\overline{F}$ centrais, que nos será útil nas aplicações que serão apresentadas no Capítulo 3 deste trabalho. Esta propriedade foi inicialmente apresentada em Mexia et al. (2011). Apesar do desenvolvimento ser feito utilizando a distribuição $\overline{F}$, a propriedade é igualmente válida para a distribuição $F$.

Vimos que, se $\chi^2_m$ e $\chi^2_n$ são independentes,

$$\mathcal{T}_{m,n} = \frac{\chi^2_m}{\chi^2_n} \sim \overline{F}(z|m,n).$$

Sendo $f_{1-\alpha,m,n}$ o $(1-\alpha)$-ésimo quantil desta distribuição, com $n < n^o$. Tem-se

$$f_{1-\alpha,m,n} > f_{1-\alpha,m,n^o}, \tag{2.2.8}$$

e consequentemente

$$\overline{F}(z|m,n) < \overline{F}(z|m,n^o). \tag{2.2.9}$$

Vamos em seguida verificar a veracidade destas desigualdades. De (2.2.5) e uma vez que

$$\int_0^\infty \frac{z^{\frac{m}{2}-1}}{(1+z)^{\frac{m+n}{2}}}dz = \frac{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}{\Gamma(\frac{m+n}{2})} \tag{2.2.10}$$

vem que

$$\overline{F}(z|m,n) = \frac{\int_0^z \frac{z^{\frac{m}{2}-1}}{(1+z)^{\frac{m+n}{2}}} dz}{\int_0^{+\infty} \frac{z^{\frac{m}{2}-1}}{(1+z)^{\frac{m+n}{2}}} dz}.$$

Assim, poder-se-á tratar $m$ e $n$ como variáveis reais e mostrar que $\frac{\partial \overline{F}(z|m,n)}{\partial n} > 0$, para $z > 0$, o que estabelece as desigualdades (2.2.8) e (2.2.9).

Começamos por obter

$$\begin{aligned} \frac{\partial}{\partial n} \int_0^{+\infty} \frac{z^{\frac{m}{2}-1}}{(1+z)^{\frac{m+n}{2}}} dz &= \frac{\partial}{\partial n} \int_0^{+\infty} \frac{z^{\frac{m}{2}-1}}{(1+z)^{\frac{m}{2}}} e^{-\frac{\ln(1+z)n}{2}} dz \\ = & -\frac{1}{2} \int_0^{+\infty} \frac{z^{\frac{m}{2}-1}}{(1+z)^{\frac{m+n}{2}}} \ln(1+z) dz. \end{aligned}$$

Então, de (2.2.10), tem-se

$$\frac{\partial}{\partial n} \left( \int_0^{+\infty} \frac{z^{\frac{m}{2}-1}}{(1+z)^{\frac{m+n}{2}}} dz \right)^{-1} = \frac{1}{2} \frac{\Gamma^2\left(\frac{m+n}{2}\right)}{\Gamma^2\left(\frac{m}{2}\right)\Gamma^2\left(\frac{n}{2}\right)} \int_0^{+\infty} \frac{z^{\frac{m}{2}-1}}{(1+z)^{\frac{m+n}{2}}} \ln(1+z) dz.$$

Analogamente obtém-se

$$\frac{\partial}{\partial n} \left( \int_0^{z} \frac{z^{\frac{m}{2}-1}}{(1+z)^{\frac{m+n}{2}}} dz \right) = -\frac{1}{2} \int_0^{z} \frac{z^{\frac{m}{2}-1}}{(1+z)^{\frac{m+n}{2}}} \ln(1+z) dz$$

vindo

$$\begin{aligned} \frac{\partial \overline{F}(z|m,n)}{\partial n} &= \frac{\partial}{\partial n} \left( \int_0^{z} \frac{z^{\frac{m}{2}-1}}{(1+z)^{\frac{m+n}{2}}} dz \times \left( \int_0^{+\infty} \frac{z^{\frac{m}{2}-1}}{(1+z)^{\frac{m+n}{2}}} dz \right)^{-1} \right) \\ = & \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \left( -\frac{1}{2} \int_0^{z} \frac{z^{\frac{m}{2}-1}}{(1+z)^{\frac{m+n}{2}}} \ln(1+z) dz \right) \\ + & \frac{1}{2} \frac{\Gamma^2\left(\frac{m+n}{2}\right)}{\Gamma^2\left(\frac{m}{2}\right)\Gamma^2\left(\frac{n}{2}\right)} \int_0^{+\infty} \frac{z^{\frac{n}{2}-1}}{(1+z)^{\frac{m+n}{2}}} \ln(1+z) dz \times \int_0^{z} \frac{z^{\frac{m}{2}-1}}{(1+z)^{\frac{m+n}{2}}} dz \\ = & \frac{1}{2} \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \left( \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \int_0^{+\infty} \frac{z^{\frac{m}{2}-1}}{(1+z)^{\frac{m+n}{2}}} \ln(1+z) dz \right. \\ \times & \left. \int_0^{z} \frac{z^{\frac{m}{2}-1}}{(1+z)^{\frac{m+n}{2}}} dz - \int_0^{z} \frac{z^{\frac{m}{2}-1}}{(1+z)^{\frac{m+n}{2}}} \ln(1+z) dz \right). \end{aligned}$$

Tem-se então que

$$\frac{\partial \overline{F}(z|m,n)}{\partial n} \underset{n \mapsto +\infty}{\longrightarrow} 0$$

e

$$\frac{\partial \overline{F}(z|m,n)}{\partial n} \underset{n \mapsto 0}{\longrightarrow} 0,$$

bastando provar que $\forall n$, $\frac{\partial \overline{F}(z|m,n)}{\partial n}$, como função de $z$ tem apenas um máximo local para $z > 0$, por forma a mostrar que $\frac{\partial \overline{F}(z|m,n)}{\partial n} > 0$. Tal demonstração pode ser consultada em Mexia et al. (2011).

Desta forma estabelecem-se as desigualdades (2.2.8) e (2.2.9) e conclui-se que

$$\overline{F}(z|m,n) < \overline{F}(z|m,n^o),$$

sempre que $n < n^o$. Fica então estabelecida a propriedade de monotonia dos quantis da distribuição $\overline{F}$ central.

## 2.3 Modelos Lineares Mistos

Consideram-se como modelos lineares os modelos que apresentam uma relação entre variáveis que seja linear nos seus parâmetros.

Um modelo linear que apresenta somente fatores de efeitos fixos, além do erro, $\varepsilon$, que é sempre aleatório, é denominado modelo de efeitos fixos. Os modelos que apresentam apenas fatores de efeitos aleatórios, excepto a média geral, $\mu$, que é sempre fixa, é denominado modelo de efeitos aleatórios. Um modelo misto é aquele que apresenta tanto fatores de efeitos fixos como fatores de efeitos aleatórios, além do erro experimental e da constante.

Estes modelos foram amplamente estudados, por exemplo, em Khuri et al. (1998) e Muller and Stewart (2006).

Nesta secção faremos uma abordagem aos modelos mistos, apresentando a sua estrutura. São considerados ainda os modelos mistos com estrutura ortogonal por blocos ($OBS$).

O modelo misto, considerando a notação matricial, pode ser escrito da seguinte forma

$$\boldsymbol{Y} = \boldsymbol{\mu} + \sum_{h=1}^{k} \boldsymbol{X}_h \boldsymbol{\beta}_h + \sum_{h=k+1}^{w} \boldsymbol{X}_h \boldsymbol{\beta}_h + \boldsymbol{\varepsilon}, \qquad (2.3.11)$$

onde

- $\boldsymbol{\mu} = \mathbf{1}_n\mu$, corresponde ao vetor das médias, sendo $\mu$ a média geral, um parâmetro fixo e desconhecido;
- $\boldsymbol{X}_h, h = 1, \cdots, w$, correspondem às matrizes de delineamento, conhecidas;
- $\boldsymbol{\beta}_h, h = 1, \cdots, k$, são vetores de efeitos fixos, $k < w$;
- $\boldsymbol{\beta}_h, h = k+1, \cdots, w$, são vetores de efeitos aleatórios e independentes;
- $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n)'$, corresponde ao vetor do erro aleatório.

É comum em alguma da bibliografia (ver por exemplo, Carvalho et al. 2015, Ferreira et al. 2017 e Khuri et al. 1998, Capítulo 2) o modelo (2.3.11) aparecer escrito da seguinte forma

$$\boldsymbol{Y} = \sum_{i=0}^{w+1} \boldsymbol{X}_i\boldsymbol{\beta}_i. \tag{2.3.12}$$

Neste caso, $\boldsymbol{\beta}_0 = \mu$, $\boldsymbol{X}_0 = \mathbf{1}_n$, $\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_k$ são fixos e $\boldsymbol{\beta}_{k+1}, \cdots, \boldsymbol{\beta}_{w+1}$ são vetores aleatórios independentes com vetores médios nulos e matrizes de covariância $\sigma^2_{k+1}\boldsymbol{I}_{c_{k+1}}, \cdots, \sigma^2_{w+1}\boldsymbol{I}_{c_{w+1}}$, com $c_i$ o número de componentes de $\boldsymbol{\beta}_i$, $i = k+1, \cdots, w+1$. Tem-se ainda, $\boldsymbol{X}_{w+1}\boldsymbol{\beta}_{w+1} = \boldsymbol{\varepsilon}$, com $\boldsymbol{X}_{w+1} = \boldsymbol{I}_n$ e ${\sigma^2}_{w+1} = \sigma^2$.

Ao considerarmos $\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_w$, fixos, $\boldsymbol{X}_0\boldsymbol{\beta}_0 = \boldsymbol{\mu}$, e $\boldsymbol{X}_{w+1}\boldsymbol{\beta}_{w+1} = \boldsymbol{\varepsilon}$, teremos

$$\boldsymbol{Y} = \boldsymbol{\mu} + \sum_{i=1}^{w} \boldsymbol{X}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}, \tag{2.3.13}$$

que corresponde a um modelo de **efeitos fixos**.

Considerando $\boldsymbol{X}_0\boldsymbol{\beta}_0 = \boldsymbol{\mu},\ \boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_w$, vetores aleatórios independentes e $\boldsymbol{X}_{w+1}\boldsymbol{\beta}_{w+1} = \boldsymbol{\varepsilon}$, teremos

$$\boldsymbol{Y} = \boldsymbol{\mu} + \sum_{i=1}^{w} \boldsymbol{X}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}, \tag{2.3.14}$$

que corresponde a um modelo de **efeitos aleatórios**.

Podemos portanto concluir que os modelos de efeitos fixos e os modelos de efeitos aleatórios podem ser tratados como casos particulares do modelo misto.

### 2.3.1 Modelos com estrutura ortogonal por blocos ($OBS$)

A $MPO$ sobre o espaço imagem de uma matriz $\boldsymbol{X}$, $\Omega = R(\boldsymbol{X})$, será

$$Q(\Omega) = Q(\boldsymbol{X}) = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{+}\boldsymbol{X}' = \boldsymbol{X}\boldsymbol{X}^{+},$$

cuja demonstração pode ser consultada, por exemplo, em Mexia (1995).

Então tem-se

$$\boldsymbol{X}^{+} = (\boldsymbol{X}'\boldsymbol{X})^{+}\boldsymbol{X}'.$$

Consideremos o modelo definido por (2.3.12), em que agora apenas $\boldsymbol{\beta}_0$ é fixo. Assim, $\boldsymbol{\beta}_1 \cdots, \boldsymbol{\beta}_{w+1}$, serão vetores aleatórios independentes com vetores médios nulos e matrizes de covariância $\sigma_i^2 \boldsymbol{I}_{c_i}, i = 1, \cdots, w+1,$ com $c_i$ o número de componentes de $\boldsymbol{\beta}_i, i = 1, \cdots, w+1.$

As matrizes de delineamento, $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_{w+1}$, são conhecidas, tendo-se

$$R([\boldsymbol{X}_1 \cdots \boldsymbol{X}_{w+1}]) = \mathbb{R}^n.$$

Portanto, o vetor médio de $\boldsymbol{Y}$ será dado por:

$$E(\boldsymbol{Y}) = \boldsymbol{X}_0 \boldsymbol{\beta}_0$$

e a matriz de covariância por

$$\Sigma = \sum_{i=1}^{w+1} \sigma^2 \boldsymbol{X}_i \boldsymbol{X'}_i = \sum_{i=1}^{w+1} \sigma_i^2 \boldsymbol{M}_i,$$

em que $\boldsymbol{M}_i = \boldsymbol{X}_i \boldsymbol{X}_i', i = 1, \cdots, w+1.$

O espaço gerado por $E(\boldsymbol{Y})$ será $R(\boldsymbol{X}_0)$ e a $MPO$ sobre $R(\boldsymbol{X}_0)$ será dada por

$$\boldsymbol{T} = \boldsymbol{X}_0(\boldsymbol{X}_0'\boldsymbol{X}_0)^{+}\boldsymbol{X}_0' = \boldsymbol{X}_0\boldsymbol{X}_0^{+}.$$

Segundo Houtman and Speed (1983) e Nelder (1965a, 1965b) tem-se a seguinte definição de modelo com estrutura ortogonal por blocos ($OBS$).

**Definição 2.21** *O modelo tem $OBS$ quando a sua matriz de covariâncias, $\Sigma$, pode ser escrita da seguinte forma*

$$\Sigma = \sum_{j=1}^{u} \gamma_j \boldsymbol{Q}_j,$$

*com* $\gamma_j = \sum_{i=1}^{w+1} b_{i,j}\sigma_i^2$, $j = 1 \cdots, u$, *e em que* $\boldsymbol{Q}_1, \cdots, \boldsymbol{Q}_u$ *são* $MPOMO$, *conhecidas, e tais que*

$$\sum_{j=1}^{u} \boldsymbol{Q}_j = \boldsymbol{I}_n.$$

Estes modelos desempenham um papel central na teoria dos delineamentos por blocos casualizados, ver e.g. Calinski and Kageyama (2000, 2003) e foram introduzidos por Nelder em 1965, ver Nelder (1965a, 1965b).

A demonstração da proposição que se segue pode ser consultada, por exemplo, em Carvalho et al. (2015).

**Proposição 2.26** *O modelo* $\boldsymbol{Y} = \sum_{i=0}^{w+1} \boldsymbol{X}_i\boldsymbol{\beta}_i$ *tem* $OBS$ *se e somente se as matrizes* $\boldsymbol{M}_1, \ldots, \boldsymbol{M}_{w+1}$ *comutam.*

Para terminar esta secção apresentamos de seguida a definição de uma classe especial de modelos com $OBS$. Tratam-se dos modelos com estrutura ortogonal comutativa por blocos, $COBS$. Estes foram introduzidos por Fonseca et al. (2008), mas têm vindo a ser considerados em muitos outros trabalhos, veja-se por exemplo, Carvalho et al. (2008), Carvalho et al. (2015), Ferreira et al. (2013), Nunes et al. (2008), Santos et al. (2007) e Santos (2012).

**Definição 2.22** *Um modelo tem* $COBS$*, se for um modelo com* $OBS$ *e se*

$$\boldsymbol{T}\boldsymbol{Q}_j = \boldsymbol{Q}_j\boldsymbol{T},\ j = 1, \cdots, u,$$

sendo $\boldsymbol{T}$ a $MPO$ sobre $R(\boldsymbol{X}_0)$.

Estabelecemos por fim a seguinte proposição, cuja demonstração pode ser vista mais uma vez em Carvalho et al. (2015).

**Proposição 2.27** *Um modelo tem* $COBS$ *se e somente se as matrizes* $\boldsymbol{M}_1, \cdots, \boldsymbol{M}_{w+1}$, *e* $\boldsymbol{T}$ *comutam.*

# Capítulo 3

# Modelos de efeitos fixos com amostras de dimensão aleatória

## 3.1 Introdução

Neste capítulo iremos considerar situações em que as dimensões das amostras não são conhecida à partida. Esta metodologia será aplicada a modelos de efeitos fixos. Muitos dos resultados apresentados foram obtidos em Capistrano (2015) e surgem neste trabalho com o intuito de introduzir esta nova abordagem.

Capistrano et al. (2015), Mexia et al. (2011), Moreira et al. (2013), Nunes et al. (2012a, 2013, 2014, 2019a, 2019b, 2019c), sugerem que sempre que se desconhecem as dimensões das amostras estas devem ser consideradas como realizações, $n_1, \cdots, n_m$, de variáveis aleatórias independentes, $N_1, \cdots, N_m$. Da mesma forma $n = \sum_{i=1}^{m} n_i$ é considerado como uma realização da variável aleatória $N = \sum_{i=1}^{m} N_i$.

Neste trabalho iremos seguir esta abordagem, considerando três situações distintas:

- situações em que a ocorrência das observações corresponde a processos de contagem, levando-nos a assumir que as variáveis aleatórias, $N_1, \cdots, N_m$, seguem uma distribuição de Poisson, ver e.g. Mexia et al. (2011) e Nunes et al. (2014, 2019a, 2019c);

- situações em que existe um limite superior para as dimensões das amostras que nem sempre é atingido uma vez que podem ocorrer falhas de observações, levando-nos a pressupor que as dimensões das amostras, $N_1, \cdots, N_m$, seguem uma distribuição Binomial, ver e.g. Nunes et al. (2015, 2019a, 2019b);

- situações em que o número de observações corresponde ao número de ocorrências até ao primeiro sucesso [$s$-ésimo sucesso], levando-nos a considerar que $N_i, i = 1, \cdots, m$, seguem uma distribuição Geométrica [Binomial Negativa]. Este caso foi considerado pela primeira vez nesta tese e em Mário et al. (2019).

Iremos considerar que

$$\Omega = \boxplus_{j=1}^{w} \overline{\bigtriangledown}_j,$$

corresponde ao espaço gerado pelo vetor das médias dos tratamentos, $\boldsymbol{\mu}$, com componentes $\mu_1, \cdots, \mu_m$, onde $\boxplus$ representa a soma direta ortogonal de subespaços, ver e.g. Scheffé (1959).

Estamos interessados em testar as hipóteses

$$H_{0,j} : \boldsymbol{\mu} \in \triangledown_j, \ j = 1, \cdots, w, \tag{3.1.1}$$

onde $\triangledown_j = (\overline{\triangledown}_j^{\perp} \cap \Omega), \ j = 1, \cdots, w$, é um subespaço do espaço paramétrico $\Omega$, em que $\overline{\triangledown}_j^{\perp}$ corresponde ao complemento ortogonal de $\overline{\triangledown}_j, \ j = 1, \cdots, w$. Estas hipóteses correspondem a hipóteses de ausência dos efeitos principais dos fatores e interações entre os fatores.

Assumindo que os vetores linha de $\mathbf{A}_j$ constituem uma base ortonormada para $\triangledown_j$, $j = 1, \cdots, w$, podemos reescrever estas hipóteses como (ver e.g. Mexia et al. 2011 e Nunes et al. 2014)

$$H_{0,j} : \mathbf{A}_j \boldsymbol{\mu} = \mathbf{0}, \ j = 1, \cdots, w.$$

No que se segue iremos construir as estatísticas de teste e obter as suas distribuições condicional (considerando as dimensões das amostras como fixas) e não condicional (assumindo que as dimensões das amostras são aleatórias). Serão apresentadas ainda duas aplicações com dados reais, uma considerando o caso equilibrado e outra aplicada ao caso não equilibrado. Terminamos o capítulo com um estudo numérico, em que são utilizadas simulações, por forma a comparar a abordagem proposta com a abordagem clássica.

## 3.2 Distribuições das dimensões das amostras

Nesta secção iremos obter as distribuições para as dimensões das amostras, considerando as três situações mencionadas anteriormente.

A forma mais comum para a truncatura das distribuições é a omissão do zero, visto que necessitamos de ter pelo menos uma observação por tratamento por forma a realizar a análise, ver e.g. Johnson e Kotz (1969). Assim assumiremos que

$$N_i \geq 1, \ i = 1, \cdots, m.$$

Por forma a realizar inferência, considerando que temos $m$ tratamentos na totalidade, é necessário assumir ainda que

$$N = \sum_{i=1}^{m} N_i > m.$$

Tomando a probabilidade condicional

$$\begin{aligned} p_{u_i,i} &= pr(N_i = u_i | N_i \geq 1) = \frac{pr(N_i = u_i)}{pr(N_i \geq 1)} \\ &= \frac{pr(N_i = u_i)}{1 - pr(N_i = 0)},\ u_i \geq 1,\ i = 1, \cdots, m, \end{aligned}$$

a função geradora de momentos de $N_i$, quando $N_i \geq 1$, $i = 1, \cdots, m$, será dada por

$$\varphi_i(t) = E(e^{tN_i}) = \sum_{u_i \in \mathcal{L}} e^{tu_i} p_{u_i,i},\ i = 1, \cdots, m, \tag{3.2.2}$$

onde $\mathcal{L}$ representa o suporte de $N_i$, $i = 1, \cdots, m$.

A função geradora de probabilidades será dada por

$$\chi_i(t) = E(t^{N_i}) = \varphi_i(\ln t) = \sum_{u_i \in \mathcal{L}} t^{u_i} p_{u_i,i},\ i = 1, \cdots, m. \tag{3.2.3}$$

Iremos considerar as variáveis $\ddot{N}_i$, $i = 1, \cdots, m$, que correspondem às variáveis truncadas $N_i, i = 1, \cdots, m$, em que $N_i \geq 1, i = 1, \cdots, m$, ver e.g. Capistrano (2015) e Nunes et al. (2019c). Como consequência teremos ainda

$$\ddot{N} = \sum_{i=1}^{m} \ddot{N}_i.$$

Ora, uma vez que $\ddot{N}_i$, $i = 1, \cdots, m$, são independentes, tem-se a função geradora de probabilidades de $\ddot{N}$ dada por

$$\ddot{\chi}(t) = \prod_{i=1}^{m} \chi_i(t). \tag{3.2.4}$$

Sabe-se que $\ddot{\chi}^{<r>}(0) = r!\ddot{p}_r$, onde $\ddot{p}_r = pr(\ddot{N} = r)$ e $< r >$ representa a derivada de ordem $r$. Assim,

$$\ddot{p}_u = pr(\ddot{N} = u) = \frac{1}{u!} \ddot{\chi}^{<u>}(0), \tag{3.2.5}$$

com

$$\ddot{\chi}^{<u>}(0) = \sum_{\boldsymbol{u} \in \mathcal{P}_u^{(m)}} \frac{(\sum_{i=1}^{m} u_i)!}{\prod_{i=1}^{m} u_i!} \prod_{i=1}^{m} \chi_i^{<u_i>}(0),\ u_i \geq 1,\ i = 1, \cdots, m, \tag{3.2.6}$$

onde $\mathcal{P}_u^{(m)}$ representa a família de partições de $u = u_1 + \cdots + u_m = \sum_{i=1}^m u_i$ com cardinal $m$ e $\boldsymbol{u} = (u_1, \cdots, u_m)'$. Tem-se portanto

$$\begin{aligned}\ddot{p}_u &= pr(\ddot{N} = u) = \\ &= \frac{1}{u!} \sum_{\boldsymbol{u} \in \mathcal{P}_u^{(m)}} \frac{u!}{\prod_{i=1}^m u_i!} \prod_{i=1}^m \chi_i^{<u_i>}(0) = \sum_{\boldsymbol{u} \in \mathcal{P}_u^{(m)}} \prod_{i=1}^m \frac{\chi_i^{<u_i>}(0)}{u_i!}.\end{aligned} \tag{3.2.7}$$

Consideremos agora que

$$s_1 + \cdots + s_m = s; \ s = 1, \cdots, m-1,$$

logo o vetor $\boldsymbol{s} = (s_1, \cdots, s_m)'$ tem uma ou mais componentes nulas. Assim, visto que $\chi_i(0) = 0$, $i = 1, \cdots, m$, teremos

$$\prod_{i=1}^m \chi_i^{<s_i>}(0) = 0$$

e consequentemente

$$\ddot{\chi}^{<s>}(0) = 0, \ s = 1, \cdots, m-1.$$

Deste modo vem

$$\ddot{p}_s = pr(\ddot{N} = s) = \frac{1}{s!} \ddot{\chi}^{<s>}(0) = 0, \ s = 1, \cdots, m-1, \tag{3.2.8}$$

e portanto

$$pr(\ddot{N} \leq m) = pr(\ddot{N} = m) = \ddot{p}_m.$$

Consideremos agora que $s = m$. O termo não nulo de $\ddot{\chi}^{<m>}(0)$ corresponde a $\boldsymbol{s} = \mathbf{1}_m$, visto que neste caso se tem $s_1 + \cdots + s_m = m$, então

$$\ddot{p}_m = pr(\ddot{N} = m) = \frac{1}{m!} \ddot{\chi}^{<m>}(0) = \prod_{i=1}^m \chi_i^{<1>}(0). \tag{3.2.9}$$

Além disso

$$pr(\ddot{N} > m) = 1 - pr(\ddot{N} \leq m) = 1 - \ddot{p}_m \tag{3.2.10}$$

e

$$\ddot{p}_{u,m} = pr(\ddot{N} = u | \ddot{N} > m) = \frac{\ddot{p}_u}{1 - \ddot{p}_m}, \ u \geq m+1. \tag{3.2.11}$$

Neste capítulo, por forma a evitar a ocorrência de casos altamente desequilibrados, iremos assumir que existe uma dimensão mínima global para as amostras, por exemplo

$n^\bullet > m$, o que significa que $\ddot{N} \geq n^\bullet$, vindo, ver e.g., Mexia et al. (2011) e Nunes et al. (2014),

$$\begin{aligned}\ddot{p}_{u,n^\bullet} &= pr(\ddot{N} = u|\ddot{N} \geq n^\bullet) = \frac{pr(\ddot{N} = u)}{pr(\ddot{N} \geq n^\bullet)} \\ &= \frac{\ddot{p}_u}{pr(\ddot{N} > m)} \frac{pr(\ddot{N} > m)}{pr(\ddot{N} \geq n^\bullet)} = \ddot{p}_{u,m} \frac{1 - \ddot{p}_m}{1 - \sum_{h=m}^{n^\bullet - 1} \ddot{p}_h}, \ u \geq n^\bullet.\end{aligned} \tag{3.2.12}$$

### 3.2.1 Processos de Contagem: Distribuição de Poisson

Nesta secção iremos assumir que a recolha das observações corresponde a processos de contagem, ver Mexia et al. (2011), Nunes et al. (2010, 2012a, 2014, 2019a, 2019c). Desta forma as variáveis aleatórias, $N_1, \cdots, N_m$, irão seguir uma distribuição de Poisson com parâmetros $\lambda_1, \cdots, \lambda_m$,

$$N_i \sim P(\lambda_i), i = 1, \cdots, m.$$

Podemos considerar como um exemplo prático a colheita de observações num estudo em que se pretende comparar várias patologias de pacientes que se deslocam às urgências de um hospital durante um determinado período fixo de tempo.

Devido à independência dos $N_i$, $i = 1, \cdots, m$, também se tem

$$N = \sum_{i=1}^{m} N_i \sim P(\lambda),$$

com $\lambda = \sum_{i=1}^{m} \lambda_i$.

Assim, neste caso vem

$$\begin{aligned}p_{u_i,i} &= pr(N_i = u_i|N_i \geq 1) = \\ &= \frac{e^{-\lambda_i}\lambda_i^{u_i}/u_i!}{1 - e^{-\lambda_i}} = \frac{e^{-\lambda_i}}{1 - e^{-\lambda_i}} \frac{\lambda_i^{u_i}}{u_i!}, \ u_i \geq 1, \ i = 1, \cdots, m,\end{aligned}$$

a função geradora de momentos de $N_i$, quando $N_i \geq 1$, $i = 1, \cdots, m$, é dada por

$$\varphi_i(t) = \sum_{u_i=1}^{\infty} \frac{e^{-\lambda_i}}{1 - e^{-\lambda_i}} \frac{\lambda_i^{u_i} e^{tu_i}}{u_i!} = \frac{e^{-\lambda_i}}{1 - e^{-\lambda_i}} (e^{\lambda_i e^t} - 1), \ i = 1, \cdots, m, \tag{3.2.13}$$

e a função geradora de probabilidade dada por

$$\chi_i(t) = \frac{e^{-\lambda_i}}{1 - e^{-\lambda_i}} (e^{\lambda_i t} - 1), \ i = 1, \cdots, m. \tag{3.2.14}$$

A função geradora de probabilidade da variável truncada $\ddot{N} = \sum_{i=1}^{m} \ddot{N}_i$ será então dada pela expressão, ver Mexia et al. (2011),

$$\begin{aligned} \ddot{\chi}(t) &= \prod_{i=1}^{m} \chi_i(t) = \left(\prod_{i=1}^{m} \frac{e^{-\lambda_i}}{1-e^{-\lambda_i}}\right) \prod_{i=1}^{m} (e^{\lambda_i t} - 1) = \\ &= \left(\prod_{i=1}^{m} \frac{e^{-\lambda_i}}{1-e^{-\lambda_i}}\right) \sum_{\mathcal{C} \subseteq \overline{\overline{m}}} (-1)^{m-\sharp(\mathcal{C})} e^{(\sum_{i \in \mathcal{C}} \lambda_i)t},\ i = 1, \cdots, m, \end{aligned} \tag{3.2.15}$$

com $\ddot{\chi}^{<u>}(0)$ definido em onde $\overline{\overline{m}} = \{1, \cdots, m\}$ e $\sharp(\mathcal{C})$ representa o cardinal de $\mathcal{C}$, que corresponde a um subconjunto de $\overline{\overline{m}}$.

Assim tem-se

$$\begin{aligned} \ddot{p}_u &= pr(\ddot{N} = u) = \frac{1}{u!} \ddot{\chi}^{<u>}(0) = \\ &= \frac{1}{u!} \left(\prod_{i=1}^{m} \frac{e^{-\lambda_i}}{1-e^{-\lambda_i}}\right) \sum_{\mathcal{C} \subseteq \overline{\overline{m}}} (-1)^{m-\sharp(\mathcal{C})} \left(\sum_{i \in \mathcal{C}} \lambda_i\right)^u,\ u \geq m, \end{aligned} \tag{3.2.16}$$

com $\ddot{\chi}^{<u>}(0)$ definido por (3.2.6), e

$$\ddot{p}_m = pr(\ddot{N} = m) = \prod_{i=1}^{m} \chi_i^{<1>}(0) = \prod_{i=1}^{m} \frac{e^{-\lambda_i} \lambda_i}{1-e^{-\lambda_i}}. \tag{3.2.17}$$

Consequentemente

$$\ddot{p}_{u,m} = \frac{\ddot{p}_u}{1-\ddot{p}_m},\ u \geq m+1. \tag{3.2.18}$$

### 3.2.2 Falhas de observações: Distribuição Binomial

Nesta secção iremos considerar situações em que existe um limite superior para a dimensão das amostras que poderá não ser atingido devido à ocorrência de falhas de observações. Esta situação pode ocorrer, por exemplo,

- quando se trabalha com pacientes que têm uma determinada patologia e existe uma probabilidade dos seus processos clínicos estarem incompletos ou mesmo não existirem;

- quando se trabalha com videiras ou árvores de frutos, e existe uma probabilidade, que pode depender do tipo de tratamento, de estas secarem;

- quando se aplica um questionário a um conjunto de pessoas e alguns dos questionários não são devolvidos.

Nestes casos consideramos que a distribuição Binomial é a escolha adequada para a distribuição das dimensões das amostras. Assim, assumimos que as variáveis aleatórias, $N_1, \cdots, N_m$, seguem uma distribuição Binomial com parâmetros $r_1, \cdots, r_m$ (limites superiores para as dimensões das amostras) e $1-p$ (onde $p$ representa a probabilidade de ocorrência de uma falha), ver Nunes et al.(2015, 2019a, 2019b). Colocamos

$$N_i \sim B(r_i, 1-p), i = 1, \cdots, m.$$

Devido à independência dos $N_i$, $i = 1, \cdots, m$, tem-se que

$$N = \sum_{i=1}^{m} N_i \sim B\left(r, 1-p\right),$$

com $r = \sum_{i=1}^{m} r_i$.

Assim, tem-se a seguinte probabilidade condicionada

$$p_{u_i,i} \quad = \quad pr(N_i = u_i | N_i \geq 1) = \frac{\begin{pmatrix} r_i \\ u \end{pmatrix} (1-p)^{u_i} p^{r_i - u_i}}{1 - p^{r_i}}, \; u_i = 1, \cdots, r_i, \; i = 1, \cdots, m.$$

A função geradora de momentos de $N_i$, quando $N_i \geq 1$, $i = 1, \cdots, m$, será dada por

$$\varphi_i(t) = \sum_{u_i=1}^{r_i} e^{tu_i} \frac{\begin{pmatrix} r_i \\ u_i \end{pmatrix} (1-p)^{u_i} p^{r_i - u_i}}{1 - p^{r_i}} = \frac{(p + (1-p)e^t)^{r_i} - p^{r_i}}{1 - p^{r_i}}, \; i = 1, \cdots, m, \tag{3.2.19}$$

e a função geradora de probabilidades por

$$\chi_i(t) = \frac{(p + (1-p)t)^{r_i} - p^{r_i}}{1 - p^{r_i}}, \; i = 1, \cdots, m. \tag{3.2.20}$$

Quanto à função geradora de probabilidade de $\ddot{N}$ virá

$$\ddot{\chi}(t) \quad = \quad \prod_{i=1}^{m} \chi_i(t) = \prod_{i=1}^{m} \frac{(p + (1-p)t)^{r_i} - p^{r_i}}{1 - p^{r_i}}. \tag{3.2.21}$$

Tem-se então

$$\begin{aligned}\ddot{p}_u &= pr(\ddot{N}=u)=\frac{1}{u!}\ddot{\chi}^{<u>}(0)\\ &= \sum_{\boldsymbol{u}\in\mathcal{P}_u^m}\prod_{i=1}^{m}\frac{\begin{pmatrix} r_i,\\ u_i\end{pmatrix}(1-p)^{u_i}p^{r_i-u_i}}{1-p^{r_i}},\ u_i=1,\cdots,r_i,\ i=1,\cdots,m, \end{aligned}\tag{3.2.22}$$

e

$$\ddot{p}_m = pr(\ddot{N}=m)=\prod_{i=1}^{m}\chi_i^{<1>}(0)=\prod_{i=1}^{m}\frac{r_i(1-p)p^{r_i-1}}{1-p^{r_i}}, \tag{3.2.23}$$

e consequentemente

$$\ddot{p}_{u,m} = \frac{\ddot{p}_u}{1-\ddot{p}_m},\ u=m+1,\cdots,r. \tag{3.2.24}$$

### 3.2.3 Ambientes protegidos: Distribuição Geométrica e Distribuição Binomial Negativa

As distribuições Geométrica e Binomial negativa são muitas vezes usadas para descrever a ocorrência de "contágios" em eventos discretos, como por exemplo o surto de doenças. Mais especificamente, pode por exemplo representar o número de dias necessários até à ocorrência da contaminação pelo escaravelho vermelho de palmeiras, numa área protegida.

Assumiremos que as variáveis aleatórias, $N_1,\cdots,N_m$, seguem uma distribuição Geométrica de parâmetro $q$ (que corresponde à probabilidade da ocorrência de um sucesso o que significa que $1-q$ corresponderá à probabilidade de ocorrência de falha), ver Mário et al. (2019). Colocamos

$$N_i\sim Geo(q), i=1,\cdots,m.$$

Teremos portanto

$$p_{u,i}=pr(N_i=u_i|N_i\geq 1) = \frac{q(1-q)^{u_i}}{1-q}=q(1-q)^{u_i-1}, u_i\geq 1, i=1,\cdots,m. \tag{3.2.25}$$

A função geradora de momentos de $N_i$, quando $N_i\geq 1$, $i=1,\cdots,m$, será

$$\varphi_i(t)=\sum_{u_i=1}^{+\infty}(e^{tu_i})q(1-q)^{u_i-1}=\frac{qe^t}{1-(1-q)e^t},\ t<ln\left(\frac{1}{1-q}\right),\ i=1,\cdots,m, \tag{3.2.26}$$

e a função geradora de probabilidades será dada por

$$\chi_i(t) = \frac{qt}{1-(1-q)t},\ t < \frac{1}{1-q},\ i = 1, \cdots, m. \tag{3.2.27}$$

Uma vez que as variáveis truncadas $\ddot{N}_i$, $i = 1, \cdots, m$, são independentes, $\ddot{N}$ seguirá uma distribuição Binomial Negativa de parâmetros $m$ (o número de sucessos até a experiência parar) e $q$,

$$\ddot{N} = \sum_{i=1}^{m} \ddot{N}_i \sim BN\,(m, q)\,.$$

A função geradora de probabilidade de $\ddot{N}$ será

$$\ddot{\chi}(t) \;=\; \prod_{i=1}^{m} \chi_i(t) = \prod_{i=1}^{m} \frac{qt}{1-(1-q)t} = \left(\frac{qt}{1-(1-q)t}\right)^m,\ t < \frac{1}{1-q}. \tag{3.2.28}$$

Assim, temos

$$\begin{aligned} \ddot{p}_u &= pr(\ddot{N} = u) = \frac{1}{u!}\ddot{\chi}^{<u>}(0) = \sum_{\boldsymbol{u}\in\mathcal{P}_u^{(m)}} \prod_{i=1}^{m} \frac{\ddot{\chi}_i^{<u_i>}(0)}{u_i!} = \\ &= \sum_{\boldsymbol{u}\in\mathcal{P}_u^{(m)}} \prod_{i=1}^{m} q(1-q)^{u_i-1},\ u_i = 1, \cdots,\ i = 1, \cdots, m, \end{aligned} \tag{3.2.29}$$

uma vez que $\ddot{\chi}_i{}^{<u_i>}(0) = u_i!(1-q)^{u_i-1} \times q$, e

$$\ddot{p}_m \;=\; pr(\ddot{N} = m) = \prod_{i=1}^{n} \chi_i^{<1>}(0) = \prod_{i=1}^{m} q = q^m. \tag{3.2.30}$$

Consequentemente

$$\ddot{p}_{u,m} \;=\; \frac{\ddot{p}_u}{1-q^m},\ u \geq m+1. \tag{3.2.31}$$

## 3.3 Distribuições das observações e estatísticas de teste

Nesta secção, vamos obter as distribuições das estatísticas de teste assumindo que as dimensões das amostras, considerando que se tem $m$ tratamentos, são realizações, $n_1, \cdots, n_m$, das variáveis aleatórias independentes, $\ddot{N}_1, \cdots, \ddot{N}_m$. Tal como definidos anteriormente, $\ddot{N}_i, i = 1, \cdots, m$, correspondem às variáveis $N_i, i = 1, \cdots, m$, truncadas em que $N_i \geq 1$, $i = 1, \cdots, m$.

Vamos começar pelo caso usual, em que as dimensões das amostras são fixas, o que significa que estamos a assumir que $\ddot{N}_i = n_i, i = 1, \cdots, m$.

Assim, temos as amostras

$$Y_{i,1}, \cdots, Y_{i,n_i}, \ i = 1, \cdots, m,$$

com médias $Y_{i,\bullet}$, $i = 1, \cdots, m$. A soma das somas dos quadrados do erro, será dada por

$$S = \sum_{i=1}^{m} \sum_{k=1}^{n_i} (Y_{i,k} - Y_{i,\bullet})^2,$$

ver por exemplo Searle et al. (1992) e Khuri et al. (1998).

Se assumirmos que as observações são normais e independentes, com variância $\sigma^2$, quando $\ddot{N}_i = n_i$, $i = 1, \cdots, m$, $S$ será o produto de $\sigma^2$ por um qui-quadrado central com $g(n) = n - m$ graus de liberdade,

$$S \sim \sigma^2 \chi^2_{g(n)}.$$

Além disso, $S$ será condicionalmente independente do vetor médio dos tratamentos, $\mathbf{Y}_\bullet$, que tem como componentes $Y_{1,\bullet}, \cdots, Y_{m,\bullet}$. O vetor $\mathbf{Y}_\bullet$ será normal com vetor médio $\boldsymbol{\mu}$ e matriz de covariâncias $\sigma^2 D(\frac{1}{n_1}, \cdots, \frac{1}{n_m})$, onde $D(\frac{1}{n_1}, \cdots, \frac{1}{n_m})$ representa uma matriz diagonal com elementos principais $\frac{1}{n_1}, \cdots, \frac{1}{n_m}$,

$$\mathbf{Y}_\bullet \sim \mathcal{N}\left(\boldsymbol{\mu}; \sigma^2 D\left(\frac{1}{n_1}, \cdots, \frac{1}{n_m}\right)\right).$$

Assim, quando $\ddot{N}_i = n_i$, $i = 1, \cdots, m$,

$$S_j = (\mathbf{A}_j \mathbf{Y}_\bullet)' \left(\mathbf{A}_j D\left(\frac{1}{n_1}, \cdots, \frac{1}{n_m}\right) \mathbf{A}'_j\right)^{-1} (\mathbf{A}_j \mathbf{Y}_\bullet), \ j = 1, \cdots, w,$$

é o produto de $\sigma^2$ por um qui-quadrado com $g_j = car(\mathbf{A}_j), \ j = 1, \cdots, w,$ graus de liberdade e parâmetro de não centralidade

$$\delta_j(n) \quad = \quad \frac{1}{\sigma^2} (\mathbf{A}_j \boldsymbol{\mu})' \left(\mathbf{A}_j D\left(\frac{1}{n_1}, \cdots, \frac{1}{n_m}\right) \mathbf{A}'_j\right)^{-1} (\mathbf{A}_j \boldsymbol{\mu}), \ j = 1, \cdots, w,$$

$S_j \sim \sigma^2 \chi^2_{g_j, \delta_j(n)}, \ j = 1, \cdots, w,$ ver e.g., Mexia (1990).

Teremos portanto a estatística de teste

$$\Im_j = \frac{g(n)}{g_j} \frac{S_j}{S}, \ j = 1, \cdots, w,$$

que segue uma distribuição $F$ não central com $g_j$ e $g(n)$ graus de liberdade e parâmetro de não-centralidade $\delta_j(n)$, $j = 1, \cdots, w$, $F(.|g_j, g(n), \delta_j(n))$. Designamos esta distribuição como distribuição condicional, já que estamos a considerar $\ddot{N} = n$, portanto as dimensões das amostras como fixas.

Podemos ainda considerar a estatística de teste

$$\tau_j = \frac{S_j}{S},\ j = 1, \cdots, w,$$

com distribuição condicional $\overline{F}(.|g_j, g(n), \delta_j(n))$, que, tal como definido no Capítulo 2, corresponde à distribuição do quociente de qui-quadrados independentes com $g_j$ e $g(n)$ graus de liberdade e parâmetro de não centralidade $\delta_j(n),\ j = 1, \cdots, w.$ Neste capítulo iremos por vezes considerar as estatísticas $\tau_j,\ j = 1, \cdots, w,$ já que as distribuições $\overline{F}$ são estatisticamente equivalentes às distribuições $F$ e, digamos que, mais "tratáveis".

Quando $H_{0,j},\ j = 1, \cdots, w$, se verifica, dado $\ddot{N} = n$, a distribuição condicional de $\Im_j$ corresponde a uma $F$ central com $g_j$ e $g(n)$ graus de liberdade, $F(.|g_j, g(n)), j = 1, \cdots, w,$ enquanto que a distribuição condicional de $\tau_j$ será uma $\overline{F}(.|g_j, g(n)),\ j = 1, \cdots, w$.

### 3.3.1 Distribuição não condicional da estatística

Nesta secção iremos descondicionar os resultados anteriormente obtidos em ordem às variáveis aleatórias $N_1, ..., N_m$. Como referido anteriormente vamos assumir que existe uma dimensão mínima global para as amostras que denotamos por $n^\bullet$.

Para obtermos a distribuição não condicional de $\Im_j$ quando $H_{0,j},\ j = 1, \cdots, w$, se verifica, destacamos que se tem, ver e.g. Nunes et al. (2014) e Capistrano (2015),

$$pr(\Im_j \leq z) = \sum_{n \in \mathcal{V}} pr(\ddot{N} = n|\ddot{N} > n^\bullet) pr(\Im_j \leq z|\ddot{N} = n),\ j = 1, \cdots, w,$$

com

$$pr(\Im_j \leq z|\ddot{N} = n) = F(z|g_j, g(n)),\ j = 1, \cdots, w,$$

e

$$pr(\ddot{N} = n|\ddot{N} \geqslant n^\bullet) = \ddot{p}_{n,n^\bullet}$$

dada por (3.2.12), que posteriormente será obtida de acordo com as distribuições de $\ddot{N}_i, i = 1, \cdots, m$. Assim, podemos escrever a distribuição não condicional de $\Im_j$ como

$$\overline{\overline{F}}_j(z) = pr(\Im_j \leq z) = \sum_{n \in \mathcal{V}} \ddot{p}_{n,n^\bullet} F(z|g_j, g(n)),\ j = 1, \cdots, w, \qquad (3.3.32)$$

onde $\mathcal{V}$ representa o suporte da variável aleatória truncada $\ddot{N}$, considerando $\ddot{N} \geqslant n^\bullet$.

Se considerarmos a estatística $\tau_j, j = 1, \cdots, w,$ a expressão da sua distribuição não condicional será obtida de forma análoga.

## 3.4 Aplicações com dados reais

Nesta secção serão apresentadas duas aplicações com dados reais, relativos ao desemprego em alguns países da união Europeia. Os dados provém do PORDATA-Base de dados de Portugal contemporâneo.

Quando se trata de um modelo com mais do que um fator, como é o nosso caso, geralmente as análises devem começar com um teste à interação entre fatores. Se a interação não é significativa a análise deve prosseguir com os testes sobre os efeitos principais dos fatores. Se a interação é significativa, a inferência é realizada para cada nível de um dos fatores. Neste trabalho não seguiremos esta abordagem uma vez que temos como intuito mostrar como estes testes podem ser realizados considerando a metodologia proposta (considerando as dimensões das amostras como aleatórias), ver e.g. Nunes et al. (2014). Consideramos primeiro o caso equilibrado e um segundo caso não equilibrado.

Como referido anteriormente, assumiremos que existe uma dimensão mínima global para as amostras, $n^\bullet$, o que significa que $\ddot{N} \geqslant n^\bullet$.

Na primeira aplicação vamos considerar a estatística $\tau_j$.

Vimos que, quando $H_{0,j}, j = 1, \cdots, w$, se verifica e $\ddot{N} = n$,

$$\tau_j = \frac{S_j}{S} \sim \overline{F}(z|g_j, n-m),\ j = 1, \cdots, w.$$

Quanto à distribuição não condicional de $\tau_j$, esta será dada por

$$\overline{\overline{F}}_j(z) = \sum_{n \in \mathcal{V}} \ddot{p}_{n,n^\bullet} \overline{F}(z|g_j, n-m),\ j = 1, \cdots, w,$$

com $g_j = car(\boldsymbol{A}_j)$ $j = 1, \cdots, w,$ a probabilidade $\ddot{p}_{n,n^\bullet}$ é dada por (3.2.12) e $\mathcal{V}$ corresponde ao suporte da variável truncada $\ddot{N}$ para o caso em que $n \geq n^\bullet$.

Ora, pelas propriedades de monotonia de $\overline{F}$, (apresentadas em Mexia et al. (2011) e detalhadas no capítulo anterior, secção 2.2.3) e considerando $n < n'$, tem-se

$$\overline{F}(z|g_j, n-m) < \overline{F}(z|g_j, n'-m),$$

e consequentemente

$$\overline{F}(z|g_j, n^\bullet - m) \leq \overline{\overline{F}}_j(z) \leq 1,$$

o que significa que $\overline{F}(z|g_j, n^\bullet - m)$ nos dará um limite inferior para $\overline{\overline{F}}_j(z)$, $j = 1, \cdots, w$. Assim, através de $\overline{F}(z|g_j, n^\bullet - m)$ podemos obter o limite superior para os quantis de

$\overline{\overline{F}}_j(z)$, $j = 1, \cdots, w$. Se utilizarmos esses limites superiores como valores críticos teremos testes com tamanhos que não excederão os valores teóricos.

Podemos usar esses limites superiores como um teste preliminar, ver e.g. Capistrano (2015) e Nunes et al. (2019c). Se o valor observado da estatística de teste exceder esse limite superior, também irá exceder o valor crítico real (obtido considerando as dimensões das amostras como aleatórias) e nesse caso, devemos rejeitar a hipótese nula. Quando o valor observado da estatística é menor do que o limite superior, devemos calcular o valor crítico real ou calcular o valor mínimo de $n^\bullet$ que leva à rejeição da hipótese nula.

Os valores críticos reais (obtidos quando usamos a distribuição não condicional) excederão os referentes à abordagem clássica. Assim, ao não rejeitarmos a hipótese pela abordagem clássica, iremos tomar a mesma decisão considerando a abordagem proposta (em que as dimensões das amostras são consideradas como aleatórias) e consequentemente considerando os limites superiores dos quantis.

Nesta secção iremos seguir esta abordagem, apresentando o cálculo dos limites superiores dos quantis para a distribuição não condicional.

Na segunda aplicação a metodologia proposta será aplicada à estatística

$$\Im_j = \frac{n-m}{g_j}\frac{S_j}{S} \sim F(z|g_j, n-m),\ j = 1, \cdots, w.$$

É importante referir que devido às elevadas dimensões das amostras, trabalhamos com distribuições normais assimptóticas, o que nos garante a possibilidade de usarmos os testes $F$, ver Ito (1980).

Todos os procedimentos computacionais foram realizados com a utilização do *software* R.

### 3.4.1 O desemprego na UE antes e durante a crise económica.

A metodologia proposta é agora aplicada à incidência do desemprego em quatro países Europeus, os países que solicitaram apoio ao fundo monetário internacional (FMI). Os dados usados provêm do PORDATA e a informação diz respeito à idade das pessoas desempregadas, Base de dados contemporâneos (2014).

Consideramos um modelo de efeitos fixos com dois fatores: *Países* e *períodos de tempo*. O primeiro fator tem quatro níveis: *Espanha*, *Grécia*, *Itália* e *Portugal*. Quanto ao segundo fator tem dois níveis: o primeiro corresponde ao período *2006-2007* (antes da crise financeira) e o segundo corresponde a *2012-2013* (durante a crise financeira). Temos portanto $m = 4 \times 2 = 8$ diferentes tratamentos. A Tabela 3.1 apresenta a média das idades dos desempregados para cada país e período de tempo considerado.

Tabela 3.1: Médias amostrais das idades

| | | Períodos de tempo (segundo fator) | |
|---|---|---|---|
| | | 2006-2007 | 2012-2013 |
| Países (primeiro fator) | Espanha | 35.23425 | 35.80489 |
| | Grécia | 30.54088 | 34.96733 |
| | Itália | 30.56569 | 31.29323 |
| | Portugal | 34.93271 | 34.43291 |

Testaremos as hipóteses

$$H_{0,j} : \mathbf{A_j}\boldsymbol{\mu} = 0,\ j = 1, 2, 3,$$

que correspondem às hipóteses de ausência dos efeitos principais dos fatores e ausência de interação entre os dois fatores.

Vamos considerar o caso equilibrado (robusto na presença da não normalidade e heterocedasticidade, ver Scheffé (1959, Capítulo 10)). Tomaremos $n_i = 6,\ i = 1, \cdots, 8$, que correspondem ao número das diferentes classes de idades dos desempregados. Assim, $n = \sum_{i=1}^{8} 6 = 48$.

Dado $\ddot{N} = n$, quando $H_{0,j},\ j = 1, 2, 3$, se verifica, tem-se

$$\tau_j = \frac{S_j}{S} \sim \overline{F}(z|g_j, n-8),\ j = 1, 2, 3,$$

com $g_j = car(\mathbf{A}_j)$ e $n - 8 = 40$.

**Primeiro fator (países)**

Para o primeiro fator temos a matriz

$$\mathbf{A_1} = \begin{bmatrix} -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \\ -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \\ \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \end{bmatrix},$$

com $g_1 = car(\mathbf{A_1})$= 3. Obtemos

$$\mathbf{A_1 y_\bullet} = \begin{bmatrix} 0.6985 \\ 1.8819 \\ 4.6095 \end{bmatrix},$$

onde o vetor $\mathbf{y_\bullet}$ tem como componentes as médias amostrais apresentadas na Tabela 3.1. Assim, para o numerador da estatística obtemos

$$S_1 \ = \ (\mathbf{A_1 y_\bullet})' \left(\mathbf{A_1 D A_1}'\right)^{-1} (\mathbf{A_1 y_\bullet}) = 151.6617, \qquad (3.4.33)$$

com $\mathbf{D} = D(\frac{1}{n_1}, \cdots, \frac{1}{n_8})$, em que $n_i = 6$, $i = 1, \cdots, 8$.

Quando $\ddot{N}_i = n_i,\ i = 1, \cdots, 8$, o denominador da estatística é o produto de $\sigma^2$ por um qui-quadrado central com $g(n)$=40 graus de liberdade, $S \sim \sigma^2 \chi^2_{40}$. Neste caso obtemos

$$S = 145.4374.$$

Assim, o valor observado da estatística, $\tau_{1,Obs}$, é igual a

$$\tau_{1,Obs} = \frac{151.6617}{145.4374} = 1.0428. \tag{3.4.34}$$

Os quantis, $z_{1-\alpha}$, da distribuição condicional da $\tau_1$, que corresponde a $\overline{F}(z|3, 40)$, são apresentados na Tabela 3.2.

Visto que $\tau_{1,Obs} > z_{1-\alpha}$, podemos concluir que se rejeita $H_{0,1}$ para os níveis usuais de significância.

Tabela 3.2: Quantis da distribuição condicional e limites superiores para os quantis de $\tau_1$ e $\tau_3$

| **Valores de** $\alpha$ | 0.1 | 0.05 | 0.01 |
|---|---|---|---|
| $z_{1-\alpha}$ | 0.16696 | 0.21291 | 0.32344 |
| $z^u_{1-\alpha}$ | 0.43014 | 0.56414 | 0.91500 |

Assumiremos que temos uma dimensão mínima global para as amostras, por exemplo $\ddot{N} \geq 25$, o que significa que $n^\bullet = 25$.

A Tabela 3.2 apresenta também os limites superiores para os quantis, $z^u_{1-\alpha}$, para a probabilidade $1 - \alpha$ da distribuição não condicional de $\tau_1$. Como $\tau_{1,Obs} > z^u_{1-\alpha}$ podemos concluir que também se rejeita $H_{0,1}$ para os níveis usuais de significância considerando a abordagem não condicional, o que significa que o primeiro fator é significativo.

### Segundo fator (período de tempo)

Para o segundo fator temos a matriz

$$\mathbf{A_2} = \left[ \begin{array}{cccccccc} -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \end{array} \right],$$

e $g_2 = car(\mathbf{A}_2)$=1. Assim obtemos

$$S_2 = (\mathbf{A}_2\mathbf{y}_\bullet)' \left(\mathbf{A}_2 \left(D\right) \mathbf{A_2}'\right)^{-1} (\mathbf{A_2 y_\bullet}) = 20.4741$$

para o numerador da estatística $\tau_2$.

Portanto, o valor observado da estatística $\tau_{2,Obs}$, é dado por

$$\tau_{2,Obs} = \frac{20.4741}{145.4374} = 0.1408.$$

Considerando a distribuição condicional de $\tau_2$, que corresponde a $\overline{F}(z|1, 40)$, obtemos os quantis $z_{1-\alpha}$ apresentados na Tabela 3.3.

Tabela 3.3: Quantis da distribuição condicional e limites superiores dos quantis para $\tau_2$

| **Valores de** $\alpha$ | 0.10 | 0.05 | 0.01 |
|---|---|---|---|
| $z_{1-\alpha}$ | 0.07088 | 0.10212 | 0.18285 |
| $z^u_{1-\alpha}$ | 0.17801 | 0.26184 | 0.49410 |

Assim, pela abordagem usual podemos concluir que se rejeita $H_{0,2}$ para $\alpha = 0.1$ e $0.05$ ($\tau_{2,Obs} > z_{1-\alpha}$) e não se rejeita para $\alpha = 0.01$.

A Tabela 3.3 mostra ainda os limites superiores para os quantis, $z^u_{1-\alpha}$, para a probabilidade $1 - \alpha$ da distribuição não condicional de $\tau_2$.

Assumindo $n^\bullet = 25$, podemos concluir que não se rejeita $H_{0,2}$ para os níveis usuais de significância ($\tau_{2,Obs} < z^u_{1-\alpha}$). Estes resultados podem portanto levar-nos a tomar uma decisão contrária à que tínhamos tomado ao usarmos a abordagem condicional, para $\alpha = 0.1$ e $0.05$.

Assumindo que os valores da estatística de teste se mantém inalterados, teremos que considerar o valor mínimo de $n^\bullet$, apresentado na Tabela 3.4, por forma a assegurar a rejeição da hipótese nula.

Tabela 3.4: Valor mínimo de $n^\bullet$ que leva à rejeição da hipótese $H_{0,2}$.

| **Valores de** $\alpha$ | 0.1 | 0.05 | 0.01 |
|---|---|---|---|
| $n^\bullet$ | 30 | 38 | 59 |

Uma vez que maiores valores de $n^\bullet$ correspondem a menores valores para os quantis, podemos considerar que $\tau_{2,Obs} > z^u_{1-\alpha}$ para todo $n^\bullet \geq 59$, o que significa que, neste caso, se rejeita $H_{0,2}$ considerando os níveis usuais de significância e portanto o segundo fator será significativo.

### Interação entre os dois fatores

Para a interação temos

$$\mathbf{A_3} = \begin{bmatrix} \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \\ \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \\ -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \end{bmatrix}.$$

Assim, $g_3 = car(\mathbf{A_3})$=3 e para o numerador da estatística obtemos

$$S_1 = (\mathbf{A_3 y_\bullet})' \left(\mathbf{A_3 D A_3}'\right)^{-1} (\mathbf{A_3 y_\bullet}) = 41.6205.$$

Logo, o valor observado da estatística, $\tau_{3,Obs}$, é dado por

$$\tau_{3,Obs} = \frac{41.6205}{145.4374} = 0.2862. \qquad (3.4.35)$$

A Tabela 3.2 apresenta os quantis da distribuição condicional, $z_{1-\alpha}$, e os limites superiores dos quantis da distribuição não condicional, $z^u_{1-\alpha}$, da estatística $\tau_3$.

Considerando a distribuição condicional, podemos concluir que se rejeita $H_{0,3}$ para $\alpha = 0.1$ e $0.05$ ($\tau_{3,Obs} > z_{1-\alpha}$) e não se rejeita para $\alpha = 0.01$.

Vamos assumir mais uma vez que $n^\bullet = 25$. Através dos resultados apresentados na Tabela 3.2, visto que $\tau_{3,Obs} < z^u_{1-\alpha}$, não se rejeita $H_{0,3}$ para os níveis usuais de significância. Portanto estes resultados podem levar-nos mais uma vez a tomar uma decisão contrária à que tínhamos tomando considerando a abordagem usual, para uma significância de $5\%$ e de $10\%$.

Assumindo que os valores da estatística de teste se mantêm inalterados, teremos que considerar o valor mínimo de $n^\bullet$ apresentado na Tabela 3.5 por forma a rejeitar a hipótese testada.

Tabela 3.5: Valor mínimo de $n^\bullet$ que leva à rejeição da hipótese $H_{0,3}$.

| **Valores de** $\alpha$ | 0.1 | 0.05 | 0.01 |
|---|---|---|---|
| $n^\bullet$ | 33 | 39 | 53 |

Temos portanto que $\tau_{3,Obs} > z^u_{1-\alpha}$, para todo o $n^\bullet \geq 53$, o que significa que neste caso se rejeita $H_{0,3}$ considerando os níveis usuais de significância e a interação entre fatores é significativa.

### Discussão dos resultados

Os resultados obtidos através desta aplicação mostram que:

- a inferência depende da abordagem considerada (abordagem clássica ou abordagem proposta). Para a interação e para o segundo fator concluímos que as duas abordagens nos podem levar a decisões contrárias. Consideramos que a abordagem proposta (não condicional) é mais realista sempre que se desconhecem os tamanhos das amostras à priori, pelo que podemos concluir que neste caso a abordagem usual pode derivar em falsas rejeições para $\alpha$= 0.1 e $\alpha$= 0.05;

- quanto ao primeiro fator, ambas as abordagens levaram à rejeição da hipótese nula. Logo este é significativo, o que significa que a idade dos desempregados é significativamente diferente entre pelo menos dois destes quatros países.

**Observação importante**: Consideramos que ocorre uma "falsa rejeição" sempre a hipótese nula é rejeitada através da abordagem clássica e não é rejeitada considerando a abordagem proposta.

### 3.4.2 O desemprego em Portugal por género e região

Nesta secção, a metodologia proposta é aplicada ao número de desempregados considerando as várias regiões de Portugal continental. Os dados usados são mais uma vez provenientes do PORDATA e a informação diz respeito à idade dos desempregados, Base de dados contemporâneos (2016).

Vamos considerar dois fatores. O primeiro fator tem cinco níveis: *Alentejo*, *Algarve*, *Área Metropolitana de Lisboa*, *Centro* ( Aveiro, Beira baixa, Beiras e Serra da Estrela, Castelo Branco, Coimbra, Leiria, Médio Tejo, Oeste, Viseu Dão-Lafões) e *Norte* (Área Metropolitana do Porto, Alto Minho, Alto Tâmega, Ave, Câvado, Douro, Tâmega e Sousa e Trás-os-Montes). Quanto ao segundo fator este corresponde ao género e tem portanto dois níveis: *homens e mulheres*. Tem-se neste caso $m = 5 \times 2 = 10$ diferentes tratamentos. A Tabela 3.6 ilustra a média das idades dos desempregados e o número de desempregados sobre 10, por região e género da pessoa.

Tabela 3.6: Médias amostrais das idades e número de desempregados

| | | Média das idades | | Número de desempregados/10 | |
|---|---|---|---|---|---|
| Género (segundo fator) | | Mulheres | Homens | Mulheres | Homens |
| Regiões (primeiro fator) | Norte | 39.3956 | 39.3958 | 13416 | 12002 |
| | Centro | 38.7373 | 38.7364 | 6076 | 5526 |
| | Área Metropolitana de Lisboa | 38.4093 | 38.4091 | 8765 | 9413 |
| | Alentejo | 38.7811 | 38.7804 | 2232 | 2164 |
| | Algarve | 38.8139 | 38.8144 | 1648 | 1830 |

Testaremos as hipóteses

$$H_{0,j} : \mathbf{A_j}\boldsymbol{\mu} = 0,\ j = 1, 2, 3,$$

que são hipóteses de ausência dos efeitos e de interação entre os dois fatores.

Dado $\ddot{N} = n$, quando $H_{0,j}$, $j = 1, 2, 3$, se verifica,

$$\Im_j = \frac{n - 10}{g_j}\frac{S_j}{S} \sim F(z|g_j, n - 10),\ j = 1, 2, 3,$$

com $g_j = car(A_j)$ e $n$= 63071 (número total de desempregados /10).

**Primeiro fator (regiões)**

Neste caso para o primeiro fator temos

$$\mathbf{A_1} = \begin{bmatrix} -\frac{1}{2} & -\frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & -\frac{1}{2} & -\frac{1}{2} & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2\sqrt{3}} & \frac{1}{2\sqrt{3}} & 0 & 0 & -\frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} & 0 & 0 & \frac{1}{2\sqrt{3}} & \frac{1}{2\sqrt{3}} \\ \frac{1}{\sqrt{15}} & \frac{1}{\sqrt{15}} & -\frac{3}{2\sqrt{15}} & -\frac{3}{2\sqrt{15}} & \frac{1}{\sqrt{15}} & \frac{1}{\sqrt{15}} & -\frac{3}{2\sqrt{15}} & -\frac{3}{2\sqrt{15}} & \frac{1}{\sqrt{15}} & \frac{1}{\sqrt{15}} \end{bmatrix},$$

com $g_1 = car(\boldsymbol{A}_1)$= 4. Obtemos

$$\mathbf{A_1 y_\bullet} = \begin{bmatrix} -0.5816 \\ 0.0438 \\ 0.8033 \\ 0.1770 \end{bmatrix},$$

onde o vetor $\mathbf{y_\bullet}$ tem como componentes as médias amostrais apresentadas na Tabela 3.6. Obtemos para o numerador da estatística

$$S_1 \quad = \quad (\mathbf{A_1 y_\bullet})' \left(\mathbf{A_1 D A_1}'\right)^{-1} (\mathbf{A_1 y_\bullet}) = 10981.78, \tag{3.4.36}$$

com $\mathbf{D} = D(\frac{1}{n_1}, \frac{1}{n_2}, \cdots, \frac{1}{n_{10}})$, onde $n_1, \cdots, n_{10}$ correspondem ao número de desempregados que constam da Tabela 3.6. O denominador da estatística é, quando $\ddot{N}_i = n_i,\ i = 1, \cdots, 10$, o produto de $\sigma^2$ por um qui-quadrado central com $g(n) = n - 10 = 60361$ graus de liberdade. Neste caso obtemos

$$S = 14094449.4217.$$

Assim, o valor observado da estatística, $\Im_{1,Obs}$, é dado por

$$\Im_{1,Obs} = \frac{60361}{4} \frac{10981.78}{14094449.4217} = 12.2836 \tag{3.4.37}$$

Os quantis, $z_{1-\alpha}$, da distribuição condicional da $\Im_1$, que corresponde a $F(z|4, n-10)$, são apresentados na Tabela 3.7.

Visto que $\Im_{1,Obs} > z_{1-\alpha}$, podemos concluir que se rejeita $H_{0,1}$ para os níveis usuais de significância.

Tabela 3.7: Quantis da distribuição condicional e os limites superiores dos quantis para $\tau_1$ e $\tau_3$

| **Valores de $\alpha$** | 0.1 | 0.05 | 0.01 |
|---|---|---|---|
| $z_{1-\alpha}$ | 1.9449 | 2.3721 | 3.3195 |
| $z^u_{1-\alpha}$ | 1.9593 | 2.3948 | 3.3676 |

Assumiremos que temos como dimensão mínima global para as amostras, por exemplo, $n^{\bullet} = 400$, o que significa que a dimensão mínima é de cerca de 40 observação por tratamento.

A Tabela 3.7 mostra os limites superiores para os quantis, $z^{u}_{1-\alpha}$, para a probabilidade $1-\alpha$ da distribuição não condicional de $\Im_1$. Assim contínua-se a rejeita $H_{0,1}$ para os diferentes valores de $\alpha$ considerados, donde se conclui que o primeiro fator é significativo.

**Segundo fator (género)**

Para o segundo fator temos

$$\mathbf{A_2} = \left[ \begin{array}{cccccccccc} -\frac{1}{\sqrt{10}} & \frac{1}{\sqrt{10}} & -\frac{1}{\sqrt{10}} & \frac{1}{\sqrt{10}} & -\frac{1}{\sqrt{\sqrt{10}}} & \frac{1}{\sqrt{10}} & -\frac{1}{\sqrt{10}} & \frac{1}{\sqrt{10}} & -\frac{1}{\sqrt{10}} & \frac{1}{\sqrt{10}} \end{array} \right],$$

e $g_2 = car(\mathbf{A}_2) = 1$. Assim obtemos

$$S_2 = (\mathbf{A}_2\mathbf{y}_{\bullet})' \left(\mathbf{A}_2\boldsymbol{D}\mathbf{A_2}'\right)^{-1} (\mathbf{A_2}\mathbf{y}_{\bullet}) = 0.0005$$

para o numerador da estatística $\Im_2$.

Portanto, o valor observado da estatística, $\Im_{2,Obs}$, é dado por

$$\Im_{2,Obs} = \frac{60361}{1}\frac{0.0005}{14094449.4217} = 2.3116e^{-06}. \tag{3.4.38}$$

Considerando a distribuição condicional de $\Im_2$, que corresponde a $F(z|1, n-10)$, obtemos os quantis apresentados na Tabela 3.8.

Tabela 3.8: Quantis da distribuição condicional e os limites superiores dos quantis para $\Im_2$

| **Valores de** $\alpha$ | 0.10 | 0.05 | 0.01 |
|---|---|---|---|
| $z_{1-\alpha}$ | 2.7056 | 3.8416 | 6.6353 |
| $z^{u}_{1-\alpha}$ | 2.7184 | 3.8654 | 6.7003 |

Considerando a abordagem condicional, concluímos que não se rejeita $H_{0,2}$, já que $\Im_{2,Obs} < z_{1-\alpha}$, para os níveis usuais de significância.

A Tabela 3.8 também apresenta os limites superiores para os quantis, $z^{u}_{1-\alpha}$, de $\Im_2$. Assumindo mais uma vez que $n^{\bullet} = 400$, podemos concluir, tal como seria expectável, que também não se rejeita $H_{0,2}$ para os níveis usuais de significância considerando a abordagem não condicional. Concluímos portanto que o segundo fator não é significativo.

### Interação entre os fatores

Para a interação entre os dois fatores temos

$$\mathbf{A_3} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{2} & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & -\frac{1}{2} & 0 & 0 & -\frac{1}{2} & \frac{1}{2} & 0 & 0 \\ -\frac{1}{2\sqrt{3}} & \frac{1}{2\sqrt{3}} & 0 & 0 & \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} & 0 & 0 & -\frac{1}{2\sqrt{3}} & \frac{1}{2\sqrt{3}} \\ -\frac{1}{\sqrt{15}} & \frac{1}{\sqrt{15}} & \frac{3}{2\sqrt{15}} & -\frac{3}{2\sqrt{15}} & -\frac{1}{\sqrt{15}} & \frac{1}{\sqrt{15}} & \frac{3}{2\sqrt{15}} & -\frac{3}{2\sqrt{15}} & -\frac{1}{\sqrt{15}} & \frac{1}{\sqrt{15}} \end{bmatrix},$$

com $g_3 = car(\mathbf{A_3}) = 4$.

Portanto para o numerador da estatística obtemos

$$S_3 = (\mathbf{A_3 y_\bullet})' \left(\mathbf{A_3 D A_3}'\right)^{-1} (\mathbf{A_3 y_\bullet}) = 0.0030.$$

Assim, o valor observado da estatística $\Im_{3,Obs}$, é dado por

$$\Im_{3,Obs} = \frac{60361}{4} \frac{0.0030}{14094449.4217} = 3.3605e^{-06}. \tag{3.4.39}$$

Se considerarmos a distribuição condicional de $\Im_3$, obtemos os quantis apresentados na Tabela 3.7. Podemos concluir que não se rejeita $H_{0,3}$ para os níveis usuais de significância ($\Im_{3,Obs} < z_{1-\alpha}$).

Assumindo uma dimensão mínima global de $n^\bullet = 400$ e através dos limites superiores dos quantis da distribuição não condicional de $\Im_3$, apresentados igualmente na Tabela 3.7, concluímos igualmente que não se rejeita $H_{0,3}$ para os níveis usuais de significância. Portanto a interação entre os dois fatores não é significativa.

### Discussão dos resultados

Através desta aplicação podemos concluir que:

- a interação entre os dois fatores não é significativa;
- em relação ao primeiro fator este é significativo, logo existem diferenças entre as médias das idades dos desempregados pelo menos entre duas das regiões de Portugal continental;
- o segundo fator não é significativo, o que significa que a idade dos desempregados não é significativamente diferente para os dois géneros;

- neste caso não foram tomadas decisões contrárias quando comparados os resultados através das duas abordagens. Tal facto deve-se, muito provavelmente aos elevados valores de $n_1, \cdots, n_{10}$.

## 3.5 Um estudo com simulações

Nesta secção, realizamos um estudo através de simulações para ilustrar a metodologia proposta e analisar a percentagem de falsas rejeições que podem ser evitadas considerando a nossa abordagem, ver Nunes et. al. (2019b).

Com este estudo tentamos "replicar" a situação descrita nas aplicações anteriormente apresentadas, considerando o caso de um modelo com dois fatores de efeitos fixos, o primeiro fator (com dois níveis) e um segundo fator (com quatro níveis). As frequências para cada uma das classes de idades, para os 8 diferentes tratamentos, foram gerados segundo distribuições uniformes no intervalo $[50; 700]$.

Foi obtido o vetor das médias amostrais e as somas dos quadrados. Em seguida, calcularam-se os valores observados da estatística de teste para os dois fatores e interação, bem como os respetivos $p$-values, considerando as abordagens condicional e não condicional.

Assumimos como dimensões mínimas globais $n^\bullet = 10$, $n^\bullet = 30$ e $n^\bullet = 50$.

A Tabela 3.9 mostra os resultados obtidos. Os valores a negrito correspondem aos casos em que a hipótese nula é rejeitada para uma significância de $5\%$.

Para uma dimensão mínima global $n^\bullet = 10$ concluímos que é tomada uma decisão contrária considerando as duas abordagens (rejeita-se a hipótese considerando a abordagem clássica e não se rejeita considerando a abordagem proposta), para o segundo fator e interação.

Tabela 3.9: Valores dos $p$-values considerando a abordagem condicional e não condicional

| | | Abordagem condicional ($p$-value) | Abordagem não condicional ($p$-value) | | |
|---|---|---|---|---|---|
| | $\Im_{j,Obs}$ | $n = 7680$ | $n^\bullet = 10$ | $n^\bullet = 30$ | $n^\bullet = 50$ |
| 1º fator | 1.3558 | 0.2443 | 0.3644 | 0.2567 | 0.2503 |
| 2º fator | 16.9934 | **<0.001** | 0.0561 | **<0.001** | **<0.001** |
| Interação | 7.8342 | **<0.001** | 0.1153 | **<0.001** | **<0.001** |

Calculou-se o número de rejeições, em 1000 simulações, considerando ambas as abordagens, bem como a percentagem de falsas rejeições que podem ser evitadas quando a metodologia proposta é considerada. Através dos resultados apresentados na Tabela 3.10 concluímos que a proporção de falsas rejeições é muito elevada para $n^\bullet = 10$ (superior a $50\%$ para os 2 fatores e interação), sendo menor para valores mais altos de $n^{\bullet}$.

Tabela 3.10: Número de rejeições e percentagens de falsas rejeições, em 1000 simulações.

| | Abordagem condicional (número de rejeições) | Abordagem não condicional (número de rejeições) | | | $\%$ falsas rejeições | | |
|---|---|---|---|---|---|---|---|
| | | $n^{\bullet}=10$ | $n^{\bullet}=30$ | $n^{\bullet}=50$ | $n^{\bullet}=10$ | $n^{\bullet}=30$ | $n^{\bullet}=50$ |
| 1º fator | 645 | 301 | 623 | 635 | 53.3$\%$ | 3.4$\%$ | 1.6$\%$ |
| 2º fator | 923 | 355 | 905 | 913 | 61.5$\%$ | 2.0$\%$ | 1.1$\%$ |
| Interação | 927 | 354 | 913 | 920 | 61.8$\%$ | 1.5$\%$ | 0.8$\%$ |

Os resultados obtidos através deste estudo corroboram com os nossos comentários anteriores sobre a relevância da nossa abordagem em evitar falsas rejeições.

Assim, sugerimos o uso desta abordagem sempre que as dimensões da amostras não são conhecidas à partida.

## 3.6 Conclusões

Neste capítulo apresentámos uma abordagem que propõe considerar a dimensão das amostras como variáveis aleatórias independentes, sempre que estas não são conhecidas à partida.

Esta abordagem foi desenvolvida considerando modelos paramétricos de efeitos fixos com mais do que um fator.

Como distribuições para as dimensões das amostras foram consideradas as distribuições Poisson, Binomial, Geométrica e Binomial Negativa.

Apresentámos duas aplicações com dados reais e uma aplicação considerando dados simulados. No caso da primeira aplicação foi considerada a incidência de desempregados nos quatro países da União Europeia que solicitaram o apoio do FMI, antes e durante a crise económica. Na segunda aplicação recorreu-se ao número de desempregados nas várias regiões de Portugal continental, por género.

Através destas concluímos que a inferência depende da abordagem utilizada, podendo a abordagem clássica conduzir-nos a falsas rejeições (rejeitarmos a hipótese considerando a abordagem clássica e não rejeitarmos ao considerar a abordagem proposta).

Tal situação é mais frequente sempre que não se consideram amostras de dimensão elevada (Aplicação 3.4.1), já que com aumento da dimensão da amostra as duas abordagens convergem para a mesma decisão (Aplicação 3.4.2), ver e.g. Capistrano et al. (2015).

Uma vez que consideramos a nossa abordagem mais realista, propomos a utilização desta sempre que as dimensões das amostras são desconhecida à partida.

# Capítulo 4

# Modelos mistos com amostras de dimensão aleatória

## 4.1 Introdução

No capítulo anterior abordámos situações em que as dimensões das amostras podem não ser conhecidas à partida. Vimos que neste caso é aconselhável considerar essas dimensões como realizações, $n_1, \cdots, n_m$, de variáveis aleatória independentes, $N_1, \cdots, N_m$.

O desenvolvimento desta abordagem foi apresentado no Capítulo 3 deste trabalho, considerando modelos de efeitos fixos, onde foram considerados três situações distintas para as dimensões das amostras:

- situações em que a ocorrência das observações corresponde a processos de contagem, em que $N_i \sim P(\lambda_i),\ i = 1, \cdots, m$;
- situações em que existe um limite superior para a dimensão das amostras que pode não ser atingido porque podem ocorrer falhas de observações, então $N_i \sim B(r_i, 1-p),\ i = 1, \cdots, m,$ em que $p$ corresponde à probabilidade de falha;
- situações em que o número de observações corresponde ao número de ocorrências até ao primeiro sucesso, então $N_i \sim Geo(q), i = 1, \cdots, m$ e $N = \sum_{i=1}^{m} \sim BN(m, q),$ em que $q$ corresponde à probabilidade de sucesso.

No presente capítulo iremos estender esta abordagem a modelos mistos. A utilização dos modelos mistos no caso em que as dimensões das amostras são aleatórios foi tida em conta em Capistrano (2015) e mais recentemente em Nunes et al. (2019c) onde a formulação do modelo foi feita recorrendo a uma classe mais ampla de modelos, as designadas extensões $L$. Estas foram desenvolvidas por Ferreira et al. (2009) e Moreira et al. (2009), por forma a lidar com a falta de ortogonalidade, originadas por situações de desequilíbrio. Nestes trabalhos (Capistrano, 2015 e Nunes et al. 2019c) foi realizada inferência apenas para a parte aleatória do modelo.

Iremos agora continuar o estudo dos modelos mistos, realizando inferência quer para a parte fixa quer para a parte aleatória do modelo, assumindo que estamos perante situações de estabilidade. Dizemos que temos estabilidade quando a estatística de teste tem a mesma distribuição, quer para parte de efeitos fixos quer para a parte de efeitos aleatórios do modelo, quando a hipótese nula se verifica, ver Ferreira (2006). Em Nunes et al. (2019b), um trabalho em co-autoria com o autor da presente tese, a formulação do modelo já foi obtida considerando situações de estabilidade.

Consideraremos esta abordagem para o caso em que se tem uma dimensão mínima para cada amostra, situação diferente da considerada no capítulo anterior em que se assumiu a existência de uma dimensão mínima global para as amostras.

Assumiremos que os $m$ tratamentos correspondem a tratamentos de um modelo mistos com estrutura ortogonal onde $j = 1, \cdots, \underline{w}$, corresponde à parte de efeitos fixos e $j = \underline{w} + 1, \cdots, w$, à parte de efeitos aleatórios do modelo.

Assim, com $\boldsymbol{\mu}_{\bullet}$ o vetor médio do vetor das médias amostrais, $\boldsymbol{Y}_{\bullet}$, tem-se

$$\boldsymbol{\mu}_{\bullet} \in \Omega = \boxplus_{j=1}^{w} \nabla_j,$$

com $\nabla_j, j = 1, \cdots, w$, um subespaço de $\Omega$.

Tem-se ainda

$$R^m = \boxplus_{j=1}^{w} \nabla_j \; .$$

Neste capítulo é apresentada a formulação do modelo misto, considerando situações de estabilidade, assumindo a normalidade. Mostramos ainda como se podem testar as hipóteses para os efeitos principais e para a interação entre os fatores. Iremos construir as estatísticas de teste e obter as suas distribuições condicional (assumindo as dimensões das amostras como fixas) e não condicional (assumindo as dimensões como aleatórias e considerando três situações distintas para as distribuições das dimensões das amostras). Serão mais uma vez apresentadas aplicações com dados reais. Terminamos o capítulo com um estudo com simulações por forma a comparar a nossa abordagem com a abordagem clássica, e uma subsecção com uma generalização de cruzamento de $u$ fatores.

## 4.2 Modelo e hipóteses

Nesta secção começaremos por assumir as dimensões das amostras como fixas, $n_1, \cdots, n_m$. O número total de observações será então $n = \sum_{i=1}^{m} n_i$. O vetor das observações terá como vetor médio

$$\boldsymbol{\mu} = D(\mathbf{1}_{n_1}, \cdots, \mathbf{1}_{n_m})\boldsymbol{\mu}_{\bullet},$$

onde $D(\mathbf{1}_{n_1}, \cdots, \mathbf{1}_{n_m})$ corresponde a uma matriz diagonal por bloco com elementos principais $\mathbf{1}_{n_1}, \cdots, \mathbf{1}_{n_m}$, em que $\mathbf{1}_{n_i}$ representa o vetor com todas as $n_i$, $i = 1, \cdots, m$, componentes iguais a $1$.

Se os $g_j$ vetores linha de uma matriz $\boldsymbol{A}_j$ constituem uma base ortonormada para $\triangledown_j$, a projeção ortogonal de um vetor $\boldsymbol{v}$ sobre $\triangledown_j$, $\boldsymbol{v}_{\triangledown_j}$, é nulo se e somente se

$$\boldsymbol{A}_j\boldsymbol{v} = \boldsymbol{0},\ j = 1, \cdots, w,$$

o que é equivalente a $\boldsymbol{v}$ pertencer ao complemento ortogonal de $\triangledown_j$, $\triangledown_j^{\perp}$.

Vamos agora considerar o vetor das observações, $\boldsymbol{Y}$, dado por

$$\boldsymbol{Y} \;=\; \boldsymbol{\mu} + D(\boldsymbol{1}_{n_1}, \cdots, \boldsymbol{1}_{n_m})\boldsymbol{Z} + \boldsymbol{e}, \tag{4.2.1}$$

onde

$$\boldsymbol{Z} = \sum_{j=\underline{w}+1}^{w} \boldsymbol{A}_j'\boldsymbol{Z}_j,$$

com $\boldsymbol{Z}_{\underline{w}+1}, \cdots, \boldsymbol{Z}_w$, normais e independentes com vetores médios nulos e matrizes de covariância $\gamma_j\boldsymbol{I}_{g_j}$,

$$\boldsymbol{Z}_j \sim N(\boldsymbol{0}; \gamma_j\boldsymbol{I}_{g_j}),\ j = \underline{w} + 1, \cdots, w.$$

Os vetores $\boldsymbol{Z}_j, j = \underline{w} + 1, \cdots, w$, são independentes de $\boldsymbol{e} \sim N(\boldsymbol{0}; \sigma^2\boldsymbol{I}_n)$.

Assim, o vetor das médias amostrais será dado por

$$\boldsymbol{Y}_\bullet = \boldsymbol{\mu}_\bullet + \boldsymbol{Z} + \boldsymbol{e}_\bullet, \tag{4.2.2}$$

onde as componentes de $\boldsymbol{e}_\bullet$ são as médias das componentes do vetor $\boldsymbol{e}$, correspondendo às diferentes amostras. Então, $\boldsymbol{e}_\bullet \sim N(\boldsymbol{0}; \sigma^2 D(\frac{1}{n_1}, \cdots, \frac{1}{n_m}))$.

Considerando

$$\boldsymbol{e}^\bullet = \boldsymbol{e} - D(\boldsymbol{1}_{n_1}, \cdots, \boldsymbol{1}_{n_m})\boldsymbol{e}_\bullet \sim N(\boldsymbol{0}; \sigma^2 D(\boldsymbol{K}_{n_1}, \cdots, \boldsymbol{K}_{n_m})),$$

com $\boldsymbol{K}_r = \boldsymbol{I}_r - \frac{1}{r}\boldsymbol{1}_r\boldsymbol{1}_r^t$, onde $\boldsymbol{I}_r$ representa a matriz identidade de ordem $r$, $\boldsymbol{Y}_\bullet$ será independente de $\boldsymbol{e}^\bullet$.

Teremos então

$$\boldsymbol{A}_j\boldsymbol{Y}_\bullet = \boldsymbol{A}_j\boldsymbol{\mu}_\bullet + \boldsymbol{A}_j\boldsymbol{e}_\bullet \sim N(\boldsymbol{A}_j\boldsymbol{\mu}_\bullet; \sigma^2\boldsymbol{A}_j D(\frac{1}{n1}, \cdots, \frac{1}{n_m})\boldsymbol{A}_j'),\ j = 1, \cdots, \underline{w},$$

e

$$\boldsymbol{A}_j\boldsymbol{Y}_\bullet = \boldsymbol{Z}_j + \boldsymbol{A}_j\boldsymbol{e}_\bullet \sim N(\boldsymbol{0}; \gamma_j\boldsymbol{I}_{g_j} + \sigma^2\boldsymbol{A}_j D(\frac{1}{n_1}, \cdots, \frac{1}{n_m})\boldsymbol{A}_j'),\ j = \underline{w} + 1, \cdots, w.$$

Assim, para a parte dos **efeitos fixos** do modelo temos as hipóteses

$$H_{f,0,j} : \boldsymbol{A}_j\boldsymbol{\mu}_\bullet = \boldsymbol{0}_{g_j},\ j = 1, \cdots, \underline{w},$$

que podem ser reescritas como

$$H_{f,0,j} : \boldsymbol{\mu}_\bullet \in \nabla_j^\perp;\ j = 1, \cdots, \underline{w}. \tag{4.2.3}$$

Enquanto que para a parte dos **efeitos aleatórios** se tem

$$H_{r,0,j} : pr(\boldsymbol{Z}_j = \boldsymbol{0}_{g_j}) = 1; j = \underline{w} + 1, \cdots, w$$

que são equivalentes a

$$H_{r,0,j} : \gamma_j = 0;\ j = \underline{w} + 1, \cdots, w. \tag{4.2.4}$$

É importante notar que se podem escrever todas as hipóteses (quer correspondam a hipóteses para a parte de efeitos fixos ou para a parte de efeitos aleatórios do modelo) do seguinte modo

$$H_{0,j} : pr(\boldsymbol{A}_j\boldsymbol{Y}_\bullet = \boldsymbol{A}_j\boldsymbol{e}_\bullet) = 1;\ j = 1, \cdots, w, \tag{4.2.5}$$

o que significa que estamos perante situações de estabilidade, ver Ferreira (2006).

## 4.3 Distribuições das dimensões das amostras

Nesta secção iremos considerar os três casos distintos para as distribuições das dimensões das amostras, também considerados no capítulo anterior.

Começaremos por assumir que a ocorrência das observações correspondem a processo de contagem, o que nos leva a considerar que as variáveis aleatórias, $N_1, \cdots, N_m$, seguem uma distribuição de Poisson. De seguida vamos lidar com situações em que podem ocorrer falhas de observações e existe um limite superior para as dimensões das amostras, induzindo-nos a considerar a distribuição Binomial. Por fim consideramos a recolha das observações até à ocorrência do primeiro contágio [$m$-ésimo contágio], levando-nos a considerar a distribuição geométrica [Binomial negativa] para as variáveis aleatórias.

Por forma a evitar a ocorrência de casos altamente desequilibrados, serão assumidas dimensões mínimas para cada uma das amostras, ver e.g. Nunes et al. (2019a). Em trabalhos anteriores, como em Mexia et al. (2011), Capistrano (2015) e Nunes et al. (2014; 2019c), foi considerada uma dimensão mínima global para as amostras (situação considerada também no capítulo anterior desta tese).

### 4.3.1 Processos de contagem: Distribuição de Poisson

Vamos assumir que a ocorrência das observações corresponde a um processo de contagem. Alguns exemplos concretos em que esta situação pode ocorrer foram mencionados no capitulo anterior.

Como vimos, nestas situações é aconselhado assumir que as dimensões das amostras, $N_1, \cdots, N_m$, seguem uma distribuição de Poisson com parâmetros, $\lambda_1, ..., \lambda_m$,

$$N_i \sim P(\lambda_i),\ i = 1, \cdots, m.$$

Entretanto $n = \sum_{i=1}^m n_i$ será uma realização da variável aleatória $N = \sum_{i=1}^m N_i$, e dada a independência dos $N_i, i = 1, \cdots, m$,

$$N \sim P(\lambda),$$

com $\lambda = \sum_{i=1}^m \lambda_i$. Para além disso o vetor $\boldsymbol{n} = (n_1, \cdots, n_m)'$ será uma realização do vetor $\boldsymbol{N} = (N_1, \cdots, N_m)'$.

Como mencionado anteriormente, assumiremos que temos uma dimensão mínima para cada uma das amostras, logo vamos assumir que $N_i \geqslant n_i^\bullet$, $i = 1, \cdots, m$. Nestes casos a dimensão mínima global será $n^\bullet = \sum_{i=1}^m n_i^\bullet$, tendo-se como vetor das dimensões mínimas $\boldsymbol{n}^\bullet = (n_1^\bullet, \cdots, n_m^\bullet)'$.

Assim, considerando os $m$ diferentes tratamentos e contemplando todas as possíveis partições de $n$ em $n_1, \cdots, n_m$, iremos considerar a probabilidade

$$\begin{aligned} p_{n^\bullet}(n) &= pr(N = n|\boldsymbol{N} \geqslant \boldsymbol{n}^\bullet) = \\ &= \sum_{n_1=n_1^\bullet}^{n-\sum_{i=2}^m n_i^\bullet} \cdots \sum_{n_l=n_l^\bullet}^{n-(\sum_{i=1}^{l-1} n_i^\bullet + \sum_{i=l+1}^m n_i^\bullet)} \cdots \\ &\cdots \sum_{n_m=n-\sum_{i=1}^{m-1} n_i}^{n-\sum_{i=1}^{m-1} n_i} pr(\boldsymbol{N} = \boldsymbol{n}|\boldsymbol{N} \geqslant \boldsymbol{n}^\bullet),\ n_i = n_i^\bullet, \cdots, i = 1, \cdots, m, \end{aligned} \quad (4.3.6)$$

onde, devido à independência dos $N_i, i = 1, \cdots, m$, se tem

$$\begin{aligned} pr(\boldsymbol{N} = \boldsymbol{n}|\boldsymbol{N} \geqslant \boldsymbol{n}^\bullet) &= \prod_{i=1}^m pr(N_i = n_i|N_i \geqslant n_i^\bullet) \\ &= \prod_{i=1}^m \frac{pr(N_i = n_i)}{pr(N_i \geqslant n_i^\bullet)} = \prod_{i=1}^m \frac{e^{-\lambda_i}(\lambda_i^{n_i}/n_i!)}{1 - \sum_{u_i=0}^{n_i^\bullet-1} e^{-\lambda_i}(\lambda_i^{u_i}/u_i!)} \\ &= \prod_{i=1}^m \frac{\lambda_i^{n_i}}{n_i!(e^{\lambda_i} - \sum_{u_i=0}^{n_i^\bullet-1} \frac{\lambda_i^{u_i}}{u_i!})},\ n_i = n_i^\bullet, \cdots,\ i = 1, \cdots, m \end{aligned} \quad (4.3.7)$$

### 4.3.2 Falhas de observações: Distribuição Binomial

Vamos agora assumir que existe um limite superior para a dimensão de cada uma das amostras, $r_i, i = 1, \cdots, m$. Estes limites superiores podem nem sempre ser atingidos, uma vez que podem ocorrer falhas de observações. Exemplos práticos a respeito destas situações já foram mencionadas no capítulo anterior.

Nestes casos consideramos que a escolha apropriada para a distribuição das dimensões das amostras é a distribuição Binomial. Assim assumiremos que as variáveis aleatórias, $N_1, \cdots, N_m$, seguem uma distribuição Binomial com parâmetros $r_1, \cdots, r_m$ e $1-p$ (onde $p$ representa a probabilidade de ocorrência de uma falha),

$$N_i \sim B(r_i, 1-p), i = 1, ..., m.$$

Além disso, de acordo com a reprodutibilidade da distribuição Binomial, temos

$$N = \sum_{i=1}^{m} N_i \sim B(r, 1-p),$$

com $r = \sum_{i=1}^{m} r_i$.

Assumindo que $N_i \geqslant n_i^\bullet, i = 1, \cdots, m$, e $n^\bullet = \sum_{i=1}^{m} n_i^\bullet$, temos $p_{n^\bullet}(n)$ definida por (4.3.6), com $n_i = n_i^\bullet, \cdots, r_i, i = 1, ..., m$, onde $pr(\boldsymbol{N} = \boldsymbol{n} | \boldsymbol{N} \geq \boldsymbol{n}^\bullet)$ é agora dada por

$$\begin{aligned} pr(\boldsymbol{N} = \boldsymbol{n} | \boldsymbol{N} \geq \boldsymbol{n}^\bullet) &= \prod_{i=1}^{m} pr(N_i = n_i | N_i \geq n_i^\bullet) \\ &= \prod_{i=1}^{m} \frac{\begin{pmatrix} r_i \\ n_i \end{pmatrix} (1-p)^{n_i} p^{r_i - n_i}}{\sum_{u_i = n_i^\bullet}^{r_i} pr(N_i = u_i)} \\ &= \prod_{i=1}^{m} \frac{\begin{pmatrix} r_i \\ n_i \end{pmatrix} (1-p)^{n_i} p^{r_i - n_i}}{\sum_{u_i = n_i^\bullet}^{r_i} \begin{pmatrix} r_i \\ u_i \end{pmatrix} (1-p)^{u_i} p^{r_i - u_i}}, n_i = n_i^\bullet, \cdots, r_i, i = 1, \cdots, m. \end{aligned} \tag{4.3.8}$$

### 4.3.3 Ambiente protegidos: Distribuição Geométrica e Binomial Negativa

Vamos agora assumir que as observações descrevem a ocorrência de "contágios". Exemplos concretos que podem justificar a ocorrência desta situações foram abordados no capítulo anterior.

Nestas situações assumiremos que as variáveis aleatórias, $N_1, \cdots, N_m$, seguem uma distribuição Geométrica de parâmetro $q$ ( onde $q$ corresponde à probabilidade de ocorrência de um sucesso),

$$N_i \sim Geo(q),\ i = 1, \cdots, m.$$

Dada a independência dos $N_i, i = 1, \cdots, m$, tem-se ainda

$$N = \sum_{i=1}^{m} N_i \sim BN(m, q).$$

Assumindo que $N_i \geqslant n_i^\bullet$, e $n^\bullet = \sum_{i=1}^m n_i^\bullet,\ i = 1, \cdots, m$, temos $p_{n^\bullet}(n)$ definida por (4.3.6), com $n_i = n_i^\bullet, \cdots,\ i = 1, \cdots, m$, onde $pr(\boldsymbol{N} = \boldsymbol{n}|\boldsymbol{N} \geq \boldsymbol{n}^\bullet)$ é dada por

$$\begin{aligned} pr(\boldsymbol{N} = \boldsymbol{n}|\boldsymbol{N} \geq \boldsymbol{n}^\bullet) &= \prod_{i=1}^{m} pr(N_i = n_i|N_i \geq n_i^\bullet) \\ &= \prod_{i=1}^{m} \frac{pr(N_i = n_i)}{pr(N_i \geq n_i^\bullet)} \\ &= \prod_{i=1}^{m} \frac{(1-q)^{n_i} q}{1 - pr(N_i < n_i^\bullet)} \\ &= \prod_{i=1}^{m} \frac{(1-q)^{n_i} q}{1 - \sum_{u_i=0}^{n_i^\bullet - 1} (1-q)^{u_i} q}, n_i = n_i^\bullet, \cdots, i = 1, \cdots, m. \end{aligned} \quad (4.3.9)$$

## 4.4 Distribuições das observações e estatísticas de teste

Nesta secção iremos obter as estatísticas de testes assim como as suas distribuições condicional e não condicional (assumindo as dimensões das amostras como aleatórias).

A soma dos quadrados para testar $H_{0,j}, j = 1, \cdots, w$, será dada por, ver e.g. Mexia (1990),

$$S_j = (\boldsymbol{A}_j \boldsymbol{Y}_\bullet)' \left( \boldsymbol{A}_j D \left( \frac{1}{n_1}, \cdots, \frac{1}{n_m} \right) \boldsymbol{A}_j')^{-1} \right) (\boldsymbol{A}_j \boldsymbol{Y}_\bullet),\ j = 1, \cdots, w. \quad (4.4.10)$$

Quando $H_{0,j}\ \ j = 1, \cdots, w$, se verifica, $S_j$ será o produto de $\sigma^2$ por um qui-quadrado central com $g_j$ graus de liberdade, $S_j \sim \sigma^2 \chi^2_{g_j}, j = 1, \cdots, w$.

No caso dos **efeitos fixos**, quando $H_{0,j},\ j = 1, \cdots, \underline{w}$, não se verifica, $S_j$ será o produto de $\sigma^2$ por um qui-quadrado não central com $g_j$ graus de liberdade e parâmetro de não centralidade

$$\delta_j = \frac{1}{\sigma^2}(\boldsymbol{A_j\mu_\bullet})' \left(\boldsymbol{A_j} D\left(\frac{1}{n_1}, ..., \frac{1}{n_m}\right)\boldsymbol{A}_j^{'}\right)^{-1}(\boldsymbol{A_j\mu_\bullet}),\ j = 1, \cdots, \underline{w},$$

$S_j \sim \sigma^2\chi^2_{g_j,\delta_j},\ j = 1, \cdots, \underline{w}$.

Por outro lado, no caso dos **efeitos aleatórios**, quando $H_{0,j}, j = \underline{w} + 1, \cdots, w$, não se verifica, tem-se

$$S_j \sim (\gamma_j + \sigma^2)\chi^2_{g_j},\ j = \underline{w} + 1, \cdots, w.$$

Visto que $\boldsymbol{Y_\bullet}$ é independente de $\boldsymbol{e}^\bullet$, $S_j,\ j = 1, \cdots, w$, será independente de

$$S = \boldsymbol{e}^{\bullet t} D(\boldsymbol{K}_{n_1}, \cdots, \boldsymbol{K}_{n_m})\boldsymbol{e}^\bullet, \tag{4.4.11}$$

em que $S \sim \sigma^2\chi^2_{n-m}$, ver e.g. Khuri et al. (1998) e Searle et al. (1992).

Assim, quando $N = n$ e $H_{0,j},\ j = 1, \cdots, w$, se verifica a distribuição condicional da estatística

$$\Im_j = \frac{n-m}{g_j}\frac{S_j}{S},\ j = 1, \cdots, w, \tag{4.4.12}$$

será uma $F$ central com $g_j,\ j = 1, \cdots, w$ e $n - m$ graus de liberdade, $F(.\mid g_j, n-m)$. Esta estatística de teste é comum quer para a parte fixa quer para a parte aleatória do modelo.

É fácil ver que no caso dos **efeitos fixos**, quando $H_{0,j},\ j = 1, \cdots, \underline{w}$, não se verifica, teremos uma distribuição $F$ não central com $g_j$ e $n - m$ graus de liberdade e parâmetro de não centralidade $\delta_j,\ F(.\mid g_j, n-m, \delta_j)$. A potência do teste $F$ aumenta com o parâmetro de não centralidade $\delta_j,\ j = 1, \cdots, \underline{w}$, pelas propriedades das distribuições $F$ não centrais, apresentadas na secção 2.2.2 deste trabalho (vem também Mexia, 1990).

Por outro lado, no caso dos **efeitos aleatórios**, quando $H_{0,j}, j = \underline{w} + 1, \cdots, w$, não se verifica, $\Im_j$ será o produto de $\frac{\gamma_j+\sigma^2}{\sigma^2}$ por uma variável aleatória com distribuição $F$ central com $g_j$ e $n-m$ graus de liberdade. Neste caso a potência do teste aumenta com $\gamma_j,\ j = \overline{w} + 1, \cdots, w$, ver e.g Lehmann and Romano (2005).

### 4.4.1 Distribuição não condicional da estatística

Iremos agora descondicionar os resultados obtidos anteriormente, assumindo que as dimensões das amostras são aleatórias, considerando as três distribuições apresentadas na secção 4.3.

Quando assumimos as dimensões das amostras como aleatórias focamo-nos no caso em que $H_{0,j},\ j = 1, ..., w$, se verifica. Acabámos de ver que nesse caso, se tem como distribuição condicional da estatística

$$\Im_j \sim F(.\mid g_j, n-m),\ j = 1, ..., w.$$

### Processos de Contagem

Assumindo que $N_i \sim P(\lambda_i),\ i = 1, \cdots, m$, a distribuição não condicional de $\Im_j$, quando $H_{0,j},\ j = 1, \cdots, w$, se verifica, será então dada por

$$\begin{aligned} \overline{\overline{F}}_j(z) &= pr(\Im_j \leqslant z) \\ &= \sum_{n=n^\bullet}^{\infty} pr(N = n|\boldsymbol{N} \geqslant \boldsymbol{n}^\bullet) pr(\Im_j \leqslant z|N = n) \\ &= \sum_{n=n^\bullet}^{\infty} p_{n^\bullet}(n) F(z \mid g_j, n-m), j = 1, ..., w, \end{aligned} \quad (4.4.13)$$

em que $n^\bullet$ corresponde à dimensão mínima global das amostras, $\boldsymbol{n}^\bullet = (n_1^\bullet, \cdots, n_m^\bullet)'$ ao vetor das dimensões mínimos e $p_{n^\bullet}(n)$ é dada por (4.3.6), com a probabilidade $pr(\boldsymbol{N} = \boldsymbol{n}|\boldsymbol{N} \geq \boldsymbol{n}^\bullet)$ definido por (4.3.7).

### Falhas de Observações

Assumindo agora que $N_i \sim B(r_i, 1-p),\ i = 1, \cdots, m$, a distribuição não condicional de $\Im_j$, quando $H_{0,j},\ j = 1, \cdots, w$, se verifica, será dada por

$$\begin{aligned} \overline{\overline{F}}_j(z) &= pr(\Im_j \leqslant z) \\ &= \sum_{n=n^\bullet}^{r} p_{n^\bullet}(n) F(z \mid g_j, n-m), j = 1, ..., w, \end{aligned} \quad (4.4.14)$$

em que $p_{n^\bullet}(n)$ é dada por (4.3.6), com a probabilidade $pr(\boldsymbol{N} = \boldsymbol{n}|\boldsymbol{N} \geq \boldsymbol{n}^\bullet)$ definida por (4.3.8).

### Ambientes Protegidos

Considerando $N_i \sim Geo(q), i = 1, \cdots, m$, ( consequentemente $N = \sum_{i=1}^{m} N_i \sim BN(m, q)$) a distribuição não condicional de $\Im_j$, quando $H_{0,j},\ j = 1, \cdots, w$, se verifica, será obtida por

$$\begin{aligned} \overline{\overline{F}}_j(z) &= pr(\Im_j \leqslant z) \\ &= \sum_{n=n^\bullet}^{\infty} p_{n^\bullet}(n) F(z \mid g_j, n-m), j = 1, ..., w, \end{aligned} \quad (4.4.15)$$

em que $p_{n^\bullet}(n)$ é dada por (4.3.6), com a probabilidade $pr(\boldsymbol{N} = \boldsymbol{n}|\boldsymbol{N} \geq \boldsymbol{n}^\bullet)$ definida por (4.3.9).

**Erro de Truncatura**

Para o caso das distribuições de Poisson e Geométrica (ou Binomial negativa) a expressão da distribuição não condicional da estatística apresenta uma série infinita (expressões (4.4.13) e (4.4.15)). De seguida exemplificamos de que forma podemos truncar essa série.

É fácil chegar à conclusão de que, para $\overline{\overline{F}}_j(z), j = 1, \cdots, w$, se tem o "encaixe", ver e.g. Nunes et al. (2019a),

$$\overline{\overline{F}}_{j,\overline{n}}(z) \leq \overline{\overline{F}}_j(z) \leq \overline{\overline{F}}^*_{j,\overline{n}}(z), \tag{4.4.16}$$

com

$$\overline{\overline{F}}_{j,\overline{n}}(z) = \sum_{n=n^\bullet}^{\overline{n}} p_{n^\bullet}(n) F_j(z|g_j, n-m) \tag{4.4.17}$$

e

$$\overline{\overline{F}}^*_{j,\overline{n}}(z) = \overline{\overline{F}}_{j,\overline{n}}(z) + \sum_{n=\overline{n}+1}^{\infty} p_{n^\bullet}(n), \tag{4.4.18}$$

onde

$$\sum_{n=\overline{n}+1}^{\infty} p_{n^\bullet}(n) = 1 - \sum_{n=n^\bullet}^{\overline{n}} p_{n^\bullet}(n). \tag{4.4.19}$$

O valor de $\overline{n}$ pode ser encarado como o limite superior necessário por forma a controlar o erro de truncatura da distribuição não condicional da estatística, $\overline{\overline{F}}_j(z), j = 1, \cdots, w.$

Para calcular os quantis da distribuição não condicional da estatística iremos truncar a série em (4.4.13) ou em (4.4.15) no caso de usarmos a distribuição Geométrica calculando $\overline{n}$ tal que

$$1 - \sum_{n=n^\bullet}^{\overline{n}} p_{n^\bullet}(n) < \varepsilon, \tag{4.4.20}$$

com $\varepsilon > 0$ pequeno. Isto significa que a série é truncada não considerando os termos para os quais $n > \overline{n}$.

## 4.5 Aplicações com dados reais

Nesta secção, a metodológica proposta é aplicada a bases de dados reais, mais uma vez provenientes do PORDATA-Base de dados de Portugal contemporâneo.

Iremos considerar duas aplicações, uma considerando o caso equilibrado e outra para o caso não equilibrado. Em ambas as aplicações vamos aplicar a nossa metodologia considerando duas das distribuições para as variáveis $N_i, i = 1, \cdots, m,$ iremos considerar $N_i \sim P(\lambda_i),\ i = 1, \cdots, m,$ e $N_i \sim B(r_i, 1-p),\ i = 1, \cdots, m.$

Devido à elevada dimensão das amostras, trabalhamos com distribuições normais assimptóticas para assegurar a utilização dos testes $F$, ver Ito (1980).

Todos os procedimentos computacionais, nomeadamente os cálculos dos quantis das distribuições, os limites superiores, $\overline{n}$, por forma a controlar o erro de truncatura da distribuição não condicional para o caso em que $N_i \sim P(\lambda_i),\ i = 1, \cdots, m,$ e a estimativa dos $r_i, i = 1, \cdots, m,$ quando $N_i \sim B(r_i, 1-p),\ i = 1, \cdots, m,$ foram realizadas através do *software R*.

### 4.5.1 O desemprego na Europa antes e durante a crise económica

Nesta primeira aplicação vamos considerar a incidência do desemprego na União Europeia. Os dados provêm do PORDATA e a informação diz respeito à idade das pessoas desempregadas, Base de dados de Portugal contemporâneos (2016).

Consideramos um modelo misto com um fator de efeitos fixos e um fator de efeitos aleatórios. O fator de efeitos fixos corresponde ao *período de tempo*, e tem dois níveis: *2006 - 2007* (antes da crise financeira) e *2012 - 2013* (durante da crise financeira) o fator de efeitos aleatórios corresponde aos países da União Europeia. Devido ao grande número de países cuja informação se encontra disponíveis na base de dados, recorremos ao método de amostragem aleatória simples para selecionar uma amostra de quatro países, ver Nunes et al. (2019b). Tem-se então $m = 4 \times 2 = 8$ diferentes tratamentos. A Tabela 4.1 apresenta a média amostral da idade dos desempregados para cada país selecionado e período de tempo.

Tabela 4.1: médias amostrais das idades

| | | Período de tempo (fator de efeitos fixos) | |
|---|---|---|---|
| | | 2006-2007 | 2012-2013 |
| Países (fator de efeitos aleatórios) | Eslováquia | 35.47186 | 33.96895 |
| | Espanha | 35.33543 | 35.88430 |
| | Grécia | 30.58468 | 34.92408 |
| | República Checa | 34.85264 | 35.03557 |

Iremos testar as hipóteses definidas por (4.2.3) e (4.2.4) que como vimos podem ser reescritas como

$$H_{0,j} : pr(\boldsymbol{A}_j\boldsymbol{Y}_\bullet = \boldsymbol{A}_j\boldsymbol{e}_\bullet) = 1,\ j = 1, 2, 3.$$

Estas correspondem às hipóteses de ausência dos efeitos e interação entre os dois fatores.

Nesta aplicação consideramos o caso equilibrado (robusto na presença da não-normalidade e heteroscedaticidade, ver Scheffé (1959, Capítulo 10)), onde $n_i = 6, i = 1, \cdots, 8$, correspondem ao número das diferentes classes das idades das pessoas desempregadas. Assim, $n = \sum_{i=1}^{8} 6 = 48$.

Dado $N = n$, quando $H_{0,j},\ j = 1, 2, 3$, se verifica, tem-se como distribuição condicional

$$\Im_j = \frac{n-m}{g_j}\frac{S_j}{S}, \sim F(z|g_j, 40),\ j = 1, 2, 3,$$

com $g_j = car(\mathbf{A}_j)$.

Geralmente as análises começam com um teste à interação entre os fatores, no entanto, como referido no capítulo anterior, iremos realizar os testes à interação mas também aos efeitos principais dos fatores.

É importante referir que a interação pertence à parte dos efeitos aleatórios do modelo.

**Fator de efeitos fixos**

Para o fator de efeitos fixos, que corresponde aos *períodos de tempo*, temos

$$\mathbf{A_1} = \begin{bmatrix} -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \end{bmatrix},$$

com $g_1 = car(\mathbf{A}_1) = 1$.

Obtemos

$$S_1 = (\mathbf{A}_1\mathbf{y}_\bullet)' \left(\mathbf{A}_1 D\left(\frac{1}{6}, \cdots, \frac{1}{6}\right)\mathbf{A_1}'\right)^{-1}(\mathbf{A_1}\mathbf{y}_\bullet) = 9.5495,$$

para o numerador da estatística $\Im_1$, onde o vetor $\mathbf{y}_\bullet$ tem como componentes as médias amostrais apresentada na Tabela 4.1.

O denominador da estatística, quando $N_i = n_i,\ i = 1, \cdots, 8$, corresponde ao produto de $\sigma^2$ por um qui-quadrado central com $g(n) = 40$ graus de liberdade, $S \sim \sigma^2\chi^2_{40}$. Neste caso obtivemos

$$S = 143.4496.$$

Portanto, o valor observado da estatística, $\Im_{1,Obs}$, é dado por

$$\Im_{1,Obs} = \frac{40}{1}\frac{9.5495}{143.4496} = 2.6628. \tag{4.5.21}$$

Considerando a distribuição condicional de $\Im_1$, que corresponde a $F(z|1,40)$, obtemos os quantis apresentados na Tabela 4.2. Assim, considerando a abordagem condicional, podemos concluir que não se rejeita $H_{0,1}$ para os níveis usuais de significância.

Tabela 4.2: Quantis da distribuição condicional de $\Im_1$

| **Valores de** $\alpha$ | 0.10 | 0.05 | 0.01 |
|---|---|---|---|
| $z_{1-\alpha}$ | 2.8354 | 4.0847 | 7.3141 |

**Fator de efeitos aleatórios**

Para o fator de efeitos aleatórios, que corresponde aos países (membros da União Europeia) temos

$$\mathbf{A_2} = \begin{bmatrix} -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \\ -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \\ \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \end{bmatrix},$$

com $g_2 = car(\mathbf{A}_2) = 3$.

Para o numerador da estatística obtém-se

$$S_2 = (\mathbf{A}_2\mathbf{y}_\bullet)' \left(\mathbf{A}_2 D\left(\frac{1}{6},\cdots,\frac{1}{6}\right)\mathbf{A_2}'\right)^{-1}(\mathbf{A_2 y_\bullet}) = 54.2951.$$

Assim, o valor observado da estatística, $\Im_{2,Obs}$, é dado por

$$\Im_{2,Obs} = \frac{40}{1}\frac{54.2951}{143.4496} = 5.0466. \tag{4.5.22}$$

Se usamos a distribuição condicional de $\Im_2$, que corresponde a $F(z|3,40)$, obteremos os quantis apresentados na Tabela 4.3. Logo, considerando a distribuição condicional, podemos concluir que se rejeita $H_{0,2}$ para os níveis usuais de significância.

Tabela 4.3: Quantis da distribuição condicional de $\Im_2$ e de $\Im_3$

| **Valores de** $\alpha$ | 0.10 | 0.05 | 0.01 |
|---|---|---|---|
| $z_{1-\alpha}$ | 2.2261 | 2.8387 | 4.3126 |

**Interação entre os fatores**

Quanto à interação tem-se

$$\mathbf{A_3} = \begin{bmatrix} \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \\ \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \\ -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \end{bmatrix},$$

com $g_1 = car(\mathbf{A_3}) = 3$.

Portanto para o numerador da estatística, obtemos

$$S_3 = (\mathbf{A}_3\mathbf{y_\bullet})' \left(\mathbf{A}_3 D\left(\frac{1}{6}, \cdots, \frac{1}{6}\right)\mathbf{A_3}'\right)^{-1} (\mathbf{A_3 y_\bullet}) = 54.7220,$$

sendo, o valor observado da estatística, $\Im_{3,Obs}$, dado por

$$\Im_{3,Obs} = \frac{40}{1}\frac{54.7220}{143.4496} = 5.0863. \qquad (4.5.23)$$

Considerando os quantis para a distribuição condicional apresentados na Tabela 4.3 concluímos que se rejeita $H_{0,3}$ para os níveis usuais de significância.

### 4.5.1.1 Processos de contagem

Vamos agora considerar que a ocorrência das observações corresponde a processos de contagem, o nos leva a assumir que $N_i \sim P(\lambda_i),\ i = 1, \cdots, m.$

Assumiremos que temos uma dimensão mínima para cada amostra, por exemplo que $N_i \geq 2,\ i = 1, \cdots, 8$, o que significa que $n_i^\bullet = 2, i = 1, \cdots, 8$. Assim, a dimensão mínima global será $n^\bullet = 16$. Neste caso a distribuição não condicional truncada de $\Im_j, j = 1, 2, 3,$ com $\overline{n}$ o limite superior por forma a controlar o erro de truncatura, será dada por

$$\overline{\overline{F}}_{j,\overline{n}}(z) \;=\; \sum_{n=16}^{\overline{n}} p_{n^\bullet}(n) F(z|g_j, n-8),\ j = 1, 2, 3, \qquad (4.5.24)$$

com

$$
\begin{aligned}
p_{\boldsymbol{n}^\bullet}(n) &= pr(N=n|\boldsymbol{N}\geq \boldsymbol{n}^\bullet)\\
&= \sum_{n_1=2}^{n-14}\cdots\sum_{n_7=2}^{n-(\sum_{i=1}^{6}n_i+2)}\sum_{n_8=n-\sum_{i=1}^{7}n_i}^{n-\sum_{i=1}^{7}n_i} pr(\boldsymbol{N}=\boldsymbol{n}|\boldsymbol{N}\geqslant \boldsymbol{n}^\bullet)\\
&= \sum_{n_1=2}^{n-14}\cdots\sum_{n_7=2}^{n-(\sum_{i=1}^{6}n_i+2)}\sum_{n_8=n-\sum_{i=1}^{7}n_i}^{n-\sum_{i=1}^{7}n_i}\prod_{i=1}^{8}\frac{\lambda_i^{n_i}}{n_i!(e^{\lambda_i}-\sum_{u_i=0}^{1}\frac{\lambda_i^{u_i}}{u_i})}. \quad (4.5.25)
\end{aligned}
$$

### Estimação dos parâmetros da Poisson

Os parâmetros $\lambda_i, i=1,\cdots,8,$ não são conhecidos então vamos mostrar de seguida como podemos estimar estes parâmetros. Em Nunes et al. (2014) os autores mostraram que a distribuição não condicional cresce com $\lambda_i,\ 1,\cdots,m$. Assim, os correspondentes quantis decrescem e utilizaremos limites inferiores para estes parâmetros. Os limites inferiores corresponderão aos valores mínimos de $\lambda_i,\ i=1,\cdots,m,$ tais que

$$
e^{-\lambda_i}\frac{\lambda_i^{n_i}}{n_i!}=\alpha,\ i=1,\cdots,m,
$$

em que $\alpha$ corresponde aos níveis usuais de significância, ver também Nunes et al. (2019a) .

Com $n_i=6, i=1,\cdots 8,$ obtivemos os limites inferiores de $\lambda_i, \lambda_{i,\alpha}, i=1,\cdots,8,$ apresentados na Tabela 4.4. Assumindo estes limites inferiores para $\lambda_i, i=1,\cdots,8,$ obtivemos também o valor mínimo, $\overline{n}$, por forma a que

$$
1-\sum_{n=16}^{\overline{n}} pr(N=n|\boldsymbol{N}\geqslant \boldsymbol{n}^\bullet)<10^{-4}.
$$

Os valores obtidos também são apresentados na Tabela 4.4. Estes permitem-nos concluir que temos um bom controlo do erro de truncatura já que não são valores muito elevados.

Tabela 4.4: Limites inferiores para $\lambda_i, i=1,\cdots,8,$ e valor mínimo $\overline{n}$

| **Valores de** $\alpha$ | 0.10 | 0.05 | 0.01 |
|---|---|---|---|
| $\lambda_{i,\alpha}, i=1,\cdots,8$ | 3.9 | 3 | 1.9 |
| $\overline{n}$ | 55 | 47 | 37 |

### Fator de efeitos fixos

Os quantis, $z_{1-\alpha}^{t}$, para a probabilidade $1-\alpha$ da distribuição não condicional truncada definida por (4.5.24), são apresentados na Tabela 4.5.

Visto que $\Im_{1,Obs} = 2.6628 < z_{1-\alpha}^{t}$ podemos concluir que não se rejeita $H_{0,1}$ para os níveis usuais de significância. Portanto concluímos que o fator de efeitos fixos não é significativo.

Tabela 4.5: Quantis da distribuição não condicional truncada de $\Im_1$

| **Valores de $\alpha$** | 0.10 | 0.05 | 0.01 |
|---|---|---|---|
| $z_{1-\alpha}^{t}$ | 2.9185 | 4.3637 | 8.8353 |

### Fator de efeitos aleatórios

Através da interpretação dos valores dos quantis, $z_{1-\alpha}^{t}$, da distribuição não condicional truncada de $\Im_2$, apresentamos na Tabela 4.6 concluímos que se rejeita a hipótese para $\alpha = 0.1$ e $\alpha = 0.05$ e não se rejeita para 0.01 ($\Im_{2,Obs} = 5.0466 < z_{1-\alpha}^{t}$).

Logo estes resultados levam-nos a tomar uma decisão contrária à que tínhamos tomado através da abordagem condicional, para $\alpha = 0.01$.

Tabela 4.6: Quantis da não condicional truncada de $\Im_2$ e $\Im_3$

| **Valores de $\alpha$** | 0.10 | 0.05 | 0.01 |
|---|---|---|---|
| $z_{1-\alpha}^{t}$ | 2.3180 | 3.1110 | 5.5452 |

### Interação entre os fatores

Através dos quantis da distribuição não condicional truncada, apresentados na Tabela 4.6, concluímos que se rejeita $H_{0,3}$ para $\alpha = 0.1$ e $\alpha = 0.05$ e não se rejeita para 0.01 $(\Im_{3,Obs} = 5.0863 < z_{1-\alpha}^{t})$, o que significa que tomamos uma decisão contrária à tomada usando a distribuição condicional, para $\alpha = 0.01$.

#### 4.5.1.2 Falhas de observações

Vamos agora assumir que podem ocorrer falhas de observações, o que nos leva a considerar que

$$N_i \sim B(r_i, 1-p), i = 1, \cdots, 8,$$

onde $r_1, \cdots, r_m$, representam os limites superiores para cada uma das amostras e $p$ a probabilidade de ocorrência de uma falha (vamos assumir que esta probabilidade é conhecida através de estudos realizados anteriormente).

### Estimativa dos parâmetros $r_i, i = 1, \cdots, 8,$ da distribuição Binomial

Para estimar os $r_i, i = 1, \cdots, 8,$ iremos considerar a regra estabelecida em Nunes et al. (2019b).

Vamos portanto assumir que se tem uma probabilidade fixa, $q$, de tomar uma observação, sendo $r_i$ o menor valor por forma a que se verifique

$$pr(N_i \geq n_i^\bullet) \geq q \Leftrightarrow \sum_{n=n_i^\bullet}^{r_i} \begin{pmatrix} r_i \\ n \end{pmatrix} (1-p)^n p^{r_i - n} \geq q. \quad (4.5.26)$$

Para tal vamos continuar a assumir que $n_i^\bullet = 2, i = 1, \cdots, 8$ (portanto $n_i^\bullet = 16$) e que a probabilidade $q = 0.95$.

A Tabela seguinte apresenta o valor mínimo de $r_i, i = 1, \cdots, 8$, e de $r = \sum_{i=1}^{8} r_i$ por forma a que (4.5.26) se verifique, considerando diferentes valores para a probabilidade de ocorrência de uma falha, $p$.

Tabela 4.7: Valores mínimos de $r_i, i = 1, \cdots, 8$, para uma probabilidade $q = 0.95$

| **Valores de** $p$ | 0.1 | 0.2 | 0.3 | 0.4 |
|---|---|---|---|---|
| $r_i, i = 1, \cdots, 8$ | 3 | 4 | 5 | 6 |
| $r = \sum_{i=1}^{8}$ | 24 | 32 | 40 | 48 |

Os resultados mostram que se obtêm limites superior razoáveis para as dimensões das amostras.

Vamos assumir que se sabe, de estudos realizados anteriormente, que $p = 0.2$. Temos então, como distribuição não condicional da estatística,

$$\overline{\overline{F}}_{j,\overline{n}}(z) = \sum_{n=16}^{32} p_{\boldsymbol{n}^\bullet}(n) F(z|g_j, n-8),\ j = 1, 2, 3, \quad (4.5.27)$$

com

$$\begin{aligned} p_{\boldsymbol{n}^\bullet}(n) &= pr(N = n|\boldsymbol{N} \geq \boldsymbol{n}^\bullet) \\ &= \sum_{n_1=2}^{n-14} \cdots \sum_{n_7=2}^{n-(\sum_{i=1}^{6} n_i + 2)} \sum_{n_8=n-\sum_{i=1}^{7} n_i}^{n-\sum_{i=1}^{7} n_i} pr(\boldsymbol{N} = \boldsymbol{n}|\boldsymbol{N} \geqslant \boldsymbol{n}^\bullet) \\ &= \sum_{n_1=2}^{n-14} \cdots \sum_{n_7=2}^{n-(\sum_{i=1}^{6} n_i + 2)} \sum_{n_8=n-\sum_{i=1}^{7} n_i}^{n-\sum_{i=1}^{7} n_i} \prod_{i=1}^{8} \frac{\begin{pmatrix} 4 \\ n_i \end{pmatrix} (0.8)^{n_i} (0.2)^{4-n_i}}{\sum_{u_i=2}^{4} \begin{pmatrix} 4 \\ u_i \end{pmatrix} (0.8)^{u_i} (0.2)^{4-u_i}} \end{aligned} \quad (4.5.28)$$

### Fator de efeitos fixos

A Tabela seguinte apresenta os quantis da distribuição não condicional de $\Im_1$. Atendendo a estes resultados tomamos a decisão de não rejeitar $H_{0,1}$ para os níveis usuais de significância ($\Im_{1,Obs} = 2.6628 < z_{1-\alpha}^{b}$).

Tabela 4.8: Quantis da distribuição não condicional de $\Im_1$

| **Valores de** $\alpha$ | 0.10 | 0.05 | 0.01 |
|---|---|---|---|
| $z^b_{1-\alpha}$ | 3.0097 | 4.4194 | 8.3039 |

**Fator de efeitos aleatórios**

No que respeita aos quantis da distribuição não condicional de $\Im_2$ estes são apresentados na Tabela 4.9. Concluímos portanto que se rejeita $H_{0,2}$ para $\alpha = 0.10$ e $0.05$ e não se rejeita para $\alpha = 0.01$ ($\Im_{2,Obs} = 5.0446 < z^b_{1-\alpha}$).

Tal como para o caso da em que $N_i \sim P(\lambda_i),\ \ i = 1, \cdots, 8,$ tomamos uma decisão contrária para $\alpha = 0.01$ ao se considerar a abordagem clássica e a abordagem proposta.

Tabela 4.9: Quantis da distribuição não condicional de $\Im_2$ e $\Im_3$

| **Valores de** $\alpha$ | 0.10 | 0.05 | 0.01 |
|---|---|---|---|
| $z^b_{1-\alpha}$ | 2.2749 | 3.1655 | 5.1075 |

**Interação entre os fatores**

Quanto à interação, pela utilização da abordagem proposta rejeita-se a hipótese nula para $\alpha = 0.10$ e $0.05$ e não se rejeita para $\alpha = 0.01$ ($\Im_{3,Obs} = 5.0863 < z^b_{1-\alpha}$), ver Tabela 4.9.

Estamos mais uma vez perante uma decisão contrária para $\alpha = 0.01$.

#### 4.5.1.3 Discussão dos resultados

Os resultados que obtivemos através desta aplicação mostram que

- da abordagem clássica provêm quantis que são um pouco mais baixos que os obtidos através da abordagem não condicional, o que nos pode levar a tomar decisões contrárias quando usamos as duas abordagens. Neste caso, para a interação e fator de efeitos aleatórios, tomamos decisões contrárias considerando uma significância de 1$\%$;

- podemos concluir que a utilização da metodologia proposta, sendo mais realista quando as dimensões das amostras não são conhecidas, nos permite evitar a ocorrência de falsas rejeições. Neste caso permite-nos evitar a ocorrência de uma falsa rejeição, para $\alpha = 0.01$;

- em relação ao fator de efeitos fixos não é significativo, o que significa que a idade dos desempregados não é significativamente diferente para estes dois períodos.

### 4.5.2 O desemprego na região das Beiras por géneros.

Nesta secção a metodologia proposta é aplicada ao número de desempregados considerando os municípios da região das Beiras e Serra da Estrela, onde foram selecionados de forma aleatória 3 dos municípios existentes (Almeida, Belmonte, Celorico de Beira, Covilhã, Figueira de Castelo Rodrigo, Fornos de Algodras, Fundão, Gouveia, Guarda, Mêda, Manteigas, Pinhal, Sabugal, Seia e Trancoso).

Os dados usados são mais uma vez provenientes do PORDATA e a informação diz respeito à idade dos desempregados, Base de dados contemporâneos (2018).

Vamos considerar dois fatores: o fator de efeitos fixos, com 2 níveis, que corresponde ao género (*masculino e feminino*) e o fator de efeitos aleatórios, com 3 níveis, que corresponde aos municípios selecionados (*Manteigas*, *Sabugal*, *Gouveia*). Tem-se portanto $m = 3 \times 2 = 6$ diferentes tratamentos.

A Tabela 4.10 ilustra a média de idades dos desempregados e o número de desempregados, por município e género da pessoa.

Tabela 4.10: Médias amostrais e número de desempregados

| | | Média das idades | | Número de desempregados | |
|---|---|---|---|---|---|
| Fator de efeitos fixos | | Mulheres | Homens | Mulheres | Homens |
| | Manteigas | 40,2778 | 41,1905 | 108 | 84 |
| Fator de efeitos aleatório | Sabugal | 36,9519 | 36,7327 | 187 | 202 |
| | Gouveia | 40,5985 | 40,6061 | 401 | 363 |

Testaremos as hipóteses definidas por (4.2.3) e (4.2.5), que, como vimos, podem ser reescritas como

$$H_{0,j} : pr(\boldsymbol{A}_j\boldsymbol{Y}_\bullet = \boldsymbol{A}_j\boldsymbol{e}_\bullet) = 1,\ j = 1, 2, 3.$$

Dado $N = n$, quando $H_{0,j}$, $j = 1, 2, 3$, se verifica, a distribuição condicional de

$$\Im_j = \frac{n-m}{g_j}\frac{S_j}{S},\ j = 1, 2, 3,$$

será uma distribuição $F$ central, com $g_j = car(\mathbf{A}_j)$ e $g(n) = n - 6 = 1339$ graus de liberdade, $F(.|g_j, 1339)$.

Iremos agora considerar os testes para os efeitos principais dos fatores e para a interação entre os dois fatores.

**Fator de efeitos fixos**

Neste caso, para o primeiro fator temos

$$\mathbf{A_1} = \left[ \begin{array}{cccccc} -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \end{array} \right],$$

com $g_1 = car(\mathbf{A_1})$= 1. Assim, obtemos

$$\mathbf{A_1 y_\bullet} = \ 0.2862,$$

onde o vetor $\mathbf{y_\bullet}$ tem como componentes as médias amostrais apresentadas na Tabela 4.10.

Obtém-se ainda, para numerador da estatística

$$S_1 = (\mathbf{A_1 y_\bullet})' \left(\mathbf{A_1 D A_1}'\right)^{-1} (\mathbf{A_1 y_\bullet}) = 13.3879,$$

com $\mathbf{D} = D(\frac{1}{n_1}, \cdots, \frac{1}{n_6})$, onde $n_1, \cdots, n_6$ correspondem ao número de desempregados que consta da tabela 4.10.

O denominador da estatística é, quando $N_i = n_i,\ i = 1, \cdots, 6$, o produto de $\sigma^2$ por um qui-quadrado central com $1339$ graus de liberdade$S \sim \chi^2_{1339}$. Neste caso, obteve-se

$$S \ = 247701.7760. \qquad (4.5.29)$$

Assim, o valor observado da estatística, $\Im_{1,Obs}$, é dado por

$$\Im_{1,Obs} \ = \ \frac{1339}{1} \frac{13.3879}{247701.7760} = 0.0724 \qquad (4.5.30)$$

Os quantis, $z_{1-\alpha}$, da distribuição condicional da $\Im_1$, que corresponde a $F(z|1, 1339)$, são apresentados na Tabela 4.11.

Visto que $\Im_{1,Obs} < z_{1-\alpha}$, podemos concluir que não se rejeita $H_{0,1}$ para $\alpha = 0.05$ e 0.01.

Tabela 4.11: Quantis da distribuição condicional de $\Im_1$

| **Valores de** $\alpha$ | 0.05 | 0.01 |
|---|---|---|
| $z_{1-\alpha}$ | 3.8484 | 6.6539 |

**Fator de efeitos aleatórios**

Para o segundo fator temos

$$\mathbf{A_2} = \left[ \begin{array}{cccccc} -\frac{1}{2} & -\frac{1}{2} & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2\sqrt{3}} & \frac{1}{2\sqrt{3}} & -\frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} & \frac{1}{2\sqrt{3}} & \frac{1}{2\sqrt{3}} \end{array} \right],$$

e $g_2 = car(\mathbf{A}_2) = 2.$

Assim obtemos

$$S_2 = (\mathbf{A}_2\mathbf{y}_\bullet)' \left(\mathbf{A}_2 (D) \mathbf{A_2}'\right)^{-1} (\mathbf{A_2 y_\bullet}) = 3956.2153$$

para o numerador da estatística de $\Im_2$.

Portanto, o valor observado da estatística, $\Im_{2,Obs}$, é dado por

$$\Im_{2,Obs} \quad = \quad \frac{1339}{2}\frac{3956.2153}{247701.7760} = 10.6930. \qquad (4.5.31)$$

Considerando a distribuição condicional de $\Im_2$, que corresponde a $F(z|2, 1339)$, obtêm-se os quantis apresentados na Tabela 4.12.

Tabela 4.12: Quantis da distribuição condicional de $\Im_2$ e $\Im_3$

| **Valores de $\alpha$** | 0.05 | 0.01 |
|---|---|---|
| $z_{1-\alpha}$ | 3.0024 | 4.6210 |

Assim, considerando a abordagem condicional, concluímos que se rejeita $H_{0,2}$, já que $\Im_{2,Obs} > z_{1-\alpha}$, para $\alpha$ =0.05 e 0.01.

**Interação entre os fatores**

Para a interação entre os dois fatores temos

$$\mathbf{A_3} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & 0 & 0 & -\frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2\sqrt{3}} & \frac{1}{2\sqrt{3}} & \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} & -\frac{1}{2\sqrt{3}} & \frac{1}{2\sqrt{3}} \end{bmatrix},$$

com $g_3 = car(\mathbf{A_3}) = 2$.

Para o numerador da estatística obtemos

$$S_3 = (\mathbf{A_3 y_\bullet})' \left(\mathbf{A_3 D A_3}'\right)^{-1} (\mathbf{A_3 y_\bullet}) = 42.4188.$$

Assim, o valor observado da estatística, $\Im_{3,Obs}$, é dado por

$$\Im_{3,Obs} \quad = \quad \frac{1339}{2}\frac{42.4188}{247701.7760} = 0.1147. \qquad (4.5.32)$$

Se considerarmos a distribuição condicional de $\Im_{3,Obs}$, obtemos os quantis apresentados na Tabela 4.12.

Assim, podemos concluir que não se rejeita $H_{0,3}$ para $\alpha = 0.05$ e $0.01$ ($\Im_{3,Obs} < z_{1-\alpha}$).

#### 4.5.2.1 Processos de contagem

Assumiremos que temos uma dimensão mínima para cada uma das amostras, por exemplo que $N_i \geq 10, i = 1, \cdots, 6$, o que significa que $n_i^\bullet = 10, i = 1, \cdots, 6$. Assim, a dimensão mínima global será $n^\bullet = 60$.

Neste caso a distribuição não condicional truncada de $\Im_j, j = 1, 2, 3$, não considerando os termos para os quais $n > \overline{n}$, será dada por

$$\overline{\overline{F}}_{j,\overline{n}}(z) = \sum_{n=60}^{\overline{n}} p_{n^\bullet}(n) F(z|g_j, n-6), \ j = 1, 2, 3, \tag{4.5.33}$$

com

$$\begin{aligned} p_{\boldsymbol{n}^\bullet}(n) &= pr(N = n | \boldsymbol{N} \geq \boldsymbol{n}^\bullet) \\ &= \sum_{n_1=10}^{n-50} \cdots \sum_{n_5=10}^{n-(\sum_{i=1}^{4} n_i + 10)} \sum_{n_6 = n - \sum_{i=1}^{5} n_i}^{n - \sum_{i=1}^{5} n_i} pr(\boldsymbol{N} = \boldsymbol{n} | \boldsymbol{N} \geqslant \boldsymbol{n}^\bullet) \\ &= \sum_{n_1=10}^{n-50} \cdots \sum_{n_5=10}^{n-(\sum_{i=1}^{4} n_i + 10)} \sum_{n_6 = n - \sum_{i=1}^{5} n_i}^{n - \sum_{i=1}^{5} n_i} \prod_{i=1}^{6} \frac{\lambda_i^{n_i}}{n_i!(e^{\lambda_i} - \sum_{u_i=0}^{9} \frac{\lambda_i^{u_i}}{u_i})}. \end{aligned} \tag{4.5.34}$$

**Estimação dos parâmetros da Poisson**

Vamos mais uma vez calcular os limites inferiores de $\lambda_i, i = 1, \cdots, 6$, $\lambda_{\alpha,i}$, de acordo com a regra estabelecida em Nunes et al. (2014) e Nunes et al. (2019a). Como vimos estes correspondem aos valores mínimos de $\lambda_i$, $i = 1, \cdots, m$, tais que

$$e^{-\lambda_i} \frac{\lambda_i^{n_i}}{n_i!} = \alpha, \ i = 1, \cdots, m,$$

em que $\alpha$ corresponde aos níveis usuais de significância e para os $n_i$ foram considerados os números de desempregados apresentados na Tabela 4.10. A Tabela 4.13 apresenta os valores obtidos.

Assumindo estes limites inferiores para $\lambda_i, i = 1, \cdots, 6$, obtivemos o valor mínimo $\overline{n}$ por forma a que

$$1 - \sum_{n=60}^{\overline{n}} pr(N = n | \boldsymbol{N} \geqslant \boldsymbol{n}^\bullet) < 10^{-4}.$$

Tabela 4.13: Limites inferiores para $\lambda_i, i = 1, \cdots, 6$, e valor mínimo $\overline{n}$

| **Valores de** $\alpha$ | $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | $\lambda_4$ | $\lambda_5$ | $\lambda_6$ | $\overline{n}$ |
|---|---|---|---|---|---|---|---|
| 0.05 | 7.01 | 4.9 | 14.33 | 15.7 | 35.9 | 31.9 | 160 |
| 0.01 | 5.02 | 3.3 | 11.01 | 12.2 | 29.15 | 25.8 | 136 |

Atendendo aos valores de $\overline{n}$ obtidos (Tabela 4.13), podemos concluir mais uma vez que temos um bom controlo do erro de truncatura.

#### Fator de efeitos fixos

Considerando a abordagem não condicional, cujos quantis são apresentados na Tabela 4.14, continuamos a não rejeitar $H_{0,1}$ para $\alpha = 0.05$ e $0.01$ ($\Im_{1,Obs} = 0.0724 < z^t_{1-\alpha}$).

Tabela 4.14: Quantis da distribuição não condicional truncada de $\Im_1$

| **Valores de** $\alpha$ | 0.05 | 0.01 |
|---|---|---|
| $z^t_{1-\alpha}$ | 3.9240 | 6.9070 |

#### Fator de efeitos aleatórios

Na Tabela 4.15 são apresentados os quantis da distribuição não condicional truncada de $\Im_2$. Concluímos que se rejeita $H_{0,2}$ considerando a abordagem não condicional ($\Im_{2,Obs} = 10.6930 < z^t_{1-\alpha}$).

Tabela 4.15: Quantis da distribuição não condicional truncada de $\Im_2$ e $\Im_3$

| **Valores de** $\alpha$ | 0.05 | 0.01 |
|---|---|---|
| $z^t_{1-\alpha}$ | 3.0757 | 4.8338 |

#### Interação entre os fatores

No que respeita à interação, cujos quantis também são apresentados na Tabela 4.15, podemos concluir que não se rejeita $H_{0,3}$ para os níveis usuais de significância ($\Im_{3,Obs} = 0.1147 < z^t_{1-\alpha}$), considerando a abordagem não condicional.

### 4.5.2.2 Falhas de observações

Vamos assumir agora que

$$N_i \sim B(r_i, 1-p), i = 1, \cdots, 6.$$

Devido à independência dos $N_i$ $i = 1, \cdots, 6$, tem-se ainda

$$N = \sum_{i=1}^{6} N_i \sim B(r, 1-p),$$

com $r = \sum_{i=1}^{6} r_i$.

#### Estimativa dos parâmetros $r_i, i = 1, \cdots, 6,$ da distribuição Binomial

Tal como na subsecção 4.5.1.2, vamos voltar a estimar $r_i, i = 1, \cdots, 6,$ considerando a regra estabelecida em Nunes et al. (2019b).

Portanto, $r_i, i = 1, \cdots, 6,$ corresponde ao valor mínimo de $r_i$ que verifica

$$pr(N_i \geq n_i^{\bullet}) \geq q \quad \Leftrightarrow \quad \sum_{n=n_i^{\bullet}}^{r_i} \begin{pmatrix} r_i \\ n \end{pmatrix} (1-p)^n p^{r_i - n} \geq q. \tag{4.5.35}$$

Vamos continuar a assumir que $n_i^{\bullet} = 10,\ i = 1, \cdots, 6,$ e que $q = 0.95.$

A Tabela seguinte apresenta os valores obtidos, considerando diferentes valores para $p$ (probabilidade de ocorrência de uma falha).

Pela análise dos resultados concluímos que se obtêm limites superior razoáveis para as dimensões das amostras.

Tabela 4.16: Valores mínimos de $r_i, i = 1, \cdots, 6$

| **Valores de** $p$ | 0.1 | 0.2 | 0.3 | 0.4 |
|---|---|---|---|---|
| $r_i, i = 1, \cdots, 6,$ | 13 | 16 | 19 | 23 |
| $r = \sum_{i=1}^{8}$ | 78 | 96 | 114 | 138 |

Vamos assumir que $p = 0.2$ (valor conhecido pela realização de estudos anteriores). Nesse caso, temos como distribuição não condicional da estatística

$$\overline{\overline{F}}_{j,\overline{n}}(z) \quad = \quad \sum_{n=60}^{96} p_{n^{\bullet}}(n) F(z|g_j, n-6), j = 1, 2, 3, \tag{4.5.36}$$

com

$$\begin{aligned} p_{\boldsymbol{n}^{\bullet}}(n) &= pr(N = n|\boldsymbol{N} \geq \boldsymbol{n}^{\bullet}) \\ &= \sum_{n_1=10}^{n-50} \cdots \sum_{n_5=10}^{n-(\sum_{i=1}^{4} n_i + 10)} \sum_{n_6 = n - \sum_{i=1}^{5} n_i}^{n - \sum_{i=1}^{5} n_i} pr(\boldsymbol{N} = \boldsymbol{n} | \boldsymbol{N} \geqslant \boldsymbol{n}^{\bullet}) \\ &= \sum_{n_1=10}^{n-50} \cdots \sum_{n_5=10}^{n-(\sum_{i=1}^{4} n_i + 10)} \sum_{n_6 = n - \sum_{i=1}^{5} n_i}^{n - \sum_{i=1}^{5} n_i} \prod_{i=1}^{6} \frac{\begin{pmatrix} 16 \\ n_i \end{pmatrix} (0.8)^{n_i} (0.2)^{16-n_i}}{\sum_{u_i=10}^{16} \begin{pmatrix} 16 \\ u_i \end{pmatrix} (0.8)^{u_i} (0.2)^{16-u_i}}. \end{aligned} \tag{4.5.37}$$

**Fator de efeitos fixos**

A Tabela que se segue apresenta os quantis da distribuição não condicional de $\Im_1$. Pela análise dos valores obtidos chegamos à conclusão que não se rejeita $H_{0,1}$ ($\Im_{1,Obs} = 0.0724 < z_{1-\alpha}^{b}$). Logo o género não é significativo.

Tabela 4.17: Quantis da não condicional truncada de $\Im_1$

| **Valores de** $\alpha$ | 0.05 | 0.01 |
|---|---|---|
| $z_{1-\alpha}^b$ | 3.9752 | 7.0043 |

#### Fator de efeitos aleatórios

A Tabela que se segue apresenta os quantis da distribuição não condicional de $\Im_2$. Os resultados mostram que devemos rejeitar $H_{0,2}$ para $\alpha = 0.05$ e $\alpha = 0.01$ ($\Im_{2,Obs} = 10.6930 > z_{1-\alpha}^b$).

Tabela 4.18: Quantis da distribuição não condicional de $\Im_2$ e $\Im_3$

| **Valores de** $\alpha$ | 0.05 | 0.01 |
|---|---|---|
| $z_{1-\alpha}^b$ | 3.1252 | 4.9159 |

#### Interação entre os fatores

No que respeita à interação entre os dois fatores, concluímos que não se rejeita $H_{0,3}$ para $\alpha = 0.01$ e $0.05$ ($\Im_{3,Obs} = 0.1147 > z_{1-\alpha}^b$). Os valores dos quantis da distribuição não condicional de $\Im_3$ também estão apresentados na Tabela 4.18.

### 4.5.2.3 Discussão dos resultados

A nossa discussão mostra que:

- quer pela abordagem clássica quer pela abordagem não condicional, não se rejeita a hipótese nula para o caso do fator de efeitos fixos e para a interação;

- o fator de efeitos fixos não é significativo, o que significa que a média da idade não é significativamente diferente para os dois géneros;

- o fator de efeitos aleatórios é significativo, o que quer dizer que a média das idades depende do município;

- Não temos situações de falsas rejeições, provavelmente devido à elevada dimensão das amostras.

## 4.6 Um estudo com simulações

Nesta secção, será apresentado um estudo com simulações para comparar a percentagem de falsas rejeições que podem ser evitadas com a aplicação da abordagem proposta, agora considerando um modelo misto, ver Nunes et al. (2019b).

Tal como para o caso das simulações apresentadas no Capítulo 3, também aqui tentámos "replicar" a situação descrita nas aplicações anteriores com dados reais, considerando agora um modelo misto com um fator de efeitos fixos (com dois níveis) e um fator de efeitos aleatórios (com três níveis).

As frequências para cada uma das classes das idades, para os 6 diferentes tratamentos, foram geradas segundo distribuições uniformes no intervalo $[10; 100]$.

Foi obtido o vetor das médias amostrais, as somas dos quadrados e os valores observados das estatísticas de teste para o fator de efeitos fixos, fator de efeitos aleatórios e interação.

Assumiu-se que as variáveis aleatórias $N_i$, $i = 1, \cdots, 6$, seguiam uma distribuição de Poisson com parâmetros $\lambda_i$, $i = 1, \cdots, 6$. Os limites inferiores deste parâmetro foram calculados considerando as regras apresentadas na secções 4.5.1.1 e 4.5.2.1, para um $\alpha = 0.05$.

Foram considerados como dimensões mínimas para as amostras $n_i^\bullet = 2$, $n_i^\bullet = 5$ e $n_i^\bullet = 10$, $i = 1, ..., 6$, o que significa que se tem como dimensões mínimas globais, respetivamente, $n^\bullet = 12$, $n^\bullet = 30$ e $n^\bullet = 60$.

Ao se assumir um erro de truncatura não superior a $10^{-4}$, a série infinita definida em (4.4.13) foi truncada não considerando os termos para os quais $n > 47$, $n > 53$ e $n > 75$, respetivamente.

A seguir, foram calculados os respetivos $p-$*values considerando ambas as abordagens (condicional e não condicional). A Tabela 4.19 mostra os valores obtidos. Os valores a negrito correspondem às rejeições da hipótese nula para uma significância de* $5\%$.

*Pela interpretação dos resultados obtidos conclui-se que, para* $n_i^\bullet = 2$, *a utilização da abordagem usual conduz a uma falsa rejeição no caso do fator de efeitos aleatórios.*

Tabela 4.19: Valores dos *p- values* considerando ambas as abordagens

| | $\Im_{j,obs}$ | Abordagem condicional ($p$-*value*) | Abordagem não condicional ($p$-*values*) | | |
|---|---|---|---|---|---|
| | | n=1099 | $n_i^\bullet = 2$ | $n_i^\bullet = 5$ | $n_i^\bullet = 10$ |
| Efeitos fixos | 0.4588 | 0.4983 | 0.5234 | 0.5047 | 0.5011 |
| Efeitos aleatórios | 4.7001 | **0.0093** | 0.0591 | **0.0189** | **0.0131** |
| Interação | 1.6427 | 0.1939 | 0.2698 | 0.2145 | 0.2030 |

*A Tabela 4.20 mostra o número de rejeições e a percentagem de falsas rejeições que podem ser evitadas usando a abordagem proposta, considerando 1000 simulações. Concluímos que a percentagem é bastante elevada para* $n_i^\bullet = 2$, *sendo menor para* $n_i^\bullet = 5$ *e* $n_i^\bullet = 10, i = 1, ..., 6.$

Tabela 4.20: Número de rejeições e percentagem de falsas rejeições, em 1000 simulações

| | Abordagem condicional Número de rejeições | Abordagem não condicional número de rejeições (% falsas rejeições) | | |
|---|---|---|---|---|
| | | $n_i^\bullet = 2, \overline{n} = 47$ | $n_i^\bullet = 5,\ \overline{n} = 53$ | $n_i^\bullet = 10, \overline{n} = 75$ |
| Efeitos fixos | 392 | 280 (28,57%) | 368 (6,12%) | 381 (2,81%) |
| Efeitos aleatórios | 549 | 356 (35,16%) | 508 (7,47%) | 530 (3,46%) |
| Interação | 553 | 379 (31,46%) | 520 (5,97%) | 538 (2,71%) |

*Os resultados obtidos corroboram com os nossos comentários anteriores sobre a relevância da nossa abordagem em evitar falsas rejeições.*

## 4.7 Cruzamento de fatores de efeitos fixos e aleatórios

*Nas secções anteriores considerámos o cruzamento de dois fatores. Assim, se cada um deles tiver $a_1$ e $a_2$, níveis, respetivamente, teremos $m = a_1 \times a_2$, tratamentos diferentes.*

*Nesta seção, pretendemos generalizar os nossos resultados considerando $u$ fatores, com $a_1, \cdots, a_u$ níveis, que cruzam. Apresentamos ainda uma generalização para o cruzamento e aninhamento de fatores.*

### 4.7.1 Cruzamento simples

*Se cruzamos $u$ fatores com $a_1, \cdots, a_u$ níveis teremos*

$$m = \prod_{i=1}^{u} a_i$$

*tratamentos. Uma vez que fatores de efeitos aleatórios não aninham fatores de efeitos fixos assumiremos que apenas os primeiros $\underline{u}$ fatores, se existirem, correspondem a fatores de efeitos fixos.*

*Os efeitos dos fatores e interações estão associados aos conjuntos $\mathcal{C} \subseteq \overline{\overline{u}} = \{1, \cdots, u\}$ , que correspondem aos índices dos fatores. Ao $\varnothing$ corresponderá a média geral, se $\sharp(\mathcal{C}) = 1$ a $\mathcal{C}$ corresponderão os efeitos dos níveis do único fator com índice em $\mathcal{C}$ e se $\sharp(\mathcal{C}) > 1$ a $\mathcal{C}$ corresponderão as interações entre os níveis dos fatores com índice em $\mathcal{C}$. Tomando*

$$j(\mathcal{C}) = 1 + \sum_{l \in \mathcal{C}} 2^{l-1},$$

*temos $j(\mathcal{C}) \leqslant 2^{\underline{u}}$ se e somente se $\mathcal{C} \subseteq \overline{\overline{\underline{u}}} = \{1, \cdots, \underline{u}\}$, isto é, se fatores com índices em $\mathcal{C}$ tiverem efeitos fixos. Assim, teremos $\underline{w} = 2^{\underline{u}}$ conjuntos contidos em $\overline{\overline{\underline{u}}}$ e $w = 2^u$ conjuntos contidos em $\overline{\overline{u}}$*

*Ao conjunto $\mathcal{C}_j$, com índice $j$, está associada a matriz, ver e.g. Ferreira et al. (2007b, 2010),*

$$\boldsymbol{A}_j = \otimes_{l=1}^{u} \boldsymbol{A}_j(l)\boldsymbol{l},\ j = 1, \cdots, w,$$

*onde $\otimes$ representa o produto de Kronecker de matrizes, com*

$$\begin{cases} \boldsymbol{A}_j(l) = \frac{1}{\sqrt{a_l}}\mathbf{1}_{a_l}^{'}, l \notin \mathcal{C}_j,\ j = 1, \cdots, w \\ \boldsymbol{A}_j(l) = \boldsymbol{L}_j(l), l \in \mathcal{C}_j,\ j = 1, \cdots, w, \end{cases},$$

*onde $\boldsymbol{L}_j(l)$ é obtida excluindo o primeira linha, igual a $\frac{1}{\sqrt{(a_l)}}\mathbf{1}_{a_l}^{'}$, de uma matriz ortogonal $a_l \times a_l$.*

*Assim*

$$g_j = car(\boldsymbol{A}_j) = \prod_{l \in \mathcal{C}_j} (a_l - 1),\ j = 1, \cdots, w,$$

*então podemos estimar as componentes da variância para os efeitos dos fatores de efeitos aleatórios e interações, nos quais elas participem, através de*

$$\tilde{\gamma}_j = \frac{\|\boldsymbol{A}_j \boldsymbol{Y}_\bullet\|^2}{g_j},\ j = w^{'} + 1, \cdots, w,$$

*com $\boldsymbol{Y}_\bullet$ o vetor de média dos tratamentos.*

*Quando $j \leq \underline{w},\ l \notin \mathcal{C}_j$, sempre que $\underline{u} < l \leq u$, assim teremos*

$$\boldsymbol{A}_j = \otimes_{l=1}^{\underline{u}} \boldsymbol{A}_j(l) \otimes \left( \otimes_{l=\underline{u}+1}^{u} \frac{1}{\sqrt{a_l}} \mathbf{1}_{a_l}^{'} \right).$$

*Então, considerando $\boldsymbol{Y}^\bullet$ o vetor da média das observações para as combinações de níveis dos fatores de efeitos fixos, temos*

$$\boldsymbol{A}_j \boldsymbol{Y}_\bullet = \sqrt{r} \left( \otimes_{l=1}^{\underline{u}} \boldsymbol{A}_j(l) \boldsymbol{Y}^\bullet \right),$$

*com*

$$r = \prod_{l=\underline{u}+1}^{u} \sqrt{a}_l.$$

*Assim, do modelo da média dos tratamentos podemos extrair um modelo no qual apenas os fatores de efeitos fixos são relevantes. Este último modelo apenas possui uma observação por tratamento, portanto é equilibrado. Então, de acordo com a discussão apresentada em Scheffé (1959) (Capítulo 10), podemos usar a ANOVA nestas situações.*

### 4.7.2 Cruzamento e aninhamento

*Vamos agora cruzar $v$ grupos de $u(1), \cdots, u(v)$ fatores que aninham. Os primeiros $\underline{u}(1), \cdots, \underline{u}(v)$ fatores para os diferentes grupos correspondem a fatores de efeitos fixos (se $\underline{u}(h) > 0$, $h = 1, ..., v$) e os restantes correspondem a fatores de efeitos aleatórios.*

*Os primeiros $l$ fatores no $h$-ésimo grupo terão*

$$b_l(h) = \prod_{t=1}^{l} a_t(h),\ l = 1, \cdots, u(h),\ h = 1, \cdots, v,$$

*níveis, uma vez que, para todo o $l > 1$, o $l-$ésimo fator tem $a_l(h)$ níveis aninhados em cada $b_{l-1}(h)$ níveis do $(l-1)$-ésimo fator do mesmo grupo, $h = 1, ..., v$.*

*Aos fatores no $h$-ésimo grupo correspondem as matrizes, ver e.g. Ferreira et al. (2007a, 2010),*

$$\begin{cases} \boldsymbol{A}_1(h) = \otimes_{l=1}^{u(h)} \left( \frac{1}{\sqrt{a_l(h)}} \mathbf{1}'_{a_l(h)} \right) \\ \boldsymbol{A}_l(h) = \left( \otimes_{l'=1}^{l} \boldsymbol{L}_{a_{l'}(h)} \right) \otimes \left( \otimes_{l'=l+1}^{u(h)} \frac{1}{\sqrt{a_{l'}(h)}} \mathbf{1}'_{a_{l'}(h)} \right),\ l = 2, \cdots, u(h) - 1 \\ \boldsymbol{A}_{u(h)}(h) = \otimes_{l'=1}^{u(h)} \boldsymbol{L}_{a'_l(h)} \end{cases} .$$

*Assim temos para este grupo as matrizes de projeção ortogonal dadas por*

$$\begin{cases} \boldsymbol{Q}_1(h) = \boldsymbol{A}'_1(h)\boldsymbol{A}_1(h) = \otimes_{l=1}^{u(h)} \left( \frac{1}{a_l(h)} \boldsymbol{J}_{a_l(h)} \right) \\ \boldsymbol{Q}_l(h) = \boldsymbol{A}'_l(h)\boldsymbol{A}_l(h) = \left( \otimes_{l'=1}^{l} \boldsymbol{K}_{a_{l'}(h)} \right) \otimes \left( \otimes_{l'=l+1}^{u(h)} \frac{1}{a_{l'}(h)} \boldsymbol{J}_{a_{l'}(h)} \right),\ l = 2, \cdots, u(h) - 1 \\ \boldsymbol{Q}_{u(h)}(h) = \otimes_{l'=1}^{u(h)} \boldsymbol{K}_{a'_l(h)} \end{cases} ,$$

*onde*

$$\begin{cases} \boldsymbol{J}_a = \mathbf{1}_a \mathbf{1}'_a \\ \boldsymbol{K}_a = \boldsymbol{I}_a - \frac{1}{a} \mathbf{1}_a \mathbf{1}'_a = \boldsymbol{L}'_a \boldsymbol{L}_a \end{cases} .$$

*Podemos ver que*

$$\begin{cases} \sum_{l'=1}^{l} \boldsymbol{Q}_{l'}(h) = \boldsymbol{I}_{b_l(h)} \otimes \left( \otimes_{l'=l+1}^{u(h)} \frac{1}{a_{l'}(h)} \boldsymbol{J}_{a_{l'}(h)} \right),\ l = 1, \cdots, u(h) - 1 \\ \sum_{l'=1}^{u(h)} \boldsymbol{Q}_{l'}(h) = \boldsymbol{I}_{b_{u(h)}(h)} \end{cases} .$$

*Quando os $v$ grupos de fatores cruzam, a média geral, os efeitos dos fatores e interações corresponderão aos vetores*

$$\boldsymbol{j} = (j(1), ..., j(v)),$$

*onde $j(h) = 0, 1, ..., u(h)$, $h = 1, ..., v$. Ao vetor nulo, $\mathbf{0}$, corresponderá a média geral, se $\boldsymbol{j}$ tem apenas uma componente não nula, $j(h')$, a $\boldsymbol{j}$ corresponderão os efeitos do $j(h')$-ésimo fator no $h'$-ésimo grupo. Por outro lado, se $\boldsymbol{j}$ tiver mais do que uma componente não nula, $\boldsymbol{j}$ corresponderá às interações entre essas componentes não nulas.*

*Seja $\Gamma$ o conjunto dos vetores $\boldsymbol{j}$. Aos vetores de $\Gamma$ correspondem as matrizes, ver Ferreira et al. (2007a, 2010),*

$$\left\{\begin{array}{l} \boldsymbol{A_j} = \otimes_{h=1}^{v} \boldsymbol{A}_{j(h)}(h),\ \boldsymbol{j} \in \Gamma \\ \boldsymbol{Q_j} = \otimes_{h=1}^{v} \boldsymbol{Q}_{j(h)}(h),\ \boldsymbol{j} \in \Gamma \end{array}\right. ,$$

*e a soma dos quadrados serão dadas por*

$$S_{\boldsymbol{j}} = \|\boldsymbol{A_j Y_\bullet}\|^2;\ \boldsymbol{j} \in \Gamma. \quad (4.7.38)$$

*É fácil ver que*

$$\sum_{\boldsymbol{j}\in\Gamma} \boldsymbol{Q_j} = \otimes_{h=1}^{v} \left( \sum_{j(h)=1}^{u(h)} \boldsymbol{Q}_{j(h)}(h) \right) = \otimes_{h=1}^{v} \boldsymbol{I}_{b_{u(h)}(h)} = \boldsymbol{I}_m, \quad (4.7.39)$$

*onde*

$$m = \prod_{h=1}^{v} b_{u(h)}(h)$$

*corresponde ao número de tratamentos no modelo global.*

### *4.7.3 Extensão às amostras de dimensão aleatória*

*Vimos portanto como obter as matrizes $\boldsymbol{A_j}$, tanto para parte dos efeitos fixos como para a parte dos efeitos aleatório, quer para modelos com cruzamento simples e quer para modelos com cruzamento e aninhamento.*

*Assim, uma vez que as somas dos quadrados, $S_j$, são obtidos através de $\boldsymbol{Y_\bullet}$, e apesar da extensão ao caso de aleatoriedade das dimensões das amostras nos levar a situações de heterocedasticidade em $\boldsymbol{Y_\bullet}$, o modelo é equilibrado, logo de acordo com Scheffé (1959) (capítulo 10), é rubusto, garantido a possibilidade de se continuar a usar a ANOVA quando se desconhecem as dimensões das amostras à priori.*

## *4.8 Conclusões*

*Neste capítulo procurámos apresentar uma nova abordagem na aplicação de modelos mistos ortogonais, em situações em que as dimensões das amostras não são previamente conhecidas.*

*Destacamos o facto interessante da distribuição da estatística de teste, sob a hipótese nula, ser a mesma para a parte de efeitos fixos e a parte de efeitos aleatórios do modelo, mesmo quando temos amostras com dimensão aleatória. Esta situação é designado por situação de estabilidade e foi abordada inicialmente em Ferreira (2006).*

*Os resultados obtidos através das duas aplicações e do estudo com dados simulados mostram a relevância da nossa abordagem em evitar falsas rejeições (caso em que a hipótese nula é rejeitada através da abordagem comum, mas não é rejeitada considerando a abordagem proposta). Tivemos oportunidade de ver ainda que tal ocorre com mais frequência quando não se trabalha com amostras de grande dimensão (Aplicação 4.5.1).*

*Terminámos este capítulo, mostrando como generalizar o tratamento apresentado (considerando as dimensões das amostras como aleatórias), ao caso do cruzamento de* $u$ *fatores, dos quais* $\underline{u},$ *com* $\underline{u}, < u,$ *correspondem a fatores de efeitos fixos e os restantes a fatores de efeitos aleatórios. Foi considerada ainda a generalização ao caso do cruzamento e aninhamento de fatores.*

# Capítulo 5

## Conclusões finais e trabalhos futuros

*O presente trabalho teve como propósito estender a teoria dos modelos mistos ortogonais a situações em que as dimensões das amostras não são previamente conhecidas.*

*Vimos que em casos como este, em que se desconhecem as dimensões das amostras à priori, é mais correto considerar estas dimensões como realizações de variáveis aleatórias independentes.*

*Consideraram-se três situações distintas para a distribuição dessas variáveis aleatórias:*

- *a distribuição de Poisson, quando a ocorrência das observações corresponde a processos de contagem;*
- *a distribuição Binomial, caso exista um limite superior para a dimensão das amostras que nem sempre é atingido devido à ocorrência de falhas de observações;*
- *a distribuição Geométrica [Binomial Negativa], quando o número de observações corresponde ao número de ocorrências até ao primeiro sucesso [$s$-ésimo sucesso].*

*Como vimos, esta metodologia tem vindo a ser aplicada ao caso da ANOVA de efeitos fixos e ANOVA de efeitos aleatórios, ver e.g. Capristrano (2015), Capistrano et al. (2015), Moreira et al. (2013), Nunes et al. (2012a, 2012b, 2014, 2015, 2019a).*

*Os modelos mistos com amostras de dimensão aleatória também já foram abordados em Capistrano (2015) e Nunes et al. (2019c), onde a formulação do modelo foi feita através do uso das extensões $L$, permitindo a realização de inferência apenas para a parte aleatória do modelo.*

*No presente trabalho a formulação do modelo misto foi realizada considerando situações de estabilidade, em que a distribuição da estatística de teste, sob a hipótese nula, é a mesma para a parte de efeitos fixos e parte de efeitos aleatórios do modelo (Ferreira, 2006), mesmo quando temos amostras de dimensão aleatória.*

*Vimos que a abordagem proposta leva a aplicações interessantes em várias áreas de investigação, nomeadamente em investigação médica e económica. Através das aplicações com dados reais e dados simulados, apresentadas ao longo do trabalho, chegamos* *à conclusão que a nossa abordagem é robusta, uma vez que a sua utilização "controla" a ocorrência de falsas rejeições, mesmo para o caso em que não se têm amostras de dimensão muito elevada. Dizemos que ocorreu uma falsa rejeição, sempre que a hipótese nula é rejeitada considerando a abordagem clássica e não é rejeitada através da aplicação da abordagem proposta.*

*É importante referir que, atendendo à lei dos grandes números, a probabilidade da ocorrência de falsas rejeições converge quase certamente para zero quando a dimensão das amostras tende para infinito. Algo que tivemos a oportunidade de verificar pelos resultados obtidos através de algumas das aplicações com dados reais. Podemos então concluir que com o aumento da dimensão das amostras, ambas as abordagens convergem para a mesma decisão, algo que foi possível de verificar em Capistrano (2015), considerando modelos de efeitos fixos, e agora neste trabalho considerando modelos mistos.*

*Portanto, é de incentivar a utilização da metodologia proposta em situações em que não se conhecem as dimensões das amostras à partida, evitando tomar decisões incorretas pela aplicação de abordagem usual.*

*Em termos de trabalhos futuros, tencionamos estender a análise de variância multivariada (MANOVA) ao caso em que se têm amostras de dimensões aleatórias. Tencionamos ainda aplicar esta abordagem a outras metodologias estatísticas, como por exemplo, ao modelo de regressão linear.*

# Bibliografia


[1] *Bailey, R.A., Ferreira, S.S., Ferreira D. and Nunes, C. (2016). Estimability of variance components when all model matrices commute. Linear Algebra and its Applications, 492, 144-160.*

[2] *Barsotti, F., Philippe, A. and Rochet, P. (2016). Hypothesis testing for markovian models with random time observations. Journal of Statistical Planning and Inference, 173, 8798.*

[3] *Bunge, J.A. and Nagaraja, H.N. (1991). The distributions of certain record statistics from a random number of observations. Stochastic Processes and Their Applications, 38 (1), 16783.*

[4] *Calinski, T. and Kageyama S. (2000). Block Designs: A Randomization Approach: Vol. I, Analysis. Lecture Note in Statistics, 150, New York : Springer-Verlag.*

[5] *Calinski, T. and Kageyama S. (2003). Block Designs: A Randomization Approach: Vol. II, Design. Lecture Note in Statistics, 170, New York : Springer-Verlag.*

[6] *Capistrano, G. (2015). Análise de variância com amostra de dimensão aleatória e suas aplicações. PhD Thesis, Universidade da Beira Interior, Covilhã Portugal.*

[7] *Capistrano, G., Nunes, C., Ferreira, D., Ferreira, S.S. and Mexia, J.T. (2015). One-way Random Effects ANOVA with Random Sample Sizes: An Application to a Brazilian Database on Cancer Registries. 12th International Conference on Numerical Analysis and Applied Mathematics. AIP Conference Proceedings 1648, 110009. doi: 10.1063/1.4912416.*

[8] *Carvalho, F., Mexia, J.T. and Oliveira, M.M. (2008). Canonic inference and commutative orthogonal block structure. Discussiones Mathematicae - Probability and Statistics, 28(2), 171-181.*

[9] *Carvalho, F., Mexia, J. T., Santos, C. and Nunes, C. (2015). Inference for types and structured families of commutative orthogonal block structures. Metrika, 78, 337-372.*

[10] *Clarke, Brenton R. (2008). Linear models: the theory and application of analysis of variance. Wiley series in probability and statistics. John Wiley & Sons, Inc., New York.*

[11] *Esquível , M.L., Mota, P.P. , and Mexia, J.T. (2016). On some statistical models with a random number of observations. Journal of Statistical Theory and Practice, 10(4), 805-823.*

[12] *Ferreira, S.S. (2006). Inferência para modelos ortogonais com segregação. PhD Thesis, Universidade da Beira Interior, Covilhã Portugal.*

[13] *Ferreira, S. S. , Ferreira, D., Fernandes, C. and Mexia, J.T. (2007a). Orthogonal Mixed Models and Perfect Families of Symmetric Matrices. In proceedings of 56th session of the International Statistical Institute. Lisboa.*

[14] *Ferreira, S. S. , Ferreira, D. and Mexia, J.T. (2007b). Cross additivity in balanced cross nested models. Journal of Statistical Theory and Practice, 1 (3-4), 377-392.*

[15] *Ferreira, S.S., Ferreira, D., Moreira, E. and Mexia, J.T. (2009). Inference for $L$ orthogonal models. Journal of Interdisciplinary Mathematics, 12(6), 815-824.*

[16] *Ferreira, S. S. , Ferreira, D., Nunes, C. and Mexia, J.T. (2010). Nesting Segregated Mixed Models. Journal of Statistical Theory and Practice, 4 (2), 233-242.*

[17] *Ferreira, S.S., Ferreira, D., Nunes, C. and Mexia, J.T. (2013). Estimation of variance components in linear mixed models with commutative orthogonal block structure. Revista Colombiana de Estadistica, 36(2), 261- 271.*

[18] *Ferreira, D., Ferreira, S. S., Nunes, C. and Mexia, J. T. (2017). Estimation in mixed models through Three Step Minimization. Communications in Statistics - Simulation and Computation, 46 (2), 1156-1166.*

[19] *Fonseca, M., Mexia, J.T. and Zmyślony, R. (2003). Estimators and tests for variance components in cross nested orthogonal designs. Discussiones Mathematicae -Probability and Statistics, 23(2), 175-201.*

[20] *Fonseca, M., Mexia, J.T. and Zmyslony, R. (2006). Binary Operations on Jordan algebras. Linear Algebra and its Applications, 117, 75-86.*

[21] *Fonseca, M., Mexia, J.T. and Zmyślony, R. (2008). Inference in normal models with commutative orthogonal block structure. Acta et Commentationes Universitatis Tartuensis de Mathematica, 12, 3-16.*

[22] *Fonseca, M., Jesus, V., Mexia, J.T. and Zmyślony R. (2009). Binary Operations and canonical forms for factorial and related models. Linear Algebra and its Applications, 430, 2781-2797.*

[23] *Graham, A. (1981). Kronecker products and calculus with applicatiom. Holsted press, John wiley and sons, New York.*

[24] *Horn, R. and Johnson, C. (1985). Matrix Analysis. Cambridge University Press.*

[25] *Houtman, A.M. and Speed, T.P. (1983). Balance in designed experiments with orthogonal block structure. Annals of Statistics, 11(4), 1069-1983.*

[26] *Ito, P.K. (1980). Robustness of ANOVA and MANOVA test procedures. In: Krishnaiah P.R. (eds) Handbook of statistics Vol. 1. North-Holland, Amsterdam, 199-236.*

[27] *Jacobson, N. (1953). Lectures in Abstract Algebra. Volume II-Linear Algebra. New York: D. Van Nostrand.*

[28] *Jesus, V., Rodrigues, P.C. and Mexia, J.T. (2007). Inference for random effects in prime basis factorials using commutative Jordan algebras. Discussiones Mathematicae -Probability and Statistics, 27, 15-25.*

[29] *Jesus, V., Fonseca, M., Mexia, J.T. and Zmyślony, R. (2009). Binary operations and canonical forms for factorial and related models. Linear Algebra and its Applications, 43, 2781-2797.*

[30] *Johnson, N.L. and Kotz, S. (1969). Discrete distributions. John Wiley & Sons, Inc., New York.*

[31] *Jordan, P., Von Neumann, J. and Wigner, E. (1934). On an algebraic generalization of the quantum mechanical formulation. Annals of Mathematical, 35 (1), 29-64.*

[32] *Khuri, A. and Sahai, A. (1985). Variance components analysis: a selective literature survey. International Statistical Review, 279300.*

[33] *Khuri, A.I., Mathew, T. and Sinha, B. K. (1998). Statistical Tests for Mixed Linear Models. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., New York.*

[34] *Khuri, A.I. and Ghosh, M. (1990). Minimal Sufficient Statistics for the Unbalanced Two- Fold Nested Model. Statistics and Probability Letters, 10, 351-353.*

[35] *Lehmann, E. L. and Romano, J. P. (2005). Testing Statistic Hypotheses. 3th ed. New York, Ny: Springer Science + Business Media.*

[36] *Mann, H. B. and Wald, A. (1943). On Stochastic Limit and Order Relationships. Annals of Mathematical Statistics, 14(3), 217-226.*

[37] *Mário, A., Nunes, C., Ferreira, D., Ferreira, S.S. and Mexia, J. T. (2019). Parametric models with random sample sizes. In recent studies on risk Analysis and Statistical Modelling. Spriger, Berlin, Heidelberg. Accepted.*

[38] *Mejza, S. (1992). On some aspects of general balance in designed experiments. Statistica, 52, 263-278.*

[39] *Mexia, J.T. (1989). Controlled Heteroscedasticity, Quocient Vector Spaces and $F$ Tests for Hypothesis on Mean Vectors. Trabalhos de investigação, nº 1, FCT/UNL.*

[40] *Mexia, J.T. (1990). Best linear unbiased estimates, duality of $F$ tests and the Scheffé multiple comparison method in presence of controled heterocedasticity. Computational Statistics & Data Analysis, 10 (3), 271-281.*

[41] *Mexia, J.T. (1991). Subnormal distributions and F tests. Trabalhos de Investiga?o, nº 1, FCT/UNL.*

[42] *Mexia, J.T. (1995). Introdução à Inferência Estatística Linear. Centro de Estudos de Matemática Aplicada. Edições Lusófonas.*

[43] *Mexia, J.T. and Moreira, E. (2010). Randomized sample size F tests for the one-way layout. 8th International Conference on Numerical Analysis and Applied Mathematics. AIP Conference Proceedings, 1281(II), 1248-1251. doi: 10.1063/1.3497917.*

[44] *Mexia, J.T., Nunes, C., Ferreira, D., Ferreira, S.S. and Moreira, E. (2011). Orthogonal fixed effects ANOVA with random sample sizes. Proceedings of the 5th International Conference on Applied Mathematics, Simulation, Modelling (ASM'11), Corfu, Greece, 84-90.*

[45] *Michalski, A. and Zmyślony, R. (1996). Testing hypothesis for variance components in mixed linear models. Statistics, 27, 297-310.*

[46] *Michalski, A. and Zmyślony, R. (1999). Testing hypothesis for linear functions of parameters in mixed linear models. Tatra Mountain Mathematical Publications, 17, 103- 110.*

[47] *Mood, M.M.; Graybill, F.A. and Boes, D.C. (1987). Introdution to the Theory of Statistics. McGraw-Hill.*

[48] *Moreira, E., Mexia J.T., Fonseca, M. and Zmyslony, R. (2009). $L$ models and multiple regressions designs. Statistical Papers, 50(4), 869-885.*

[49] *Moreira E.E., Mexia J.T. and Minder C.E. (2013). F tests with random sample size.*

*Theory and applications. Statistics & Probability Letters, 83 (6), 1520-1526.*

*[50] Muller, K.E., and Stewart, P.W. (2006). Linear Model Theory; Univariate, Multivariate, and Mixed Models. John Wiley & Sons, Inc. Hoboken, New Jersey.*

*[51] Nelder, J.A. (1965a). The analysis of randomized experiments with ortogonal block structure I. Block structure and null analysis of variance. Proceedings of the Royal Society, Series A, 283, 147-162.*

*[52] Nelder, J.A. (1965b). The analysis of randomized experiments with orthogonal block structure II. Treatment structure and general analysis of variance. Proceedings of the Royal Society, Series A, 283, 163-178.*

*[53] Nguyen Bac Van, On the statistical analysis of a random number of observations. Acta Mathematica Vietnamica, 13(1) (1988), 5561.*

*[54] Nunes, C. (2005). Testes F e relacionados em modelos mistos com cross-nesting ortogonal. PhD Thesis, Universidade da Beira Interior, Covilhã Portugal.*

*[55] Nunes, C. and Mexia, J.T. (2006). Non-central generelized F distributions. Discussiones Mathematicae -Probability and Statistics, 26(I), 297-310.*

*[56] Nunes, C., Santos, C. and Mexia, J.T. (2008). Relevant statistics for models with commutative orthogonal block structure and unbiased estimator for variance components. Journal of Interdisciplinary Mathematics. 11(4), 553-564.*

*[57] Nunes, C., Ferreira, D., Ferreira, S.S. and Mexia, J. T. (2010). F Tests with Random Sample Sizes. 8th International Conference on Numerical Analysis and Applied Mathematics. AIP Conference Proceedings, 1281(II), 1241-1244. doi: 10.1063/1.3497911.*

*[58] Nunes, C., Ferreira, D., Ferreira, S.S. and Mexia, J. T. (2012a). $F$-tests with a rare pathology. Journal of Applied Statistics, 39(3), 551–561.*

*[59] Nunes, C., Ferreira, D., Ferreira, S.S., Oliveira, M.M. and Mexia, J.T. (2012b). One-way Random Effects ANOVA: An Extension to Samples with Random Size. 10th International Conference on Numerical Analysis and Applied Mathematics. AIP Conference Proceedings, 1479, 1678-1681. doi: 10.1063/1.4756492.*

*[60] Nunes, C., Ferreira, D., Ferreira, S.S. and Mexia, J. T.(2012c). Control of The Truncation errors for generalized F distributions. Journal of Statistical Computation and Simulation, 82(2), 165-171.*

[61] *Nunes, C., Capistrano, G., Ferreira, D. and Ferreira, S.S. (2013). ANOVA with Random Sample Sizes: An Application to a Brazilian Database on Cancer Registries. 11th International Conference on Numerical Analysis and Applied Mathematics. AIP Conference Proceedings, 1558, 825-828. doi: 10.1063/1.4825623.*

[62] *Nunes, C., Ferreira , D., Ferreira, S.S. and Mexia, J.T. (2014). Fixed effects ANOVA: an extension to samples with random size. Journal of Statistical Computation and Simulation, 84(11), 2316-2328.*

[63] *Nunes, C., Capistrano, G., Ferreira, D., Ferreira, S.S. and Mexia, J.T. (2015). One-way Fixed effects ANOVA with Missing observations. 12th International Conference on Numerical Analysis and Applied Mathematics. AIP Conference Proceedings, 1648, 110008. doi: 10.1063/1.4912415.*

[64] *Nunes, C., Capistrano, G., Ferreira, D., Ferreira, S.S. and Mexia, J.T. (2019a). Exact critical values for one-way fixed effects models with random sample sizes. Journal of Statistical Computation and Applied Mathematics, 354, 112-122, doi: 10.1016/j.cam.2018.05.057.*

[65] *Nunes, C., Mário, A., Ferreira, D., Ferreira, S.S. and Mexia, J.T. (2019b). Random sample sizes in orthogonal mixed models with stability. Computational and Mathematics Methods, 1(5). doi: 10.1002/cmm4.1050.*

[66] *Nunes, C., Moreira, E.E., Ferreira, D., Ferreira, S.S. and Mexia, J.T. (2019c). Considering the samples sizes as truncated Poisson random variables in mixed effects models. Journal of Applied Statistical. doi: 10.1080/02664763.2019.1641188.*

[67] *Pestana, D.D. e Velosa, S.F. (2006). Introdução à Probabilidade e à Estatística. Volune I, 2ª Edição. Lisboa: Fundação Calouste Gulbenkian.*

[68] *Pollock, D.S.G. (1979). The Algebra of Econometrics. John Wiley & Sons, Inc., New York.*

[69] *PORDATA - BASE DE DADOS DE PORTUGAL CONTEPORÂNEOS, Fundação Francisco Manuel dos Santos. Available at: http://www.pordata.pt/home. Accessed 1 October 2014.*

[70] *PORDATA - BASE DE DADOS DE PORTUGAL CONTEPORÂNEOS, Fundação Francisco Manuel dos Santos. Available at: http://www.pordata.pt/home. Accessed 1 October 2016.*

[71] *PORDATA - BASE DE DADOS DE PORTUGAL CONTEPORÂNEOS, Fundação Francisco Manuel dos Santos. Available at: http://www.pordata.pt/home. Accessed*

*20 November 2018.*

*[72] Rao, C.R. (1973). Linear Statistical Inference and Its Applications. Second Edition; John Wiley & Sons, Inc., New York.*

*[73] Rao, C., Kleve, J. (1988). Estimation of variance components and applications. North Holland, Elsevier - Amsterdam.*

*[74] Robbins, H. (1948). Mixture of distribution. The Annals of Mathematical Statistics, 19, 360-369.*

*[75] Robbins, H. and Pitman, E.J.G. (1949). Application of the method of mixtures to quadratic forms in normal variates. The Annals of Mathematical Statistics, 20, 552-560.*

*[76] Rodrigues, P.C. and Mexia, J.T. (2006). ANOVA using commutative Jordan algebras. Discussiones Mathematicae -Probability and Statistics, 26, 179-191.*

*[77] Sahai, Hardeo, Ojeda, Mario M. (2004). Analysis of Variance for Random Models, Volume I: Balanced Data Theory, Methods, Applications and Data Analysis. Birkhäuser Applied Probability and Statistics, Basel, Berlin.*

*[78] Sahai, Hardeo, Ojeda, Mario M. (2004). Analysis of Variance for Random Models, Volume II: Unbalanced Data Theory, Methods, Applications and Data Analysis. Birkhäuser Applied Probability and Statistics, Basel, Berlin.*

*[79] Santos, C., Nunes, C. and Mexia, J.T. (2007). OBS, COBS and Mixed Models associated to commutative Jordan Algebra. In proceedings of 56th session of the International Statistical Institute. Lisboa.*

*[80] Santos, C. (2012). Error orthogonal models: Structure, Operations and Inference. PhD thesis. Universidade da Beira Interior, Covilhã Portugal.*

*[81] Salvador, D. (2013). Modelação de Matrizes Estocásticas Simétricas, Operadores do tipo vec. PhD Thesis, FCT - Universidade Nova Lisboa, Lisboa, Portugal.*

*[82] Scheffé, H. (1959). The analysis of variance, Wiley series in Probability and Statistics, John Wiley & Sons, New York.*

*[83] Schott, J.R. (1997). Matrix Analysis for Statistics. Wiley - Interscience. John Wiley & Sons, Inc., New York.*

*[84] Searle, S.R., Casella, G. e McCulloch, C.E. (1992). Variance Components. Wiley series in Probability and statistics. John Wiley & Sons, Inc., New York.*

[85] *Searle, S. (1995). An overview of variance component estimation. Metrika, 42(1), 215-230.*

[86] *Seely, J. (1970a). Linear spaces and unbiased estimation. The Annals of Mathematical Statistics, (41), 1725- 1734.*

[87] *Seely, J. (1970b). Linear spaces and unbiased estimation. An aplication to the mixed linear model. The Annals of Mathematical Statistics, 41, 1735-1748.*

[88] *Seely, J. (1971). Quadratic subspaces and completeness. The Annals of Mathematical Statistics, 42, 710-721.*

[89] *Seely, J. and Zyskind, G. (1971). Linear spaces and minimum variance estimators. The Annals of Mathematical Statistics, 42(2), 691-703.*

[90] *Seely, J. (1977). Minimal sufficient statistics and completeness for multivariate normal families. Sankhya, 39, 170-185.*

[91] *Singh, M. and Gupta, V. K. (1980). On the estimation of mean parameter with random number of observations on each unit. Biometrical Journal, 22 (1), 41145.*

[92] *Steeb, W.H. (1991). Kronecker Products of Matrices and Applications. BI-Wissenschaftsvlg.*

[93] *Steeb, W.H. and Hardy, Y. (2011).Matrix calculus and kronecker product a pratical approch to linear and multilinear algebra. Second edition. World scientific publishing company.*

[94] *The R Project for Statistical Computing. Available at: http://www.r-project.org*

[95] *VanLeeuwen, D.M., Seely, J.F. and Birkes, D.S. (1998). Sufficient conditions for orthogonal designs in mixed linear models. Journal of Statistical Planning and Inference, 73, 373-389.*

[96] *VanLeeuwen, D.M., Birkes, D.S. and Seely, J.F. (1999). Balance and orthogonality in designs for mixed classification models. The Annals of Statistics, 27 (6), 1927-1947.*