UNIVERSIDADE DA BEIRA INTERIOR
Ciências

# Análise de Variância com Amostras de Dimensão Aleatória e suas Aplicações

**Gilberto Capistrano Cunha de Andrade**

Tese para obtenção do Grau de Doutor em
**Matemática Aplicada**
(3º ciclo de estudos)

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

**Covilhã, Outubro de 2015**

# Dedicatória

Gilberto Mendes Andrade e Maria Cristina Capistrano Cunha de Andrade, meus pais, a vossa presença significou segurança e certeza de que não estive sozinho nessa caminhada.

# Agradecimentos

Faço questão de agradecer, mesmo que direta ou indiretamente, a todos os que ajudaram na conclusão deste trabalho. Nome por nome e não importando a ordem.

Professora Célia Nunes, minha orientadora, penso que esta é a oportunidade ideal para agradecer por tudo aquilo que você fez e faz por mim, por tudo o que me ensina e, também, por tudo de bom que a sua postura séria, honesta e ética me sugere. Muito obrigado, eu jamais vou esquecer seu gesto e sempre que puder vou tentar retribuir com muita satisfação e zelo. Tenho o maior orgulho de ser seu aluno.

Professor Dário Ferreira, meu coorientador, com a sua calma do costume, me deu grandes oportunidades e incentivos. Obrigado por todo o apoio oferecido para concluir o meu trabalho.

Professora Sandra Ferreira, pelo seu otimismo e disponibilidade para colaborar e trocar ideias. Professor João Tiago Mexia, pela atenção e pela ajuda, principalmente pela suas prestimosas orientações na elaboração deste trabalho.

Professor Ricardo Campos, pela atenção e pela força dadas através do companheirismo, sempre que precisei. Professor Alberto Simões, pela sua alegria, vivência académica e pelos momentos de lazer que foram essênciais neste percurso onde rimos e nos ajudamos mutuamente.

Departamento de Matemática através do seu Presidente, Professor Helder Vilarinho, ao caríssimo Diretor do Curso Professor Mário Bessa e ao Centro de Matemática e Aplicações da Universidade da Beira Interior (CMA-UBI), onde sou colaborador. A incansável, Secretária do Departamento de Matemática, Filipa Raposo por poder "contar" sempre com ela nos momentos que precisei, pela atenção e disponibilidade. Sr. Jorge Madeira, sempre disponível para esclarecer as minhas dúvidas perante os Serviços Academicos da Universidade da Beira Interior (UBI).

Sr. Arlindo Campos e Srª Maria Suzel Campos, que me levaram e ensinaram um pouco da história da Covilhã e Portugal. Jamais esquecerei a Quinta onde residem, na frequesia do Ferro, os passeios na Travessa do Cotovelo, na Rua da Formosura, na Rua da Estrela entre muitas outras.

Minha adorada esposa, Maria Regina Cantelmo Silva Cunha de Andrade, por seu amor e estímulo. Sinto-me profundamente grato pela sua compreensão e apoio durante muitas tardes, noites e finais de semana passamos juntos pesquisando e fazendo cálculos para chegar ao final deste trabalho. Aos meus irmãos, Guilherme, Giuliana, Gisèlly, meu filho, Gabriel Leão Junho Cunha de Andrade, meus cunhados, minha cunhada Marília Pereira Andrade e sobrinhos que torceram por mim mesmo longe e me apoiaram nesta jornada. Ao Instituto Nacional do Câncer(INCA), por ter disponibilizado os dados para a minha pesquisa. E por fim ao Centro Universitário de Itajubá (FEPI), o apoio, incentivo e oportunidade.

A todos vocês, o meu muito obrigado!

# Resumo

A Análise de variância (ANOVA) é utilizada em muitas áreas de investigação, nomeadamente em investigação médica, agricultura ou psicologia, para citar apenas algumas, onde as dimensões das amostras podem não ser previamente conhecidas. Esta situação ocorre com frequência quando o intervalo de tempo para a recolha das observações é fixado à partida. Um bom exemplo corresponde à recolha de observações para um estudo onde se pretende comparar várias patologias de pacientes que chegam às urgências de um hospital num determinado período de tempo.

Neste trabalho iremos estender a ANOVA, com um e mais fatores, ao caso em que as dimensões das amostras são desconhecidas, devendo ser tratadas como realizações de variáveis aleatórias. Esta abordagem deve ser baseada na escolha adequada da distribuição destas variáveis. No presente trabalho são consideradas duas situações distintas:

- No primeiro caso assumiremos que as variáveis aleatórias seguem distribuições de Poisson, situação em que a ocorrência das observações corresponde a processos de contagem e não existem limites superiores para as dimensões das amostras (tal como ilustrado no exemplo anterior, referente à comparação de patologias);

- No segundo caso, consideraremos a distribuição Binomial, quando existe um limite superior para as dimensões das amostras, que nem sempre é atingido uma vez que podem ocorrer falhas nas observações.

Como resultados, serão obtidas as estatísticas de teste e suas distribuições, condicional e não condicional assumindo as dimensões das amostras como aleatórias, para modelos de efeitos fixos, modelos de efeitos aleatórios e modelos mistos.

Adicionalmente, serão apresentadas várias aplicações com registos do cancro no Brasil que nos permitirão ilustrar a utilidade da nossa abordagem assim como comparar os resultados obtidos com os da ANOVA usual.

# Palavras-chave

Testes $F$, amostras de dimensão aleatória, processos de contagem, falhas de observações, registros do cancro no Brasil.

# Abstract

The analysis of variance (ANOVA) is routinely used in several research areas, namely in medical research, agriculture or psychology, to name just a few, where the sample sizes may not be previously known. This often occurs when there is a fixed time span for collecting the observations. An illustrative example of this corresponds to the collection of observations during a given time period for the comparison of pathologies from patients arriving at a hospital.

In this work, we aim to extend the one-way and multi-way ANOVA to the case where the sample sizes are unknown. We will assume the sample sizes as realizations of random variables. This approach must be based on an adequate choice of the distribution of these variables. For this, we will consider two distinct situations:

- In the first case, we will assume the Poisson distribution when the occurrence of the observations corresponds to a counting process and there is no upper bound for the sample sizes (as illustrated in the example concerning the comparison of pathologies);

- In the second case, we will consider the Binomial distribution if there is an upper bound for the sample sizes, which is not always achieved since we may have observations failures.

As results, we will obtain the test statistics and their conditional distribution and unconditional distribution under the assumption that we have random sample sizes, for fixed effects models, random effects models and mixed models.

This new approach will be illustrated through several applications on cancer registries from Brazil. This will enable us to show the usefulness of our approach as well as to compare the obtained results with the usual ANOVA results.

# Keywords

$F$-tests, random sample sizes, counting processes, observations failures, cancer registries from Brazil.

# Índice

# Lista de Figuras

# Lista de Tabelas

# Lista de Acrónimos

# Capítulo 1

# Introdução

A análise de variância (ANOVA) é atualmente um dos métodos estatísticos mais usados em aplicações práticas nas mais diversas áreas da ciência. Historicamente, esta técnica foi introduzida por Ronald A. Fisher em 1918 quando estudava problemas na área da agricultura, ver Scheffé (1959).

A ANOVA é uma técnica estatística que tem como principal objetivo a comparação de mais do que dois grupos no que respeita à localização. Visa analisar as observações que dependem de vários tipos de efeitos que atuam em simultâneo para decidir quais os efeitos importantes e estimar esses efeitos.

É essencialmente um processo baseado na decomposição de variação total dos dados (variância total ou soma dos quadrados total) em partes que podem ser atribuídas a causas conhecidas (variação entre grupos, soma dos quadrados dos tratamentos) e numa parte devido a causas desconhecidas (variação dentro dos grupos, soma dos quadrados do erro).

Foi portanto R. A. Fisher (1918, 1925, 1935) que introduziu os termos "variância" e "análise de variância" na estatística. O último termo, que parece mais apropriado para os modelos de efeitos aleatórios, pode ter constituído o caminho pelo qual Fisher originalmente abordou o assunto, ver Scheffé (1959).

O exemplo mais simples é o caso em que existem vários grupos de observações classificados através de um só fator (por exemplo, grupos de pacientes sujeitos a diferentes tratamentos para uma mesma patologia). Tem-se portanto a análise de variância com um fator (*One-way* ANOVA). Só é legítimo considerar este fator como sendo a causa das diferenças entre as médias se puder admitir a homogeneidade das populações em relação a todos os outros fatores que poderiam ser relevantes para a explicação do fenómeno. Em muitas situações porém poderá haver mais do que um fator a influenciar os resultados das observações. Neste caso estamos perante uma análise de variância com mais do que um fator (*Multi-way* ANOVA). Consideremos, por exemplo, que se pretende avaliar a eficácia de diferentes medicamentos, em homens e mulheres, no tratamento de uma determinada patologia. Coloca-se a questão se o tratamento é influenciado pelo tipo de medicamento administrado e pelo género do paciente. Neste caso estamos perante uma ANOVA com dois fatores.

A análise de variância tem tantos níveis ou efeitos quantos tratamentos (grupos) distintos forem considerados. Quando os níveis são fixados à partida diz-se que se tem uma ANOVA com efeitos fixos. Se estes forem selecionados aleatoriamente de um conjunto alargado de níveis, não sendo possível estudá-los a todos, teremos uma ANOVA com efeitos aleatórios.

Em inúmeras situações práticas, onde se utiliza a ANOVA, pode não ser possível saber previamente as dimensões das amostras. Esta situação ocorre, por exemplo, quando o tempo para a recolha das observações é fixado à partida. Um bom exemplo é a recolha de observações durante um determinado período de tempo para a comparação de patologias de pacientes que chegam às urgências de um hospital, ver por exemplo Moreira et al. (2013), Nunes et al. (2013, 2014). Um outro exemplo surge quando uma das patologias é rara. Neste caso, o número de pacientes com essa patologia pode não ser conhecido, ver Nunes et al. (2012a).

O objetivo desta tese é estender a ANOVA, com um e mais fatores, ao caso em que as dimensões das amostras não são conhecidas, devendo ser tratadas como realizações de variáveis aleatórias. Assim consideraremos as dimensões das amostras como realizações, $n_1, \ldots, n_m$, das variáveis aleatórias independentes, $N_1, \ldots, N_m$, ver por exemplo Mexia et al. (2011), Nunes et al. (2013, 2014, 2015). Esta abordagem deverá basear-se na escolha adequada das distribuições de $N_1, \ldots, N_m$. Iremos assumir duas situações distintas:

- consideraremos que a ocorrência das observações corresponde a processos de contagem, o que nos leva a assumir que $N_1, \ldots, N_m$ seguem distribuições de Poisson;

- assumiremos a distribuição Binomial caso os limites superiores para as dimensões das amostras sejam conhecidos e possam ocorrem falhas nas observações.

Uma vez apresentado o tema em que se insere o trabalho, de seguida falaremos um pouco da forma como este está estruturado.

No Capítulo 2, são apresentados alguns conceitos e resultados preliminares importantes para os capítulos seguintes, nomeadamente, alguns conceitos e resultados algébricos com aplicação à estatística. São apresentadas ainda distribuições teóricas importantes para as dimensões das amostras e da ANOVA e faz-se uma breve introdução aos modelos lineares. No que diz respeito aos modelos mistos, é considerada a estrutura ortogonal por blocos ($OBS$). Por fim é feita uma pequena abordagem às extensões $L$, uma classe de modelos que será utilizada na formulação dos modelos mistos considerando a dimensão das amostras como aleatórias.

No Capítulo 3, vamos estender a ANOVA, com um e mais fatores, ao caso em que as dimensões das amostras não são conhecidas. Assumimos que a ocorrência das observações corresponde a processos de Poisson. São obtidas as estatísticas de teste e as suas distribuições condicional e não condicional. A distribuição condicional é obtida assumindo-se que $N_i = n_i,\ i = 1, \ldots, m$, ou seja, que as dimensões das amostras são fixas, correpondendo portanto à abordagem usual. Quando descondicionamos esta distribuição em ordem a $N_i,\ i = 1, \ldots, m$, obtemos a distribuição não condicional da estatística e portanto as dimensões das amostras são consideradas como variáveis aleatórias. Mediante a situação prática em questão, a distribuição não condicional será obtida assumindo que se tem uma dimensão mínima global para as amostras, situação já considerada em alguns trabalhos anteriormente publicados, ver por exemplo Capistrano et al. (2015), Mexia et al. (2011) e Nunes et al. (2013, 2014), ou que a dimensão mínima é requerida para cada uma das amostras, ver Nunes et al. (2015). Apresentamos ainda uma forma alternativa para a obtenção dos valores críticos considerando apenas um fator. Neste capítulo consideramos modelos de efeitos fixos, modelos de efeitos aleatórios e modelos mistos.

Quanto ao Capítulo 4, vamos alargar o tratamento apresentado no Capítulo 3 ao caso em que podem ocorrer falhas nas observações. Assumiremos portanto que $N_1, \ldots, N_m$ seguem distribuições Binomiais. Consideramos modelos de efeitos fixos e modelos mistos.

Por fim, no Capítulo 5, são apresentadas as principais conclusões obtidas no decorrer deste trabalho. São ainda referidos alguns desenvolvimentos que pretendemos realizar no futuro.

É importante ainda referir que ao longo dos Capítulos 3 e 4 são apresentadas várias aplicações com dados reais, referentes a registos do Cancro no Brasil, por forma a mostrar a aplicabilidade desta nova abordagem. Os dados utilizados foram disponibilizados pelo Instituto Nacional do Câncer José Alencar Gomes da Silva (INCA). Os valores observados das estatísticas assim como os quantis das distribuições condicional e não condicional foram obtidos com recurso ao *software* R.

# Capítulo 2

# Conceitos e resultados auxiliares

Este capítulo contém uma breve abordagem a alguns conceitos importantes e bem conhecidos na área da estatística. Inicialmente são apresentados alguns resultados algébricos sobre matrizes inversas generalizadas, matrizes de projeção ortogonal, produto de Kronecker de matrizes e álgebras de Jordan comutativas. Seguidamente apresentamos resultados bem conhecidos referentes às distribuições teóricas usadas neste trabalho. Finalmente é feita uma breve introdução aos modelos lineares, onde se inserem os modelos mistos com estrutura ortogonal por blocos e as extensões $L$.

Estes conceitos serão úteis para o desenvolvimento e obtenção dos resultados dos capítulos seguintes e grande parte deles podem ser encontrados, por exemplo, em Schott (1997), Horn e Johnson (1985) e Rao (1973).

## 2.1 Resultados algébricos

**Definição 2.1** *A característica de uma matriz $\boldsymbol{A}$ é igual ao número máximo de linhas (ou colunas) linearmente independentes e denota-se por $car(\boldsymbol{A})$.*

**Definição 2.2** *Consideremos a matriz quadrada $\boldsymbol{A}$ de ordem $n$ com coeficientes reais. Designa-se por raio espectral de $\boldsymbol{A}$, denotando-se por $\rho(\boldsymbol{A})$,*

$$\rho(\boldsymbol{A}) = max\{|\theta_i|,\ i = 1, \ldots, n\},$$

*com $\theta_i,\ i = 1, \ldots, n$, os valores próprios de $\boldsymbol{A}$.*

**Definição 2.3** *Uma matriz $\boldsymbol{A}$ de ordem $n$ cujas colunas (ou linhas) formam um conjunto ortonormado de vetores é designada por matriz ortogonal. Assim, $\boldsymbol{A}$ é ortogonal se e somente se*

$$\boldsymbol{A}^{'}\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^{'} = \boldsymbol{I}_n,$$

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

**Proposição 2.1** *Se $\boldsymbol{A}$ é uma matriz simétrica com valores próprios $\theta_1, ..., \theta_n$, e vetores próprios normalizados $\gamma_1, ..., \gamma_n$ tem-se*

$$\boldsymbol{A} = \boldsymbol{PD}(\theta_1, ..., \theta_n)\boldsymbol{P}^{'} = \sum_{i=1}^{n} \theta_i \gamma_i \gamma_i^{'},$$

*onde $\boldsymbol{P}$ é uma matriz cujas colunas são os vetores $\gamma_1, ..., \gamma_n$, $\boldsymbol{P}'$ representa a transposta da matriz $\boldsymbol{P}$ e $\boldsymbol{D}(\theta_1, ..., \theta_n)$ é uma matriz diagonal cujos elementos principais são $\theta_1, ..., \theta_n$. Ao somatório, $\sum_{i=1}^{n} \theta_i \gamma_i \gamma_i^{'}$, chama-se decomposição espectral de $\boldsymbol{A}$.*

**Dem:** Uma vez que $\boldsymbol{P}$ é uma matriz cujas colunas são os vetores normalizados $\gamma_1, ..., \gamma_n$, então $\boldsymbol{P}$ é ortogonal. Temos portanto

$$\boldsymbol{PP}^{'} = \boldsymbol{I}_n \Leftrightarrow \boldsymbol{APP'} = \boldsymbol{A} \Leftrightarrow \boldsymbol{A} \left[\gamma_1 \dots \gamma_n\right] \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_n \end{bmatrix} = \boldsymbol{A} \Leftrightarrow$$

$$\Leftrightarrow \left[\boldsymbol{A}\gamma_1 \dots \boldsymbol{A}\gamma_n\right] \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_n \end{bmatrix} = \boldsymbol{A} \Leftrightarrow \left[\theta_1\gamma_1 \dots \theta_n\gamma_n\right] \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_n \end{bmatrix} = \boldsymbol{A} \Leftrightarrow$$

$$\Leftrightarrow \boldsymbol{PD}\left(\theta_1, \dots, \theta_n\right)\boldsymbol{P'} = \boldsymbol{A} \Leftrightarrow \sum_{i=1}^{n} \theta_i \gamma_i \gamma_i^{'} = \boldsymbol{A}.$$

■

**Corolário 2.1.1** *Seja $\boldsymbol{A}$ uma matriz simétrica, então a matriz ortogonal $\boldsymbol{P}$ diagonaliza $\boldsymbol{A}$, isto é,*

$$\boldsymbol{P}^{'}\boldsymbol{AP} = \boldsymbol{D}(\theta_1, ..., \theta_n).$$

**Proposição 2.2** *Se $\boldsymbol{A}$ é uma matriz com valores próprios $\theta_1, ..., \theta_n$, então:*

$$det(\boldsymbol{A}) = \prod_{i=1}^{n} \theta_i;$$

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

onde $det(\boldsymbol{A})$ e $tr(\boldsymbol{A})$ representam, respectivamente, o determinante e o traço da matriz $\boldsymbol{A}$.

### 2.1.1 Matrizes Inversas Generalizadas

Uma matriz $\boldsymbol{A}$ diz-se singular se não existir a sua inversa, $\boldsymbol{A}^{-1}$. No entanto pode ser calculada uma inversa generalizada, que tal como o nome indica, é uma generalização da matriz inversa, ver Schott (1997).

**Definição 2.4** *A matriz inversa generalizada de uma matriz $\boldsymbol{A}$ do tipo $r \times s$, que vamos representar por $\boldsymbol{A}^{-}$, será uma matriz do tipo $s \times r$ que verifica a seguinte igualdade*

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

**Definição 2.5** *Qualquer que seja a matriz $\boldsymbol{A}$ existe uma e uma só matriz $\boldsymbol{A}^{+}$ tal que:*

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

A matriz $\boldsymbol{A}^{+}$ é chamada de inversa de Moore-Penrose de $\boldsymbol{A}$, ver por exemplo Pollock (1979). Se $\boldsymbol{A}$ for uma matriz regular então $\boldsymbol{A}^{+} = \boldsymbol{A}^{-1}$.

Note-se que se $\boldsymbol{A}$ for uma matriz simétrica de ordem $k$, com

$$car(\boldsymbol{A}) = l < k,$$

podem-se ordenar os vetores linha de uma matriz $\boldsymbol{P}$ ortogonal, diagonalizadora de $\boldsymbol{A}$, de forma a ter-se

$$\boldsymbol{P}^{'}\boldsymbol{A}\boldsymbol{P} = \boldsymbol{D}(\theta_1, ..., \theta_l, 0, ..., 0),$$

com $\theta_1, ..., \theta_l$ os valores próprios não nulos de $\boldsymbol{A}$. Temos portanto

$$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}(\theta_1, ..., \theta_l, 0, ..., 0)\boldsymbol{P}^{'}$$

e

$$\boldsymbol{A}^{+} = \boldsymbol{P}\boldsymbol{D}(\theta_1^{+}, ..., \theta_l^{+}, 0, ..., 0)\boldsymbol{P}^{'},$$

onde $\theta_i^{+} = \theta_i^{-1},\ i = 1, \ldots, l.$

### 2.1.2 Matrizes de projeção ortogonal

Seja agora $S$ um subespaço vetorial do espaço vetorial $E$.

**Definição 2.6** *O complemento ortogonal de $S$, denotado por $S^{\perp}$, é o conjunto de todos os vetores de $E$ que são ortogonais a cada vetor de $S$. Então*

$$S^{\perp} = \{\boldsymbol{x} : \boldsymbol{x} \in E \quad e\, \boldsymbol{x}^{'}\boldsymbol{y} = 0, \quad para\ todo\, \boldsymbol{y} \in S\}.$$

**Obervação 2.1** *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 $E$ de dimensão $r$, então $S^{\perp}$ é um subespaço vetorial de $E$ de dimensão $m-r$.*

**Definição 2.7** *Diz-se que uma base $B = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m\}$ do espaço vetorial $E$ é ortogonal se o conjunto dos seus vetores for ortogonal.*

**Definição 2.8** *Uma base $B = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_m\}$ de um espaço vetorial $E$ diz-se ortonormada, se $B$ é uma base ortogonal e todos os seus vetores são unitários, $||\boldsymbol{v}_i|| = 1$, $i = 1, \ldots, m$, ou seja,*

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

**Proposição 2.3** *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 $E$. Se $\boldsymbol{x} \in E$, então a projeção ortogonal de $\boldsymbol{x}$ em $S$ é dada por $\boldsymbol{Z}_1\boldsymbol{Z}_1^{'}\boldsymbol{x}$.*

De forma análoga tem-se

**Proposição 2.4** *Suponhamos 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 $m-r$). Se $\boldsymbol{x} \in E$, então a projeção ortogonal de $\boldsymbol{x}$ em $S^{\perp}$ é dada por $\boldsymbol{Z}_2\boldsymbol{Z}_2^{'}\boldsymbol{x}$.*

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 $S^{\perp}$. Note-se que

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

é a $MPO$ sobre o espaço vetorial $E$.

Assim, visto que $\boldsymbol{Z}\boldsymbol{Z}^{'} = \boldsymbol{I}_m$, tem-se

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

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

**Definição 2.9** *O Espaço Imagem de uma matriz $\boldsymbol{A}$, representado por $R(\boldsymbol{A})$, é dado por*

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

**Definição 2.10** *O espaço nulidade de uma matriz* $\boldsymbol{A}$*,* $\mathcal{N}(\boldsymbol{A})$*, é o espaço que contém todas as soluções da equação* $\boldsymbol{Ax} = \boldsymbol{0}$*, ou seja*

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

Sendo $\boldsymbol{P}$ a $MPO$ sobre $S$, sabe-se que $R(\boldsymbol{P}) = S$, pois

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

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

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

A soma direta de subespaços vetoriais, $\oplus$ , pode também ser representada pelo espaço imagem de matrizes de projeção ortogonais. As proposições que se seguem mostram isso.

**Proposição 2.5** *Sejam as* $MPO$*,* $\boldsymbol{P}_1, ..., \boldsymbol{P}_k$*, tais que* $\boldsymbol{P}_i\boldsymbol{P}_j = \boldsymbol{0}$*, com* $i \neq j,\ i, j = 1, \ldots, k$*, tem-se*

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

**Proposição 2.6** *Seja* $\boldsymbol{P}$ *uma* $MPO$ *associada a um subespaço vetorial* $S$*. Suponhamos que* $S$ *é uma soma direta de subespaços, isto é,* $S = \oplus_{i=1}^{k} S_i$*. Então existem* $MPO$ *únicas,* $\boldsymbol{P}_1, ..., \boldsymbol{P}_k$ *tais que*

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

*e* $\boldsymbol{P}_i\boldsymbol{P}_j = \boldsymbol{0}$*, com* $i \neq j$*.*

**Proposição 2.7** *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* $\boldsymbol{A}(\boldsymbol{A}^{'}\boldsymbol{A})^{+}\boldsymbol{A}^{'}$ *e* $(\boldsymbol{A}^{'}\boldsymbol{A})^{+}(\boldsymbol{A}^{'}\boldsymbol{A})$*, respectivamente.*

Temos agora a seguinte proposição

**Proposição 2.8** *A matriz* $\boldsymbol{P}$ *é uma matriz de projeção ortogonal* $MPO$*, se e somente se for simétrica e idempotente.*

**Dem:** Dado um espaço vetorial $E$, qualquer vetor $\boldsymbol{x} \in E$, pode ser expresso como $\boldsymbol{x} = \boldsymbol{x}_1 + \boldsymbol{x}_2$, onde $\boldsymbol{x}_1$ pertence a um subespaço $S \subseteq E$ e $\boldsymbol{x}_2$ ao seu complemento ortogonal, $S^{\perp}$. Se $\boldsymbol{P}$ é $MPO$ de $\boldsymbol{x}$ sobre $S$, $\boldsymbol{Px} = \boldsymbol{x}_1$ e $\boldsymbol{Px}_1 = \boldsymbol{x}_1$ o que significa que outras projeções sobre $S$ não devem ter efeito em $\boldsymbol{Px}_1$. Assim $\boldsymbol{Px} = \boldsymbol{x}_1 = \boldsymbol{Px}_1 = \boldsymbol{P}(\boldsymbol{Px}) = \boldsymbol{P}^2\boldsymbol{x}$, logo $(\boldsymbol{P} - \boldsymbol{P}^2)\boldsymbol{x} = 0$. Uma vez que $\boldsymbol{x}$ é arbitrário, temos $\boldsymbol{P} = \boldsymbol{P}^2$, o que significa que $\boldsymbol{P}$ é uma matriz idempotente.

Sendo $\boldsymbol{x}_1$ a projeção ortogonal de $\boldsymbol{x}$, observando que $\boldsymbol{x}_2 = \boldsymbol{x} - \boldsymbol{x}_1 = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{x}$, temos que $0 = (\boldsymbol{P}\boldsymbol{x})^{'}(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{x} = \boldsymbol{x}^{'}\boldsymbol{P}(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{x} = \boldsymbol{x}_1^{'}\boldsymbol{x}_2 = \boldsymbol{x}^{'}\boldsymbol{P}'(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{x}$, por isso $\boldsymbol{P}'(\boldsymbol{I} - \boldsymbol{P}) = 0$ e então $\boldsymbol{P}' = \boldsymbol{P}'\boldsymbol{P}$ e $\boldsymbol{P} = \boldsymbol{P}\boldsymbol{P}'$ o que significa que $\boldsymbol{P}$ é simétrica.

Por outro lado, se $\boldsymbol{P}$ é uma matriz simétrica e idempotente,

$$\begin{aligned} \boldsymbol{x}_1^{'}\boldsymbol{x}_2 = \boldsymbol{x}^{'}\boldsymbol{P}'(\boldsymbol{x} - \boldsymbol{x}_1) &= \boldsymbol{x}^{'}\boldsymbol{P}'(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{x} \\ &= \boldsymbol{x}^{'}\boldsymbol{P}(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{x} \\ &= \boldsymbol{x}^{'}(\boldsymbol{P}\boldsymbol{x} - \boldsymbol{P}^2\boldsymbol{x}) \\ &= \boldsymbol{x}^{'}(\boldsymbol{P} - \boldsymbol{P}^2)\boldsymbol{x} \\ &= \boldsymbol{x}^{'}(\boldsymbol{P} - \boldsymbol{P})\boldsymbol{x} = 0, \end{aligned}$$

portanto, $\boldsymbol{P}$ é uma $MPO$. ■

**Proposição 2.9** *Se $\boldsymbol{P}$ é uma $MPO$, os seus valores próprios serão iguais a $0$ ou $1$.*

**Dem:** Sendo $\boldsymbol{P}$ uma $MPO$ e $\boldsymbol{x}$ um vetor próprio da matriz $\boldsymbol{P}$ para o valor próprio $\theta$, temos

$$\theta\boldsymbol{x} = \boldsymbol{P}\boldsymbol{x} = \boldsymbol{P}^2\boldsymbol{x} = \boldsymbol{P}(\boldsymbol{P}\boldsymbol{x}) = \boldsymbol{P}(\theta\boldsymbol{x}) = \theta\boldsymbol{P}\boldsymbol{x} = \theta^2\boldsymbol{x}.$$

Vindo $\theta\boldsymbol{x} = \theta^2\boldsymbol{x} \Leftrightarrow \theta(1-\theta)\boldsymbol{x} = 0 \Leftrightarrow \theta = 0 \vee \theta = 1$. ■

Da proposição (2.8) vem que, se $\boldsymbol{P}$ é uma $MPO$ então

$$\boldsymbol{P}^{+} = \boldsymbol{P}.$$

**Definição 2.11** *Duas matrizes de projeção ortogonais, $\boldsymbol{P}_1$ e $\boldsymbol{P}_2$, são mutuamente ortogonais, $\boldsymbol{P}_1 \perp \boldsymbol{P}_2$, se*

$$\boldsymbol{P}_2\boldsymbol{P}_1 = \boldsymbol{0},$$

*onde $\boldsymbol{0}$ representa a matriz nula.*

**Proposição 2.10** *Se $\boldsymbol{P}_1$ e $\boldsymbol{P}_2$ são matrizes de projeção ortogonal mutuamente ortogonais, $MPOMO$, então $\boldsymbol{P}_1 + \boldsymbol{P}_2$ é uma $MPO$.*

**Dem:** Uma vez que $\boldsymbol{P}_1$ e $\boldsymbol{P}_2$ são simétricas, idempotentes e mutuamente ortogonais tem-se

$$(\boldsymbol{P}_1 + \boldsymbol{P}_2)(\boldsymbol{P}_1 + \boldsymbol{P}_2) = \boldsymbol{P}_1\boldsymbol{P}_1 + \boldsymbol{P}_1\boldsymbol{P}_2 + \boldsymbol{P}_2\boldsymbol{P}_1 + \boldsymbol{P}_2\boldsymbol{P}_2 = \boldsymbol{P}_1 + \boldsymbol{P}_2,$$

logo $\boldsymbol{P}_1 + \boldsymbol{P}_2$ é idempotente. Além disso a soma de matrizes simétricas é uma matriz simétrica. ■

### 2.1.3 Produto de Kronecker de matrizes

Nesta seção apresentamos algumas noções e propriedades sobre o produto de Kronecker de matrizes, que representaremos por $\otimes$. Ao contrário do produto usual de matrizes este está defi-

nido para quaisquer tipos de matrizes. Esta operação foi amplamente estudada por exemplo por Steeb (1991).

**Definição 2.12** *Considerando a matriz* $\boldsymbol{A} = [a_{i,j}]$ *de ordem* $r \times s$*, e uma matriz* $\boldsymbol{B}$ *qualquer, tem-se*

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

Seja $R_j$ o espaço imagem de $\boldsymbol{B}_j$, $j = 1, 2$, $R(\boldsymbol{B}_j)$, $j = 1, 2$, tem-se que

$$R_1 \otimes R_2 = R(\boldsymbol{B}_1 \otimes \boldsymbol{B}_2).$$

O produto de Kronecker não é comutativo mas satisfaz a propriedade associativa, tendo-se a proposição seguinte.

**Proposição 2.11** *Quaisquer que sejam as matrizes* $\boldsymbol{A}$*,* $\boldsymbol{B}$ *e* $\boldsymbol{C}$*,*

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

**Proposição 2.12** *Sejam* $\boldsymbol{A}$ *e* $\boldsymbol{B}$ *matrizes de ordem* $r \times s$*, e* $\boldsymbol{C}$ *e* $\boldsymbol{D}$ *matrizes de ordem* $v \times t$*, tem-se*

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

*o que significa que o produto de Kronecker satisfaz a propriedade distributiva.*

**Proposição 2.13** *Quaisquer que sejam as matrizes* $\boldsymbol{A}$ *e* $\boldsymbol{B}$*, tem-se*

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

Se o produto usual de matrizes , $\boldsymbol{A}_h\boldsymbol{B}_h$, $h = 1, 2$, estiver definido ter-se-á, com $\boldsymbol{A}_1 = [a_{i,j}]$, $\boldsymbol{A}_2$, $\boldsymbol{B}_1 = [b_{j,k}]$ e $\boldsymbol{B}_2$ matrizes do tipo $r \times s$, $p \times q$, $s \times t$, e $q \times v$, respectivamente,

$$
\begin{aligned}
(\boldsymbol{A}_1 \otimes \boldsymbol{A}_2)(\boldsymbol{B}_1 \otimes \boldsymbol{B}_2) &= \begin{bmatrix} a_{1,1}\boldsymbol{A}_2 & \cdots & a_{1,s}\boldsymbol{A}_2 \\ \vdots & \ddots & \vdots \\ a_{r,1}\boldsymbol{A}_2 & \cdots & a_{r,s}\boldsymbol{A}_2 \end{bmatrix} \begin{bmatrix} b_{1,1}\boldsymbol{B}_2 & \cdots & b_{1,t}\boldsymbol{B}_2 \\ \vdots & \ddots & \vdots \\ b_{s,1}\boldsymbol{B}_2 & \cdots & b_{s,t}\boldsymbol{B}_2 \end{bmatrix} \\
&= \begin{bmatrix} (\sum_{j=1}^{s} a_{1,j}b_{j,1})\boldsymbol{A}_2\boldsymbol{B}_2 & \cdots & (\sum_{j=1}^{s} a_{1,j}b_{j,t})\boldsymbol{A}_2\boldsymbol{B}_2 \\ \vdots & \ddots & \vdots \\ (\sum_{j=1}^{s} a_{r,j}b_{j,1})\boldsymbol{A}_2\boldsymbol{B}_2 & \cdots & (\sum_{j=1}^{s} a_{r,j}b_{j,t})\boldsymbol{A}_2\boldsymbol{B}_2 \end{bmatrix} \\
&= \begin{bmatrix} (\sum_{j=1}^{s} a_{1,j}b_{j,1}) & \cdots & (\sum_{j=1}^{s} a_{1,j}b_{j,t}) \\ \vdots & \ddots & \vdots \\ (\sum_{j=1}^{s} a_{r,j}b_{j,1}) & \cdots & (\sum_{j=1}^{s} a_{r,j}b_{j,t}) \end{bmatrix} \otimes (\boldsymbol{A}_2\boldsymbol{B}_2) \\
&= (\boldsymbol{A}_1\boldsymbol{B}_1) \otimes (\boldsymbol{A}_2\boldsymbol{B}_2). \qquad (2.1.1)
\end{aligned}
$$

Da definição de $\otimes$ resulta, com $\boldsymbol{A} = [a_{i,j}]$ uma matriz do tipo $r \times s$, que

$$
(\boldsymbol{A} \otimes \boldsymbol{B})^{'} = \begin{bmatrix} a_{1,1}\boldsymbol{B}^{'} & \cdots & a_{r,1}\boldsymbol{B}^{'} \\ \vdots & \ddots & \vdots \\ a_{1,s}\boldsymbol{B}^{'} & \cdots & a_{r,s}\boldsymbol{B}^{'} \end{bmatrix} = \boldsymbol{A}^{'} \otimes \boldsymbol{B}^{'}. \quad (2.1.2)
$$

**Proposição 2.14** *Com $\boldsymbol{P}_j$ diagonalizadora ortogonal de $\boldsymbol{A}_j$, $j = 1, 2$, $\boldsymbol{P}_1 \otimes \boldsymbol{P}_2$ será diagonalizadora ortogonal de $\boldsymbol{A}_1 \otimes \boldsymbol{A}_2$ e os valores próprios de $\boldsymbol{A}_1 \otimes \boldsymbol{A}_2$ serão o produto dos valores próprios de $\boldsymbol{A}_1$ pelos valores próprios de $\boldsymbol{A}_2$.*

**Dem:** Como $\boldsymbol{P}_j$ é diagonalizadora ortogonal de $\boldsymbol{A}_j,\ j = 1, 2$, tem-se $\boldsymbol{P}_j\boldsymbol{A}_j\boldsymbol{P}_j^{'} = \boldsymbol{D}_j$, com $\boldsymbol{D}_j,\ j = 1, 2$ uma matriz diagonal. Aplicando (2.1.1) e (2.1.2), obtém-se

$$
\begin{aligned}
(\boldsymbol{P}_1 \otimes \boldsymbol{P}_2)(\boldsymbol{A}_1 \otimes \boldsymbol{A}_2)(\boldsymbol{P}_1 \otimes \boldsymbol{P}_2)^{'} &= (\boldsymbol{P}_1 \otimes \boldsymbol{P}_2)(\boldsymbol{A}_1 \otimes \boldsymbol{A}_2)(\boldsymbol{P}_1^{'} \otimes \boldsymbol{P}_2^{'}) \\
&= (\boldsymbol{P}_1\boldsymbol{A}_1\boldsymbol{P}_1^{'}) \otimes (\boldsymbol{P}_2\boldsymbol{A}_2\boldsymbol{P}_2^{'}) \\
&= \boldsymbol{D}_1 \otimes \boldsymbol{D}_2,
\end{aligned}
$$

logo $\boldsymbol{P}_1 \otimes \boldsymbol{P}_2$ é diagonalizadora ortogonal de $\boldsymbol{A}_1 \otimes \boldsymbol{A}_2$, visto que $\boldsymbol{D}_1 \otimes \boldsymbol{D}_2$ é uma matriz diagonal. Para terminar a demonstração basta observar que os elementos principais de $\boldsymbol{D}_1 \otimes \boldsymbol{D}_2$ são o produto dos elementos principais de $\boldsymbol{D}_1$ pelos de $\boldsymbol{D}_2$. ■

Da aplicação desta proposição resulta que o número de valores próprios não nulos de $\boldsymbol{B}_1\boldsymbol{B}_1' \otimes \boldsymbol{B}_2\boldsymbol{B}_2'$ é o produto do número de valores próprios relativos a $\boldsymbol{B}_1\boldsymbol{B}_1'$ e $\boldsymbol{B}_2\boldsymbol{B}_2'$. Como esses números correspondem às características das matrizes $\boldsymbol{B}_1\boldsymbol{B}_1' \otimes \boldsymbol{B}_2\boldsymbol{B}_2'$, $\boldsymbol{B}_1\boldsymbol{B}_1'$ e $\boldsymbol{B}_2\boldsymbol{B}_2'$ tem-se que

$$\begin{aligned} car(\boldsymbol{B}_1 \otimes \boldsymbol{B}_2) &= car[(\boldsymbol{B}_1 \otimes \boldsymbol{B}_2)(\boldsymbol{B}_1 \otimes \boldsymbol{B}_2)^{'}] \\ &= car[(\boldsymbol{B}_1 \otimes \boldsymbol{B}_2)(\boldsymbol{B}_1^{'} \otimes \boldsymbol{B}_2^{'})] \\ &= car[(\boldsymbol{B}_1\boldsymbol{B}_1^{'}) \otimes (\boldsymbol{B}_2\boldsymbol{B}_2^{'})] \\ &= car(\boldsymbol{B}_1\boldsymbol{B}_1^{'})car(\boldsymbol{B}_2\boldsymbol{B}_2^{'}) \\ &= car(\boldsymbol{B}_1)car(\boldsymbol{B}_2). \end{aligned}$$

Como vimos atrás, $R_j = R(\boldsymbol{B}_j)$, $j = 1, 2$, e $R_1 \otimes R_2 = R(\boldsymbol{B}_1 \otimes \boldsymbol{B}_2)$, vindo

$$\begin{aligned} dim(R_1 \otimes R_2) &= car(\boldsymbol{B}_1 \otimes \boldsymbol{B}_2) \\ &= car(\boldsymbol{B}_1)car(\boldsymbol{B}_2) \\ &= dim(R_1)dim(R_2). \end{aligned}$$

**Proposição 2.15** *O produto de Kronecker de $MPO$ é uma $MPO$.*

**Dem:** Sendo $\boldsymbol{G}_1$ e $\boldsymbol{G}_2$ $MPO$ tem-se $\boldsymbol{G}_i^{'} = \boldsymbol{G}_i$ e $\boldsymbol{G}_i\boldsymbol{G}_i = \boldsymbol{G}_i,\ i = 1, 2$, logo

$$(\boldsymbol{G}_1 \otimes \boldsymbol{G}_2)^{'} = \boldsymbol{G}_1^{'} \otimes \boldsymbol{G}_2^{'} = \boldsymbol{G}_1 \otimes \boldsymbol{G}_2,$$

assim como

$$(\boldsymbol{G}_1 \otimes \boldsymbol{G}_2)(\boldsymbol{G}_1 \otimes \boldsymbol{G}_2) = (\boldsymbol{G}_1\boldsymbol{G}_1) \otimes (\boldsymbol{G}_2\boldsymbol{G}_2) = \boldsymbol{G}_1 \otimes \boldsymbol{G}_2,$$

logo $\boldsymbol{G}_1 \otimes \boldsymbol{G}_2$ é simétrica e idempotente o que significa que é $MPO$. ■

### 2.1.4 Álgebras de Jordan Comutativas

As Álgebras de Jordan foram introduzidas por Pascual Jordan, em parceria com John von Neumann e Eugene Wigner, para resolver problemas de mecânica quântica, ver Jordan et al. (1934). Com Seely estas estruturas algébricas deram origem a uma linha de pesquisa com desenvolvimentos relevantes em inferência estatística linear e foram designadas por Espaços Vetoriais Quadráticos, ver Seely (1970a, 1970b, 1971, 1977) e Seely e Zyskind (1971).

Mais tarde, Michalski e Zmślony em (1996) e (1999), usaram as Álgebras de Jordan primeiro para testar hipóteses sobre componentes de variância e, depois para funções lineares dos parâmetros em modelos mistos.

As Álgebras de Jordan continuam a ser muito utilizadas, veja-se por exemplo Vanleuwen et al. (1998, 1999), Fonseca et al. (2006, 2007, 2008, 2009), Rodrigues e Mexia (2006), Jesus et al. (2007, 2009), Ferreira et al. (2013), Carvalho et al. (2015).

Interessa-nos fundamentalmente as álgebras Jordan comutativas de matrizes simétricas, $AJCS$. Estas são espaços lineares constituídos por matrizes simétricas que comutam e contêm os quadrados das suas matrizes.

Seely (1971) mostrou que toda a $AJCS$, $\mathcal{A}$, tem uma única base, a base principal, $bp(\mathcal{A})$, constituída por $MPOMO$.

Seja $\underline{\boldsymbol{Q}} = \{\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_m\}$ a base principal, $bp(\mathcal{A})$, da $AJCS$ $\mathcal{A}$ . Dada $\boldsymbol{M}$ uma matriz pertencente à $AJCS$ $\mathcal{A}$, tem-se

$$\boldsymbol{M} = \sum_{j=1}^{m} b_j \boldsymbol{Q}_j = \sum_{j \in \mathcal{C}(\boldsymbol{M})} b_j \boldsymbol{Q}_j,$$

com $\mathcal{C}(\boldsymbol{M}) = \{j : b_j \neq 0\}$.

A inversa de Moore-Penrose de $\boldsymbol{M}$ será dada por

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

onde $b_j^{+} = 0$, se $b_j = 0$ e $b_j^{+} = b_j^{-1}$, qualquer que seja $b_j \neq 0$, $j = 1, \ldots, m$, e

$$\mathcal{C}(\boldsymbol{M}^{+}) = \mathcal{C}(\boldsymbol{M}).$$

Assim, podemos concluir que uma $AJCS$ contém as inversas de Moore-Penrose de quaisquer das suas matrizes.

Com

$$\mathcal{R}_j = R(\boldsymbol{Q}_j),\ j = 1, \ldots, m$$

e

$$g_j = car(\boldsymbol{Q}_j),\ j = 1, \ldots, m,$$

tem-se

- $R(\boldsymbol{M}) = \oplus_{j \in \mathcal{C}(\boldsymbol{M})} \mathcal{R}_j$;
- $car(\boldsymbol{M}) = \sum_{j \in \mathcal{C}(\boldsymbol{M})} g_j,$

onde $\oplus$ representa a soma direta ortogonal de subespaços. Além disso, a $MPO$, sobre $R(\boldsymbol{M})$ será

$$\boldsymbol{Q}(\boldsymbol{M}) = \sum_{j \in \mathcal{C}(\boldsymbol{M})} \boldsymbol{Q}_j.$$

**Proposição 2.16** *As $MPO$ pertencentes a uma $AJCS$, $\mathcal{A}$, são somas de matrizes da $bp(\mathcal{A})$.*

**Dem:** Seja $\boldsymbol{Q}$ uma $MPO$ pertencente a $\mathcal{A}$, tem-se

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

Como $\boldsymbol{Q}$ é idempotente e $\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_m$ são idempotentes e mutuamente ortogonais,

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

logo teremos $b_j^2 = b_j$ o que significa que $b_j = 0$ ou $b_j = 1, j = 1, \ldots, m$. Então, com $\mathcal{C}(\boldsymbol{Q}) = \{j : b_j \neq 0\}$, tem-se

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

■

Suponhamos agora que as matrizes de $\mathcal{A}$ são matrizes do tipo $n \times n$.

**Definição 2.13** *Uma $AJCS$, $\mathcal{A}$, diz-se completa se contém matrizes invertíveis.*

Se $\mathcal{A}$ contém matrizes invertíveis então teremos

$$\sum_{j=1}^{m} \boldsymbol{Q}_j = \boldsymbol{I}_n, \quad (2.1.3)$$

uma vez que se tem $\sum_{j=1}^{m} g_j = n$, o que significa que a soma das matrizes da base principal de uma $AJCS$ completa é igual a $\boldsymbol{I}_n$.

Seja $\boldsymbol{M}$ uma matriz pertencente a uma $AJCS$, $\boldsymbol{M}$ é regular se e somente se $\mathcal{C}(\boldsymbol{M}) = \{j : b_j \neq 0\} = \{1, \ldots, m\}$. Assim, dada

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

com $b_j \neq 0, j = 1, \ldots, m$, os $b_j, j = 1, \ldots, m$, serão os valores próprios de $\boldsymbol{M}$ com multiplicidades $g_j, j = 1, \ldots, m$. Ter-se-á então

$$det(\boldsymbol{M}) = \prod_{j=1}^{m} b_j^{g_j}$$

e

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

Dada a família de matrizes $\underline{\boldsymbol{M}} = \{\boldsymbol{M}_1, \ldots, \boldsymbol{M}_w\}$ de $\mathcal{A}$, tem-se

$$\boldsymbol{M}_i = \sum_{j=1}^{m} b_{i,j}\boldsymbol{Q}_j,\ i = 1, \ldots, w,$$

sendo $\boldsymbol{B} = [b_{i,j}]$ a matriz de transição entre $\underline{\boldsymbol{M}}$ e $\underline{\boldsymbol{Q}}$. As matrizes de $\boldsymbol{M}$ são linearmente independentes se e só se os vetores linha de $\boldsymbol{B}$ são linearmente independentes. Como $dim(\mathcal{A}) = m$, se $w = m$ e as matrizes $\boldsymbol{M}_1, \ldots, \boldsymbol{M}_m$ são linearmente independentes, os $m$ vetores linha de $\boldsymbol{B}$ serão linearmente independentes. Assim $\boldsymbol{B}$ será uma matriz $m \times m$ com $car(\boldsymbol{B}) = m$. Então $\boldsymbol{B}$ será invertível, e com $\boldsymbol{B}^{-1} = [b_{l,h}]$, teremos

$$\boldsymbol{Q}_l = \sum_{h=1}^{m} b_{l,h}\boldsymbol{M}_h,\ l = 1, \ldots, m,$$

e $\underline{\boldsymbol{M}} = \{\boldsymbol{M}_1, \ldots, \boldsymbol{M}_w\}$ será uma base de $\mathcal{A}$.

As matrizes de $\underline{\boldsymbol{M}} = \{\boldsymbol{M}_1, \ldots, \boldsymbol{M}_w\}$ comutam se e somente se forem diagonalizadas pela mesma matriz, $\boldsymbol{P}^{\circ}$, ver Schott (1997, pg 155).

Temos, então,

$$\underline{\boldsymbol{M}} \subset \boldsymbol{F}(\boldsymbol{P}^{\circ}),$$

com $\boldsymbol{F}(\boldsymbol{P}^{\circ})$ a família de matrizes diagonalizadas por $\boldsymbol{P}^{\circ}$. Uma vez que $\boldsymbol{F}(\boldsymbol{P}^{\circ})$ é uma $AJCS$, vemos que uma família de matrizes simétricas $n \times n$ está contida numa $AJCS$ se, e somente se elas comutarem. Como a intersecção de $AJCS$ resulta numa $AJCS$ existirá uma $AJCS$ mínima que contém $\underline{\boldsymbol{M}}$ , cujas matrizes comutam. Será a $AJCS$ gerada por $\underline{\boldsymbol{M}}$, $\mathcal{A}(\underline{\boldsymbol{M}})$, ver por exemplo Jacobson (1953).

## 2.2 Algumas Distribuições teóricas

Nessa seção apresentamos alguns resultados bem conhecidos sobre algumas distribuções que iremos utilizar nos capítulos seguintes, nomeadamente as distribuições discretas, Binomial e Poisson, e as distribuições contínuas, Normal, Qui-quadrado, Quocientes de Qui-quadrado e distribuição $F$. Não serão apresentadas demonstrações por serem resultados triviais e fáceis de encontrar na mais diversificada bibliografia, veja-se por exemplo Pestana e Velosa (2006).

### 2.2.1 Distribuição Binomial

Seja $X$ o número de sucessos obtidos na realização de $n$ provas de Bernoulli. $X$ tem distribuição Binomial com parâmetros $n$ e $p$, em que $p$ é a probabilidade de sucesso em cada prova, se a sua função de probabilidade for dada por, ver por exemplo Meyer (1970) e Pestana e Velosa (2006),

$$P(X = k) = \begin{pmatrix} n \\ k \end{pmatrix} p^k (1-p)^{n-k},\ k = 0, \ldots, n. \quad (2.2.4)$$

Usaremos a notação $X \sim B(n,p)$.

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

$$\begin{aligned}\varphi_X(t) = E\left(e^{tX}\right) &= \sum_{k=0}^{n} e^{tk} \left( \begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k} \\ &= \sum_{k=0}^{n} \left( \begin{array}{c} n \\ k \end{array} \right) (e^t p)^k (1-p)^{n-k} \\ &= (pe^t + 1 - p)^n.\end{aligned}$$

em que, na última igualdade, foi utilizada a fórmula do binómio de Newton. Como é bem conhecido pode-se encontrar o valor esperado de $X$, $E(X)$, e a variância de $X$, $Var(X)$, da seguinte forma:

$$\begin{aligned}\frac{d}{dt}\varphi_X(t) &= \frac{d}{dt}(pe^t + 1 - p)^n \\ &= n(pe^t + 1 - p)^{n-1} pe^t\end{aligned}$$

e, como $E(X) = \frac{d}{dt}\varphi_X(0)$, então

$$E(X) = n(p + 1 - p)^{n-1} p = np.$$

Por outro lado

$$\begin{aligned}\frac{d^2}{dt^2}\varphi_X(t) &= \frac{d^2}{dt^2}(pe^t + 1 - p)^n \\ &= n(n-1)(pe^t + 1 - p)^{n-2}(pe^t)^2 + npe^t(pe^t + 1 - p)^{n-1}.\end{aligned}$$

E, portanto,

$$E(X^2) = \frac{d^2}{dt^2}\varphi_X(0) = n(n-1)p^2 + pn,$$

e consequentemente, a $Var(X)$ será

$$\begin{aligned}Var(X) &= E(X^2) - E^2(X) \\ &= [n(n-1)p^2 + pn] - (np)^2 \\ &= n^2p^2 - np^2 + np - n^2p^2 = np(1-p).\end{aligned}$$

Por definição, a função geradora de probabilidades, será dada por

$$\begin{aligned}\psi_X(t) = E(t^X) &= \varphi_X(\ln t) \\ &= (pt + 1 - p)^n.\end{aligned}$$

Apresentamos agora a:

**Proposição 2.17** *Considerando as variáveis aleatórias independentes* $X \sim B(m,p)$ *e* $Y \sim B(n,p)$, *tem-se*

$$X + Y \sim B(m+n, p).$$

### 2.2.2 Distribuição de Poisson

Em situações em que o número $n$ de provas é demasiado grande ($n \to \infty$) e a probabilidade de sucesso $p$ é pequeno ($p \to 0$), a utilização do modelo Binomial, embora adequado, leva-nos a algumas dificuldades em termos de cálculo.

Observe-se que a expressão (2.2.4) pode ser reescrita da seguinte forma

$$\begin{aligned}P(X = k) &= \frac{n!}{k!(n-k)!} p^k \frac{n^k}{n^k} \left(1 - \frac{np}{n}\right)^{n-k} \\ &= \frac{n!}{k!(n-k)!} \frac{(np)^k}{n^k} \left(1 - \frac{np}{n}\right)^{n-k}.\end{aligned}$$

Tomando $\lambda = np$, tem-se

$$\begin{aligned}P(X = k) &= \frac{n(n-1)\cdots(n-k+1)}{k!} \frac{\lambda^k}{n^k} \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{\lambda^k}{k!} 1 \left(1 - \frac{1}{n}\right) \ldots \left(1 - \frac{k-1}{n}\right) \left(1 - \frac{\lambda}{n}\right)^{n-k}.\end{aligned}$$

Se considerarmos o limite quando $n \to \infty$, obtemos

$$\lim_{n\to\infty} 1 \left(1 - \frac{1}{n}\right) \ldots \left(1 - \frac{k-1}{n}\right) = 1$$

e

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{n-k}=\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{n}=e^{-\lambda},\ k=0,1,\ldots.$$

Assim obtemos

$$\lim_{n\to\infty}P(X=k)=\frac{e^{-\lambda}\lambda^{k}}{k!},$$

o que corresponde à função de probabilidade da distribuição de Poisson.

Tem-se então

**Definição 2.14** *Uma variável aleatória $X$ segue uma distribuição de Poisson com parâmetro $\lambda$, $\lambda > 0$, se a sua função de probabilidade for dada por*

$$P(X=k)=\frac{e^{-\lambda}\lambda^{k}}{k!},\ k=0,1,\ldots.$$

*Utilizamos a notação $X \sim P(\lambda)$.*

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

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

Tem-se que

$$\frac{d}{dt}\varphi_X(t)=\frac{d}{dt}e^{\lambda(e^{t}-1)}=\lambda e^{\lambda(e^{t}-1)+t}.$$

Então,

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

Como

$$\frac{d^2}{dt^2}\varphi(t) = \lambda(e^{\lambda(e^t-1)+t}(\lambda e^t + 1)),$$

tem-se

$$E(X^2) = \frac{d^2}{dt^2}\varphi(0) = \lambda(\lambda + 1)$$

donde

$$\text{Var}(X) = E(X^2) - (E(X))^2 = \lambda(\lambda + 1) - \lambda^2 = \lambda.$$

A função geradora de probabilidades é dada por

$$\begin{aligned} \psi_X(t) = E(t^X) &= \varphi_X(\ln t) \\ &= e^{\lambda(t-1)}. \end{aligned}$$

**Proposição 2.18** *Considerando as variáveis aleatórias independentes* $X \sim P(\lambda_1)$ *e* $Y \sim P(\lambda_2)$, *então*

$$X + Y \sim P(\lambda_1 + \lambda_2).$$

### 2.2.3 Distribuição Normal

Diz-se que a variável aleatória $X$ segue uma lei Normal com valor esperado $\mu$ e desvio padrão $\sigma$, e escreve-se $X \sim \mathcal{N}(\mu, \sigma)$, se a sua função densidade de probabilidade for dada por

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

Como é bem conhecido a variável $Z = \frac{X-\mu}{\sigma}$ terá vetor médio nulo e variância 1, $Z \sim \mathcal{N}(0, 1)$, e consequentemente função densidade

$$n(x|0, 1) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}},\ -\infty < x < +\infty.$$

**Proposição 2.19** *Consideremos as variáveis aleatórias independentes* $X_i \sim \mathcal{N}(\mu_i, \sigma_i),\ i = 1, 2, ..., n$. *Então*

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

**Corolário 2.19.1** *Se as variáveis aleatórias independentes* $X_i \sim \mathcal{N}(\mu, \sigma)$*,* $i = 1, ..., n$*, então*

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}}).$$

### 2.2.4 Distribuição Qui-quadrado

Como é sabido, ver por exemplo Mood et al. (1987), a soma de $m$ variáveis independentes, $Z_i \sim \mathcal{N}(0,1),\ i = 1, \ldots, m$ ao quadrado é um qui-quadrado central com $m$ graus de liberdade, $\chi^2_m$ , isto é,

$$X = \sum_{i=1}^{m} Z_i^2 \sim \chi^2_m.$$

A sua função densidade é dada por:

$$g(x|m) = \frac{1}{2\Gamma(\frac{m}{2})}(\frac{x}{2})^{\frac{m}{2}-1} e^{\frac{-x}{2}},\ x > 0,$$

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

Vamos em seguida obter a função geradora de momentos, $\varphi_X(t)$, considerando a substituição $x = \frac{2}{1-2t}u$. Assim tem-se

$$\begin{aligned}
\varphi_X(t) &= \int_0^{+\infty} e^{tx} \frac{1}{2\Gamma(\frac{m}{2})}(\frac{x}{2})^{\frac{m}{2}-1} e^{\frac{-x}{2}} dx \\
&= \frac{1}{2\Gamma(\frac{m}{2})} \int_0^{+\infty} (\frac{u}{1-2t})^{\frac{m}{2}-1} e^{-u} \frac{2}{1-2t} du \\
&= \frac{1}{(1-2t)^{\frac{m}{2}}} \frac{1}{\Gamma(\frac{m}{2})} \int_0^{+\infty} u^{\frac{m}{2}-1} e^{-u} du \\
&= \frac{1}{(1-2t)^{\frac{m}{2}}}\ ; t < \frac{1}{2},
\end{aligned}$$

uma vez que $\int_0^{+\infty} u^{\frac{m}{2}-1} e^{-u} du = \Gamma(\frac{m}{2})$.

Considerando o logarítmo da função $\varphi_X(t)$, tem-se

$$\eta(t) = \ln \varphi_X(t) = \frac{-m}{2} \ln(1-2t).$$

O valor médio e a variância da variável serão obtidos por $\frac{d}{dt}\eta(0)$ e $\frac{d^2}{dt^2}\eta(0)$ pelo que teremos

$$\begin{cases} E(\chi^2_m) = m, \\ Var(\chi^2_m) = 2m. \end{cases}$$

Segue-se a

**Proposição 2.20** *Consideremos as variáveis aleatórias independentes* $X_i \sim \chi^2_{m_i},\ i = 1, \ldots, k$, *então*

$$\sum_{i=1}^{k} X_i \sim \chi^2_{\sum_{i=1}^{k} m_i}.$$

#### 2.2.4.1 Qui-quadrados não centrais

Vamos agora considerar qui-quadrados não centrais, ver por exemplo Mexia (1995) e Nunes (2005). Para tal, consideramos a variável aleatória $X \sim N(\mu, 1)$. A função geradora de momentos de $X^2$ será

$$\varphi_{X^2}(t) = \int_{-\infty}^{+\infty} e^{tx^2} \frac{e^{\frac{-(x-\mu)^2}{2}}}{\sqrt{2\pi}} dx.$$

Uma vez que se tem

$$tx^2 - \frac{1}{2}(x-\mu)^2 = \frac{t}{1-2t}\mu^2 - \frac{1}{2\left(\sqrt{\frac{1}{1-2t}}\right)^2}\left(x - \frac{\mu}{1-2t}\right)^2,$$

então

$$\varphi_{X^2}(t) = \frac{e^{\frac{t}{1-2t}\mu^2}}{\sqrt{1-2t}} \int_{-\infty}^{+\infty} \frac{e^{\frac{-1}{2\left(\sqrt{\frac{1}{1-2t}}\right)^2}\left(x-\frac{\mu}{1-2t}\right)^2}}{\sqrt{2\pi}\sqrt{\frac{1}{1-2t}}} dx = \frac{e^{(\frac{t}{1-2t})\mu^2}}{\sqrt{1-2t}},$$

uma vez que $\frac{e^{\frac{-1}{2\left(\sqrt{\frac{1}{1-2t}}\right)^2}\left(x-\frac{\mu}{1-2t}\right)^2}}{\sqrt{2\pi}\sqrt{\frac{1}{1-2t}}}$ corresponde à função densidade de uma distribuição $\mathcal{N}(\frac{\mu}{1-2t}, \sqrt{\frac{1}{1-2t}})$. Assumindo que $X_1, ..., X_m$ são variáveis aleatórias independentes, com densidades $n(x|\mu_j, 1)$, $j = 1, ..., m$, ter-se-á

$$\varphi_{\sum_{j=1}^{m} X_j^2}(t) = \prod_{j=1}^{m} \varphi_{X_j^2}(t) = \frac{e^{\frac{\delta t}{1-2t}}}{(1-2t)^{\frac{m}{2}}}, \tag{2.2.5}$$

onde $\delta = \sum_{j=1}^{m} \mu_j^2$, corresponde ao parâmetro de não centralidade. Assim $\sum_{j=1}^{m} X_j^2$ será um qui-quadrado não central com $m$ graus de liberdade e parâmetro de não centralidade $\delta$, que representaremos por $\chi^2_{m,\delta}$. Como se tem

$$\eta_{\chi^2_{m,\delta}}(t) = \ln \varphi_{\sum_{j=1}^{m} X_j^2}(t) = \frac{\delta t}{1-2t} - \frac{m}{2}\ln(1-2t),$$

obtém-se

$$\begin{cases} E(\chi^2_m, \delta) = m + \delta \\ Var(\chi^2_m, \delta) = 2m + 4\delta. \end{cases}$$

Considerando $\chi^2_{m_1,\delta_1}$ e $\chi^2_{m_2,\delta_2}$ independentes, facilmente se verifica que

$$\begin{aligned} \varphi_{\chi^2_{m_1,\delta_1}+\chi^2_{m_2,\delta_2}}(t) &= \varphi_{\chi^2_{m_1,\delta_1}}(t)\varphi_{\chi^2_{m_2,\delta_2}}(t) \\ &= \frac{e^{\frac{(\delta_1+\delta_2)t}{1-2t}}}{(1+2t)^{\frac{m_1+m_2}{2}}} \\ &= \varphi_{\chi^2_{(m_1+m_2),(\delta_1+\delta_2)}}(t). \end{aligned}$$

Segue-se a

**Proposição 2.21** *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*

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

Em seguida iremos mostrar que a densidade $g(x|m,\delta)$ de um $\chi^2_{m,\delta}$ é uma combinação linear de densidades de qui-quadrados centrais. Como os coeficientes dessa combinação são não negativos, com soma um, ter-se-á uma mistura, ver Robbins (1948) e Robbins and Pitman (1949). Portanto, visto que $\frac{\delta t}{1-2t} = \frac{\delta}{2}(\frac{1}{1-2t} - 1)$, de (2.2.5) teremos

$$\begin{aligned} \varphi_{\chi^2_{m,\delta}}(t) &= \frac{e^{\frac{-\delta}{2}}}{(1-2t)^{\frac{m}{2}}} e^{\frac{\delta}{2(1-2t)}} \\ &= \frac{e^{\frac{-\delta}{2}}}{(1-2t)^{\frac{m}{2}}} \sum_{j=0}^{+\infty} \frac{1}{j!} \frac{(\frac{\delta}{2})^j}{(1-2t)^j} \\ &= e^{\frac{-\delta}{2}} \sum_{j=0}^{+\infty} \frac{(\frac{\delta}{2})^j}{j!} \frac{1}{(1-2t)^{\frac{(m+2j)}{2}}} \\ &= e^{\frac{-\delta}{2}} \sum_{j=0}^{+\infty} \frac{(\frac{\delta}{2})^j}{j!} \varphi_{\chi^2_{m+2j}}(t). \end{aligned}$$

Como a expressão obtida corresponde à função geradora de momentos da mistura

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

teremos, ver Robbins (1948) e Robbins and Pitman (1949),

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

o que corresponde à densidade de uma variável $X \sim \chi^2_{m,\delta}$.

### 2.2.5 Quocientes de Qui-quadrados independentes e Distribuição $F$

Nesta subseção iremos considerar as distribuições $F$ e $\overline{F}$, onde esta última corresponde ao quociente de qui-quadrados independentes, ver por exemplo Nunes (2005).

Considerando 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, que representamos por $F(z|m,n)$. A sua densidade será

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

Por outro lado a densidade de

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

será

$$\bar{f}(z|m,n) = \frac{\Gamma\frac{(m+n)}{2} z^{\frac{m}{2}-1}}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})(1+z)^{\frac{m+n}{2}}}, z>0. \quad (2.2.6)$$

Diremos que $\mathcal{T}_{m,n}$ tem distribuição $\overline{F}(z|m,n)$.

Assim, uma vez que

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

tem-se

$$\int_0^{+\infty} \frac{z^{\frac{m}{2}-1}}{(1+z)^{\frac{m+n}{2}}} dz = \frac{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}{\Gamma(\frac{m+n}{2})}, \quad (2.2.7)$$

e o valor esperado de $\mathcal{T}_{m,n}$ será dado por,

$$\begin{aligned} E(\mathcal{T}_{m,n}) &= \int_0^{+\infty} z \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \frac{z^{\frac{m}{2}-1}}{(1+z)^{\frac{(m+n)}{2}}} dz \\ &= \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \int_0^{+\infty} \frac{z^{\frac{m+2-1}{2}}}{(1+z)^{\frac{(m+2)+(n-2)}{2}}} dz \\ &= \frac{\Gamma(\frac{m+n}{2})\Gamma(\frac{m+2}{2})\Gamma(\frac{n-2}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})\Gamma(\frac{m+n}{2})} \\ &= \frac{\frac{m}{2}}{\frac{n-2}{2}} = \frac{m}{n-2}, \ n > 2, \end{aligned}$$

visto que $\Gamma(x+1) = x\Gamma(x)$. Tem-se então

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

Consideremos agora

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

que seguirá uma distribuição $\overline{F}$ com $m$ e $n$ graus de liberdade e parâmetros de não centralidade $\delta$ e $0$, $\overline{F}(z|m,n,\delta)$.

Como vimos anteriormente, a densidade de $\chi^2_{m,\delta}$ é uma mistura, logo a densidade de $\mathcal{T}_{m,n,\delta}$ será, ver Robbins (1948) e Robbins e Pitman (1949), uma mistura com os mesmos coeficientes, vindo

$$\bar{f}(z|m,n,\delta) = e^{\frac{-\delta}{2}} \sum_{i=0}^{+\infty} \frac{(\frac{\delta}{2})^i}{i!} \bar{f}(z|m+2i,n). \quad (2.2.8)$$

Seja $Y$ uma variável indicatriz com distribuição de Poisson de parâmetro $\frac{\delta}{2}$. Quando $Y = i$, $\chi^2_{m,\delta}$ distribui-se como um $\chi^2_{m+2i}$ e $\overline{F}(z|m,n,\delta)$ como $\overline{F}(z|m+2i,n)$, ver Mexia (1995) e Nunes (2005). Portanto, a distribuição de $\mathcal{T}_{m,n,\delta}$ será

$$\begin{aligned}
\overline{F}(z|m,n,\delta) &= pr(\mathcal{T}_{m,n,\delta} \leq z) \\
&= \sum_{i=0}^{+\infty} pr(Y=i) pr(\mathcal{T}_{m,n,\delta} \leq z|Y=i) \\
&= \sum_{i=0}^{+\infty} e^{\frac{-\delta}{2}} \frac{\frac{\delta}{2}^i}{i!} \overline{F}(z|m+2i,n) \\
&= e^{\frac{-\delta}{2}} \sum_{i=0}^{+\infty} \frac{\frac{\delta}{2}^i}{i!} \overline{F}(z|m+2i,n). \qquad (2.2.9)
\end{aligned}$$

Derivando $\overline{F}(z|m,n,\delta)$ em ordem a $\delta$, obtém-se, com $j = i-1$,

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

Seja $F_l$ a distribuição de $U_l$, $l = 1, 2$. Se $U_1 < U_2$, então $U_2 \leq u$ implica que $U_1 \leq u$, logo tem-se, ver Nunes (2005),

$$F_2(u) \leq F_1(u).$$

Assim, considerando os qui-quadrados independentes, $\chi^2_{m,\delta}, \chi^2_n, \chi^2_2$, como

$$pr\left(\frac{\chi^2_{m,\delta}}{\chi^2_n} < \frac{\chi^2_{m,\delta} + \chi^2_2}{\chi^2_n}\right) = 1,$$

e uma vez que estas frações seguem distribuições $\overline{F}(z|m,n,\delta)$ e $\overline{F}(z|m+2,n,\delta)$ respectivamente, vem, para $z > 0$, $\overline{F}(z|m+2,n,\delta)$<$\overline{F}(z|m,n,\delta)$ e consequentemente,

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

Por outro lado

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

terá distribuição

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

e densidade

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

Como vamos ver nos próximos capítulos optámos por trabalhar com a distribuição $\overline{F}$ por ser mais "tratável" e estatisticamente equivalente à distribuição $F$.

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

Nesta subseção apresentamos uma propriedade de monotonia das distribuições $F$ e $\overline{F}$ centrais, que nos será útil nas aplicações que serão apresentadas neste trabalho, ver Mexia et al. (2011). O desenvolvimento será feito utilizando a distribuição $\overline{F}$ no entanto a propriedade é igualmente válida para a distribuição $F$. Como foi visto anteriormente, com $\chi^2_r$ e $\chi^2_s$ independentes,

$$\mathcal{T}_{r,s} = \frac{\chi^2_r}{\chi^2_s}$$

tem distribuição $\overline{F}(z|r,s)$. Com $f_{1-\alpha,r,s}$ o $(1-\alpha)$-ésimo quantil desta distribuição e $s < s^0$, ter-se-á

$$f_{1-\alpha,r,s} > f_{1-\alpha,r,s^o}, \quad (2.2.10)$$

o que sugere que, para $z$ suficientemente grande, se tem

$$\overline{F}(z|r,s) < \overline{F}(z|r,s^o). \quad (2.2.11)$$

Vamos em seguida verificar a veracidade destas desigualdades. De (2.2.6) e (2.2.7) vem que

$$\overline{F}(z|r,s) = \frac{\int_0^z \frac{x^{\frac{r}{2}-1}}{(1+x)^{\frac{r+s}{2}}}dx}{\int_0^{+\infty} \frac{x^{\frac{r}{2}-1}}{(1+x)^{\frac{r+s}{2}}}dx}.$$

Esta expressão permite-nos tratar $r$ e $s$ como variáveis reais e mostrar que $\frac{\partial \overline{F}(z|r,s)}{\partial s} > 0$, para $z > 0$, o que estabelece as desigualdades (2.2.10) e (2.2.11). Começamos por obter

$$\begin{aligned}
\frac{\partial}{\partial s}\int_0^{+\infty} \frac{x^{\frac{r}{2}-1}}{(1+x)^{\frac{r+s}{2}}}dx &= \frac{\partial}{\partial s}\int_0^{+\infty} \frac{x^{\frac{r}{2}-1}}{(1+x)^{\frac{r}{2}}} e^{-\frac{\ln(1+x)s}{2}}dx \\
= & -\int_0^{+\infty} \frac{x^{\frac{r}{2}-1}}{(1+x)^{\frac{r}{2}}} e^{-\frac{\ln(1+x)s}{2}} \frac{\ln(1+x)}{2}dx \\
= & -\frac{1}{2}\int_0^{+\infty} \frac{x^{\frac{r}{2}-1}}{(1+x)^{\frac{r+s}{2}}} \ln(1+x)dx.
\end{aligned}$$

Assim, de (2.2.7), tem-se

$$\begin{aligned}
&\frac{\partial}{\partial s}\left(\int_0^{+\infty}\frac{x^{\frac{r}{2}-1}}{(1+x)^{\frac{r+s}{2}}}dx\right)^{-1}\\
=\;&\frac{1}{2}\left(\int_0^{+\infty}\frac{x^{\frac{r}{2}-1}}{(1+x)^{\frac{r+s}{2}}}dx\right)^{-2}\int_0^{+\infty}\frac{x^{\frac{r}{2}-1}}{(1+x)^{\frac{r+s}{2}}}\ln(1+x)dx\\
=\;&\frac{1}{2}\frac{\Gamma^2\left(\frac{r+s}{2}\right)}{\Gamma^2\left(\frac{r}{2}\right)\Gamma^2\left(\frac{s}{2}\right)}\int_0^{+\infty}\frac{x^{\frac{r}{2}-1}}{(1+x)^{\frac{r+s}{2}}}\ln(1+x)dx.
\end{aligned}$$

Analogamente

$$\begin{aligned}
&\frac{\partial}{\partial s}\left(\int_0^{z}\frac{x^{\frac{r}{2}-1}}{(1+x)^{\frac{r+s}{2}}}dx\right)\\
=\;&-\frac{1}{2}\int_0^{z}\frac{x^{\frac{r}{2}-1}}{(1+x)^{\frac{r+s}{2}}}\ln(1+x)dx
\end{aligned}$$

vindo

$$\begin{aligned}
&\frac{\partial\overline{F}(z|r,s)}{\partial s}=\frac{d}{ds}\left(\int_0^{z}\frac{x^{\frac{r}{2}-1}}{(1+x)^{\frac{r+s}{2}}}dx\times\left(\int_0^{+\infty}\frac{x^{\frac{r}{2}-1}}{(1+x)^{\frac{r+s}{2}}}dx\right)^{-1}\right)\\
=\;&\frac{\Gamma\left(\frac{r+s}{2}\right)}{\Gamma\left(\frac{r}{2}\right)\Gamma\left(\frac{s}{2}\right)}\left(-\frac{1}{2}\int_0^{z}\frac{x^{\frac{r}{2}-1}}{(1+x)^{\frac{r+s}{2}}}\ln(1+x)dx\right)\\
+\;&\frac{1}{2}\frac{\Gamma^2\left(\frac{r+s}{2}\right)}{\Gamma^2\left(\frac{r}{2}\right)\Gamma^2\left(\frac{s}{2}\right)}\int_0^{+\infty}\frac{x^{\frac{r}{2}-1}}{(1+x)^{\frac{r+s}{2}}}\ln(1+x)dx\times\int_0^{z}\frac{x^{\frac{r}{2}-1}}{(1+x)^{\frac{r+s}{2}}}dx.\\
=\;&\frac{1}{2}\frac{\Gamma\left(\frac{r+s}{2}\right)}{\Gamma\left(\frac{r}{2}\right)\Gamma\left(\frac{s}{2}\right)}\left(\frac{\Gamma\left(\frac{r+s}{2}\right)}{\Gamma\left(\frac{r}{2}\right)\Gamma\left(\frac{s}{2}\right)}\int_0^{+\infty}\frac{x^{\frac{r}{2}-1}}{(1+x)^{\frac{r+s}{2}}}\ln(1+x)dx\right.\\
\times\;&\left.\int_0^{z}\frac{x^{\frac{r}{2}-1}}{(1+x)^{\frac{r+s}{2}}}dx-\int_0^{z}\frac{x^{\frac{r}{2}-1}}{(1+x)^{\frac{r+s}{2}}}\ln(1+x)dx\right).
\end{aligned}$$

Tem-se então

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

e

$$\frac{\partial\overline{F}(z|r,s)}{\partial s}\underset{z\mapsto 0}{\longrightarrow}0.$$

Para estabelecer as desigualdades, ou seja, mostrar que $\frac{\partial\overline{F}(z|r,s)}{\partial s}>0$, para $z>0$, será suficiente provar que $\forall s$, $\frac{\partial\overline{F}(z|r,s)}{\partial s}$, como função de $z$, tem apenas um máximo local para $z>0$.

Consideremos $g(z) = \frac{\partial \overline{F}(z|r,s)}{\partial s}$, tem-se então

$$\begin{aligned}
& g'(z) = \frac{\partial^2 \overline{F}(z|r,s)}{\partial s \partial z} = \frac{\partial^2 \overline{F}(z|r,s)}{\partial z \partial s} \\
= & \frac{\partial}{\partial s}\left(\frac{\partial}{\partial z}\left[\int_0^z \frac{x^{\frac{r}{2}-1}}{(1+x)^{\frac{r+s}{2}}}dx\left(\int_0^{+\infty}\frac{x^{\frac{r}{2}-1}}{(1+x)^{\frac{r+s}{2}}}dx\right)^{-1}\right]\right) \\
= & \frac{\partial}{\partial s}\left[\left(\int_0^{+\infty}\frac{x^{\frac{r}{2}-1}}{(1+x)^{\frac{r+s}{2}}}dx\right)^{-1}\frac{z^{\frac{r}{2}-1}}{(1+z)^{\frac{r+s}{2}}}\right] \\
= & \frac{1}{2}\frac{\Gamma^2\left(\frac{r+s}{2}\right)}{\Gamma^2\left(\frac{r}{2}\right)\Gamma^2\left(\frac{s}{2}\right)}\int_0^{+\infty}\frac{x^{\frac{r}{2}-1}}{(1+x)^{\frac{r+s}{2}}}\ln(1+x)dx \\
\times & \frac{z^{\frac{r}{2}-1}}{(1+z)^{\frac{r+s}{2}}} - \frac{1}{2}\frac{\Gamma\left(\frac{r+s}{2}\right)}{\Gamma\left(\frac{r}{2}\right)\Gamma\left(\frac{s}{2}\right)}\frac{z^{\frac{r}{2}-1}}{(1+z)^{\frac{r+s}{2}}}\ln(1+z) \\
= & \frac{1}{2}\frac{\Gamma\left(\frac{r+s}{2}\right)}{\Gamma\left(\frac{r}{2}\right)\Gamma\left(\frac{s}{2}\right)}\frac{z^{\frac{r}{2}-1}}{(1+z)^{\frac{r+s}{2}}}\left(\frac{\Gamma\left(\frac{r+s}{2}\right)}{\Gamma\left(\frac{r}{2}\right)\Gamma\left(\frac{s}{2}\right)}\right. \\
\times & \left.\int_0^{+\infty}\frac{x^{\frac{r}{2}-1}}{(1+x)^{\frac{r+s}{2}}}\ln(1+x)dx - \ln(1+z)\right) \\
= & u(z,s)\left[v(s) - \ln(1+z)\right],
\end{aligned}$$

com

$$\begin{cases} u(z,s) = \frac{1}{2}\frac{\Gamma\left(\frac{r+s}{2}\right)}{\Gamma\left(\frac{r}{2}\right)\Gamma\left(\frac{s}{2}\right)}\frac{z^{\frac{r}{2}-1}}{(1+z)^{\frac{r+s}{2}}} > 0,\ z > 0 \\ v(s) = \frac{\Gamma\left(\frac{r+s}{2}\right)}{\Gamma\left(\frac{r}{2}\right)\Gamma\left(\frac{s}{2}\right)}\int_0^{+\infty}\frac{x^{\frac{r}{2}-1}}{(1+x)^{\frac{r+s}{2}}}\ln(1+x)dx > 0. \end{cases}$$

Como

$$\ell(z) = \ln(1+z)$$

cresce com $z$, $\ell(0) = 0$ e

$$\begin{array}{ccc} \ell(z) & \longrightarrow & +\infty \\ & z \mapsto +\infty & \end{array},$$

a equação

$$\ell(z) = v(s)$$

terá uma única raiz $z_0$, para $z > 0$. Assim $z_0$ será a única raiz de $g'(z)$, para $z > z_0$, tendo-se a situação descrita na Figura 2.1. Concluímos então que $\overline{F}(z|r,s) < \overline{F}(z|r,s^o)$ sempre que $s < s^o$, o que estabelece a propriedade de monotonia dos quantis da distribuição $\overline{F}$ central.

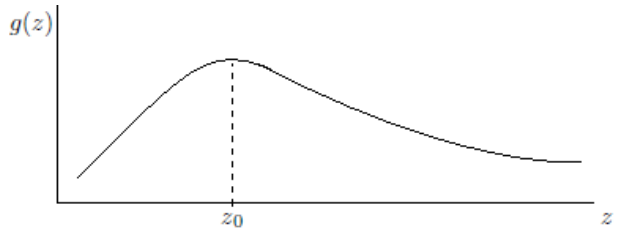


Figura 2.1: Representação gráfica de $g(z)$.

## 2.3 Modelos Lineares

Os modelos lineares são muito usados pelos estatísticos na análise de dados e desenvolvimento de novos métodos estatísticos. Estes apresentam uma relação entre variáveis que é linear nos seus parâmetros. Nesta seção faremos uma pequena abordagem a estes modelos, apresentando a sua estrutura para fatores de efeitos fixos e efeitos aleatórios. São considerados ainda os modelos mistos com estrutura ortogonal por blocos ($OBS$) e apresentados alguns resultados sobre extensões $L$. Estes modelos foram amplamente estudados, por exemplo em Scheffé (1959), Searle et al. (1992), Rao (1973), Khuri et al. (1998) e Muller and Stewart (2006).

### 2.3.1 Modelos de efeitos fixos

Um modelo linear que apresenta somente fatores de efeitos fixos, para além do erro que é sempre aleatório, designa-se por modelo de efeitos fixos. O modelo mais simples é o modelo com apenas um fator, que pode ser descrito da seguinte forma, ver por exemplo Scheffé (1959),

$$Y_{i,j} = \mu + \alpha_i + \varepsilon_{i,j},\ j = 1, ..., n_i,\ i = 1, ..., r, \quad (2.3.12)$$

onde

- $Y_{i,j}$, representa a *j*-ésima observação do nível $i$ do fator;
- $\mu$, representa a média geral, sendo um parâmetro fixo desconhecido;
- $\alpha_i$, representa o efeito do nível $i$ do fator, que é fixo;
- $\varepsilon_{i,j}$, corresponde ao erro aleatório.

À semelhança do que acontece na maioria da literatura, utilizaremos a seguinte notação:

- $Y_{i\bullet} = \dfrac{\sum_{j=1}^{n_i} Y_{ij}}{n_i}$, é a média das observações do nível $i$ do fator;

- $Y_{\bullet\bullet} = \dfrac{\sum_{i=1}^{r}\sum_{j=1}^{n_i} Y_{ij}}{n}$, é a média geral das observações;

- $n = \sum_{i=1}^{r} n_i$, é o número total de observações.

Considerando a notação matricial, o modelo (2.3.12) pode ser escrito, ver por exemplo Searle et al. (1992),

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{D}(\mathbf{1}_{n_1}, ..., \mathbf{1}_{n_r})\boldsymbol{\alpha} + \boldsymbol{\varepsilon},$$

onde

- $\boldsymbol{Y} = (Y_{1,1}, Y_{1,2}, ..., Y_{1,n_1}, Y_{2,1}, ..., Y_{2,n_2}, ..., Y_{r,1}, ..., Y_{r,n_r})'$, corresponde ao vetor das observações;

- $\boldsymbol{\mu} = \mathbf{1}_n\mu$, sendo $\mathbf{1}_n$ o vetor com todas as $n$ componentes iguais a $1$;

- $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_r)'$, o vetor de efeitos fixos;

- $\boldsymbol{D}(\mathbf{1}_{n_1}, ..., \mathbf{1}_{n_r})$, indica uma matriz diagonal por blocos com, $\mathbf{1}_{n_1}, ..., \mathbf{1}_{n_r}$, ao longo dos blocos;

- $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)'$, corresponde ao vetor dos erros aleatórios.

Se considerarmos um modelo de efeitos fixos com dois fatores e interação, teremos

$$Y_{i,j,k} = \mu + \alpha_i + \tau_j + (\alpha\tau)_{i,j} + \varepsilon_{i,j,k},\ i = 1, ..., r; j = 1, ..., n_i; k = 1, ..., n_{i,j},$$

onde

- $Y_{i,j,k}$, representa a $k$-ésima observação do nível $i$ do primeiro fator e nível $j$ do segundo fator;

- $\mu$, representa a média geral dos dados;

- $\alpha_i$, representa o efeito do nível $i$ do primeiro fator;

- $\tau_j$, representa o efeito do nível $j$ do segundo fator;

- $(\alpha\tau)_{i,j}$, representa a interação, entre o nível $i$ do primeiro fator e o nível $j$ do segundo fator;

- $\varepsilon_{i,j,k}$, $i = 1, ..., r; j = 1, ..., n_i; k = 1, ..., n_{i,j}$, corresponde ao erro aleatório.

Em notação matricial e em termos gerais teremos, ver por exemplo Searle et al. (1992),

$$\boldsymbol{Y} = \boldsymbol{\mu} + \sum_{h=1}^{c} \boldsymbol{X}_h \boldsymbol{\alpha}_h + \boldsymbol{\varepsilon},$$

onde

- $\boldsymbol{\mu} = \mathbf{1}_n \mu$;
- $\boldsymbol{X}_h, h = 1, ..., c$, correspondem às matrizes de delineamento, conhecidas;
- $\boldsymbol{\alpha}_h = h = 1, ..., c$, correspondem aos vetores de efeitos fixos;
- $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)'$, será o vetor dos erros aleatórios.

### 2.3.2 Modelos de efeitos aleatórios

Se um fator tem um grande número (ou mesmo uma infinidade) de possíveis níveis, não sendo possível estudá-los a todos, teremos que estudar apenas uma amostra aleatória de níveis do fator. Neste caso, os níveis selecionados aleatoriamente para o estudo terão efeitos que são descritos por variáveis aleatórias, e não por constantes.

Consideremos o modelo de efeitos aleatórios mais simples, que será o que apresenta apenas um fator, ver por exemplo Scheffé (1959),

$$Y_{i,j} = \mu + \beta_i + \varepsilon_{i,j},\ i = 1, ..., r, j = 1, ..., n_i, \quad (2.3.13)$$

onde

- $Y_{i,j}$, representa a j-ésima observação do nível $i$ do fator;
- $\mu$, representa a média geral dos dados;
- $\beta_i$, representa o efeito do nível $i$ do fator, que é aleatório;
- $\varepsilon_{i,j}$, $i = 1, ..., r; j = 1, ..., n_i$, corresponde ao erro aleatório.

O modelo (2.3.13) pode ser escrito em notação matricial da seguinte forma, ver por exemplo Khuri et al. (1998),

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{D}(\mathbf{1}_{n_1}, ..., \mathbf{1}_{n_r})\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

onde

- $\boldsymbol{Y} = (Y_{1,1}, Y_{1,2}, ..., Y_{1,n_1}, Y_{2,1}, ..., Y_{2,n_2}, ..., Y_{r,1}, ..., Y_{r,n_r})'$, é o vetor das observações;

- $\boldsymbol{\mu} = \mathbf{1}_n \mu$;
- $\boldsymbol{\beta} = (\beta_1, ..., \beta_r)'$, é o vetor de efeitos aleatórios;
- $\boldsymbol{D}(\mathbf{1}_{n_1}, ..., \mathbf{1}_{n_r})$, indica uma matriz diagonal por blocos com, $\mathbf{1}_{n_1}, ..., \mathbf{1}_{n_r}$ ao longo dos blocos;
- $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)'$, é o vetor dos erros aleatórios.

Para o caso geral ter-se-á, em notação matricial, ver mais uma vez por exemplo Searle et al. (1992) e Khuri et al. (1998),

$$\boldsymbol{Y} = \boldsymbol{\mu} + \sum_{h=1}^{w} \boldsymbol{X}_h \boldsymbol{\beta}_h + \boldsymbol{\varepsilon},$$

onde

- $\boldsymbol{\mu} = \mathbf{1}_n \mu$, corresponde ao vetor das médias;
- $\boldsymbol{X}_h, h = 1, ..., w$, são as matrizes de delineamento, conhecidas;
- $\boldsymbol{\beta}_h, i = 1, ..., w$, são os vetores de efeitos aleatórios, independentes;
- $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)'$, corresponde ao vetor dos erros aleatórios.

### 2.3.3 Modelos mistos

Um modelo misto é um modelo que contém tanto fatores de efeitos fixos como fatores de efeitos aleatórios. Estes modelos têm vindo a ser aplicados nas mais diversas áreas, nomeadamente na investigação biológica e médica, agricultura ou indústria. Detalhe referente, à teoria de modelos mistos podem ser consultados, por exemplo, em Khuri et al. (1998).

O modelo misto com um fator fixo, e outro aleatório e com interação pode ser escrito da seguinte forma, ver por exemplo Searle et al. (1992),

$$Y_{i,j,k} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{i,j} + \varepsilon_{i,j,k},\ i = 1, ..., r, j = 1, ..., n_i, k = 1, ..., n_{i,j},$$

onde

- $Y_{i,j,k}$, representa a $k$-ésima observação do nível $i$ do primeiro fator e nível $j$ do segundo fator;
- $\mu$, representa a média geral;

- $\alpha_i, i = 1, ..., r$, corresponde ao efeito do nível $i$ do primeiro fator, que é fixo;
- $\beta_j, j = 1, ..., n_i$, corresponde ao efeito do nível $j$ do segundo fator, que é aleatório;
- $(\alpha\beta)_{i,j}, i = 1, ..., r; j = 1, ..., n_i$, representa a interação, entre o nível $i$ do primeiro fator e o nível $j$ do segundo fator;
- $\varepsilon_{i,j,k}, i = 1, ..., r; j = 1, ..., n_i; k = 1, ..., n_{i,j}$, corresponde ao erro aleatório.

Em notação matricial e em termos gerais, ter-se-á

$$\boldsymbol{Y} = \boldsymbol{\mu} + \sum_{h=1}^{w} \boldsymbol{X}_h \boldsymbol{\beta}_h + \boldsymbol{\varepsilon}, \quad (2.3.14)$$

onde

- $\boldsymbol{\mu} = \mathbf{1}_n \mu$;
- $\boldsymbol{X}_h, h = 1, ..., w$, correspondem às matrizes de delineamento, conhecidas;
- $\boldsymbol{\beta}_h, h = 1, ..., c$, são vetores de efeitos fixos, $c < w$;
- $\boldsymbol{\beta}_h, h = c + 1, ..., w$, são vetores de efeitos aleatórios e independentes;
- $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)'$, corresponde ao vetor dos erros aleatórios.

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

O modelo (2.3.14) aparece em muita literatura, ver por exemplo Carvalho et al. (2015), Khuri et al. (1998) e Santos (2012), escrito da seguinte forma

$$\boldsymbol{Y} = \sum_{i=0}^{w+1} \boldsymbol{X}_i \boldsymbol{\beta}_i, \quad (2.3.15)$$

em que $\boldsymbol{\beta}_0$ é fixo e $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{w+1}$ são vetores aleatórios independentes com vetores médios nulos e matrizes de covariância $\sigma_1^2 \boldsymbol{I}_{c_1}, \ldots, \sigma_{w+1}^2 \boldsymbol{I}_{c_{w+1}}$, com $c_i$ o número de componentes de $\boldsymbol{\beta}_i$, $i = 1, \ldots, w + 1$. O modelo misto assim definido será considerado no capítulo 3 e 4 do presente trabalho. As matrizes $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{w+1}$ são conhecidas e tais que

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

O vetor médio de $\boldsymbol{Y}$ é dado por

$$\boldsymbol{\mu} = E\left(\sum_{i=0}^{w+1} \boldsymbol{X}_i\boldsymbol{\beta}_i\right) = \boldsymbol{X}_0\boldsymbol{\beta}_0 + \sum_{i=1}^{w+1} \boldsymbol{X}_i E(\boldsymbol{\beta}_i) = \boldsymbol{X}_0\boldsymbol{\beta}_0,$$

e a matriz de covariâncias por

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}\left(\sum_{i=1}^{w+1} \boldsymbol{X}_i\boldsymbol{\beta}_i\right) = \sum_{i=1}^{w+1} \sigma_i^2 \boldsymbol{X}_i\boldsymbol{X}_i^{'} = \sum_{i=1}^{w+1} \sigma_i^2 \boldsymbol{M}_i,$$

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

Podemos considerar, para o modelo (2.3.15), $\sigma_{w+1}^2 = \sigma^2, \boldsymbol{X}_{w+1} = \boldsymbol{I}_n$ e $\boldsymbol{\beta}_{w+1} = \boldsymbol{\varepsilon}$. Os modelos de efeitos fixos e de efeitos aleatórios abordados anteriormente podem ser tratados como casos particulares do modelo misto. Para se ter um modelo de efeitos aleatórios considera-se $\boldsymbol{X}_0 = \mathbf{1}_n$ e $\boldsymbol{\beta}_0 = \boldsymbol{\mu}$ enquanto que para o modelo de efeitos fixos se tem, $w+1 = 1$ e $\boldsymbol{X}_{w+1}\boldsymbol{\beta}_{w+1} = \boldsymbol{\varepsilon}$.

O espaço gerado por $\boldsymbol{\mu}$ será $R(\boldsymbol{X}_0)$. Assim, de acordo com a proposição 2.8, página 9, 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 Nelder (1965a, 1965b), Houtman and Speed (1983) e Mejza (1992) tem-se

**Definição 2.15** *Um modelo misto tem estrutura ortogonal por blocos ($OBS$) quando a matriz de covariância, $\boldsymbol{\Sigma}$, pode ser escrita como a seguinte combinação linear*

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

*com $\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_m$ $MPOMO$ conhecidas e tais que*

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

Estes modelos foram introduzidos por J. A. Nelder, ver Nelder (1965a, 1965b), e continuam a desempenhar um papel central na teoria do delineamento em blocos casualizados, ver Calinski and Kageyama (2000, 2003).

Estabelecemos agora a seguinte proposição.

**Proposição 2.22** *O modelo misto $\boldsymbol{Y} = \sum_{i=0}^{w+1} \boldsymbol{X}_i\boldsymbol{\beta}_i$ tem $OBS$ se e somente se, as matrizes $\boldsymbol{M}_1, \ldots, \boldsymbol{M}_{w+1}$ comutam.*

**Dem:** Se as matrizes $\boldsymbol{M}_1, \ldots, \boldsymbol{M}_{w+1}$ comutam, geram uma $AJCS$, $\mathcal{A}$, com base principal $bp(\mathcal{A}) = \{\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_m\}$. Assim teremos

$$\boldsymbol{M}_i = \sum_{j=1}^{m} b_{i,j}\boldsymbol{Q}_j,\ i = 1, \ldots, w,$$

e

$$\boldsymbol{\Sigma} = \sum_{i=1}^{w+1} \sigma_i^2 \boldsymbol{M}_i = \sum_{j=1}^{m} \gamma_j \boldsymbol{Q}_j,$$

com $\gamma_j = \sum_{i=1}^{w+1} b_{i,j}\sigma_i^2$, $j = 1 \ldots, m$. A tese fica estabelecida visto que a matriz invertível $\sum_{i=1}^{w+1} \boldsymbol{M}_i \in \mathcal{A}$, então $\mathcal{A}$ é completa e $\sum_{j=1}^{m} \boldsymbol{Q}_j = \boldsymbol{I}_n$.

A demonstração da implicação no sentido inverso pode ser vista em Carvalho et al. (2015). ■

Apresentaremos agora uma classe especial de modelos com $OBS$.

**Definição 2.16** *Um modelo misto tem estrutura ortogonal por blocos comutativa, $COBS$, se tem $OBS$ e além disso,*

$$\boldsymbol{T}\boldsymbol{Q}_j = \boldsymbol{Q}_j\boldsymbol{T},\ j = 1, \ldots, m,$$

sendo $\boldsymbol{T}$ a $MPO$ sobre o espaço gerado pelo vetor médio $\boldsymbol{\mu}$ e $\{\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_m\}$ as matrizes da $bp(\mathcal{A})$.

Os modelos com $COBS$, foram introduzidos em Fonseca et al. (2008). Os mesmos têm vindo a ser estudados, veja-se por exemplo Santos et al. (2007), Nunes et al. (2008), Carvalho et al. (2008), Ferreira et al. (2013) e Carvalho et al. (2015).

Estabelecemos agora a seguinte proposição.

**Proposição 2.23** *O modelo tem $COBS$ se e só se as matrizes $\boldsymbol{M}_1, \ldots, \boldsymbol{M}_{w+1}$ e $\boldsymbol{T}$ comutarem.*

**Dem:** A demonstração pode ser vista, por exemplo, em Carvalho et al. (2015). ■

### 2.3.5 Extensões $L$

Nesta seção, apresentaremos alguns resultados importantes sobre extensões $L$, as quais têm sido usadas para resolver certas questões sobre a falta de ortogonalidade em modelos de efeitos fixos e modelos mistos, ver Ferreira et al. (2009) e Moreira et al. (2009). Utilizaremos esta classe de modelos nos próximos dois capítulos, na formulação dos modelos mistos com amostras de dimensão aleatória.

Consideremos um modelo linear com $m$ tratamentos e $n_1, \ldots, n_m$ observações por tratamento e seja

$$\boldsymbol{L} = \boldsymbol{D}(\mathbf{1}_{n_1}, \ldots, \mathbf{1}_{n_m})$$

uma matriz diagonal por blocos, com blocos principais $\mathbf{1}_{n_1}, \ldots, \mathbf{1}_{n_m}$.

Segundo Ferreira et al. (2009) e Moreira et al. (2009), o modelo

$$\boldsymbol{Y} = \boldsymbol{L}\boldsymbol{Y}^o + \boldsymbol{\varepsilon},$$

em que $\boldsymbol{\varepsilon}$ corresponde ao vetor dos erros com vetor médio nulo e matriz de covariância $\sigma^2\boldsymbol{I}_n$, é uma extensão $L$ de

$$\boldsymbol{Y}^o = \sum_{i=0}^{w} \boldsymbol{X}_i\boldsymbol{\beta}_i,$$

onde $\boldsymbol{\beta}_0$ é fixo e $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_w$ são aleatórios e independentes com vetores médios nulos e matrizes de covariância $\sigma_1^2\boldsymbol{I}_{c_1}, \ldots, \sigma_w^2\boldsymbol{I}_{c_w}$, onde $c_i, i = 1, \ldots, w$, é o número de componentes de $\boldsymbol{\beta}_i, i = 1, \ldots, w$. Os modelos $\boldsymbol{Y}$ e $\boldsymbol{Y}^0$ são modelos mistos.

Vamos assumir que $\boldsymbol{Y}^o$ tem estrutura ortogonal por blocos, $OBS$, com vetor médio e matriz de covariância, respectivamente,

- $\boldsymbol{\mu}^o = \boldsymbol{X}_0\boldsymbol{\beta}_0$,
- $\boldsymbol{V}^o = \sum_{j=1}^{\ell} \gamma_j\boldsymbol{K}_j$,

onde $\boldsymbol{K}_1, ..., \boldsymbol{K}_\ell$ são $MPOMO$ e $\gamma_j$, $j = 1, ..., \ell$, correspondem às chamadas componentes de variância canónicas, ver por exemplo, Ferreira et al. (2013). Consideremos as matrizes $\boldsymbol{A}_j$, $j = 1, ..., \ell$, cujos vetores linha constituem uma base ortonormada para o espaço gerado por $\boldsymbol{K}_j$, $R(\boldsymbol{K}_j)$, $j = 1, ..., \ell$. Teremos portanto

- $\boldsymbol{K}_j = \boldsymbol{A}_j^{'}\boldsymbol{A}_j,\ j = 1, ..., \ell,$
- $\boldsymbol{I}_{g_j} = \boldsymbol{A}_j\boldsymbol{A}_j^{'},\ j = 1, ..., \ell,$

com $g_j = car(\boldsymbol{K}_j)$.

As $MPO$ sobre $\overline{\Omega} = R(\boldsymbol{L})$ e sobre o seu complemento ortogonal, $\overline{\Omega}^{\perp}$, serão respectivamente, ver Schott (1997),

- $P(\boldsymbol{L}) = \boldsymbol{L}\boldsymbol{L}^{+}$,
- $Q(\boldsymbol{L}) = \boldsymbol{I}_n - P(\boldsymbol{L})$.

Note-se, que com $\boldsymbol{L} = D(\mathbf{1}_{n_1}, ..., \mathbf{1}_{n_m})$, e $n = \sum_{i=1}^{m} n_i$, teremos

$$\boldsymbol{L}^{+} = D(\frac{1}{n_1}\mathbf{1}_{n_1}^{'}, ..., \frac{1}{n_m}\mathbf{1}_{n_m}^{'}).$$

Como os vetores coluna de $\boldsymbol{L}$ são linearmente independentes, ver Schott (1997), temos

$$\boldsymbol{L}^{+}\boldsymbol{L} = \boldsymbol{I}_n.$$

Quando $\boldsymbol{Y}^o$ é independente de $\boldsymbol{\varepsilon}$, com $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$, $P(\boldsymbol{L})\boldsymbol{\varepsilon}$ e $Q(\boldsymbol{L})\boldsymbol{\varepsilon}$ também serão independentes, uma vez que têm distribuição conjunta normal e matrizes de covariância nulas. Assim, ter-se-á

$$P(\boldsymbol{L})\boldsymbol{Y} = P(\boldsymbol{L})\boldsymbol{L}\boldsymbol{Y}^o + P(\boldsymbol{L})\boldsymbol{\varepsilon} = \boldsymbol{L}\boldsymbol{Y}^o + P(\boldsymbol{L})\boldsymbol{\varepsilon}$$

e

$$Q(\boldsymbol{L})\boldsymbol{Y} = Q(\boldsymbol{L})\boldsymbol{\varepsilon}$$

independentes.

Podemos então considerar, uma vez que $\boldsymbol{L}^{+}P(\boldsymbol{L}) = \boldsymbol{L}^{+}$,

$$\boldsymbol{Y}^{oo} = \boldsymbol{L}^{+}\boldsymbol{Y} = \boldsymbol{Y}^o + \boldsymbol{L}^{+}\boldsymbol{\varepsilon} = \boldsymbol{Y}^o + \boldsymbol{L}^{+}P(\boldsymbol{L})\boldsymbol{\varepsilon},$$

independente de $Q(\boldsymbol{L})\boldsymbol{Y} = Q(\boldsymbol{L})\boldsymbol{\varepsilon}$, ver Ferreira et al. (2009), e portanto, independente de

$$S = \|Q(\boldsymbol{L})\boldsymbol{Y}\|^2 = \|Q(\boldsymbol{L})\boldsymbol{\varepsilon}\|^2, \quad (2.3.16)$$

que será o produto de $\sigma^2$ por um qui-quadrado central com

$$g(n) = n - m$$

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

Observemos que $\boldsymbol{Y}^{oo}$ tem vetor médio e matriz de covariância dados por

- $\boldsymbol{\mu}^{oo} = \boldsymbol{\mu}^o = \boldsymbol{X}_0\boldsymbol{\beta}_0$,
- $\boldsymbol{V}^{oo} = \boldsymbol{V}^o + \sigma^2(\boldsymbol{L}^{+}(\boldsymbol{L}^{+})') = \sum_{j=1}^{m}\gamma_j\boldsymbol{K}_j + \sigma^2(\boldsymbol{L}^{+}(\boldsymbol{L}^{+})')$.

Com $\boldsymbol{L} = D(\boldsymbol{1}_{n_1}, ..., \boldsymbol{1}_{n_m})$ teremos

$$\boldsymbol{L}^{+}(\boldsymbol{L}^{+})' = D(n_1^{-1}, ..., n_m^{-1})$$

e, com $\boldsymbol{X}_j^o = \boldsymbol{A}_j\boldsymbol{X}_0$, $j = 1, ..., \ell$, ter-se-á

$$\boldsymbol{Y}_j^{oo} = \boldsymbol{A}_j\boldsymbol{Y}^{oo},\ j = 1, ..., \ell,$$

com vetor médio e a matriz de covariância dados por

- $\boldsymbol{\mu}_j^o = \boldsymbol{A}_j\boldsymbol{\mu}^o = \boldsymbol{X}_j^o\boldsymbol{\beta}_0, \quad j = 1, ..., \ell,$
- $\boldsymbol{V}_j^{oo} = \gamma_j\boldsymbol{I}_{g_j} + \sigma^2\boldsymbol{A}_j(\boldsymbol{L}^{+}(\boldsymbol{L}^{+})')\boldsymbol{A}_j',\ j = 1, ..., \ell.$

Sendo $\boldsymbol{P}_j$ e $\boldsymbol{Q}_j$ as $MPO$ sobre $R(\boldsymbol{X}_j^o)$ e $R(\boldsymbol{X}_j^o)^\perp$, com características $p_j$ e $f_j = g_j - p_j$, $j = 1, ..., \ell$, respectivamente, e $\boldsymbol{T}_j$ e $\boldsymbol{W}_j$ as matrizes cujos vetores linha constituem uma base ortonormada para $R(\boldsymbol{X}_j^o)$ e $R(\boldsymbol{X}_j^o)^\perp$, $j = 1, ..., \ell$, teremos

- $\boldsymbol{P}_j = \boldsymbol{T}_j'\boldsymbol{T}_j, \quad j = 1, ..., \ell,$
- $\boldsymbol{Q}_j = \boldsymbol{W}_j'\boldsymbol{W}_j, \ j = 1, ..., \ell.$

A seguir consideramos os testes de hipóteses para as componentes de variância canónicas, $\gamma_1, ..., \gamma_\ell$. Assumimos que

$$p_j < g_j, \ j = z+1, ...., \ell,$$

com $0 \leq z < \ell$. Consideremos

$$^{oo}\boldsymbol{Y}_j = \boldsymbol{W}_j\boldsymbol{Y}_j^{oo}, \ j = z+1, ...., \ell$$

os quais têm vetores médios nulos e matrizes de covariância $\gamma_j\boldsymbol{I}_{f_j} + \sigma^2\boldsymbol{B}_j, j > z$, com

$$\boldsymbol{B}_j = \boldsymbol{W}_j\boldsymbol{A}_j\boldsymbol{L}^+(\boldsymbol{L}^+)'\boldsymbol{A}_j'\boldsymbol{W}_j', j > z. \quad (2.3.17)$$

Como $\boldsymbol{Y}_j^{oo}$ é independente de $S$, $^{oo}\boldsymbol{Y}_j$ também é independente de $S$, $j > z$. Por outro lado,

$$^{oo}\boldsymbol{Y}_j = {}^{o}\boldsymbol{Y}_j + \boldsymbol{W}_j\boldsymbol{A}_j\boldsymbol{L}^+\boldsymbol{\varepsilon}, \ j > z, \quad (2.3.18)$$

com

$$^{o}\boldsymbol{Y}_j = \boldsymbol{W}_j\boldsymbol{A}_j\boldsymbol{Y}^{o}, \ j > z.$$

Quando a hipótese

$$H_{0,j} : \gamma_j = 0, \ j > z \quad (2.3.19)$$

se verifica, temos

$$pr(^{o}\boldsymbol{Y}_j = \boldsymbol{0}) = 1, \ j > z,$$

vindo

$$pr(^{oo}\boldsymbol{Y}_j = \boldsymbol{W}_j\boldsymbol{A}_j\boldsymbol{L}^+\boldsymbol{\varepsilon}) = 1, \ j > z.$$

Assim, $^{oo}\boldsymbol{Y}_j$ terá vetor médio nulo e matriz de covariância $\sigma^2\boldsymbol{B}_j, \ j > z$.

Devido à independência entre $^{oo}\boldsymbol{Y}_j$ e $S$, $j > z$, quando $H_{0,j}$ se verifica, a estatística

$$\mathcal{T}_j = \frac{(^{oo}\boldsymbol{Y}_j)'(\boldsymbol{B}_j^{-1})^{oo}\boldsymbol{Y}_j}{S}, \ j > z \quad (2.3.20)$$

seguirá uma distribuição $\overline{F}$ central com $f_j$, $j > z$, e $g(n)$ graus de liberdade, $\overline{F}(z|f_j, g(n))$, e pode ser utilizada como estatística de teste.

Vamos estabelecer

**Proposição 2.24** *Os testes com a estatística $\mathcal{T}_j$, $j > z$, são não distorcidos.*

**Dem:** Como ${}^{o}\boldsymbol{Y}_j$ é independente de $\boldsymbol{W}_j\boldsymbol{A}_j\boldsymbol{L}^{+}\varepsilon$ pode-se trabalhar com probabilidades condicionais, definidas a partir dos valores de $\|{}^{o}\boldsymbol{Y}_j\|^2$. As hipóteses $H_{0,j}$ verificam-se se e somente se $pr(\|{}^{o}\boldsymbol{Y}_j\|^2 = 0) = 1$ e quando $\|{}^{o}\boldsymbol{Y}_j\|^2 = \delta_j$ a distribuição de $\mathcal{T}_j$ é uma distribuição $\overline{F}$ não central com parâmetro de não-centralidade $\delta_j$, $\overline{F}(.|f_j, g(n), \delta_j),\ j > z$. Para completar a demonstração, basta observar que esta distribuição diminui com $\delta_j,\ j > z$, como vimos na seção (2.2.5) (ver também por exemplo Mexia (1990) e Nunes (2005)). ■

# Capítulo 3

# Testes $F$ com amostras de dimensão aleatória. Processos de contagem

Neste Capítulo iremos estender a ANOVA usual, com um e mais fatores, ao caso em que as dimensões das amostras não são previamente conhecidas. Estas serão tratadas como realizações de variáveis aleatórias, ver Capistrano et al. (2014), Mexia et al. (2011), Moreira et al. (2013), Nunes et al. (2012a, 2013, 2014).

Comecemos por supor que temos $m$ tratamentos e que:

- a distribuição do vetor das dimensões das amostras, $\boldsymbol{N} = (N_1, ..., N_m)^{'}$, é conhecida a menos de certos parâmetros,
- a distribuição condicional do vetor das observações, $\boldsymbol{Y}$, dado $\boldsymbol{N} = \boldsymbol{n}$, com $\boldsymbol{n} = (n_1, ..., n_m)^{'}$, é igualmente conhecida a menos de certos parâmetros,

então podemos formular um modelo afim de realizar inferência aquando da recolha das observações.

Assim, quando as dimensões das amostras não são previamente conhecidas, é mais correto considerá-las como realizações, $n_1, ..., n_m$, das variáveis aleatórias independentes, $N_1, ..., N_m$. O vetor $\boldsymbol{n} = (n_1, ..., n_m)^{'}$ será uma realização do vetor $\boldsymbol{N} = (N_1, ..., N_m)^{'}$.

Esta abordagem deve ser baseada na escolha apropriada da distribuição dos $N_1, ..., N_m$. No que se segue presume-se que a ocorrência das observações, para cada uma das amostras, corresponde a processos de contagem. A contagem é interrompida no final do período de tempo definido à partida. Isso leva-nos a considerar que as variáveis aleatórias $N_1, ..., N_m$ seguem uma distribuição de Poisson com parâmetros $\lambda_1, ..., \lambda_m$.

Consideremos ainda

$$n = \sum_{i=1}^{m} n_i,$$

como realização da variável aleatória

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

Devido à independência dos $N_1, ..., N_m$, a variável $N$ terá distribuição de Poisson com parâmetro

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

De seguida apresentamos um exemplo de uma situação prática em que, quanto a nós, o mais correto será a utilização desta abordagem. Suponhamos que se pretende realizar um estudo para comparar várias patologias de pacientes que chegam às urgências de um Hospital durante um determinado intervalo de tempo. O número de pacientes não é conhecido à partida e se as contagens forem repetidas num outro intervalo de tempo, com a mesma duração, o mais provável é obtermos um número diferente de pacientes para essas patologias. Assim, se pretendermos realizar apenas um estudo é mais correto considerar as dimensões das amostras como realizações de variáveis aleatórias.

Neste capítulo consideramos modelos de efeitos fixos, efeitos aleatórios e modelos mistos. São apresentados aplicações com dados reais, nomeadamente com pacientes com diferentes tipos de cancro no Brasil, para ilustrar a utilidade da nossa abordagem.

## 3.1 Distribuição de Poisson Truncada

Nesta seção iremos obter a expressão da função de probabilidade da Poisson truncada assumindo inicialmente que se tem uma dimensão mínima global para as amostras e depois assumindo uma dimensão mínina para cada uma das amostras. Ao assumirmos a existência de dimensões mínimas pretendemos evitar a ocorrência de casos altamente desequilibrados, veja-se por exemplo Mexia et al. (2011) e Nunes et al. (2014). Continuaremos a assumir que existem $m$ diferentes tratamentos no total. Os resultados obtidos nesta seção serão utilizados neste capítulo por forma a obter as distribuições não condicionais das estatísticas de teste.

### 3.1.1 Dimensão mínima global para as amostras

A forma mais comum para a distribuição de Poisson truncada é a omissão do valor zero como valor da variável, como se pode ver por exemplo em Johnson and Kotz (1969). Iremos portanto assumir que $N_i \geq 1,\ i = 1, ..., m$, uma vez que é necessário ter-se pelo menos uma observação por tratamento. Por forma a podermos realizar inferência assumiremos ainda que $N = \sum_{i=1}^{m} N_i > m$.

Como foi dito anteriormente consideremos que as variáveis $N_i, i = 1, ..., m$ seguem uma distribuição de Poisson com parâmetros $\lambda_i, i = 1, ..., m, N_i \sim P(\lambda_i)$, $i = 1, ..., m$. Assim, $N$ terá distribuição de Poisson com parâmetro $\lambda$, $N \sim P(\lambda)$. Teremos

$$\begin{aligned} p_{u,i} &= pr(N_i = u | N_i \geq 1) \\ &= \frac{pr(N_i = u)}{pr(N_i \geq 1)} = \frac{pr(N_i = u)}{1 - p(N_i = 0)} \\ &= \frac{e^{-\lambda_i}\lambda_i^u/u!}{1 - e^{-\lambda_i}} = \frac{e^{-\lambda_i}}{1 - e^{-\lambda_i}} \frac{\lambda_i^u}{u!}, \ u = 1, ...; \ i = 1, ..., m. \end{aligned}$$

A função geradora de momento dos $N_i$, quando $N_i \geq 1$, $i = 1, ..., m$, será dada por

$$\begin{aligned} \varphi_i(t) &= \sum_{u=1}^{\infty} \frac{e^{ut}e^{-\lambda_i}}{1 - e^{-\lambda_i}} \frac{\lambda_i^u}{u!} \\ &= \frac{e^{-\lambda_i}}{1 - e^{-\lambda_i}}(e^{\lambda_i e^t} - 1), \ i = 1, ..., m, \end{aligned}$$

e a sua função geradora de probabilidade dada por

$$\psi_i(t) = \varphi_i(\ln t) = \frac{e^{-\lambda_i}}{1 - e^{-\lambda_i}}(e^{\lambda_i t} - 1), \ i = 1, ..., m.$$

Vamos assumir agora que $\ddot{N}_i$, $i = 1, ..., m$, correspondem às variáveis truncadas $N_i$, $i = 1, ..., m$, quando $N_i \geq 1$, e que

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

Obtém-se então a função geradora de probabilidades de $\ddot{N}$ dada por

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

visto que

$$\prod_{i=1}^{m} (e^{\lambda_i t} - 1) = \sum_{\mathcal{C} \subseteq \overline{\overline{m}}} (-1)^{m - \sharp(\mathcal{C})} e^{(\sum_{i \in \mathcal{C}} \lambda_i)t},$$

onde $\overline{\overline{m}} = \{1, ..., m\}$ e $\sharp(\mathcal{C})$ corresponde ao cardinal de $\mathcal{C}$, sendo $\mathcal{C}$ um qualquer subconjunto de $\overline{\overline{m}}$.

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

Assim ter-se-á

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

com

$$\ddot{\psi}^{<u>}(0)=\sum_{\boldsymbol{u}\in P_u^{(m)}}\frac{(\sum_{i=1}^m u_i)!}{\prod_{i=1}^m u_i!}\prod_{i=1}^m\psi_i^{<u_i>}(0),\ u_i=1,\ldots,i=1,\ldots,m,$$

onde $P_u^m$ representa a família das partições com cardinal $m$ de $u=u_1+\ldots+u_m$ e $\boldsymbol{u}=(u_1,\ldots,u_m)'$.

Vamos considerar que

$$j_1+...+j_m=s;\ s=1,...,m-1,$$

logo $\boldsymbol{j}=(j_1,...,j_m)'$ tem pelo menos uma componente nula. Uma vez que $\psi_i(0)=0, i=1,\ldots,m$, ter-se-á

$$\prod_{i=1}^m\psi_i^{<j_i>}(0)=0$$

e consequentemente $\ddot{\psi}^{<s>}(0)=0, s=1,\ldots,m-1$. Assim, uma vez que $\ddot{\psi}^{<s>}(0)=s!\ddot{p}_s$, obtém-se

$$\ddot{p}_s=\frac{1}{s!}\ddot{\psi}^{<s>}(0)=0,\ s=1,\ldots,m-1.$$

Consideremos agora $s=m$. Neste caso o único termo não nulo de $\ddot{\psi}^{<m>}(0)$ corresponde a $\boldsymbol{j}=\mathbf{1}_m$, uma vez que $j_1+...+j_m=m$, logo

$$\begin{aligned}&\ddot{p}_m=pr(\ddot{N}=m)\\ =\ &\frac{1}{m!}\ddot{\psi}^{<m>}(0)=\prod_{i=1}^m\psi_i^{<1>}(0)\\ =\ &\prod_{i=1}^m\frac{e^{-\lambda_i}\lambda_i}{1-e^{-\lambda_i}}\end{aligned}$$

e consequentemente

$$\begin{aligned}&pr(\ddot{N}>m)=1-pr(\ddot{N}\leq m)\\ =\ &1-\ddot{p}_m=1-\prod_{i=1}^m\frac{e^{-\lambda_i}\lambda_i}{1-e^{-\lambda_i}}.\end{aligned}$$

Considerando $\ddot{\boldsymbol{N}}$ o vetor aleatório com componentes $\ddot{N}_1,...,\ddot{N}_m$, teremos $\ddot{N}>m$, o que significa que existe pelo menos um $\ddot{N}_i>1,\ i=1,...,m$, se e somente se $\ddot{\boldsymbol{N}}>\mathbf{1}_m$, então

$$pr(\ddot{\boldsymbol{N}}>\mathbf{1}_m)=p(\ddot{N}>m)=1-\ddot{p}_m.$$

Obtemos portanto

$$\begin{aligned}\ddot{p}_{u,m+1} &= pr(\ddot{N}=u|\ddot{N}\geq m+1)=\frac{pr(\ddot{N}=u)}{pr(\ddot{N}>m)}\\ &= \frac{pr(\ddot{N}=u)}{1-pr(\ddot{N}\leq m)}=\frac{\ddot{p}_u}{1-\ddot{p}_m},\ u=m+1,\ldots.\end{aligned}$$

Assumindo que se tem uma dimensão mínima global para as amostras, por exemplo $n^\bullet$, com $n^\bullet \geq m+1$, teremos portanto $\ddot{N}\geq n^\bullet$, donde se pode considerar

$$\begin{aligned}&\ddot{p}_{u,n^\bullet}=pr(\ddot{N}=u|\ddot{N}\geq n^\bullet)=\frac{pr(\ddot{N}=u)}{pr(\ddot{N}\geq n^\bullet)}\\ =\ &\frac{\ddot{p}_u}{pr(\ddot{N}>m)}\frac{pr(\ddot{N}>m)}{pr(\ddot{N}\geq n^\bullet)}\\ =\ &\ddot{p}_{u,m+1}\frac{1-\ddot{p}_m}{1-\sum_{\ell=m}^{n^\bullet-1}\ddot{p}_\ell},\ u=n^\bullet,\ldots.\end{aligned} \tag{3.1.1}$$

### 3.1.2 Dimensão mínima para cada uma das amostras

Como vimos na subseção anterior apenas podemos considerar $m$ parcelas para as partições, uma vez que estamos a considerar $m$ tratamentos na totalidade. Iremos assumir agora que existe uma dimensão mínima para cada uma das amostras, ver por exemplo Nunes et al. (2015). Consideremos então que $N_i\geq n_i^\bullet, i=1,\ldots,m$, o que significa que $\boldsymbol{N}\geq \boldsymbol{n}^\bullet$, com $\boldsymbol{n}^\bullet=(n_1^\bullet,\ldots,n_m^\bullet)'$ e portanto se tem como dimensão mínima global $n^\bullet=\sum_{i=1}^m n_i^\bullet$. Assim poderemos tomar

$$\begin{aligned}&p_{n,n_i^\bullet}=pr(N=n|\boldsymbol{N}\geq\boldsymbol{n}^\bullet)\\ =\ &\sum_{n_1=n_1^\bullet}^{n-\sum_{i=2}^m n_i^\bullet}\ldots\sum_{n_\ell=n_\ell^\bullet}^{n-(\sum_{i=1}^\ell n_i+\sum_{i=\ell+1}^m n_i^\bullet)}\ldots\sum_{n_m=n-\sum_{i=1}^{m-1}n_i}^{n-\sum_{i=1}^{m-1}n_i} pr(\boldsymbol{N}=\boldsymbol{n}|\boldsymbol{N}\geq\boldsymbol{n}^\bullet),\ n_i=n_i^\bullet,\ldots,\ i=1,\ldots,m,\end{aligned} \tag{3.1.2}$$

onde devido, à independência dos $N_i, i=1,\ldots,m$, se terá

$$\begin{aligned}&pr(\boldsymbol{N}=\boldsymbol{n}|\boldsymbol{N}\geq\boldsymbol{n}^\bullet)=\prod_{i=1}^m pr(N_i=n_i|N_i\geq n_i^\bullet)\\ =\ &\prod_{i=1}^m\frac{pr(N_i=n_i)}{pr(N_i\geq n_i^\bullet)}=\prod_{i=1}^m\frac{e^{-\lambda_i}(\lambda_i^{n_i}/n_i!)}{1-\sum_{u_i=0}^{n_i^\bullet-1}e^{-\lambda_i}(\lambda_i^{u_i}/u_i!)}\\ =\ &\prod_{i=1}^m\frac{\lambda_i^{n_i}}{n_i!(e^{\lambda_i}-\sum_{u_i=0}^{n_i^\bullet-1}\frac{\lambda_i^{u_i}}{u_i!})},\ n_i=n_i^\bullet,\ldots,\ i=1,\ldots,m.\end{aligned} \tag{3.1.3}$$

## 3.2 Modelos de efeitos fixos

Nesta seção iremos abordar a ANOVA de efeitos fixos e estender essa abordagem ao caso em que as dimensões das amostras são desconhecidas à partida. Começamos por considerar apenas um fator e em seguida mais que um fator de efeitos fixos.

### 3.2.1 Um fator com apenas um nível com dimensão aleatória

Vamos assumir que temos apenas um fator de efeitos fixos com $m$ níveis e que:

- a dimensão do $m$-ésimo nível é desconhecido à partida;
- $N$ é a variável aleatória que representa a dimensão do $m$-ésimo nível do fator;
- $n$ é a realização da variável aleatória $N$;
- as dimensões dos restantes $m-1$ níveis são consideradas fixas e iguais a $r$, com $r \geq n$.

Estamos interessados em testar a hipótese

$$H_{0,F} : \mu_1 = ... = \mu_m,$$

que pode ser escrita na forma

$$H_{0,F} : \boldsymbol{A\mu} = \mathbf{0},$$

onde $\boldsymbol{A} = [\boldsymbol{I}_{m-1}| - \mathbf{1}_{m-1}]$, com $\mathbf{1}_{m-1}$ o vetor com todos as $m-1$ componentes iguais a $1$ e $\boldsymbol{\mu}$ o vetor médio com componentes $\mu_1, \ldots, \mu_m$.

Como referido anteriormente vamos considerar que $N$ segue uma distribuição de Poisson truncada de parâmetro $\lambda$, $N \sim P(\lambda)$. Precisamos de ter pelo menos uma observação para o $m$-ésimo nível, o que significa que $N \geq 1$. Além disso como estamos a considerar que $r \geq n$, ter-se-á $n$ com valores entre $1$ e $r$. Esta situação é justificável para o caso em que o $m$-ésimo nível corresponde, por exemplo, a uma patologia rara, ver Nunes et al. (2010, 2012a). Assim tomaremos

$$\begin{aligned} p_n &= pr(N = n | 1 \leq N \leq r) = \frac{pr(N = n)}{pr(1 \leq N \leq r)} \\ &= \frac{e^{-\lambda}\lambda^n/n!}{\sum_{j=1}^{r} e^{-\lambda}\lambda^j/j!} = \frac{\lambda^n}{n!\sum_{j=1}^{r} \lambda^j/j!}, \ n = 1, ..., r. \end{aligned} \tag{3.2.4}$$

#### 3.2.1.1 Estatística de teste e suas distribuições

Quando $N = n$ teremos as amostras

$$Y_{i,1}, ..., Y_{i,r},\ i = 1, ..., m-1,$$

e

$$Y_{m,1}, ..., Y_{m,n},$$

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

$$S = \sum_{i=1}^{m-1}\sum_{j=1}^{r}(Y_{i,j} - Y_{i,\bullet})^2 + \sum_{j=1}^{n}(Y_{m,j} - Y_{m,\bullet})^2,$$

como se pode ver, por exemplo, em Searle et al. (1992) e Khuri et al. (1998). Ao assumirmos que as observações são normais e independentes com variância $\sigma^2$ e valores médios $\mu_i$, $i = 1, ..., m$, quando $N = n$, $S$ será o produto de $\sigma^2$ por um qui-quadrado central com

$$g(n) = (m-1)(r-1) + n - 1$$

graus de liberdade,

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

Sendo $n$ uma realização de $N$, $g(n)$ será uma realização de $g(N)$, logo teremos graus de liberdade aleatórios para os erros.

Além disso $S$ será condicionalmente independente do vetor $\boldsymbol{Y}_\bullet$, com componentes $Y_{1,\bullet}, ..., Y_{m,\bullet}$. Quando $N = n$, o vetor $\boldsymbol{Y}_\bullet$ será condicionalmente normal com vetor médio $\boldsymbol{\mu}$, com componentes $\mu_1, \ldots, \mu_m$, e matriz de covariância $\sigma^2 D\left(\frac{1}{r}, ..., \frac{1}{r}, \frac{1}{n}\right)$, onde $D\left(\frac{1}{r}, ..., \frac{1}{r}, \frac{1}{n}\right)$ representa a matriz diagonal com elementos principais $\frac{1}{r}, ..., \frac{1}{r}, \frac{1}{n}$. Colocamos

$$\underset{(N=n)}{\boldsymbol{Y}_\bullet \sim} \mathcal{N}\left(\boldsymbol{\mu}, \sigma^2 D\left(\tfrac{1}{r}, ..., \tfrac{1}{r}, \tfrac{1}{n}\right)\right).$$

Assim ter-se-á

$$\boldsymbol{A}\boldsymbol{Y}_\bullet \sim \mathcal{N}\left(\boldsymbol{A}\boldsymbol{\mu}, \sigma^2 \boldsymbol{A} D\left(\frac{1}{r}, ..., \frac{1}{r}, \frac{1}{n}\right)\boldsymbol{A}'\right),$$

vindo, ver por exemplo Mexia (1990),

$$S_{num} = (\boldsymbol{A}\boldsymbol{Y}_\bullet)' \left(\boldsymbol{A} D\left(\tfrac{1}{r}, ..., \tfrac{1}{r}, \tfrac{1}{n}\right)\boldsymbol{A}'\right)^{-1} (\boldsymbol{A}\boldsymbol{Y}_\bullet) \sim \sigma^2 \chi^2_{g,\delta(n)},$$

com $g = m - 1$ e

$$\delta(n) = \frac{1}{\sigma^2}(\boldsymbol{A}\boldsymbol{\mu})' \left(\boldsymbol{A} D\left(\frac{1}{r}, ..., \frac{1}{r}, \frac{1}{n}\right)\boldsymbol{A}'\right)^{-1} (\boldsymbol{A}\boldsymbol{\mu}).$$

Note-se que $\delta(n)$ será uma realização de $\delta(N)$, quando $N = n$, o que nos leva a admitir que o parâmetro de não centralidade é aleatório.

Observemos que

$$\boldsymbol{A}D\left(\frac{1}{r},...,\frac{1}{r},\frac{1}{n}\right)\boldsymbol{A}^{'}=\frac{1}{r}\boldsymbol{I}_g+\frac{1}{n}\boldsymbol{J}_g,$$

com $\boldsymbol{J}_r=\mathbf{1}_r\mathbf{1}_r^{'}$.

Considerando $\boldsymbol{K}_g=\boldsymbol{I}_g-\frac{1}{g}\boldsymbol{J}_c$ vamos obter

$$\boldsymbol{A}D\left(\frac{1}{r},...,\frac{1}{r},\frac{1}{n}\right)\boldsymbol{A}^{'}=\frac{1}{r}\boldsymbol{K}_g+\left(\frac{g}{n}+\frac{1}{r}\right)\frac{1}{g}\boldsymbol{J}_g.$$

Uma vez que $\boldsymbol{K}_g$ e $\frac{1}{g}\boldsymbol{J}_g$ são matrizes ortogonais entre si, teremos

$$\left(\boldsymbol{A}D\left(\frac{1}{r},...,\frac{1}{r},\frac{1}{n}\right)\boldsymbol{A}^{'}\right)^{-1}=r\boldsymbol{K}_g+\frac{nr}{gr+n}\frac{1}{g}\boldsymbol{J}_g.$$

Então, os valores próprios distintos de $\left(\boldsymbol{A}D\left(\frac{1}{r},...,\frac{1}{r},\frac{1}{n}\right)\boldsymbol{A}^{'}\right)^{-1}$ serão $r$ e $\frac{nr}{gr+n}$, logo o raio espectral desta matriz será $r$.

Assim

$$\delta(n)\leq r\frac{\|\boldsymbol{A\mu}\|^2}{\sigma^2},$$

e uma vez que $\boldsymbol{A\mu}$ tem componentes $\mu_i-\mu_m$, $i=1,...,g$, temos

$$\|\boldsymbol{A\mu}\|^2=\sum_{i=1}^{m-1}(\mu_i-\mu_m)^2,$$

vindo

$$\delta(n)\leq\frac{r}{\sigma^2}\sum_{i=1}^{m-1}(\mu_i-\mu_m)^2=r\Delta, \quad (3.2.5)$$

com

$$\Delta=\frac{1}{\sigma^2}\sum_{i=1}^{m-1}(\mu_i-\mu_m)^2,$$

onde $\Delta$ é invariante para a escala.

Quando $N=n$, tem-se a estatística de teste,

$$\Im_F=\frac{(\boldsymbol{AY}_\bullet)^{'}\left(\boldsymbol{A}D\left(\frac{1}{r},...,\frac{1}{r},\frac{1}{n}\right)\boldsymbol{A}^{'}\right)^{-1}(\boldsymbol{AY}_\bullet)}{S}=\frac{S_{num}}{S}, \quad (3.2.6)$$

com distribuição condicional $\overline{F}(.|g,g(n),\delta(n))$, que, como vimos no capítulo anterior, corresponde à distribuição do quociente de qui-quadrados independentes com $g$ e $g(n)$ graus de liberdade e parâmetros de não centralidade $\delta(n)$ e 0.

De acordo com o teorema das probabilidades totais a distribuição não condicional da estatística será dada por, ver Mexia et al. (2011) e Nunes et al. (2012a),

$$\overline{\overline{F}}(z) = pr(\Im_F \leq z) = \sum_{n=1}^{r} pr(N = n|1 \leq N \leq r) pr(\Im_F \leq z|N = n),$$

com

$$pr(\Im_F \leq z|N = n) = \overline{F}(z|g, g(n), \delta(n))$$

e

$$pr(N = n|1 \leq N \leq n) = p_n.$$

Assim ter-se-á

$$\begin{aligned} \overline{\overline{F}}(z) &= \sum_{n=1}^{r} p_n \overline{F}(z|g, g(n), \delta(n)) \\ &= \sum_{n=1}^{r} p_n e^{-\delta(n)/2} \sum_{j=0}^{\infty} \frac{\delta(n)^j}{2^j j!} \overline{F}(z|g + 2j, g(n)), \end{aligned} \tag{3.2.7}$$

uma vez que

$$\overline{F}(z|g, g(n), \delta(n)) = e^{\frac{-\delta(n)}{2}} \sum_{j=0}^{\infty} \frac{\delta(n)^j}{2^j j!} \overline{F}(z|g + 2j, g(n)), \tag{3.2.8}$$

ver por exemplo Nunes and Mexia (2006) e Nunes et al. (2012c).

Quando $H_{0,F}$ se verifica e $N = n$, $\delta(n) = 0$ e consequentemente a estatística $\Im_F$ terá como distribuição condicional

$$\overline{F}(.|g, g(n)).$$

Neste caso a distribuição não condicional poderá ser reescrita da seguinte forma

$$\overline{\overline{F}}(z) = \sum_{n=1}^{r} p_n \overline{F}(z|g, g(n)),$$

com

$$p_n = \frac{\lambda^n}{n! \sum_{i=1}^{r} \frac{\lambda^j}{j!}},\ n = 1, \ldots, r,$$

como definido em (3.2.4).

Uma aplicação a dados reais, considerando apenas um nível do fator com dimensão aleatória, pode ser consultada em Nunes et al. (2012a), em que o referido nível corresponde a uma patologia rara.

#### 3.2.1.2 Erro de truncatura

Como vimos na subseção anterior quando a hipótese nula não se verifica, as estatísticas de teste têm distribuições condicionais $\overline{F}$ não centrais dando origem, após descondicionamento, à expressão para $\overline{\overline{F}}(z)$ obtida em (3.2.7). Nesta seção vamos considerar esta expressão da distribuição não condicional e truncar a série nela existente, por forma a tornar mais fáceis os cálculos quando for necessário utilizá-la, ver Nunes et al. (2012a).

Consideremos

$$\begin{aligned}\overline{\overline{F}}_J(z) &= \sum_{n=1}^{r} p_n e^{-\delta(n)/2} \sum_{j=0}^{J} \frac{\delta(n)^j}{2^j j!} \overline{F}(z|g+2j, g(n)) \\ &= \sum_{n=1}^{r} p_n \overline{F}_J(z|g, g(n), \delta(n)), \end{aligned} \tag{3.2.9}$$

com

$$\overline{F}_J(z|g, g(n), \delta(n)) = e^{\frac{-\delta(n)}{2}} \sum_{j=0}^{J} \frac{(\delta(n))^j}{2^j j!} \overline{F}(z|g+2j, g(n)).$$

Uma vez que

$$0 \leq \overline{F}(z|g+2j, g(n)) \leq 1,\ j = 0, ..., \infty,\ n = 1, ..., r,$$

tem-se

$$\overline{F}_J(z|g, g(n), \delta(n)) < \overline{F}(z|g, g(n), \delta(n)) < \overline{F}_J(z|g, g(n), \delta(n)) + \varepsilon_J(n),$$

com

$$\varepsilon_J(n) = e^{-\delta(n)/2} \sum_{j=J+1}^{\infty} \frac{\delta(n)^j}{2^j j!} = 1 - e^{-\delta(n)/2} \sum_{j=0}^{J} \frac{\delta(n)^j}{2^j j!},\ n = 1, ..., r.$$

Considerando

$$f(u) = 1 - e^{-u} \sum_{j=0}^{v} \frac{u^j}{j!},$$

tem-se

$$\frac{df(u)}{du} = e^{-u} \sum_{j=0}^{v} \frac{u^j}{j!} - e^{-u} \sum_{j=1}^{v} \frac{u^{j-1}}{(j-1)!} = e^{-u} \frac{u^v}{v!} > 0.$$

Consequentemente, ver em Nunes et al. (2010) e Nunes et al. (2012a),

$$\varepsilon_J(n) < \varepsilon_J,$$

com

$$\varepsilon_J = 1 - e^{-r\Delta/2} \sum_{j=0}^{J} \frac{(r\Delta)^j}{2^j j!},$$

em que, como vimos anteriormente em (3.2.5), $r\Delta = \frac{r}{\sigma^2}\sum_{i=1}^{m-1}(\mu_i - \mu_m)^2$ é um majorante de $\delta(n)$.

Este limite superior, $\varepsilon_J$, não depende de $n$, podemos portanto considerar

$$\overline{\overline{F}}_J(z) < \overline{\overline{F}}(z) < \overline{\overline{F}}_J(z) + \varepsilon_J.$$

Calculámos, através do *software* R, o valor mínimo de $J$ de modo a que $\varepsilon_J < w$, com $w = 10^{-6}$. Considerámos diferentes valores de $r\Delta$ e obtivemos os valores apresentados na Tabela 3.1.

Tabela 3.1: Valores mínimos de $J$.

| **Valores de** $r\Delta$ | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1/50 | 1/10 | 1/5 | 1 | 2 | 5 | 10 | 15 | 20 | 30 | 40 |
| **Valores Mínimos de** $J$ | 2 | 3 | 4 | 7 | 9 | 13 | 19 | 24 | 28 | 37 | 45 |

Desta tabela podemos concluir que não é necessário considerar um valor muito elevado de $J$ de modo a que o erro de truncatura, $\varepsilon_J$, seja inferior a $10^{-6}$. Assim podemos afirmar que temos um excelente controle do erro de truncatura.

### 3.2.2 Um fator com todos os níveis com dimensões aleatórias

Nesta seção continuamos a considerar apenas um fator com $m$ níveis em que agora a dimensão das amostras são desconhecidas à partida para esses $m$ níveis.

Vamos assumir então que:

- $N_1, ..., N_m$ são variáveis aleatórias que representam as dimensões das amostras;
- $n_1, ..., n_m$ são as realizações dessas variáveis aleatórias;
- as variáveis independentes $N_1, ..., N_m$ têm distribuições de Poisson com parâmetros $\lambda_1, ..., \lambda_m, N_i \sim P(\lambda_i), i = 1, ..., m$;
- $n = \sum_{i=1}^{m} n_i$ é a realização da variável aleatória $N = \sum_{i=1}^{m} N_i$;
- devido à independência dos $N_i,\ i = 1, \ldots, m, N$ terá distribuição de Poisson com parâmetro $\lambda = \sum_{i=1}^{m} \lambda_i, N \sim P(\lambda)$.

À semelhança do caso anterior pretendemos testar a hipótese

$$H_{0,F} : \mu_1 = ... = \mu_m,$$

que pode ser reescrita

$$H_{0,F} : \boldsymbol{A\mu} = \mathbf{0},$$

onde $\boldsymbol{\mu}$ corresponde ao vetor médio com componentes $\mu_1, ..., \mu_m$, e $\boldsymbol{A} = [\boldsymbol{I}_{m-1}| - \mathbf{1}_{m-1}]$.

#### 3.2.2.1 Estatística de teste e suas distribuições

Considerando $N_i = n_i$, $i = 1, ..., m$, temos as amostras $Y_{i,1}, ..., Y_{i,n_i}$, $i = 1, ..., m$, com médias $Y_{i,\bullet}$, $i = 1, ..., m$. A soma das somas dos quadrados dos erros será neste caso dada por, ver mais uma vez por exemplo Khuri et al. (1998) e Searle et al. (1992),

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

Considerando as observações normais e independente com variância $\sigma^2$, quando $N_i = n_i$, $i = 1, ..., m$, $S$ será o produto por $\sigma^2$ de um qui-quadrado central com $g(n) = n - m$ graus de liberdade, $S \sim \sigma^2 \chi^2_{g(n)}$.

Adicionalmente, quando $N_i = n_i$, $i = 1, \ldots, m$, $S$ será condicionalmente independente do vetor das médias dos tratamentos, $\boldsymbol{Y}_\bullet$, que será normalmente distribuido com vetor médio $\boldsymbol{\mu}$ e matriz de covariância $\sigma^2 D(\frac{1}{n_1}, ..., \frac{1}{n_m})$. Assim,

$$S_{num} = (\boldsymbol{A}\boldsymbol{Y}_\bullet)' \left( \boldsymbol{A} D\left(\tfrac{1}{n_1}, ..., \tfrac{1}{n_m}\right) \boldsymbol{A}' \right)^{-1} (\boldsymbol{A}\boldsymbol{Y}_\bullet)$$

será o produto de $\sigma^2$ por um qui-quadrado não central com $g = m - 1$ graus de liberdade e parâmetro de não-centralidade

$$\delta(n) = \frac{1}{\sigma^2} (\boldsymbol{A}\boldsymbol{\mu})' \left( \boldsymbol{A} D\left(\frac{1}{n_1}, ..., \frac{1}{n_m}\right) \boldsymbol{A}' \right)^{-1} (\boldsymbol{A}\boldsymbol{\mu}),$$

$S_{num} \sim \sigma^2 \chi^2_{g,\delta(n)}$. Quando $H_{0,F}$ se verifica, $\delta(n) = 0$ e $S_{num} \sim \sigma^2 \chi^2_g$.

Assim, quando $N = n$ e $H_{0,F}$ se verifica, a distribuição condicional da estatística

$$\Im_F = \frac{S_{num}}{S}$$

será $\overline{F}(.|g, g(n))$.

Afim de realizarmos inferência, vamos supor que temos uma dimensão mínima para cada uma das amostras. Vamos portanto considerar que $N_i \geq n_i^\bullet$, $i = 1, ..., m$, o que significa que teremos como dimensão mínima global $n^\bullet = \sum_{i=1}^m n_i^\bullet$. Assim tomaremos, com $\boldsymbol{n}^\bullet = (n_1^\bullet, ..., n_m^\bullet)'$,

$$\begin{aligned} & p_{n,n_i^\bullet} = pr(N = n | \boldsymbol{N} \geq \boldsymbol{n}^\bullet) \\ = & \sum_{n_1 = n_1^\bullet}^{n - \sum_{i=2}^m n_i^\bullet} \cdots \sum_{n_\ell = n_\ell^\bullet}^{n - (\sum_{i=1}^{\ell} n_i + \sum_{i=\ell+1}^{m} n_i^\bullet)} \cdots \sum_{n_m = n - \sum_{i=1}^{m-1} n_i}^{n - \sum_{i=1}^{m-1} n_i} pr(\boldsymbol{N} = \boldsymbol{n} | \boldsymbol{N} \geq \boldsymbol{n}^\bullet), \end{aligned}$$

onde

$$pr(\boldsymbol{N}=\boldsymbol{n}|\boldsymbol{N}\geq\boldsymbol{n}^{\bullet})=\prod_{i=1}^{m}\frac{\lambda_i^{n_i}}{n_i!(e^{\lambda_i}-\sum_{u_i=0}^{n_i^{\bullet}-1}\frac{\lambda_i^{u_i}}{u_i!})},\ n_i=n_i^{\bullet},\ldots,\ i=1,\ldots,m$$

conforme definido em (3.1.3).

A distribuição não condicional de $\Im_F$, quando $H_{0,F}$ se verifica, será dada por, ver por exemplo Mexia et. al (2011) e Nunes et. al (2014),

$$\overline{\overline{F}}(z)=\sum_{n=n^{\bullet}}^{\infty}pr(N=n|\boldsymbol{N}\geq\boldsymbol{n}^{\bullet})\overline{F}(z|g,g(n))=\sum_{n=n^{\bullet}}^{\infty}p_{n,n_i^{\bullet}}\overline{F}(z|g,g(n)).\quad(3.2.10)$$

Quando $N\leq\overline{n}$, podemos desprezar em (3.2.10) os termos para os quais $n>\overline{n}$, tendo-se o "encaixe"

$$\overline{\overline{F}}_{\overline{n}}(z)<\overline{\overline{F}}(z)<\overline{\overline{F}}^{*}_{\overline{n}}(z),\quad(3.2.11)$$

onde

$$\overline{\overline{F}}_{\overline{n}}(z)=\sum_{n=n^{\bullet}}^{\overline{n}}pr(N=n|\boldsymbol{N}\geq\boldsymbol{n}^{\bullet})\overline{F}(z|g,g(n))\quad(3.2.12)$$

e

$$\overline{\overline{F}}^{*}_{\overline{n}}(z)=\overline{\overline{F}}_{\overline{n}}(z)+\sum_{n=\overline{n}+1}^{\infty}pr(N=n|\boldsymbol{N}\geq\boldsymbol{n}^{\bullet}).\quad(3.2.13)$$

Observemos que

$$\sum_{n=\overline{n}+1}^{\infty}pr(N=n|\boldsymbol{N}\geq\boldsymbol{n}^{\bullet})=1-\sum_{n=n^{\bullet}}^{\overline{n}}pr(N=n|\boldsymbol{N}\geq\boldsymbol{n}^{\bullet}).\quad(3.2.14)$$

Este tipo de encaixe também se verifica para as distribuições não centrais. Vimos que $S_{num}\sim\sigma^2\chi^2_{g,\delta(n)}$ e portanto, quando $\boldsymbol{N}=\boldsymbol{n}$, tem-se

$$\Im_F=\frac{S_{num}}{S}\sim\overline{F}(z|g,g(n),\delta(n)),$$

obtendo-se a distribuição não condicional

$$\overline{\overline{F}}^{\circ}(z)=\sum_{n=n^{\bullet}}^{\infty}pr(N=n|\boldsymbol{N}\geq\boldsymbol{n}^{\bullet})\overline{F}(z|g,g(n),\delta(n)).$$

Ter-se-á agora o "encaixe"

$$\overline{\overline{F}}^{\circ}_{\overline{n}}(z) < \overline{\overline{F}}^{\circ}(z) < \overline{\overline{F}}^{\circ *}_{\overline{n}}(z),$$

com

$$\overline{\overline{F}}^{\circ}_{\overline{n}}(z) = \sum_{n=n^{\bullet}}^{\overline{n}} pr(N = n|\boldsymbol{N} \geq \boldsymbol{n}^{\bullet})\overline{F}(z|g, g(n), \delta(n))$$

e

$$\overline{\overline{F}}^{\circ *}_{\overline{n}}(z) = \overline{\overline{F}}^{\circ}_{\overline{n}}(z) + \sum_{n=\overline{n}+1}^{\infty} pr(N = n|\boldsymbol{N} \geq \boldsymbol{n}^{\bullet}).$$

Facilmente se conclui que

$$\overline{\overline{F}}^{\circ *}_{\overline{n}}(z) - \overline{\overline{F}}^{\circ}_{\overline{n}}(z) = \overline{\overline{F}}^{*}_{\overline{n}}(z) - \overline{\overline{F}}_{\overline{n}}(z) = \sum_{n=\overline{n}+1}^{\infty} pr(N = n|\boldsymbol{N} \geq \boldsymbol{n}^{\bullet}),$$

pelo que o valor de $\overline{n}$ a utilizar pode ser o mesmo do caso central.

#### 3.2.2.2 Uma aplicação a dados do cancro

Os dados utilizados nesta aplicação são referentes à cidade de São Paulo, 2010, e dizem respeito à idade de deteção da doença. O fator é o tipo de cancro e terá três níveis: *Tecidos moles do tórax*, *Trato intestinal* e *Cavidade nasal*. As Tabelas do Anexo 1 mostram as frequências destes três tipos de cancro, agrupadas por idade. A Tabela seguinte ilustra o número de pacientes afetados por estes tipos de cancro.

Tabela 3.2: Tipos de cancro e número de pacientes.

| Tipos de cancro | Número de pacientes |
|---|---|
| Tecidos moles do tórax | 18 |
| Trato intestinal | 22 |
| Cavidade nasal | 25 |

Neste caso vamos testar a hipótese

$$H_{0,F} : \boldsymbol{A\mu} = 0,$$

com

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix}.$$

Quando $H_{0,F}$ se verifica, o numerador da estatística $\Im_F$ é dado por

$$S_{num} = (\boldsymbol{AY}_{\bullet})' \left(\boldsymbol{A}D\left(\tfrac{1}{18}, \tfrac{1}{22}, \tfrac{1}{25}\right)\boldsymbol{A}'\right)^{-1} (\boldsymbol{AY}_{\bullet}) \sim \sigma^2\chi^2_g \ ,$$

com $g = m - 1 = 2$. Vamos então obter

$$\left(\boldsymbol{A}D\left(\frac{1}{18}, \frac{1}{22}, \frac{1}{25}\right)\boldsymbol{A}'\right)^{-1} = \begin{bmatrix} 13.0154 & -6.0923 \\ -6.0923 & 14.5538 \end{bmatrix} e \ \boldsymbol{A}\boldsymbol{y}_{\bullet} = \begin{bmatrix} -12.9000 \\ -0.6273 \end{bmatrix},$$

em que as médias amostrais, componentes do vetor $\boldsymbol{y}_{\bullet}$, são respectivamente

- $y_{1,\bullet} = 49.50$;
- $y_{2,\bullet} = 61.7727$;
- $y_{3,\bullet} = 62.40$.

Obtemos então para o numerador da estatística

$$S_{num} = 2073.021.$$

O denominador da estatística, quando $N = n$, é o produto de $\sigma^2$ por um qui-quadrado central com $g(n) = n - 3$ graus de liberdade, $S \sim \sigma^2\chi^2_{n-3}$. Neste caso vamos obter

$$S = \sum_{j=1}^{18}(y_{1,j} - \overline{y}_{1,\bullet})^2 + \sum_{j=1}^{22}(y_{2,j} - \overline{y}_{2,\bullet})^2 + \sum_{j=1}^{25}(y_{3,j} - \overline{y}_{3,\bullet})^2 = 26632.364.$$

O valor observado da estatística, $\Im_{F,Obs}$, será

$$\Im_{F,Obs} = \frac{2073.021}{26632.364} = 0.07784.$$

Quando $N = n$, e $H_{0,F}$ se verifica, a distribuição condicional de $\Im_F$ é uma distribuição $\overline{F}$ central com $g = 2$ e $g(n) = 65 - 3 = 62$ graus de liberdade, $\overline{F}(z|2, 62)$. Os quantis, $z_{1-\alpha}$, da distribuição condicional são dados na Tabela 3.3. Estes quantis são obtidos considerando $z_{1-\alpha} = \frac{2}{63} \times f_{1-\alpha,2,62}$, onde $f_{1-\alpha,2,62}$ corresponde ao quantil $(1 - \alpha)$ de uma distribuição $F$ central com $2$ e $62$ graus de liberdade. Assim, podemos concluir que se rejeita $H_{0,F}$ para $\alpha = 0.1$, pois $\Im_{F,Obs} > z_{1-\alpha}$, e não se rejeita para $\alpha = 0.05$ e $0.01$.

Tabela 3.3: Os quantis da distribuição condicional e não condicional de $\Im_F$

| **Valores de** $\alpha$ | 0.10 | 0.05 | 0.01 |
|---|---|---|---|
| $z_{1-\alpha}$ | 0.07711 | 0.10146 | 0.16016 |
| $z^e_{1-\alpha}$ | 0.07856 | 0.10341 | 0.16341 |

Por forma a realizar os cálculos vamos assumir que $\lambda_i, \ i = 1, 2, 3$, correspondem ao número médio de casos por ano. Assim, teremos $\lambda_1 = 18$; $\lambda_2 = 22$ e $\lambda_3 = 25$. Vamos supor ainda que se tem no mínimo cinco observações para cada um dos níveis do fator, o que significa que $n_i^{\bullet} = 5, \ i = 1, 2, 3, n^{\bullet} = \sum_{i=1}^{3} n_i = 15$ e consequentemente $\boldsymbol{n}^{\bullet} = (5, 5, 5)'$.

Afim de calcularmos os quantis da distribuição não condicional de $\Im_F$, iremos determinar o valor de $\overline{n}$ da expressão (3.2.12) de modo a que

$$\sum_{n=\overline{n}+1}^{\infty} pr(N=n|\boldsymbol{N}\geq\boldsymbol{n}^{\bullet}) = 1-\sum_{n=n^{\bullet}}^{\overline{n}} pr(N=n|\boldsymbol{N}\geq\boldsymbol{n}^{\bullet}) < 10^{-4}. \quad (3.2.15)$$

Obteve-se o valor mínimo de $\overline{n}=97$ para que (3.2.15) seja satisfeita. Assim, quando $H_{0,F}$ se verifica, ter-se-á

$$\begin{aligned}\overline{\overline{F}}_{\overline{n}}(z) &= \sum_{n=n^{\bullet}}^{97} p_{n,n_i^{\bullet}}\overline{F}(z|2,n-3)\\ &= \sum_{n=15}^{97}\sum_{n_1=5}^{n-10}\sum_{n_2=5}^{n-(n_1+5)}\sum_{n_3=n-(n_1+n_2)}^{n-(n_1+n_2)} pr(\boldsymbol{N}=\boldsymbol{n}|\boldsymbol{N}\geq\boldsymbol{n}^{\bullet})\overline{F}(z|2,n-3),\end{aligned}$$

com

$$pr(\boldsymbol{N}=\boldsymbol{n}|\boldsymbol{N}\geq\boldsymbol{n}^{\bullet}) = \prod_{i=1}^{3}\frac{\lambda_i^{n_i}/n_i!}{e^{\lambda_i}-\sum_{u_i=0}^{4}\frac{\lambda_i^{u_i}}{u_i!}}.$$

Os quantis, $z^e_{1-\alpha}$, para a probabilidade $1-\alpha$, são apresentados na Tabela 3.3. Visto que $\Im_{Obs,F} < z^e_{1-\alpha}$, conclui-se que não se rejeita $H_0$ para os níveis usuais de significância. Concluímos portanto, que neste caso para $\alpha=0.1$ a abordagem não condicional nos leva a tomar uma decisão contrária à que tínhamos tomado considerando a abordagem clássica.

#### 3.2.2.3 Cálculo dos valores críticos

Nesta subseção iremos apresentar uma outra forma para calcular os valores críticos. Notemos que a utilização de valores críticos incorretos nos podem conduzir à utilização de níveis incorretos para os testes de hipóteses.

Vimos que a distribuição não condicional de $\Im_F$, quando $H_0$ se verifica, é dada por

$$\overline{\overline{F}}(z) = \sum_{n=n^{\bullet}}^{\infty} pr(N=n|\boldsymbol{N}\geq\boldsymbol{n}^{\bullet})\overline{F}(z|g,g(n)) = \sum_{n=n^{\bullet}}^{\infty} p_{n,n_i^{\bullet}}\overline{F}(z|g,g(n)),$$

com $p_{n,n_i^{\bullet}}$ conforme definido em (3.1.2).

Então, considerando (3.2.11) tem-se

$$\overline{\overline{F}}_{\overline{n}}(z) < \overline{\overline{F}}(z) < \overline{\overline{F}}^{*}_{\overline{n}}(z),$$

e, consequentemente,

$$f^{*}_{\overline{n},1-\alpha} < f_{1-\alpha} < f_{\overline{n},1-\alpha},$$

com $f_{\overline{n},1-\alpha}$, $f_{1-\alpha}$ e $f^{*}_{\overline{n},1-\alpha}$ os $(1-\alpha)$-ésimo quantis para estas distribuições, como podemos ver na Figura 3.1 (ver também por exemplo Mexia et al. (2011)).

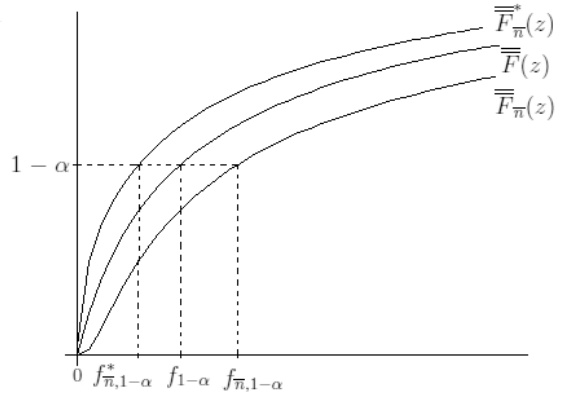


Figura 3.1: Relação entre as distribuições e os seus quantis.

Assim, o valor aproximado do quantil pode ser considerado

$$\widetilde{f}_{1-\alpha} = \frac{f_{\overline{n},1-\alpha} + f^{*}_{\overline{n},1-\alpha}}{2}, \quad (3.2.16)$$

que pode ser usado como um valor crítico, para os valores usuais de $\alpha$, ver Nunes et al. (2014).

Uma vez que os parâmetros, $\lambda_i$, $i = 1, ..., m$, são desconhecidos, iremos agora mostrar como lidar com estes, por forma a calcular os valores críticos. Para tal vamos calcular os limites inferiores para estes parâmetros.

**Cálculo dos limites inferiores dos parâmetros das distribuições de Poisson**

A distribuição não condicional cresce com $\lambda_i$, $i = 1, ..., m$, logo os correspondentes quantis decrescem, portanto vamos usar limites inferiores para estes parâmetros, ver Nunes et al. (2014).

Vamos considerar que como esses limites inferiores são os valores mínimos de $\lambda_i$, $i = 1, ..., m$, tal que

$$e^{-\lambda_i}\frac{\lambda_i^{n_i}}{n_i!} = \alpha,\ i = 1, ..., m, \quad (3.2.17)$$

com $\alpha$ os níveis usuais de significância.

Assim, considerando

$$g_{n_i}(\lambda_i) = \frac{1}{n_i!} e^{-\lambda_i} \lambda_i^{n_i},$$

obtemos

$$\frac{dg_{n_i}(\lambda_i)}{d\lambda_i} = -\frac{1}{n_i!} e^{-\lambda_i} \lambda_i^{n_i} + \frac{1}{n_i!} e^{-\lambda_i} n_i \lambda_i^{n_i - 1} = \frac{1}{n_i!} e^{-\lambda_i} \lambda_i^{n_i - 1} (-\lambda_i + n_i),$$

o que significa que $g_{n_i}(\lambda_i)$ aumenta com $\lambda_i$, quando $\lambda_i < n_i,\ i = 1, ..., m$. Vamos portanto considerar a solução da equação (3.2.17), tal que $\lambda_i < n_i,\ i = 1, ..., m$, que designamos por $\lambda_{\alpha,i}$, por forma a obtermos o intervalo de confiança a $(1-\alpha)\%$ para $\lambda_i$, $i = 1, ..., m$,

$$[\lambda_{\alpha,i}; +\infty[.$$

**Obtenção dos resultados**

Vamos considerar os dados utilizados na aplicação da seção 3.2.2.2, assumir que temos a mesma dimensão mínima para cada amostra $(n_i^\bullet = 5,\ i = 1, 2, 3)$ e calcular os valores críticos. Os valores de $n_i$, $i = 1, 2, 3$, correspondem ao número de pacientes apresentados na Tabela 3.2.

Os limites inferiores, $\lambda_{\alpha,i}$, para $\lambda_i$, $i = 1, 2, 3$, obtidos considerando os níveis usuais de significância, são apresentados na Tabela 3.4.

Tabela 3.4: Os limites inferiores para $\lambda_i$, $i = 1, 2, 3$.

| Valores de $\alpha$ | 0.05 | 0.01 |
|---|---|---|
| $\lambda_{\alpha,1}$ | 13.5 | 10.5 |
| $\lambda_{\alpha,2}$ | 17.5 | 13.75 |
| $\lambda_{\alpha,3}$ | 20.5 | 16.25 |

Recorrendo a (3.2.12), tem-se

$$\overline{\overline{F}}_{\overline{n}}(z) = \sum_{n=15}^{\overline{n}} p_{n,n_i^\bullet} \overline{F}(z|2, n-3)$$

e

$$\begin{aligned} \overline{\overline{F}}^*_{\overline{n}}(z) &= \overline{\overline{F}}_{\overline{n}}(z) + \left(1 - \sum_{n=15}^{\overline{n}} p_{n,n_i^\bullet}\right) \\ &= \overline{\overline{F}}_{\overline{n}}(z) + q, \end{aligned}$$

onde

$$q = 1 - \sum_{n=15}^{\overline{n}} p_{n,n_i^\bullet},$$

não depende de $z$. Assumindo os limites inferiores para $\lambda_i$, $i = 1, 2, 3$, dados na Tabela 3.4, obtemos $\overline{n} = 80$ [$\overline{n} = 66$] tal que $q < 10^{-4}$ para $\alpha = 0.05$ [$\alpha = 0.01$]. Calculámos os quantis de $\overline{\overline{F}}^*_{\overline{n}}(z)$ substituindo $1 - \alpha$ por $(1 - \alpha) - q$, assumindo que $q = 10^{-4}$.

Os valores críticos, $\widetilde{f}_{1-\alpha}$, definidos em (3.2.16), são apresentados na Tabela 3.5.

Tabela 3.5: Valores críticos corretos.

| Valores de $\alpha$ | 0.05 | 0.01 |
|---|---|---|
| $\widetilde{f}_{1-\alpha}$ | 0.13476 | 0.28704 |

Comparando estes valores críticos com os obtidos na seção 3.2.2.2 verificamos que estes são ligeiramente mais elevados. Ao utilizarmos este método para o cálculo dos valores críticos estaremos a considerar um modelo "mais completo", pelo que aumenta a robustez. Além disso, a utilização de um limite inferior para os parâmetros $\lambda_i$, $i = 1, 2, 3$, diminui os valores considerados para $\overline{n}$.

### 3.2.3 Mais do que um fator de efeitos fixos

Vamos agora admitir que temos mais do que um fator de efeitos fixos com $m$ tratamentos na totalidade, ver Nunes et al. (2014), e que:

- as dimensões das amostras para os diferentes $m$ tratamentos são desconhecidas à partida;
- $N_1, ..., N_m$ são variáveis aleatórias independentes que representam as dimensões das amostras;
- $n_1, ..., n_m$ são as realizações destas variáveis aleatórias;
- as variáveis $N_1, ..., N_m$ seguem distribuições de Poisson com parâmetros $\lambda_1, ..., \lambda_m, N_i \sim P(\lambda_i), i = 1, ..., m$;
- $n = \sum_{i=1}^{m} n_i$ é a realização da variável aleatória $N = \sum_{i=1}^{m} N_i$;
- devido à independência dos $N_i, i = 1, \ldots, m$, $N$ tem distribuição de Poisson com parâmetro $\lambda = \sum_{i=1}^{m} \lambda_i, N \sim P(\lambda)$.

Consideremos $\Omega = \bigoplus_{j=1}^{\tau} \overline{w}_j$ o espaço gerado pelo vetor das médias dos tratamentos $\boldsymbol{\mu}$. Neste caso estamos interessados em testar as hipóteses

$$H_{0,j} : \boldsymbol{\mu} \in w_j, j = 1, \ldots, \tau,$$

onde $w_j = (w_j^{\perp} \cap \Omega), j = 1, \ldots, \tau$, é um subespaço do espaço paramétrico $\Omega$, com $\overline{w}_j^{\perp}$ o complemento ortogonal de $\overline{w}_j, j = 1, \ldots, \tau$. Assumindo que os vetores linha das matrizes $\boldsymbol{A}_j$ constituem uma base ortonormada para $w_j, j = 1, \ldots, \tau$, podemos reescrever as hipóteses anteriores da seguinte forma

$$H_{0,j} : \boldsymbol{A}_j \boldsymbol{\mu} = 0, j = 1, \ldots, \tau,$$

as quais correspondem às hipóteses de ausência de efeitos e interação entre os fatores.

#### 3.2.3.1 Estatística de teste e suas distribuições

Para obtermos a distribuição não condicional das estatísticas assumiremos neste caso que se tem uma dimensão mínima global para as amostras. Por essa razão utilizaremos as probabilidades obtidas em 3.1.1 e vamos considerar as variáveis aleatórias $\ddot{N}_i, i = 1, \ldots, m$, que, tal como definidas anteriormente, correspondem às variáveis truncadas $N_i, i = 1, \ldots, m$, quando $N_i \geq 1$. Recordemos que $\ddot{N} = \sum_{i=1}^{m} \ddot{N}_i$.

Assim, quando $\ddot{N}_i = n_i$, $i = 1, ..., m$, temos as amostras

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

com médias $Y_{i,\bullet}$, $i = 1, ..., m$. A soma das somas dos quadrados dos erros será mais uma vez dada por

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

Se assumirmos que as observações são normais, independentes, com variância $\sigma^2$, tem-se

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

com

$$g(n) = n - m.$$

Além disso $S$ será condicionalmente independentes do vetor $\boldsymbol{Y}_\bullet$, e ter-se-á

$$\underset{(\ddot{N} = n)}{\boldsymbol{Y}_\bullet} \sim \mathcal{N}\left(\boldsymbol{\mu}, \sigma^2 D(\tfrac{1}{n_1}, ..., \tfrac{1}{n_m})\right).$$

Quando $\ddot{N}_i = n_i$, $i = 1, ..., m$,

$$S_j = (\boldsymbol{A}_j \boldsymbol{Y}_\bullet)' \left(\boldsymbol{A}_j D\left(\frac{1}{n_1}, ..., \frac{1}{n_m}\right) \boldsymbol{A}_j'\right)^{-1} (\boldsymbol{A}_j \boldsymbol{Y}_\bullet) \sim \sigma^2 \chi^2_{g_j, \delta_j(n)},\ j = 1, ..., \tau,$$

com

$$g_j = car(\boldsymbol{A}_j),\ j = 1, ..., \tau$$

e

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

Assim, com $\ddot{N}_i = n_i$, $i = 1, ..., m$ tem-se a estatística de teste

$$\Im_j = \frac{S_j}{S},\ j = 1, ..., \tau,$$

com distribuição condicional $\overline{F}(z|g_j, g(n), \delta_j(n))$, ver por exemplo Nunes e Mexia (2006) e Nunes et al. (2012c).

Vamos assumir agora que a dimensão mínima global para as amostras corresponde a $n^\bullet$, o que significa que $\ddot{N} \geq n^\bullet$. A distribuição não condicional da estatística será dada por

$$\overline{\overline{F}}_j(z) = \sum_{n=n^\bullet}^{\infty} pr(\ddot{N} = n|\ddot{N} \geq n^\bullet)\overline{F}(z|g, g(n), \delta_j(n)) = \sum_{n=n^\bullet}^{\infty} \ddot{p}_{n,n^\bullet}\overline{F}((z|g, g(n), \delta_j(n)),$$

onde

$$\ddot{p}_{n,n^\bullet} = \ddot{p}_{n,m+1}\frac{1-\ddot{p}_m}{1-\sum_{\ell=m}^{n^\bullet-1}\ddot{p}_\ell},\ u = n^\bullet, \ldots,$$

como definido em (3.1.1). Quando $\ddot{N} = n$, e $H_{0,j}$ se verifica, $\delta_j(n) = 0, j = 1, \ldots, \tau$ e a distribuição condicional de $\Im_j$, passará a ser

$$\overline{F}(z|g_j, g(n)),\ j = 1, ..., \tau,$$

enquanto a distribuição não condicional de $\Im_j,\ j = 1, ..., \tau$, poderá ser reescrita da seguinte forma, ver por exemplo Mexia et al. (2011) e Nunes et al. (2014),

$$\overline{\overline{F}}_j(z) = \sum_{n=n^\bullet}^{\infty} \ddot{p}_{n,n^\bullet}\overline{F}(z|g_j, g(n)), j = 1, ..., \tau.$$

#### 3.2.3.2 Uma aplicação a dados do cancro

Nesta subseção vamos aplicar a nossa abordagem a pacientes com cancro no Brasil. Os dados são referentes à cidade de São Paulo, 2010, e dizem respeito à idade de deteção da doença, ver Nunes et al. (2013).

Consideramos dois fatores, *Tipo de Cancro* e *Género*. O primeiro fator tem quatro níveis: *Amígdala*; *Cavidade nasal e ouvido médio*; *Timo* e *Coração, Mediasto e Pleura*. O segundo fator tem obviamente dois níveis: *Masculino* e *Feminino*. Daqui resultam $m = 4\times2 = 8$ diferentes tratamentos. As tabelas do Anexo 2 mostram as frequências desses quatro tipos de cancro, agrupados por idade e género. A Tabela 3.6 ilustra o número de pacientes por género e tipo de cancro.

Tabela 3.6: Número de pacientes por tipo de cancro e género.

| | | Género (segundo fator) | |
|---|---|---|---|
| | | Masculino | Feminino |
| Tipos de cancro (primeiro fator) | Amígdala | 51 | 22 |
| | Cavidade nasal e ouvido médio | 13 | 16 |
| | Timo | 7 | 9 |
| | Coração, mediasto e pleura | 18 | 13 |

Iremos então testar as hipóteses

$$H_{0,j} : \boldsymbol{A}_j\boldsymbol{\mu} = 0, j = 1, 2, 3.$$

Como vimos anteriormente, dado $\ddot{N}_i = n_i, i = 1, \ldots, m$, quando $H_{0,j}$ se verifica a distribuição condicional de

$$\Im_j = \frac{S_j}{S}, j = 1, 2, 3,$$

é uma distribuição $\overline{F}$ central com $g_j = car(\boldsymbol{A_j})$, $i = 1, 2, 3$, e $g(n) = n - 8$ graus de liberdade,

$$\overline{F}(.|g_j, n - 8).$$

Quanto à distribuição não condicional será dada por

$$\overline{\overline{F}}_j(z) = \sum_{n=n^\bullet}^{\infty} \ddot{p}_{n,n^\bullet} \overline{F}(z|g_j, n - 8),\ j = 1, 2, 3,$$

com $\ddot{p}_{n,n^\bullet} = \ddot{p}_{n,9} \frac{1-\ddot{p}_8}{1-\sum_{u=8}^{n^\bullet-1} \ddot{p}_u},\ n = n^\bullet, \ldots.$

Devido às propriedades de monotonia da distribuição $\overline{F}$, apresentadas na subseção 2.2.5.1, quando $n < n^\circ$, tem-se

$$\overline{F}(z|g_j, n - 8) < \overline{F}(z|g_j, n^\circ - 8),$$

então

$$\overline{F}(z|g_j, n^\bullet - 8) \leq \overline{\overline{F}}_j(z) \leq 1,$$

o que significa que $\overline{F}(z|g_j, n^\bullet - 8)$ nos dá um limite inferior para $\overline{\overline{F}}_j(z)$, $j = 1, 2, 3$. Assim, a partir de $\overline{F}(z|g_j, n^\bullet - 8)$ podemos obter limites superiores para os quantis de $\overline{\overline{F}}_j(z)$, $j = 1, 2, 3$. Se usarmos esses limites superiores como valores críticos teremos testes com tamanhos que não irão exceder os valores teóricos.

Ao usarmos esta abordagem, considerando os limites superiores, é necessário tomar atenção a algumas "questões" :

- Se o valor observado da estatística exceder o limite superior, também excede o valor crítico real (obtido quando se considera o tamanho das amostras como aleatório) e, neste caso, rejeita-se a hipótese nula;

- Para os casos em que o valor da estatística é menor do que o limite superior, devemos calcular os valores críticos reais (considerando o tamanho das amostras como aleatório) ou calcular o valor mínimo de $n^\bullet$ que leva à rejeição da hipótese nula;

- É de esperar que os quantis obtidos considerando as dimensões das amostras como aleatórias, sejam superiores aos quantis clássicos (obtidos ao se usar a abordagem condicional, com dimensões fixas para as amostras), uma vez que na expressão da distribuição é inserida uma nova fonte de variação. Assim, quando não se rejeita a hipótese nula usando os quantis clássicos é de se esperar que se tome a mesma decisão ao se utilizarem os quantis considerando as amostras com dimensão aleatória e, consequentemente, a abordagem usando os limites superiores.

Em seguida passamos aos cálculos usando os dados da Tabela 3.6 e das Tabelas do Anexo 2. Quando se tem mais do que um fator usualmente a análise começa com o teste à interação entre os fatores. Se a interação não é significativa, em seguida realizam-se os testes aos efeitos principais dos fatores. Caso a interação seja significativa, a inferência é realizada para cada nível de um dos fatores. Nomeadamente, se um dos fatores tiver apenas dois níveis, como acontece neste caso, podem ser obtidos intervalos de confiança para a diferença das médias, ver Nunes et al. (2014). Iremos seguir esta abordagem testando primeiro a interação entre os dois fatores. No entanto, seja qual for o resultado deste teste iremos também testar os efeitos principais dos fatores uma vez que pretendemos mostrar como estes testes podem ser realizados considerando o descondicionamento. Assim iremos ordenar os índices das hipóteses $H_{0,j}, j = 1, 2, 3,$ e das matrizes $\boldsymbol{A_j}, j = 1, 2, 3,$ da seguinte forma:

- Índice $1$- interação;
- Índice $2$- primeiro fator;
- Índice $3$- segundo fator.

Neste caso obtiveram-se as médias amostrais, componentes do vetor $\boldsymbol{y_\bullet}$,

- $y_{1,\bullet} = 47.0000;\ y_{2,\bullet} = 33.1364;$
- $y_{3\bullet} = 55.0769;\ y_{4,\bullet} = 63.8750;$
- $y_{5\bullet} = 55.5714;\ y_{6,\bullet} = 55.8889;$
- $y_{7,\bullet} = 53.9444;\ y_{8,\bullet} = 63.5385.$

A Figura 3.2 parece indicar a existência de interação entre os dois fatores, uma vez que as linhas não são paralelas. No entanto esta análise não é completamente conclusiva pois na verdade não passa de uma análise meramente descritiva já que as médias consideradas são as médias amostrais.

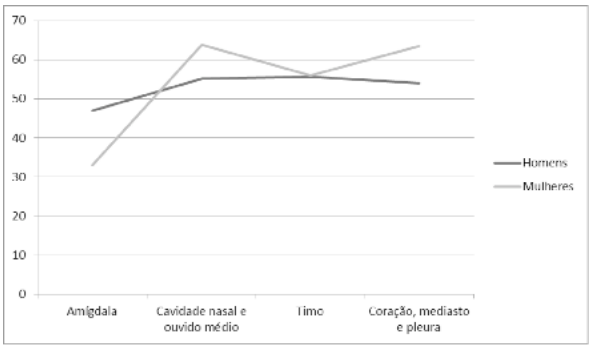


Figura 3.2: Interação entre os dois fatores.

Assim, construímos intervalos de confiança a $(1 - \alpha)\%$ para a diferença das médias entre os dois géneros, para cada um dos tipos de cancro. Esses intervalos são apresentados na Tabela 3.7. Como podemos constatar, à exceção da amígdala, para $1 - \alpha = 0.90$ e $0.95$, a origem está

contida em todos os intervalos. Os resultados podem ser igualmente observados pela análise da Figura 3.3, considerando o grau de confiança de $95\%$.

Tabela 3.7: Intervalos de confiança para a diferença das médias

| Valores de $1-\alpha$ | 0.90 | 0.95 | 0.99 |
|---|---|---|---|
| Amígdala | [3.2394; 24.4878] | [1.1454; 26.5673] | [-3.0096; 30.7368] |
| Cavidade nasal e ouvido médio | [-22.1056; 4.5094] | [-24.8287; 7.2325] | [-30.4450; 12.8487] |
| Timo | [-22.1560; 21.5210] | [-26.9107; 26.2757] | [-37.2273; 36.5923] |
| Coração, mediastino e pleura | [-22.0300; 2.8418] | [-24.5632; 5.3750] | [-29.7681; 10.5799] |

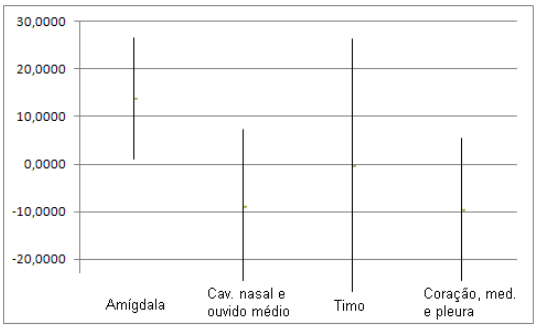


Figura 3.3: Intervalo de confiança a 95% para a diferenças das médias entre os dois géneros.

Começemos então pela **interação** entre os dois fatores. Neste caso

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

logo $g_1 = car(\boldsymbol{A_1}) = 3$. Obtemos

$$\left(\boldsymbol{A_1} D\left(\frac{1}{51}, \frac{1}{22}, \frac{1}{13}, \frac{1}{16}, \frac{1}{7}, \frac{1}{9}, \frac{1}{18}, \frac{1}{13}\right) \boldsymbol{A_1}'\right)^{-1} = \begin{bmatrix} 0.0739 & -0.0245 & 0.0227 \\ -0.0245 & 0.0739 & -0.0059 \\ 0.0227 & -0.0059 & 0.0739 \end{bmatrix},$$

e

$$\boldsymbol{A_1 y_\bullet} = \begin{bmatrix} 11.2919 \\ 5.2952 \\ -4.7324 \end{bmatrix}.$$

Assim, tem-se para o numerador da estatística $\Im_1$

$$S_1 = (\boldsymbol{A_1 y_\bullet})' \left(\boldsymbol{A_1} D\left(\frac{1}{51}, \frac{1}{22}, \frac{1}{13}, \frac{1}{16}, \frac{1}{7}, \frac{1}{9}, \frac{1}{18}, \frac{1}{13}\right) \boldsymbol{A_1}'\right)^{-1} (\boldsymbol{A_1 y_\bullet}) = 4033.3460.$$

Para o denominador da estatística, obtemos

$$\begin{aligned} S &= \sum_{k=1}^{51}(y_{1,k}-y_{1,\bullet})^2+\sum_{k=1}^{22}(y_{2,k}-y_{2,\bullet})^2+\sum_{k=1}^{13}(y_{3,k}-y_{3,\bullet})^2 \\ &+ \sum_{k=1}^{16}(y_{4,k}-y_{4\bullet})^2+\sum_{k=1}^{7}(y_{5,k}-y_{5,\bullet})^2+\sum_{k=1}^{9}(y_{6,k}-y_{6,\bullet})^2 \\ &+ \sum_{k=1}^{18}(y_{7,k}-y_{7,\bullet})^2+\sum_{k=1}^{13}(y_{8,k}-y_{8,\bullet})^2 = 76368.0424. \end{aligned}$$

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

$$\Im_{1,Obs} = \frac{4033.3460}{76368.0424} = 0.0528.$$

Se considerarmos a distribuição condicional de $\Im_1$, que corresponde a $\overline{F}(z|3,141)$ uma vez que $n-8 = 149-8 = 141$, obtemos os quantis, $z_{1-\alpha}$, representados na Tabela 3.8. Neste caso vamos rejeitar $H_{0,1}$ para $\alpha = 0.10$, uma vez que $\Im_{1,Obs} > z_{1-\alpha}$, e não rejeitar para $\alpha = 0.05$ e $0.01$.

Suponhamos agora que $n^\bullet \geq 52$, digamos que $n^\bullet = 52$, o que corresponde a menos de dez observações por tratamento. Para a distribuição não condicional da estatística de teste, temos os limites superiores para os quantis, $z^u_{1-\alpha}$, representados também na Tabela 3.8. Como $\Im_{1,Obs} < z^u_{1-\alpha}$, significa que estes valores nos podem levar a não rejeitar $H_{0,1}$, considerando os níveis usuais de significância. Podemos portanto tomar uma decisão contrária à tomada quando usámos a abordagem condicional para $\alpha = 0.1$.

Tabela 3.8: Os quantis da distribuição condicional e limites superiores para os quantis de $\Im_1$ e $\Im_2$

| **Valores de** $\alpha$ | 0.10 | 0.05 | 0.01 |
|---|---|---|---|
| $z_{1-\alpha}$ | 0.0452 | 0.0568 | 0.0835 |
| $z^u_{1-\alpha}$ | 0.1509 | 0.1920 | 0.2905 |

Vamos então calcular a dimensão mínima que leva à rejeição de $H_{0,1}$. Assumindo que os valores da estatística de teste permanecem inalterados, teremos que ter o valor mínimo de $n^\bullet$ apresentado na Tabela 3.9 por forma a rejeitar a hipótese nula.

Tabela 3.9: Valor minimo de $n^\bullet$ que leva à rejeição da hipótese $H_{0,1}$

| **Valores de** $\alpha$ | 0.10 | 0.05 | 0.01 |
|---|---|---|---|
| $n^\bullet$ | 130 | 160 | 228 |

Uma vez que para valores maiores de $n^\bullet$ vamos obter valores menores para os quantis, podemos concluir que $\Im_{1,Obs} > z^u_{1-\alpha}$ para todo $n^\bullet \geq 228$ e neste caso rejeitar a hipótese $H_{0,1}$ para os níveis usuais de significância. O valor $228$ como dimensão mínima global não é demasiado elevado, atendendo a que temos 8 tratamentos, podendo-se portanto concluir que nestas condições a interação entre os dois fatores é significativa.

Agora, para o **primeiro fator**, temos

$$\boldsymbol{A}_2 = \begin{bmatrix} -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \\ -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \\ \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \end{bmatrix},$$

logo $g_2 = car(\boldsymbol{A}_2) = 3$. Assim

$$\left(\boldsymbol{A}_2 D\left(\frac{1}{51}, \frac{1}{22}, \frac{1}{13}, \frac{1}{16}, \frac{1}{7}, \frac{1}{9}, \frac{1}{18}, \frac{1}{13}\right)\boldsymbol{A}_2'\right)^{-1} = \begin{bmatrix} 0.0739 & -0.0245 & 0.0227 \\ -0.0245 & 0.0739 & -0.0059 \\ 0.0227 & -0.0059 & 0.0739 \end{bmatrix},$$

e

$$\boldsymbol{A}_2\boldsymbol{y}_\bullet = \begin{bmatrix} 15.8527 \\ 10.5553 \\ -11.5941 \end{bmatrix}.$$

Obtém-se para o numerador da estatística

$$S_2 = (\boldsymbol{A}_2\boldsymbol{y}_\bullet)'\left(\boldsymbol{A}_2 D\left(\frac{1}{51}, \frac{1}{22}, \frac{1}{13}, \frac{1}{16}, \frac{1}{7}, \frac{1}{9}, \frac{1}{18}, \frac{1}{13}\right)\boldsymbol{A}_2'\right)^{-1}(\boldsymbol{A}_2\boldsymbol{y}_\bullet) = 11463.1200.$$

Atendendo a que $S = 76368.0424$, o valor observado da estatística, $\Im_{2,Obs}$, será dado por

$$\Im_{2,Obs} = \frac{11463.1200}{76368.0424} = 0.1501.$$

Os quantis, $z_{1-\alpha}$, da distribuição condicional de $\Im_2$, que corresponde a $\overline{F}(z|3, 141)$, são apresentados na Tabela 3.8. Uma vez que $\Im_{2,Obs} > z_{1-\alpha}$, podemos concluir que se rejeita $H_{0,2}$ para os níveis usuais de significância.

Vamos considerar mais uma vez que $n^\bullet = 52$. A Tabela 3.8 mostra os limites superiores para os quantis, $z^u_{1-\alpha}$, para a probabilidade $1-\alpha$ da distribuição não condicional de $\Im_2$. Como $\Im_{2,Obs} < z^u_{1-\alpha}$ estes resultados podem levar-nos a tomar uma decisão contrária à tomada considerando a distribuição condicional da estatística. Assumindo que os valores da estatística de teste permanecem inalterados devemos considerar o tamanho mínimo global para as amostras apresentado na Tabela 3.10 por forma a rejeitar a hipótese.

Tabela 3.10: Valor mínimo de $n^\bullet$ que leva à rejeição da hipótese $H_{0,2}$

| **Valores de** $\alpha$ | 0.10 | 0.05 | 0.01 |
|---|---|---|---|
| $n^\bullet$ | 53 | 64 | 89 |

Neste caso, por exemplo para $\alpha = 0.05$, teremos que ter pelo menos $64$ observações para que a decisão seja a mesma considerando as duas abordagens. Podemos concluir que para $n^\bullet \geq 89$ rejeitamos $H_{0,2}$, considerando os níveis usuais de significância, o que significa que o primeiro fator será significativo.

Para o **segundo fator** temos

$$\boldsymbol{A}_3 = \begin{bmatrix} -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \end{bmatrix},$$

logo $g_3 = car(\boldsymbol{A}_3) = 1$. Assim, obtemos

$$\left(\boldsymbol{A}_3 D\left(\frac{1}{51}, \frac{1}{22}, \frac{1}{13}, \frac{1}{16}, \frac{1}{7}, \frac{1}{9}, \frac{1}{18}, \frac{1}{13}\right)\boldsymbol{A}_3'\right)^{-1} = \ 0.0739 \ ,$$

e

$$\boldsymbol{A}_3\boldsymbol{y}_\bullet = \ 1.7139 \ ,$$

e consequentemente

$$S_3 = (\boldsymbol{A}_3\boldsymbol{y}_\bullet)'\left(\boldsymbol{A_3} D\left(\frac{1}{51}, \frac{1}{22}, \frac{1}{13}, \frac{1}{16}, \frac{1}{7}, \frac{1}{9}, \frac{1}{18}, \frac{1}{13}\right)\boldsymbol{A}_3'\right)^{-1}(\boldsymbol{A}_3\boldsymbol{y}_\bullet) = 39.7388$$

para o numerador da estatística $\Im_3$.

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

$$\Im_{3,Obs} = \frac{39.7388}{76368.0424} = 0.0005.$$

Considerando a distribuição condicional de $\Im_3$, que corresponde a $\overline{F}(z|1, 141)$, obtemos os quantis, $z_{1-\alpha}$, apresentados na Tabela 3.11. Como $\Im_{3,Obs} < z_{1-\alpha}$, podemos concluir que não rejeitamos $H_{0,3}$ para os níveis usuais de significância.

Tabela 3.11: Os quantis da distribuição condicional e limites superiores para os quantis de $\Im_3$

| **Valores de** $\alpha$ | 0.10 | 0.05 | 0.01 |
|---|---|---|---|
| $z_{1-\alpha}$ | 0.0194 | 0.0277 | 0.0484 |
| $z^u_{1-\alpha}$ | 0.0642 | 0.0923 | 0.1647 |

Para a distribuição não condicional da estatística para o segundo fator, considerando-se $n^\bullet = 52$, temos os limites superiores para os quantis apresentados também na Tabela 3.11. Como era esperado estes resultados levam-nos a tomar a mesma decisão que tomámos considerando a abordagem condicional, isto é, não rejeitar $H_{0,3}$. Poderíamos pensar que não temos uma dimensão suficientemente grande para as amostras por forma a detectar a significância do segundo fator. Assumindo que os valores da estatística de teste permanecem inalterados seria necessário ter o tamanho mínimo das amostras apresentados na Tabela 3.12 por forma a rejeitar $H_{0,3}$.

Tabela 3.12: Valor minimo de $n^\bullet$ que leva à rejeição da hipótese $H_{0,3}$

| **Valores de** $\alpha$ | 0.10 | 0.05 | 0.01 |
|---|---|---|---|
| $n^\bullet$ | 5421 | 7694 | 13282 |

Como estes valores são demasiado elevados somos levados a concluir que o segundo fator não é significativo, isto é, que não existe diferença significativa entre os géneros quanto à idade de deteção da doença. Se confrontarmos este resultado com os intervalos de confiança apresentados na Tabela 3.7 tiraríamos a mesma conclusão. É importante salientar que o fato dos intervalos de confiança para a amígdala, considerando $\alpha = 0.1$ e $\alpha = 0.05$, não conterem o zero é justificado pela existência de interação entre os fatores.

## 3.3 Modelos de efeitos aleatórios

É muito comum existir um grande número de diferentes patologias, isto na área de investigação médica, ou um grande número de diferentes castas de videiras, isto na área de agricultura. Nestes casos é usual optar-se por se selecionar aleatoriamente apenas alguns níveis do fator por forma a realizar o estudo estatístico, considerando portanto o fator como aleatório.

Nesta seção abordaremos a ANOVA de efeitos aleatórios com amostras de dimensões aleatórias.

Consideraremos apenas um fator com $m$ níveis. Vamos assumir que:

- as dimensões das amostras para os $m$ níveis não são conhecidas à partida;
- $N_1, ..., N_m$ são variáveis aleatórias que representam as dimensões das amostras;
- $n_1, ..., n_m$ são as realizações das variáveis aleatórias $N_1, ..., N_m$;
- as variáveis independentes $N_1, ..., N_m$ seguem distribuições de Poisson com parâmetros $\lambda_1, ..., \lambda_m, N_i \sim P(\lambda_i), i = 1, ..., m$;
- $n = \sum_{i=1}^{m} n_i$ é a realização da variável aleatória $N = \sum_{i=1}^{m} N_i$;
- $N$ segue uma distribuição de Poisson com parâmetro $\lambda = \sum_{i=1}^{m} \lambda_i, N \sim P(\lambda)$.

Como vimos no Capítulo 2, seção 2.3.2, o modelo com um fator de efeitos aleatórios pode ser escrito na seguinte forma:

$$Y_{i,j} = \mu + \alpha_i + \varepsilon_{i,j},\ j = 1, \ldots, n_i, i = 1, \ldots, m$$

com $\mu$ fixo e desconhecido e $\alpha_i$ e $\varepsilon_{i,j},\ j = 1, \ldots, n_i, i = 1, \ldots, m$, aleatórios. Vamos assumir que $\alpha_i$ e $\varepsilon_{i,j},\ j = 1, \ldots, n_i, i = 1, \ldots, m$, são normais, independentes, com valores médios nulos e variâncias $\sigma_\alpha^2$ e $\sigma^2$, respectivamente, $\alpha_i \sim \mathcal{N}(0, \sigma_\alpha^2)$ e $\varepsilon_{i,j} \sim \mathcal{N}(0, \sigma^2),\ j = 1, \ldots, n_i, i = 1, \ldots, m$.

Este modelo pode ser escrito em notação matricial da seguinte forma

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{D}(\mathbf{1}_{n_i}, \ldots, \mathbf{1}_{n_m})\boldsymbol{\alpha} + \boldsymbol{\varepsilon},$$

onde $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)^{'}$, $\boldsymbol{D}(\mathbf{1}_{n_i}, \ldots, \mathbf{1}_{n_m})$ representa uma matriz diagonal por blocos, com blocos principais $\mathbf{1}_{n_i}, \ldots, \mathbf{1}_{n_m}$, $\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \sigma_\alpha^2 \boldsymbol{I}_m)$ e $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$.

Quando $N_i = n_i, i = 1, \ldots, m$, o vetor $\boldsymbol{Y}$, será condicionalmente normal com vetor médio $\boldsymbol{\mu} = \mu \mathbf{1}_m$ e matriz de covariância dada por, ver por exemplo Khuri et al. (1998) e Searle et al. (1992),

$$\not{\Sigma} = \sigma_\alpha^2 D\left(\boldsymbol{J}_{n_1}, ..., \boldsymbol{J}_{n_m}\right) + \sigma^2 \boldsymbol{I}_n,$$

onde $\boldsymbol{J}_m = \mathbf{1}_m \mathbf{1}'_m$,

$$\underset{(N=n)}{\boldsymbol{Y}} \sim \mathcal{N}\left(\boldsymbol{\mu}, \sigma^2_\alpha D\left(\boldsymbol{J}_{n_1}, ..., \boldsymbol{J}_{n_m}\right) + \sigma^2 \boldsymbol{I}_n\right).$$

Neste caso estamos interessados em testar as hipóteses

$$H_{0,R} : \sigma^2_\alpha = 0 \; vs \; H_{1,R} : \sigma^2_\alpha > 0.$$

### 3.3.1 Estatística de teste e suas distribuições

Nesta seção apresentamos a estatística de teste e as suas distribuições condicional e não condicional. Quanto à expressão de estatística de teste optámos por considerar a obtida em Nunes et al. (2012b). Mais uma vez assumiremos que temos uma dimensão global mínima para as amostras pelo que iremos considerar as variáveis aleatórias $\ddot{N}_i, i = 1, \ldots, m$, que, como referido, correspondem às variáveis $N_i, i = 1, \ldots, m$, em que $N_i \geq 1, \; i = 1, \ldots, m$.

Quando $\ddot{N}_i = n_i$, $i = 1, ..., m$, temos as médias $Y_{i,\bullet}$, $i = 1, ..., m$, para as amostras

$$Y_{i,1}, ..., Y_{i,n_i}, \; i = 1, ..., m.$$

A soma das somas dos quadrados do erro será dada por

$$S = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (Y_{i,j} - Y_{i,\bullet})^2 = \sum_{i=1}^{m} \sum_{j=1}^{n_i} Y^2_{i,j} - \sum_{i=1}^{m} \frac{T^2_i}{n_i},$$

com $T_i = \sum_{j=1}^{n_i} Y_{i,j}$, $i = 1, ..., m$, $S$ será, quando $\ddot{N}_i = n_i$, $i = 1, ..., m$, o produto por $\sigma^2$ de um qui-quadrado central com

$$g(n) = n - m$$

graus de liberdade, $S \sim \sigma^2 \chi^2_{g(n)}$. Como $g(n)$ é uma realização de $g(\ddot{N})$, temos, como nos casos anteriores, graus de liberdade aleatórios para os erros. Como podemos ver em Searle et al. (1992), $S$ pode ser escrito usando a notação matricial, tendo-se

$$S = \boldsymbol{Y}' \left(\boldsymbol{I}_n - D\left(\frac{\boldsymbol{J}_{n_1}}{n_1}, ..., \frac{\boldsymbol{J}_{n_m}}{n_m}\right)\right) \boldsymbol{Y}.$$

Quando $\ddot{N}_i = n_i$, a média das amostras

$$Y_{i,\bullet} = \mu + \alpha_i + \varepsilon_{i,\bullet}, \; i = 1, ..., m,$$

têm valores médios $\mu$ e variâncias $\sigma^2_\alpha + \frac{\sigma^2}{n_i}$, $i = 1, ..., m$. Portanto, quando a hipótese $H_{0,R}$ se verifica, tem-se

$$Y_{i,\bullet} \sim \mathcal{N}\left(\mu, \tfrac{\sigma^2}{n_i}\right), \; , \; i = 1, ..., m,$$

e teremos a média geral

$$Y_{\bullet,\bullet} = \tfrac{1}{n} \textstyle\sum_{i=1}^{m} n_i Y_{i,\bullet} \sim \mathcal{N}\left(\mu, \tfrac{\sigma^2}{n}\right).$$

Vamos considerar

$$\boldsymbol{Z} = \boldsymbol{Y}_{\bullet} - \boldsymbol{Y}_{\bullet,\bullet} = \boldsymbol{B}\boldsymbol{Y}_{\bullet},$$

onde $\boldsymbol{Y}_{\bullet} = (Y_{1,\bullet}, ..., Y_{m,\bullet})'$, $\boldsymbol{Y}_{\bullet,\bullet} = (Y_{\bullet,\bullet}, ..., Y_{\bullet,\bullet})'$ e

$$\boldsymbol{B} = \boldsymbol{I}_m - \begin{bmatrix} \frac{n_1}{n} & \dots & \frac{n_m}{n} \\ \vdots & \ddots & \vdots \\ \frac{n_1}{n} & \dots & \frac{n_r}{n} \end{bmatrix} = \begin{bmatrix} \frac{n-n_1}{n} & \dots & \frac{-n_m}{n} \\ \vdots & \ddots & \vdots \\ \frac{-n_1}{n} & \dots & \frac{n-n_m}{n} \end{bmatrix} = \boldsymbol{I}_m - \frac{1}{n}\mathbf{1}_m\boldsymbol{n}'.$$

Quando $\ddot{N}_i = n_i$, $i = 1, ..., m$, $\boldsymbol{Z}$ seguirá uma distribuição normal com vetor médio nulo e matriz de covariância $\sigma^2\boldsymbol{V}$, $\boldsymbol{Z} \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{V})$, onde

$$\boldsymbol{V} = \boldsymbol{B}D\left(\frac{1}{n_1}, ..., \frac{1}{n_m}\right)\boldsymbol{B}' = D\left(\frac{1}{n_1}, ..., \frac{1}{n_m}\right) - \frac{1}{n}\boldsymbol{J}_m.$$

O numerador da estatística de teste será definido por, ver Nunes et al. (2012b),

$$S_{num} = \boldsymbol{Z}'\boldsymbol{V}^{-}\boldsymbol{Z},$$

onde $\boldsymbol{V}^{-}$ representa uma matriz inversa generalizada de $\boldsymbol{V}$. Quando $H_{0,R}$ se verifica

$$S_{num} = \boldsymbol{Z}'\boldsymbol{V}^{-}\boldsymbol{Z} \sim \sigma^2\chi^2_g,$$

com $g = car(\boldsymbol{V}) = m - 1$.

Assim, quando temos $\ddot{N} = n$ e $H_{0,R}$ se verifica, a distribuição condicional da estatística

$$\Im_R = \frac{S_{num}}{S} = \frac{\boldsymbol{Z}'\boldsymbol{V}^{-}\boldsymbol{Z}}{S}$$

será uma $\overline{F}$ central com $g$ e $g(n)$ graus de liberdade, $\overline{F}(z|g, g(n))$.

Vamos admitir, à semelhança do que considerámos na seção anterior, que existe uma dimensão mínima global, $n^{\bullet}$, para as amostras. Teremos então $\ddot{N} \geq n^{\bullet}$ e a distribuição não condicional de $\Im_R$, quando $H_{0,R}$ se verifica, será dada por

$$\begin{aligned} \overline{\overline{F}}(z) &= \sum_{n=n^{\bullet}}^{\infty} pr(\ddot{N} = n|\ddot{N} \geq n^{\bullet})\overline{F}(z|g, g(n)) \\ &= \sum_{n=n^{\bullet}}^{\infty} \ddot{p}_{n,n^{\bullet}}\overline{F}(z|g, g(n)), \end{aligned}$$

com

$$\ddot{p}_{u,n^{\bullet}} = \ddot{p}_{u,m+1}\frac{1 - \ddot{p}_m}{1 - \sum_{\ell=m}^{n^{\bullet}-1}\ddot{p}_{\ell}},\ u = n^{\bullet}, \dots.$$

definida pela expressão (3.1.1).

### 3.3.2 Uma aplicação a dados do cancro

Os dados usados nesta aplicação foram mais uma vez disponibilizados pelo INCA e reportam-se a pacientes com cancro no Brasil, mais precisamente, a pacientes de São Paulo, 2010, e dizem respeito à idade de deteção dos diferentes tipos de cancro. Recorreu-se ao método de amostragem aleatória simples, por forma a selecionar seis diferentes tipos de cancro, ver Capistrano et al. (2015). A Tabela 3.13 mostra os tipos de cancro obtidos através desta seleção e os respectivos números de pacientes. As tabelas do Anexo 3 apresentam as frequências desses seis tipos de cancro, agrupadas por idade.

Tabela 3.13: Diferentes tipos de cancro selecionados e número de pacientes

| **Tipos de cancro** | **Número de pacientes** |
|---|---|
| Corpo do estômago | 91 |
| Medula espinhal e outras partes S.N.C. | 42 |
| Melanoma malígno do tronco | 107 |
| Encéfalo | 93 |
| Cólon ascendente | 201 |
| Lobo superior, brônquios ou pulmão | 155 |

Quando $\ddot{N} = n$ e $H_{0,R}$ se verifica, a distribuição condicional de $\Im_R$ é uma distribuição $\overline{F}$ central com $g = 5$ e $g(n) = n-6$ graus de liberdade, $\overline{F}(z|5, n-6)$. Enquanto a distribuição não condicional da estatística será dada por

$$\overline{\overline{F}}(z) = \sum_{n=n_\bullet}^{\infty} \ddot{p}_{n,n^\bullet} \overline{F}(z|5, n-6).$$

Mais um vez recorrendo às propriedades de monotomia da distribuição $\overline{F}$, apresentados na subseção 2.2.5.1, com $n < n^\circ$, tem-se

$$\overline{F}(z|g, n-6) < \overline{F}(z|g, n^\circ - 6),$$

donde se conclui que

$$\overline{F}(z|5, n^\bullet - 6) \leq \overline{\overline{F}}(z) \leq 1,$$

então, a partir de $\overline{F}(z|5, n^\bullet - 6)$, podemos obter limites superiores para os quantis da distribuição não condicional da estatística, $\overline{\overline{F}}(z)$.

Neste caso obtemos

$$\begin{aligned} S &= \sum_{j=1}^{91}(y_{1,j} - y_{1,\bullet})^2 + \sum_{j=1}^{42}(y_{2,j} - y_{2,\bullet})^2 + \sum_{j=1}^{107}(y_{3,j} - y_{3,\bullet})^2 \\ &+ \sum_{j=1}^{93}(y_{4,j} - y_{4,\bullet})^2 + \sum_{j=1}^{201}(y_{5,j} - y_{5,\bullet})^2 + \sum_{j=1}^{155}(y_{6,j} - y_{6,\bullet})^2 \\ &= 209005.8252, \end{aligned}$$

com as médias das amostras, componentes do vetor $\boldsymbol{y}_\bullet$,

- $y_{1,\bullet} = 62.05;\ y_{2,\bullet} = 46.17;$

- $y_{3,\bullet} = 67.89;\ y_{4,\bullet} = 49.90;$

- $y_{5,\bullet} = 71.10;\ y_{6,\bullet} = 66.26.$

O numerador da estatística $\Im_R$ é definido por $S_{num} = \boldsymbol{Z}'\boldsymbol{V}^{-}\boldsymbol{Z}$ que, quando a hipótese se verifica, $S_{num} \sim \sigma^2\chi^2_5$. Neste caso temos

$$\boldsymbol{Z} = \boldsymbol{B}\boldsymbol{Y}_{\bullet} = \begin{bmatrix} -1.9654 \\ -17.8537 \\ 3.8675 \\ -14.9881 \\ 7.0842 \\ 2.2377 \end{bmatrix}$$

e

$$\boldsymbol{V}^{-} = \begin{bmatrix} 88.6759 & 0.9077 & -4.3802 & -2.5542 & -26.4091 & -13.5043 \\ 0.9077 & 43.3330 & 0.30705 & 0.8450 & -7.8147 & -2.8597 \\ -4.3802 & 0.3069 & 99.9126 & -4.6869 & -34.6912 & -18.6847 \\ -2.5542 & 0.8450 & -4.6869 & 90.2067 & -27.3849 & -14.1060 \\ -26.4091 & -7.8147 & -34.6912 & -27.3849 & 95.6745 & -66.0669 \\ -13.5044 & -2.8597 & -18.6847 & -14.1060 & -66.0669 & 115.7391 \end{bmatrix}.$$

Assim, obtemos $S_{num} = \boldsymbol{Z}'\boldsymbol{V}^{-}\boldsymbol{Z} = 47067.08$. Portanto, o valor observado da estatística, $\Im_{R,Obs}$, é dado por

$$\Im_{R,Obs} = \frac{47067.0800}{209005.8252} = 0.2252.$$

Se usarmos a distribuição condicional de $\Im_R$, que corresponde a $\overline{F}(z|5,683)$, visto que $g(n) = n - 6 = 689 - 6$, obteremos os quantis apresentados na Tabela 3.14.

Tabela 3.14: Os quantis da distribuição condicional e limites superiores para o quantis de $\Im_R$

| **Valores de** $\alpha$ | 0.10 | 0.05 | 0.01 |
|---|---|---|---|
| $z_{1-\alpha}$ | 0.0135 | 0.0163 | 0.0222 |
| $z^u_{1-\alpha}$ | 0.1848 | 0.2254 | 0.3192 |

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

Vamos agora supor que $n^{\bullet} \geq 59$, digamos que $n^{\bullet} = 59$, o que significa que temos aproximadamente 10 observações por tratamento. A Tabela 3.14 mostra os limites superiores para os quantis, $z^u_{1-\alpha}$, para a probabilidade $1-\alpha$ da distribuição não condicional $\overline{\overline{F}}(z)$. Estes resultados podem levar-nos a tomar uma decisão contrária à que tínhamos tomado ao usarmos a distribuição condicional das estatísticas, para $\alpha = 0.05$ e $\alpha = 0.01$.

Assumindo que os valores da estatística de teste permanecem inalterados, teremos que ter o tamanho total das amostras apresentado na Tabela 3.15, para que possamos rejeitar a hipótese nula.

Tabela 3.15: Valor minimo de $n^\bullet$ que leva à rejeição da hipótese $H_{0,R}$

| **Valores de** $\alpha$ | 0.10 | 0.05 | 0.01 |
|---|---|---|---|
| $n^\bullet$ | 51 | 60 | 79 |

Uma vez que, para valores maiores de $n^\bullet$ obteríamos valores menores para os quantis, ter-se-á $\Im_{R,Obs} > z^u_{1-\alpha}$ para todo $n^\bullet \geq 79$, o que significa, que neste caso, rejeitaríamos $H_{0,R}$ considerando os níveis usuais de significância. Nestas condições o tipo de cancro tem um efeito aleatório significativo e portanto a idade de deteção da doença pode diferir consoante o tipo de cancro.

## 3.4 Modelos mistos

Nas duas seções anteriores a ANOVA de efeitos fixos e de efeitos aleatórios foi abordada considerando as dimensões das amostras como aleatórias. Vamos agora estender os resultados aos modelos mistos.

A formulação de modelos mistos fica facilitada se usarmos as extensões $L$, ver Ferreira et al. (2009) e Moreira et al. (2009). Suponhamos que o vetor $\boldsymbol{Y}^o$ tem $m$ componentes as quais correspondem aos tratamentos de um modelo linear e que se tem a matriz

$$\boldsymbol{L} = D(\mathbf{1}_{n_1}, ..., \mathbf{1}_{n_m}) = \boldsymbol{L}(\boldsymbol{n}), \quad (3.4.18)$$

com $\boldsymbol{n} = (n_1, ..., n_m)'$. Então

$$\boldsymbol{Y} = \boldsymbol{L}\boldsymbol{Y}^o + \boldsymbol{\varepsilon}$$

corresponderá a um modelo com amostras $n_1, ..., n_m$, onde $\boldsymbol{\varepsilon}$ é o vetor dos erros com vetor médio nulo e matriz de covariância $\sigma^2\boldsymbol{I}_n$.

Se considerarmos

$$\boldsymbol{Y}^o = \sum_{i=0}^{w} \boldsymbol{X}_i\boldsymbol{\beta}_i, \quad (3.4.19)$$

com $\boldsymbol{\beta}_0$ fixo e $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_w$ aleatórios e independentes, com vetores médios nulos e matrizes de covariância $\sigma^2_1\boldsymbol{I}_{c_1}, ..., \sigma^2_w\boldsymbol{I}_{c_w}$, onde $c_i$, $i = 1, ..., w$, representam o número de componentes de $\boldsymbol{\beta}_i$, $i = 1, ..., w$, $\boldsymbol{Y}^o$ e $\boldsymbol{Y}$ serão modelos mistos, ver por exemplo Khuri et al. (1998).

Vamos supor mais uma vez que:

- as dimensões das amostras são desconhecidas à partida para os $m$ tratamentos, logo consideradas como variáveis aleatórias independentes $N_1, \ldots, N_m$;

- $n_1, \ldots, n_m$ são as realizações destas variáveis aleatórias;
- $\boldsymbol{n}$ é uma realização do vetor aleatório $\boldsymbol{N} = (N_1, \ldots, N_m)^{'}$;
- a recolha das observações correspondem a processos de contagem, o que nos leva a considerar mais uma vez que $N_1, \ldots, N_m$ seguem distribuições de Poisson com parâmetros $\lambda_1, \ldots, \lambda_m, N_i \sim P(\lambda_i),\ i = 1, \ldots, m.$

Assim, passaremos a considerar

$$\boldsymbol{L}(\boldsymbol{N}) = D(\mathbf{1}_{N_1}, ..., \mathbf{1}_{N_m}).$$

Para obtermos esta matriz teremos que descondicionar em ordem a $\boldsymbol{N}$ a matriz definida em (3.4.18).

Dado $\boldsymbol{N} = \boldsymbol{n}$, suponhamos que se pretende testar a hipótese

$$H_0 : \boldsymbol{\theta} = \mathbf{0},$$

e o teste é não distorcido qualquer que seja $\boldsymbol{n}$. Então, representando por $pr_{\boldsymbol{n},\boldsymbol{\theta}}(Rej)$ [$pr_{\boldsymbol{n},\mathbf{0}}(Rej)$] a probabilidade de rejeitar $H_0$, dados $\boldsymbol{n}$ e o parâmetro $\boldsymbol{\theta}$ [a probabilidade de rejeitar $H_0$, dado $\boldsymbol{n}$ e $\boldsymbol{\theta} = \mathbf{0}$], vamos obter

$$pr_{\boldsymbol{n},\boldsymbol{\theta}}(Rej) > pr_{\boldsymbol{n},\mathbf{0}}(Rej),$$

que descondicionando em ordem a $\boldsymbol{N}$ dará

$$pr_{\boldsymbol{\theta}}(Rej) > pr_{\mathbf{0}}(Rej),$$

logo o teste continua a ser não distorcido.

### 3.4.1 Estatística de teste e suas distribuições

Iremos assumir mais uma vez uma dimensão mínima global para as amostras, por isso trabalharemos com as variáveis $\ddot{N}_i, i = 1, \ldots, m$, e consequentemente com o vetor $\ddot{\boldsymbol{N}} = (\ddot{N}_1, \ldots, \ddot{N}_m)$.

Nesta seção iremos considerar as hipóteses e as estatísticas de teste apresentadas na seção 2.3.5, sobre extensões $L$.

Segundo a proposição 2.24, os testes para as hipóteses

$$H_{0,j,M} : \gamma_j = 0,\ j > z,$$

onde $\gamma_j,\ j = 1, \ldots, \ell$, correspondem às componentes da variância canónicas, são não distorcidos, qualquer que seja o $\boldsymbol{n}$. Então, descondicionando em ordem a $\ddot{\boldsymbol{N}}$, estes continuarão a ser não distorcidos.

Como vimos na seção 2.3.5, a soma dos quadrados do erro é dada por

$$S = \|Q(\boldsymbol{L})\boldsymbol{Y}\|^2 \sim \sigma^2\chi^2_{g(n)},$$

com $Q(\boldsymbol{L}) = \boldsymbol{I}_n - \boldsymbol{L}\boldsymbol{L}^+$ e $g(n) = n - m$, que é independente de

$$({}^{oo}\boldsymbol{Y}_j)'(\boldsymbol{B}_j^{-1})^{oo}\boldsymbol{Y}_j,$$

com $\boldsymbol{B}_j$ e ${}^{oo}\boldsymbol{Y}_j$ definidas em (2.3.17) e (2.3.18), respetivamente. Quando $H_{0,j,M}$ se verifica, $({}^{oo}\boldsymbol{Y}_j)'(\boldsymbol{B}_j^{-1})^{oo}\boldsymbol{Y}_j \sim \sigma^2\chi^2_{f_j}$, com $f_j = car(\boldsymbol{Q}_j), j > z$, em que $\boldsymbol{Q}_j$ são $MPO$ sobre $R(\boldsymbol{A}_j\boldsymbol{X}_0)^\perp$. Assim, considerando $\ddot{N} = n$ e que $H_{0,j,M}$ se verifica, a distribuição condicional da estatística, definida em (2.3.20),

$$\mathcal{T}_j = \frac{({}^{oo}\boldsymbol{Y}_j)'(\boldsymbol{B}_j^{-1})^{oo}\boldsymbol{Y}_j}{S},\ j > z,$$

será uma $\overline{F}$ central com $f_j$ e $g(n)$ graus de liberdade, $\overline{F} = (z|f_j, g(n))$. A distribuição não condicional de $\mathcal{T}_j$, $j > z$, considerando a dimensão mínima global para as amostras, $n^\bullet$, portanto que $\ddot{N} \geq n^\bullet$, será dada por, ver por exemplo Mexia et al. (2011) e Nunes et al. (2014),

$$\begin{aligned}\overline{\overline{F}}_j(z) &= \sum_{n=n^\bullet}^{\infty} pr(\ddot{N} = n|\ddot{N} \geq n^\bullet)F(z|f_j, g(n)) \\ &= \sum_{n=n^\bullet}^{\infty} \ddot{p}_{n,n^\bullet}F(z|f_j, g(n)),\ j > z,\end{aligned}$$

com $\ddot{p}_{n,n^\bullet} = \ddot{p}_{n,m+1}\frac{1-\ddot{p}_m}{1-\sum_{u=m}^{n^\bullet-1}\ddot{p}_u},\ n = n^\bullet, \ldots,$ como definido em (3.1.1).

### 3.4.2 Uma aplicação a dados do cancro

Nesta seção, continuamos a considerar dados de São Paulo, Brasil, de 2010 referentes à idade de deteção da doença. Vamos considerar um modelo misto com um fator de efeitos fixos e outro de efeitos aleatório. O fator de efeitos fixos será o *Género*, com dois níveis: *Masculino* e *Feminino*, e o fator de efeitos aleatórios será o *Tipo de Cancro*. Recorreu-se ao método de amostragem aleatória simples para selecionar três diferentes tipos de cancro. Isto leva-nos a $m = 2 \times 3 = 6$ diferentes tratamentos. A Tabela 3.16 ilustra o número de pacientes por género e pelos três tipos de cancro que foram selecionados.

Tabela 3.16: Tipos de cancro selecionados e número de pacientes

| | | Género (segundo fator) | |
|---|---|---|---|
| | | Masculino | Feminino |
| Tipos de cancro (primeiro fator) | Ossos e articulações dos membros | 34 | 25 |
| | Medula espinhal e outras partes S.N.C. | 19 | 23 |
| | Linfomas de células T cutâneas e periféricas | 34 | 27 |

As Tabelas de frequências dos três tipos de cancro, para homens e mulheres, são apresentadas no Anexo 4.

Neste caso, de (3.4.19), teremos o modelo

$$\boldsymbol{Y}^o = \boldsymbol{X}_0\boldsymbol{\beta}_0 + \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2, \quad (3.4.20)$$

onde $\boldsymbol{\beta}_0$ é fixo e $\boldsymbol{\beta}_1$ e $\boldsymbol{\beta}_2$ são independentes e aleatórios, correspondendo, respectivamente, ao fator de efeitos aleatórios e à interação entre os dois fatores. Para a obtenção do numerador e denominador da estatística de teste iremos utilizar os resultados apresentados na seção 2.3.5 sobre extensões $L$. Neste caso, teremos então

$$\begin{cases} \boldsymbol{X}_0 = \boldsymbol{I}_2 \otimes \mathbf{1}_3 \\ \boldsymbol{X}_1 = \mathbf{1}_2 \otimes \boldsymbol{I}_3 \\ \boldsymbol{X}_2 = \boldsymbol{I}_2 \otimes \boldsymbol{I}_3 \end{cases}.$$

As matrizes $\boldsymbol{A}_j$, $j = 1, 2$ serão dadas por

$$\begin{cases} \boldsymbol{A}_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{\sqrt{2}} \end{bmatrix} \\ \boldsymbol{A}_2 = \begin{bmatrix} -\frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & -\frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & -\frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{\sqrt{2}} \end{bmatrix} \end{cases}$$

e

$$\begin{cases} \boldsymbol{X}_1^0 = \boldsymbol{A}_1\boldsymbol{X}_0 = \frac{1}{\sqrt{2}}\mathbf{1}_2' \otimes \mathbf{1}_3 \\ \boldsymbol{X}_2^0 = \boldsymbol{A}_2\boldsymbol{X}_0 = \left[-\frac{1}{\sqrt{2}} \ \frac{1}{\sqrt{2}}\right] \otimes \mathbf{1}_3 \end{cases}.$$

As matrizes $\boldsymbol{Q}_j$, $j = 1, 2$, que são $MPO$ sobre $R(\boldsymbol{X}_j^0)^{\perp}$, $j = 1, 2$, serão dadas por

$$\begin{cases} \boldsymbol{Q}_1 = \boldsymbol{W}_1'\boldsymbol{W}_1 = \boldsymbol{I}_3 - \frac{1}{3}\boldsymbol{J}_3 \\ \boldsymbol{Q}_2 = \boldsymbol{W}_2'\boldsymbol{W}_2 = \boldsymbol{I}_3 - \frac{1}{3}\boldsymbol{J}_3 \end{cases}$$

e

$$\boldsymbol{W}_1 = \boldsymbol{W}_2 = \begin{bmatrix} -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \end{bmatrix}.$$

Além disso, $f_1 = car(\boldsymbol{Q}_1) = 3$, $f_2 = car(\boldsymbol{Q}_2) = 3$ e as $MPO$ sobre $R(\boldsymbol{X}_j^0)$, $j = 1, 2$, correspondem a

$$\begin{cases} \boldsymbol{P}_1 = \boldsymbol{X}_1^0(\boldsymbol{X}_1^0)^+ = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} \\ \boldsymbol{P}_2 = \boldsymbol{X}_2^0(\boldsymbol{X}_2^0)^+ = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} \end{cases}.$$

Como foi referido queremos testar as hipóteses

$$H_{0,j,M} : \gamma_j = 0,\ j = 1, 2,$$

que correspondem, respectivamente, às hipóteses de ausência de efeitos do segundo fator e da interação entre os dois fatores. Quando $\ddot{N} = n$ e $H_{0,j,M}$, $j = 1, 2$ se verifica, a distribuição condicional de

$$\mathcal{T}_j = \frac{(\boldsymbol{^{oo}Y}_j)'(\boldsymbol{B}_j^{-1})\boldsymbol{^{oo}Y}_j}{S},\ j = 1, 2$$

é uma $\overline{F}$ central com $f_j = rank(\boldsymbol{Q}_j) = 3$, $j = 1, 2$, e $g(n) = n - 6$ graus de liberdade, $\overline{F}(.|3, n-6)$.

Suponhamos que

$$\sum_{n=0}^{n^\bullet - 1} \ddot{p}_{n,n^\bullet} \simeq 0,$$

o que significa que, com grande probabilidade, se tem $\ddot{N} \geq n^\bullet$, sendo portanto $n^\bullet$ a dimensão mínima global para as amostras. A distribuição não condicional das estatísticas será então dada por

$$\overline{\overline{F}}_j(z) = \sum_{n=n^\bullet}^{\infty} \ddot{p}_{n,n^\bullet} F(z|3, n-6),\ j = 1, 2.$$

Pelas propriedades de monotonia da distribuição $\overline{F}$, apresentadas na subseção 2.2.5.1, tem-se

$$\overline{F}(z|3, n^\bullet - 6) \leq \overline{\overline{F_j}}(z) \leq 1,$$

e portanto, a partir de $\overline{F}(z|3, n^\bullet - 6)$, podemos obter limites superiores para os quantis da distribuição não condicional, $\overline{\overline{F_j}}(z)$. Apesar de realizarmos primeiro o teste à interação entre os dois fatores, tal como referido na seção 3.2.3, neste caso iremos manter os índices das matrizes $\boldsymbol{X}_j, j = 0, 1, 2$, definidas no modelo (3.4.20) e das hipóteses $H_{0,j,M}$, $j = 1, 2$. Assim

- Índice $0$- primeiro fator;
- Índice $1$- segundo fator;
- Índice $2$- interação.

Para a **interação**, teremos então

$$^{oo}\boldsymbol{Y}_2 = \boldsymbol{W}_2\boldsymbol{A}_2\boldsymbol{L}^+\boldsymbol{Y} = \begin{bmatrix} 10.0788 \\ -9.7891 \end{bmatrix},$$

onde

$$\boldsymbol{L}^+ = D\left(\frac{1}{34}\mathbf{1}'_{34}, \frac{1}{25}\mathbf{1}'_{25}, \frac{1}{19}\mathbf{1}'_{19}, \frac{1}{23}\mathbf{1}'_{23}, \frac{1}{34}\mathbf{1}'_{34}, \frac{1}{27}\mathbf{1}'_{27}\right)$$

e $\boldsymbol{L}^+\boldsymbol{Y}$, é o vetor das médias amostrais com componentes

- $27.5882$; • $42.8000$;
- $46.4737$; • $45.9130$;
- $57.7353$; • $47.1852$.

Obtemos

$$\boldsymbol{B}_2 = \boldsymbol{W}_2\boldsymbol{A}_2\boldsymbol{L}^+(\boldsymbol{L}^+)'\boldsymbol{A}'_2\boldsymbol{W}'_2 = \begin{bmatrix} 0.0406 & 0.0024 \\ 0.0024 & 0.0367 \end{bmatrix}$$

e para o numerador da estatística $\mathcal{T}_2$,

$$(^{oo}\boldsymbol{Y}_2)'(\boldsymbol{B}_2^{-1})^{oo}\boldsymbol{Y}_2 = 5453.8120.$$

Como previamente definido, quando $\ddot{N} = n$, $S \sim \sigma^2\chi^2_{n-6}$. Neste caso, obtemos

$$S = \|Q(\boldsymbol{L})\boldsymbol{Y}\|^2 = 87095.4899.$$

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

$$\mathcal{T}_{2,Obs} = \frac{5453.8120}{87095.4899} = 0.0626.$$

Se usarmos a distribuição condicional de $\mathcal{T}_2$, que corresponde a $\overline{F}(z|3, 156)$, uma vez que $n = 162$, vamos obter os quantis indicados na Tabela 3.17. Podemos concluir que se rejeita $H_{0,2,M}$ para $\alpha = 0.1$ e $\alpha = 0.05$ e não se rejeita para $\alpha = 0.01$.

Vamos supor que $n^\bullet \geq 24$, digamos que $n^\bullet = 24$, o que significa que temos quatro observações por tratamento.

Tabela 3.17: Os quantis da distribuição condicional e limites superiores para os quantis de $\mathcal{T}_1$ e $\mathcal{T}_2$

| **Valores de** $\alpha$ | 0.10 | 0.05 | 0.01 |
|---|---|---|---|
| $z_{1-\alpha}$ | 0.0408 | 0.0512 | 0.0752 |
| $z^u_{1-\alpha}$ | 0.4027 | 0.5267 | 0.8486 |

A Tabela 3.17 mostra os limites superiores dos quantis para a probabilidade $1-\alpha$, $z^u_{1-\alpha}$, da distribuição não condicional. Os resultados desta tabela podem levar-nos a tomar uma decisão

contrária à que tomámos ao usarmos a distribuição condicional da estatística para $\alpha = 0.1$ e $\alpha = 0.05$.

Assumindo que os valores da estatística de teste permanecem inalterados, teremos que ter o tamanho mínimo $n^\bullet$ apresentado na Tabela 3.18 por forma a garantirmos a rejeição da hipótese nula.

Tabela 3.18: Valor minimo de $n^\bullet$ que leva à rejeição da hipótese $H_{0,2,M}$

| **Valores de** $1-\alpha$ | 0.10 | 0.05 | 0.01 |
|---|---|---|---|
| $n^\bullet$ | 109 | 135 | 193 |

Uma vez que para valores maiores de $n^\bullet$ obteríamos valores menores para os quantis, temos $\mathcal{T}_{2,Obs} > z^u_{1-\alpha}$ para todo $n^\bullet \geq 193$. Neste caso, rejeitaríamos $H_{0,2,M}$ considerando os níveis usuais de significância, o que significa que a interação entre os fatores seria significativa. Pela Figura 3.4 podemos ver que não existe paralelismo entre as linhas para ambos os géneros, o que aponta para a existência de interação entre os dois fatores.

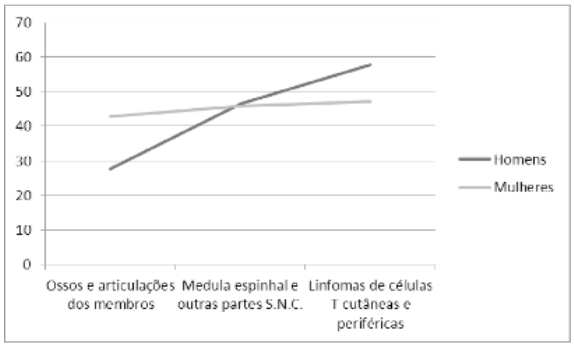


Figura 3.4: Interação entre os fatores.

Agora, para o **segundo fator**, teremos

$$ {}^{oo}\boldsymbol{Y}_1 = \boldsymbol{W}_1\boldsymbol{A}_1\boldsymbol{L}^+\boldsymbol{Y} = \begin{bmatrix} -8.8067 \\ -3.1276 \end{bmatrix} $$

e

$$ \boldsymbol{B}_1 = \boldsymbol{W}_1\boldsymbol{A}_1\boldsymbol{L}^+(\boldsymbol{L}^+)'\boldsymbol{A}_1'\boldsymbol{W}_1' = \begin{bmatrix} 0.0406 & 0.0024 \\ 0.0024 & 0.0367 \end{bmatrix}. $$

Assim, para o numerador da estatística $\mathcal{T}_1$, obtemos

$$ ({}^{oo}\boldsymbol{Y}_1)'(\boldsymbol{B}_1^{-1}){}^{oo}\boldsymbol{Y}_1 = 2093.8120. $$

Portanto, o valor da estatística, $\mathcal{T}_{1,Obs}$, é dado por

$$ \mathcal{T}_{1,Obs} = \frac{2093.8120}{87095.4899} = 0.0240. $$

Se usarmos a distribuição condicional de $\mathcal{T}_1$, que corresponde a $\overline{F}(z|3,156)$, obteremos os quantis indicados na Tabela 3.17.

Como $\mathcal{T}_{1,Obs} < z_{1-\alpha}$, podemos concluir que não se rejeita $H_{0,1,M}$ para os níveis usuais de significância.

A Tabela 3.17 também mostra os limites superiores para os quantis, $z^u_{1-\alpha}$, da distribuição não condicional, $\overline{\overline{F}}_1(z)$, considerando $n^\bullet = 24$. Concluímos, como era de se esperar, que também não se rejeita $H_{0,1,M}$ considerando a abordagem não condicional. Portanto o segundo fator não é significativo, o que significa que a idade de deteção da doença não depende do tipo de cancro.

## 3.5 Conclusões e discussão dos resultados obtidos

Pretendemos com este capítulo apresentar uma nova abordagem, que nos parece mais realista que a usual, quando as dimensões das amostras não são conhecidas à partida. Em situações como esta as referidas dimensões devem ser consideradas como realizações de variáveis aleatórias.

Em todo o capítulo supusemos que a ocorrência das observações correspondia a processos de contagem o que nos levou a assumir que as variáveis aleatórias seguiam distribuições de Poisson. Através das aplicações apresentadas concluímos que os valores críticos da nova abordagem podem exceder os da abordagem clássica. Assim, esta abordagem será mais robusta no sentido em que diminui a probabilidade de falsas rejeições.

# Capítulo 4

# Testes $F$ com amostras de dimensão aleatória. Falhas de observações

No capítulo anterior considerámos o caso em que a ocorrência das observações correspondia a processos de contagem. Assumimos que as variáveis aleatórias $N_1, ..., N_m$ seguiam distribuições de Poisson de parâmetros $\lambda_1, \ldots, \lambda_m$ e portanto as dimensões das amostras não tinham limites superiores. O caso em que este limite superior existe corresponde a uma situação diferente. Por exemplo, pode ter-se um limite superior comum conhecido, $r$, que não é sempre atingido, uma vez que podem ocorrer falhas.

Tal situação pode ocorrer, por exemplo, quando:

- se está a trabalhar com um determinado número de pacientes e existe uma probabilidade, que pode depender do tipo de doença, de alguns dos processos clínicos dos pacientes serem incompletos ou mesmo não existirem;

- se está a trabalhar com videiras e existe uma probabilidade, que pode depender do tratamento, da videira secar;

- se envia um determinado número de questionários, por forma a que o "projeto" de pesquisa seja válido, porém nem todos são enviados de volta.

A distribuição Binomial é a escolha adequada para estes casos. Considerando que existem $m$ diferentes tratamentos, vamos então supor que as variáveis aleatórias independentes, $N_1, ..., N_m$, seguem uma distribuição Binomial, com parâmetros:

- $r_1, ..., r_m$, os limites superiores para as dimensões das $m$ amostras, os quais nem sempre são atingidos uma vez que podem ocorrer falhas de observações;

- $1 - p$, onde $p$ representa a probabilidade da ocorrência de uma falha.

Colocamos $N_i \sim B(r_i, 1-p),\ i = 1, \ldots, m$ e devido à independência dos $N_i, i = 1, \ldots, m$, ter-se-á

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

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

À semelhança do que foi feito no capítulo anterior, onde assumimos a distribuição de Poisson para as dimensões das amostras, vamos agora alargar esta abordagem considerando que as dimensões das amostras seguem distribuições Binomiais. Serão considerados modelos de efeitos fixos e modelos mistos e serão apresentados algumas aplicações com dados do cancro no Brasil.

## 4.1 Distribuição Binomial Truncada

À semelhança do que foi apresentado no capítulo anterior, nesta seção vamos apresentar alguns resultados sobre a Binomial Truncada, que nos serão úteis na obtenção das distribuições não condicionais das estatísticas. Tal como no capítulo anterior, continuaremos a assumir dimensões mínimas para as amostras por forma a evitar casos altamente desequilibrados.

### 4.1.1 Dimensão mínima global para as amostras

Tal como acontece com a distribuição de Poisson, a forma mais comum para a distribuição Binomial Truncada é a omissão do valor zero como valor da variável, uma vez que necessitamos de ter pelo menos uma observação por tratamento, ver Johnson and Kotz (1969). Por forma a realizar inferência, assumiremos então que as seguintes condições são satisfeitas:

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

- $N > m.$

Como referido anteriormente, supondo que podem ocorrer falhas nas observações, vamos assumir que as variáveis $N_1, ..., N_m$ têm distribuição Binomial com parâmetros $r_1, ..., r_m$ e $1 - p$, $N_i \sim B(r_i, 1-p)$. Devido à independência de $N_1, ..., N_m$ tem-se $N = \sum_{i=1}^{m} \sim B(r, 1-p)$ em que $r = \sum_{i=1}^{m} r_i$.

Temos então

$$\begin{aligned} p_{u,i} &= pr(N_i = u | N_i \geq 1) = \frac{pr(N_i = u)}{pr(N_i \geq 1)} \\ &= \frac{pr(N_i = u)}{1 - pr(N_i = 0)} = \frac{\left( \begin{array}{c} r_i \\ u \end{array} \right) (1-p)^u p^{r_i - u}}{1 - p^{r_i}},\ u = 1, ..., r_i, i = 1, ..., m. \end{aligned}$$

Portanto, a função geradora de momentos de $N_i$, quando $N_i \geq 1$, $i = 1, ..., m$, será

$$\varphi_i(t) = \sum_{u=1}^{r_i} e^{tu} \frac{\binom{r_i}{u} (1-p)^u p^{r_i - u}}{1 - p^{r_i}}$$
$$= \frac{(p + (1-p)e^t)^{r_i} - p^{r_i}}{1 - p^{r_i}},\ i = 1, ..., m,$$

e sua função geradora de probabilidade

$$\psi_i(t) = \varphi_i(\ln t) = \frac{(p + (1-p)t)^{r_i} - p^{r_i}}{1 - p^{r_i}},\ i = 1, ..., m.$$

Vamos supor mais uma vez que $\ddot{N}_i$, $i = 1, ..., m$, correspondem às variáveis truncadas $N_i$, $i = 1, ..., m$, em que $N_i \geq 1$, e que

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

Obtém-se então a função geradora de probabilidade de $\ddot{N}$

$$\ddot{\psi}(t) = \prod_{i=1}^{m} \psi_i(t) = \prod_{i=1}^{m} \frac{(p + (1-p)t)^{r_i} - p^{r_i}}{1 - p^{r_i}}.$$

Portanto, sendo $\mathcal{P}_u^{(m)}$ a família de partições de $u_1 + ... + u_m = u$, de tal modo que $1 \leq u_i \leq r_i,\ i = 1, ..., m$ e $\boldsymbol{u} = (u_1, ..., u_m)'$, tem-se

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

Vamos agora considerar que

$$j_1 + ... + j_m = s;\ s = 1, ..., m-1,$$

logo $\boldsymbol{j} = (j_1, ..., j_m)'$ tem pelo menos uma componente nula. Com $\mathcal{P}_s^{(m)}$ a família de partições de $s$ com cardinal $m$, tem-se

$$\ddot{\psi}^{<s>}(0) = \sum_{\boldsymbol{j} \in \mathcal{P}_s^{(m)}} \frac{(\sum_{i=1}^{m} j_i)!}{\prod_{i=1}^{m} j_i!} \prod_{i=1}^{m} \psi_i^{<j_i>}(0),\ s = 1, ..., r.$$

Se $s < m$, qualquer que seja $\boldsymbol{j} \in \mathcal{P}_s^{(m)}$,

$$\prod_{i=1}^{m} \psi_i^{<j_i>}(0) = 0,$$

uma vez que $\boldsymbol{j}$ tem pelo menos uma componente nula e $\chi_i(0) = 0$, $i = 1, ..., m$, e consequentemente

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

Assim, visto que $\ddot{\psi}^{<s>}(0) = s!\ddot{p}_s$, obtém-se

$$\ddot{p}_s = pr(\ddot{N} = s) = \frac{1}{s!}\ddot{\psi}^{<s>}(0) = 0,\ s = 1, \ldots, m-1.$$

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

Vamos agora considerar que $s = m$. O único termo não nulo de $\ddot{\psi}^{<m>}(0)$ corresponde a $\boldsymbol{j} = \mathbf{1}_m$, uma vez que $j_1 + ... + j_m = m$, assim

$$\begin{aligned} \ddot{p}_m &= pr(\ddot{N} = m) = \frac{1}{m!}\ddot{\psi}^{<m>}(0) \\ &= \prod_{i=1}^{m} \psi_i^{<1>}(0) = \prod_{i=1}^{m} \frac{r_i(1-p)p^{r_i-1}}{1-p^{r_i}}. \end{aligned}$$

Além disso

$$\begin{aligned} pr(\ddot{N} > m) &= 1 - pr(\ddot{N} \leq m) = 1 - \ddot{p}_m \\ &= 1 - \prod_{i=1}^{m} \frac{r_i(1-p)p^{r_i-1}}{1-p^{r_i}}, \end{aligned}$$

vindo

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

Assumindo que se tem uma dimensão mínima global para as amostras, por exemplo $n^\bullet$, com $n^\bullet \geq m+1$, tem-se $\ddot{N} \geq n^\bullet$ e

$$\begin{aligned} & \ddot{p}_{u,n^\bullet} = pr(\ddot{N} = u | \ddot{N} \geq n^\bullet) = \frac{pr(\ddot{N} = u)}{pr(\ddot{N} \geq n^\bullet)} \\ = & \frac{\ddot{p}_u}{pr(\ddot{N} > m)} \frac{pr(\ddot{N} > m)}{pr(\ddot{N} \geq n^\bullet)} \\ = & \ddot{p}_{u,m+1} \frac{1-\ddot{p}_m}{1-\sum_{\ell=m}^{n^\bullet-1} \ddot{p}_\ell},\ u = n^\bullet, \ldots, r. \end{aligned} \tag{4.1.1}$$

### 4.1.2 Dimensão mínima para cada uma das amostras

Como vimos na subseção anterior, ao assumirmos que existem $m$ diferentes tratamentos, apenas podemos considerar $m$ parcelas para as partições. Iremos agora considerar que se tem uma dimensão mínima para cada uma das amostras, ver Nunes et al. (2015). Teremos então $N_i \geq n_i^\bullet, i = 1, \ldots, m$, o que equivale a ter-se $\boldsymbol{N} \geq \boldsymbol{n}^\bullet$, com $\boldsymbol{N} = (N_1, \ldots, N_m)'$ e $\boldsymbol{n}^\bullet = (n_1^\bullet, \ldots, n_m^\bullet)'$. Significa portanto que a dimensão mínima global será $n^\bullet = \sum_{i=1}^m n_i^\bullet$. Assim, ter-se-á a probabilidade

$$\begin{aligned} & p_{n,n_i^\bullet} = pr(N = n | \boldsymbol{N} \geq \boldsymbol{n}^\bullet) \\ = & \sum_{n_1=n_1^\bullet}^{n-\sum_{i=2}^m n_i^\bullet} \ldots \sum_{n_\ell=n_\ell^\bullet}^{n-(\sum_{i=1}^{\ell} n_i + \sum_{i=\ell+1}^{m} n_i^\bullet)} \ldots \sum_{n_m=n-\sum_{i=1}^{m-1} n_i}^{n-\sum_{i=1}^{m-1} n_i} pr(\boldsymbol{N} = \boldsymbol{n} | \boldsymbol{N} \geq \boldsymbol{n}^\bullet), \\ & n_i = n_i^\bullet, ..., \ i = 1, ..., m, \end{aligned} \tag{4.1.2}$$

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

$$\begin{aligned} & pr(\boldsymbol{N} = \boldsymbol{n} | \boldsymbol{N} \geq \boldsymbol{n}^\bullet) \\ = & \prod_{i=1}^m pr(N_i = n_i | N_i \geq n_i^\bullet) \\ = & \prod_{i=1}^m \frac{\binom{r_i}{n_i} (1-p)^{n_i} p^{r_i - n_i}}{\sum_{u_i=n_i^\bullet}^{r_i} pr(N_i = u_i)} \\ = & \prod_{i=1}^m \frac{\binom{r_i}{n_i} (1-p)^{n_i} p^{r_i - n_i}}{\sum_{u_i=n_i^\bullet}^{r_i} \binom{r_i}{u_i} (1-p)^{u_i} p^{r_i - u_i}}, \ n_i = n_i^\bullet, ..., r_i; \ i = 1, ..., m. \end{aligned} \tag{4.1.3}$$

## 4.2 Modelos de efeitos fixos

Nesta seção consideraremos a ANOVA de efeitos fixos com um e mais fatores assumindo que as dimensões das amostras seguem distribuições Binomiais.

### 4.2.1 Um fator com apenas um nível com dimensão aleatória

Nesta seção, vamos considerar que temos apenas um fator de efeitos fixos com $m$ níveis. Vamos supor que $r$ é o limite superior para a dimensão de cada uma das $m$ amostras. Supomos ainda que a dimensão das amostras é fixa para todos os níveis, à exceção do $m$-ésimo onde podem ocorrer falhas. Assim será mais correto considerar a dimensão do $m$-ésimo nível como aleatório.

Vamos então assumir que:

- $N$ é a variável aleatória que representa a dimensão do $m$-ésimo nível;
- $n$ é uma realização desta variável.

Então $N$ seguirá uma distribuição Binomial com parâmetros $r$ e $1-p$, $N \sim B(r, 1-p)$, onde $p$ corresponde à probabilidade da ocorrência de uma falha.

Estamos interessados em testar a hipótese

$$H_{0,F} : \mu_1 = ... = \mu_m,$$

que, como vimos, pode ser reescrita

$$H_{0,F} : \boldsymbol{A\mu} = \mathbf{0},$$

onde $\boldsymbol{\mu}$ é o vetor médio com componentes $\mu_1, ..., \mu_m$, e $\boldsymbol{A} = [\boldsymbol{I}_{m-1} | -\mathbf{1}_{m-1}]$.

Para realizar a inferência precisamos de ter pelo menos uma observação para o $m$-ésimo nível, o que significa que $N \geq 1$, tomando-se

$$\begin{aligned} p_n &= pr(N = n | N \geq 1) = \frac{pr(N = n)}{pr(N \geq 1)} \\ &= \frac{\left( \begin{array}{c} r \\ n \end{array} \right) (1-p)^n p^{r-n}}{1 - p^r}, \ n = 1, ..., r. \end{aligned} \tag{4.2.4}$$

#### 4.2.1.1 Estatística de teste e suas distribuições

Quando $N = n$ temos as amostras

$$Y_{i,1}, ..., Y_{i,r}, \ i = 1, ..., m-1,$$

e

$$Y_{m,1}, ..., Y_{m,n}$$

com médias $Y_{i,\bullet}$, $i = 1, ..., m$. Portanto, a soma das somas dos quadrados dos erros será dada por, ver por exemplo Khuri et al. (1998) e Searle et al. (1992),

$$S = \sum_{i=1}^{m-1} \sum_{j=1}^{r} (Y_{i,j} - Y_{i,\bullet})^2 + \sum_{j=1}^{n} (Y_{m,j} - Y_{m,\bullet})^2.$$

Se assumirmos que as observações são normais, independentes, com variância $\sigma^2$ e valores médios $\mu_i$, $i = 1, ..., m$, $S$ será, quando $N = n$, o produto de $\sigma^2$ por um qui-quadrado central com

$$g(n) = (m-1)(r-1) + n - 1$$

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

O vetor $\boldsymbol{Y}_\bullet$, com componentes $Y_{1,\bullet}, ..., Y_{m,\bullet}$, será condicionalmente independente de $S$ e, com $N = n$, condicionalmente normal com vetor médio $\boldsymbol{\mu}$ e matriz de covariância $\sigma^2 D(\frac{1}{r}, ..., \frac{1}{r}, \frac{1}{n})$,

$$\underset{(N=n)}{\boldsymbol{Y}_\bullet \sim} \mathcal{N}\left(\boldsymbol{\mu}, \sigma^2 D\left(\tfrac{1}{r}, ..., \tfrac{1}{r}, \tfrac{1}{n}\right)\right).$$

Além disso, quando $N = n$, a distribuição condicional de $\boldsymbol{AY}_\bullet$ será $\mathcal{N}\left(\boldsymbol{A\mu}, \sigma^2 \boldsymbol{A} D(\frac{1}{r}, ..., \frac{1}{r}, \frac{1}{n})\boldsymbol{A}'\right)$. Então, ver por exemplo Mexia (1990),

$$S_{num} = (\boldsymbol{AY}_\bullet)' \left(\boldsymbol{A} D\left(\tfrac{1}{r}, ..., \tfrac{1}{r}, \tfrac{1}{n}\right) \boldsymbol{A}'\right)^{-1} (\boldsymbol{AY}_\bullet) \underset{(N=n)}{\sim} \sigma^2 \chi^2_{g,\delta(n)},$$

com $g = m - 1$, e

$$\delta(n) = \frac{1}{\sigma^2} (\boldsymbol{A\mu})' \left(\boldsymbol{A} D\left(\frac{1}{r}, ..., \frac{1}{r}, \frac{1}{n}\right) \boldsymbol{A}'\right)^{-1} (\boldsymbol{A\mu}).$$

Portanto, a distribuição condicional da estatística de teste

$$\Im_F = \frac{S_{num}}{S}$$

será $\overline{F}(.|g, g(n), \delta(n))$. Quando $N = n$ e $H_{0,F}$ se verifica, $\delta(n) = 0$ e a distribuição condicional de $\Im_F$ será $\overline{F}(.|g, g(n))$.

Neste caso a distribuição não condicional de $\Im_F$ será dada por, ver por exemplo Mexia et al. (2011) e Nunes et al. (2012a),

$$\begin{aligned} \overline{\overline{F}}(z) &= \sum_{n=1}^{r} pr(N = n | N \geq 1) pr(\Im \leq z | N = n) \\ &= \sum_{n=1}^{r} p_n \overline{F}(z | g, g(n)), \end{aligned}$$

com

$$p_n = \frac{\binom{r}{n} (1-p)^n p^{r-n}}{1 - p^r}, \; n = 1, ..., r.$$

como definido em (4.2.4).

#### 4.2.1.2 Uma aplicação a dados do cancro

Os dados utilizados nesta aplicação são referentes à cidade de São Paulo, Brasil, de 2010, e dizem respeito à idade de deteção da doença. Vamos considerar como fator o *Tipo de cancro*, com três níveis: *Ossos longos do membro inferior, Parede lateral da bexiga urinária* e *Corpo do pâncreas.* As Tabelas do Anexo 5 mostram as frequências destes três tipos de cancro, agrupadas por idade. A Tabela 4.1 ilustra o número de pacientes para os diferentes tipos de cancro.

Tabela 4.1: Diferentes tipos de cancro e número de pacientes.

| Tipos de cancro | Número de pacientes |
|---|---|
| Ossos longos dos membros inferiores | 32 |
| Parede lateral da bexiga urinária | 32 |
| Corpo do pâncreas | 29 |

Vamos testar a hipótese:

$$H_{0,F} : \boldsymbol{A\mu} = 0,$$

onde

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix}.$$

O numerador da estatística $\Im_F$, é neste caso dado por

$$S_{num} = (\boldsymbol{AY}_\bullet)' \left(\boldsymbol{A}D\left(\tfrac{1}{32}, \tfrac{1}{32}, \tfrac{1}{29}\right)\boldsymbol{A}'\right)^{-1} (\boldsymbol{AY}_\bullet),$$

que, quando $H_{0,F}$ se verifica, corresponde ao produto de $\sigma^2$ por um qui-quadrado central $g = m - 1 = 2$ com graus de liberdade.

Assim, obtém-se

$$\boldsymbol{A}D\left(\frac{1}{32}, \frac{1}{32}, \frac{1}{29}\right)\boldsymbol{A}' = \begin{bmatrix} 0.0657 & 0.0345 \\ 0.0345 & 0.0657 \end{bmatrix}$$

e

$$\left(\boldsymbol{A}D\left(\frac{1}{32}, \frac{1}{32}, \frac{1}{29}\right)\boldsymbol{A}'\right)^{-1} = \begin{bmatrix} 20.9893 & -11.0108 \\ -11.0108 & 20.9893 \end{bmatrix}.$$

Uma vez que

$$\boldsymbol{Ay}_\bullet = \begin{bmatrix} -33.6045 \\ -0.4795 \end{bmatrix},$$

em que $\boldsymbol{y}_\bullet$ tem como componentes

- $y_{1,\bullet} = 31.8438$;
- $y_{2,\bullet} = 65.9688$;
- $y_{3,\bullet} = 65.4483$;

obtém-se

$$S_{num} = 23352.35.$$

O denominador da estatística $\Im_F$, quando $N = n$, é o produto de $\sigma^2$ por um qui-quadrado central com $g(n) = 2 \times 31 + n - 1 = 61 + n$ graus de liberdade. Neste caso,

$$S = \sum_{j=1}^{32}(y_{1,j} - y_{1,\bullet})^2 + \sum_{j=1}^{32}(y_{2,j} - y_{2,\bullet})^2 + \sum_{j=1}^{29}(y_{3,j} - y_{3,\bullet})^2 = 28022.3599.$$

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

$$\Im_{F,Obs} = \frac{23352.35}{28022.3599} = 0.8333.$$

Considerando $N = n$, quando $H_{0,F}$ se verifica a distribuição condicional de $\Im_F$ será uma $\overline{F}$ central com $g = 2$ e $g(n) = 61 + 29 = 90$ graus de liberdade, $\overline{F}(.|2, 90)$. Os quantis da distribuição condicional, $z_{1-\alpha}$, são apresentados na Tabela 4.2. Como $\Im_{F,Obs} > z_{1-\alpha}$, podemos concluir que se rejeita $H_{0,F}$ para os níveis usuais de significância.

Tabela 4.2: Os quantis da distribuição condicional e não condicional de $\Im_F$

| **Valores de** $\alpha$ | 0.10 | 0.05 | 0.01 |
|---|---|---|---|
| $z_{1-\alpha}$ | 0.05250 | 0.06884 | 0.10776 |
| $z^b_{1-\alpha}$ | 0.05264 | 0.06902 | 0.10805 |

Suponhamos que o limite superior para a dimensão da amostra é $r = 32$ para cada tipo de cancro. No contexto desta aplicação podemos assim assumir que:

- a probabilidade de ocorrer uma falha para os dois primeiros tipos de cancro é aproximadamente nula, portanto as dimensões das amostras nestes casos são consideradas como fixas;

- a probabilidade de ocorrer uma falha para o terceiro tipo de cancro pode ser considerada como $p = 0.1$ logo $1 - p = 0.9$, uma vez que $\frac{29}{32} \simeq 0.9$, considerando a aproximação a uma casa decimal. Neste caso a dimensão da amostra deve ser considerada como aleatória.

Podemos portanto aplicar o nosso modelo a esta situação assumindo que $N \sim B(32, 0.9)$. Obtém-se assim

$$p_n = \frac{\begin{pmatrix} 32 \\ n \end{pmatrix}(0.9)^n 0.1^{32-n}}{1 - 0.2^{32}},\ n = 1, ..., 32,$$

sendo a distribuição não condicional da estatística $\Im_F$, quando $H_{0,F}$ se verifica, dada por

$$\overline{\overline{F}}(z) = \sum_{n=1}^{32} p_n \overline{F}(z|2, 61 + n).$$

Os quantis $z^b_{1-\alpha}$, para a probabilidade $1-\alpha$ desta distribuição são apresentados na Tabela 4.2. Uma vez que $\Im_{F,Obs} > z^b_{1-\alpha}$, concluímos que neste caso se seguirmos a abordagem não condicional continuamos a rejeitar $H_{0,F}$ para o níveis usuais de significância. Significa portanto que as idades com que a doença é detectada para estes três tipos de cancro são significativamente diferentes.

É importante referir que os valores dos quantis referentes à distribuição não condicional são poucos superiores aos da distribuição condicional. Esta tão reduzida diferença deve-se ao fato de neste exemplo a probabilidade de ocorrência de falha ser pequena $(p = 0.1)$. Situações em que $p$ é maior a diferença entre os quantis é mais significativa.

### 4.2.2 Um fator com todos os níveis com dimensões aleatórias

Nesta seção vamos assumir que temos apenas um fator com $m$ níveis e que podem ocorrer falhas em todos os níveis do fator. Assumimos ainda que $r_1, \ldots, r_m$ correspondem aos limites superiores para as dimensões das amostras para cada um dos níveis do fator. Estes limites superiores nem sempre são atingidos uma vez que podem ocorrer falhas de observações.

Suponhamos então que:

- $n_1, ..., n_m$ correspondem às realizações das variáveis aleatórias $N_1, ..., N_m$, que representam as dimensões das amostras;
- $n = \sum_{i=1}^m n_i$ é uma realização da variável aleatória $N = \sum_{i=1}^m N_i$;
- $N_1, ..., N_m$ seguem distribuições Binomiais com parâmetros $r_1, ..., r_m$ e $1-p$, $N_i \sim B(r_i, 1-p)$, $i = 1, ..., m$, onde $p$ corresponde à probabilidade da ocorrência de uma falha;
- devido à independência dos $N_i, i = 1 \ldots, m, N \sim B\left(r, 1-p\right)$, com $r = \sum_{i=1}^m r_i$.

Pretendemos portanto testar a hipótese

$$H_{0,F} : \boldsymbol{A\mu} = \mathbf{0},$$

com $\boldsymbol{A} = [\boldsymbol{I}_{m-1} | -\mathbf{1}_{m-1}]$ e $\boldsymbol{\mu}$ o vetor médio.

#### 4.2.2.1 Estatística de teste e suas distribuições

Considerando-se $N_i = n_i$, $i = 1, ..., m$, temos as amostras $Y_{i,1}, ..., Y_{i,n_i}$, $i = 1, ..., m$, com médias $Y_{i,\bullet}$, $i = 1, ..., m$. Como vimos anteriormente, a soma das somas dos quadrados do erro será dada por

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

Considerando que as observações são normais e independentes com variância $\sigma^2$, quando $N_i = n_i$, $i = 1, ..., m$, $S$ será o produto de $\sigma^2$ por um qui-quadrado central com $g(n) = n - m$ graus de liberdade, $S \sim \sigma^2 \chi^2_{g(n)}$.

Quando $N_i = n_i$, $i = 1, ..., m$ o vetor das médias dos tratamentos, $\boldsymbol{Y}_\bullet$, será normal com vetor médio $\boldsymbol{\mu}$ e matriz de covariância $\sigma^2 D(\frac{1}{n_1}, ..., \frac{1}{n_m})$, que será condicionalmente independente de $S$. Assim, se $N_i = n_i$, $i = 1, ..., m$,

$$S_{num} = (\boldsymbol{A}\boldsymbol{Y}_\bullet)' \left( \boldsymbol{A} D\left(\frac{1}{n_1}, ..., \frac{1}{n_m}\right) \boldsymbol{A}' \right)^{-1} (\boldsymbol{A}\boldsymbol{Y}_\bullet) \sim \sigma^2 \chi^2_{g,\delta(n)}$$

com $g = m - 1$, e

$$\delta(n) = \frac{1}{\sigma^2} (\boldsymbol{A}\boldsymbol{\mu})' \left( \boldsymbol{A} D\left(\frac{1}{n_1}, ..., \frac{1}{n_m}\right) \boldsymbol{A}' \right)^{-1} (\boldsymbol{A}\boldsymbol{\mu}).$$

Quando $H_{0,F}$ se verifica, $\delta(n) = 0$ e ter-se-á $S_{num} \sim \sigma^2 \chi^2_g$.

Assim, quando $N = n$ e $H_{0,F}$ se verifica, a distribuição condicional de

$$\Im_F = \frac{S_{num}}{S}$$

será $\overline{F}(.|g, g(n))$.

Vamos supor que existe uma dimensão mínima para cada amostra, à semelhança do que considerámos no capítulo anterior, ver Nunes et al. (2015). Consideremos que $N_i \geq n_i^\bullet$, $i = 1, ..., m$, e que portanto a dimensão mínima global será $n^\bullet = \sum_{i=1}^m n_i^\bullet$. Assim teremos, com $\boldsymbol{n}^\bullet = (n_1^\bullet, \ldots, n_m^\bullet)'$,

$$\begin{aligned} & p_{n,n_i^\bullet} = pr(N = n | \boldsymbol{N} \geq \boldsymbol{n}^\bullet) \\ = & \sum_{n_1 = n_1^\bullet}^{n - \sum_{i=2}^m n_i^\bullet} \ldots \sum_{n_\ell = n_\ell^\bullet}^{n - (\sum_{i=1}^{\ell} n_i + \sum_{i=\ell+1}^{m} n_i^\bullet)} \ldots \sum_{n_m = n - \sum_{i=1}^{m-1} n_i}^{n - \sum_{i=1}^{m-1} n_i} pr(\boldsymbol{N} = \boldsymbol{n} | \boldsymbol{N} \geq \boldsymbol{n}^\bullet), \end{aligned}$$

onde

$$pr(\boldsymbol{N} = \boldsymbol{n} | \boldsymbol{N} \geq \boldsymbol{n}^\bullet) = \prod_{i=1}^{m} \frac{\binom{r_i}{n_i} (1-p)^{n_i} p^{r_i - n_i}}{\sum_{u_i = n_i^\bullet}^{r_i} \binom{r_i}{u_i} (1-p)^{u_i} p^{r_i - u_i}}, \; n_i = n_i^\bullet, ..., r_i; \; i = 1, ..., m,$$

como definido em (4.1.3).

A distribuição não condicional de $\Im_F$, quando a hipótese $H_{0,F}$ se verifica, será dada por, ver por exemplo Mexia et. al (2011) e Nunes et. al (2014),

$$\begin{aligned} & \overline{\overline{F}}(z) \\ = & \sum_{n=n^\bullet}^{r} pr(N=n|\boldsymbol{N} \geq \boldsymbol{n}^\bullet)\overline{F}(z|g,g(n)) \\ = & \sum_{n=n^\bullet}^{r} p_{n,n_i^\bullet}\overline{F}(z|g,g(n)). \end{aligned}$$

#### 4.2.2.2 Uma aplicação a dados do cancro

Os dados utilizados nesta aplicação são referentes à cidade de São Paulo, 2010, já considerados na subseção 3.2.2.2, e dizem respeito à idade de deteção da doença. Vamos considerar como fator o *Tipo de cancro*, com três os níveis: *Tecidos moles do tórax*, *Trato intestinal* e *Cavidade nasal*. As Tabelas do Anexo 1 mostram as frequências destes três tipos de cancro, agrupadas por idade. A Tabela seguinte ilustra o número de pacientes para cada tipo de cancro, já apresentado na Tabela 3.2.

Tabela 4.3: Tipos de cancro e número de pacientes.

| Tipos de cancro | Número de pacientes |
|---|---|
| Tecidos moles do tórax | 18 |
| Trato intestinal | 22 |
| Cavidade nasal | 25 |

Como vimos a hipótese a testar será

$$H_{0,F} : \boldsymbol{A\mu} = 0,$$

com

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix}.$$

Quando $H_{0,F}$ se verifica, o numerador da estatística $\Im_F$ será dado por

$$S_{num} = (\boldsymbol{AY}_\bullet)' \left(\boldsymbol{A}D\left(\tfrac{1}{18}, \tfrac{1}{22}, \tfrac{1}{25}\right)\boldsymbol{A}'\right)^{-1} (\boldsymbol{AY}_\bullet) \sim \sigma^2\chi^2_g \ ,$$

com $g = m - 1 = 2$. Vamos obter

$$S_{num} = 2073.021.$$

em que as médias amostrais são iguais a

- $y_{1,\bullet} = 49.50$;

- $y_{2,\bullet} = 61.7727$;
- $y_{3,\bullet} = 62.40$.

Quanto ao denominador da estatística, quando $N = n$, $S \sim \sigma^2\chi^2_{g(n)}$ com $g(n) = n - 3$. Neste caso obtemos $S = 26632.364$.

Então o valor observado da estatística, $\Im_{F,Obs}$, será

$$\Im_{F,Obs} = \frac{2073.021}{26632.364} = 0.07784.$$

Dado $N = n$, quando $H_{0,F}$ se verifica, a distribuição condicional de $\Im_F$ corresponde a uma distribuição $\overline{F}$ central com $g = 2$ e $g(n) = 65 - 3 = 62$ graus de liberdade, $\overline{F}(z|2, 62)$. Os quantis desta distribuição, $z_{1-\alpha}$, são apresentados na Tabela 4.4. Podemos concluir que se rejeita $H_{0,F}$ para $\alpha = 0.1$, pois $\Im_{F,Obs} > z_{1-\alpha}$, e não se rejeita para $\alpha = 0.05$ e $\alpha = 0.01$.

Tabela 4.4: Os quantis da distribuição condicional e não condicional de $\Im_F$

| **Valores de** $\alpha$ | 0.10 | 0.05 | 0.01 |
|---|---|---|---|
| $z_{1-\alpha}$ | 0.07711 | 0.10146 | 0.16016 |
| $z^b_{1-\alpha}$ | 0.07871 | 0.10361 | 0.16365 |

Para calcular os quantis da distribuição não condicional de estatística vamos assumir que os limites superiores para as dimensões das amostras são $r_1 = 22$, $r_2 = 27$ e $r_3 = 31$. Neste caso a probabilidade de ocorrência de uma falha será $p = 0.2$, tendo-se portanto $1 - p = 0.8$, uma vez que $\frac{n_i}{r_i} \simeq 0.8,\ i = 1, 2, 3$, considerando a aproximação a uma casa decimal. Isto significa que $N_1 \sim B(22, 0.8)$, $N_2 \sim B(27, 0.8)$ e $N_3 \sim B(31, 0.8)$. Devido à independência dos $N_i$, $i = 1, 2, 3$, tem-se $N = \sum_{i=1}^{3} N_i \sim B(80, 0.8)$.

Vamos supor, tal como na subseção 3.2.2.2, que existem pelo menos 5 observações por nível, assim $n_i^\bullet = 5$, $i = 1, 2, 3$, $\boldsymbol{n}^\bullet = (5, 5, 5)'$ e $n^\bullet = 15$.

A distribuição não condicional de $\Im_F$, quando a hipótese $H_{0,F}$ se verifica, será neste caso dada por

$$\overline{\overline{F}}(z) \quad = \quad \sum_{n=15}^{80} \sum_{n_1=5}^{n-10} \sum_{n_2=5}^{n-(n_1+5)} \sum_{n_3=n-(n_1+n_2)}^{n-(n_1+n_2)} pr(\boldsymbol{N} = \boldsymbol{n}|\boldsymbol{N} \geq \boldsymbol{n}^\bullet)\overline{F}(z|2, n-3),$$

com

$$pr(\boldsymbol{N} = \boldsymbol{n}|\boldsymbol{N} \geq \boldsymbol{n}^\bullet) = \prod_{i=1}^{3} \frac{\begin{pmatrix} r_i \\ n_i \end{pmatrix} (0.8)^{n_i}(0.2)^{r_i - n_i}}{\sum_{u_i=5}^{r_i} \begin{pmatrix} r_i \\ u_i \end{pmatrix} (0.8)^{u_i}(0.2)^{r_i - u_i}}.$$

Os quantis obtidos para a probabilidade $1 - \alpha$ desta distribuição, $z^b_{1-\alpha}$, são apresentados na Tabela 4.4. Concluímos portanto que não se rejeita a hipótese para os níveis usuais de signifi-

cância. Neste caso, para $\alpha = 0.1$, a abordagem não condicional leva-nos a tomar uma decisão contrária à que tinhamos tomado pela abordagem clássica (decisão similar à apresentada na subseção 3.2.2.2).

### 4.2.3 Mais do que um fator de efeitos fixos

Vamos agora assumir que se tem mais que um fator de efeitos fixos com $m$ tratamentos na totalidade. Consideremos ainda que os limites superiores para as dimensões das amostras correspondem a $r_1, ..., r_m$, os quais podem não ser atingidos uma vez que podem ocorrer falhas.

Assim, vamos assumir que:

- $N_1, ..., N_m$ são variáveis aleatórias independentes que representam as dimensões das amostras;
- $n_1, ..., n_m$ são as realizações das variáveis aleatórias $N_1, ..., N_m$;
- $n = \sum_{i=1}^{m} n_i$ é uma realização da variável aleatória $N = \sum_{i=1}^{m} N_i$;
- $N_1, ..., N_m$ seguem distribuições Binomiais com parâmetros $r_1, ..., r_m$ e $1-p$, $N_i \sim B(r_i, 1-p)$, $i = 1, ..., m$, onde $p$ representa a probabilidade da ocorrência de uma falha;
- devido à independência dos $N_i, i = 1 \ldots, m, N \sim B\left(r, 1-p\right)$, com $r = \sum_{i=1}^{m} r_i$.

Pretendemos testar as hipóteses

$$H_{0,j} : \boldsymbol{\mu} \in w_j,\ j = 1, ..., \tau,$$

onde $w_j = (\overline{w}_j^{\perp} \cap \Omega)$, $j = 1, ..., \tau$, é um subespaço do espaço paramétrico $\Omega$. Como vimos no capítulo anterior, se assumirmos que os vetores linha de $\boldsymbol{A_j}$ constituem uma base ortonormada para $w_j$, $j = 1, ..., \tau$, podemos reescrever estas hipóteses da seguinte forma

$$H_{0,j} : \boldsymbol{A_j \mu} = \mathbf{0},\ j = 1, ..., \tau.$$

Estas hipóteses correspondem a hipóteses de ausências de efeitos e de interação entre os fatores.

#### 4.2.3.1 Estatística de teste e suas distribuições

Nesta seção, obteremos as distribuições das estatísticas, assumindo que as dimensões das amostras para os $m$ tratamentos são realizações de variáveis aleatórias independentes com distribuição Binomial. Para obtermos as distribuições não condicionais das estatísticas consideraremos uma dimensão mínima global para as amostras. Por esse motivo vamos considerar $\ddot{N}_1, ..., \ddot{N}_m$ como sendo as variáveis correspondentes às dimensões das amostras e que

$\ddot{N}_i \sim B(r_i, 1-p), i = 1, \ldots, m$. Recordemos que $\ddot{N}_i$ correspondem às variáveis truncadas $N_i$, com $N_i \geq 1, i = 1, \ldots, m$.

Assim, quando $\ddot{N}_i = n_i$, $i = 1, ..., m$, temos as amostras

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

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

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

tendo-se, assumindo a normalidade e a independência das observações com variância $\sigma^2$,

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

com $g(n) = n - m$.

Além disso, $S$ será condicionalmente independente do vetor das médias dos tratamentos, $\boldsymbol{Y}_\bullet$, que será normal com média $\boldsymbol{\mu}$ e matriz de covariância $\sigma^2 D(\frac{1}{n_1}, ..., \frac{1}{n_m})$. Tendo-se, quando $\ddot{N}_i = n_i$, $i = 1, ..., m$,

$$S_j = (\boldsymbol{A}_j \boldsymbol{Y}_\bullet)' \left( \boldsymbol{A}_j D\left( \frac{1}{n_1}, ..., \frac{1}{n_m} \right) \boldsymbol{A}_j' \right)^{-1} (\boldsymbol{A}_j \boldsymbol{Y}_\bullet) \sim \sigma^2 \chi^2_{g_j, \delta_j(n)},\ j = 1, ..., \tau,$$

com

$$g_j = car(\boldsymbol{A}_j),\ j = 1, ..., \tau,$$

e

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

Portanto, como vimos no capítulo anterior, a distribuição condicional da estatística de teste

$$\Im_j = \frac{S_j}{S},\ j = 1, ..., \tau,$$

será $\overline{F}(.|g_j, g(n), \delta_j(n)), j = 1, ..., \tau$.

Quando $\ddot{N} = n$, e $H_{0,j}$, $j = 1, ..., \tau$, se verifica, $\delta_j(n) = 0,\ j = 1, \ldots, \tau$ e a distribuição condicional de $\Im_j$ será uma distribuição $\overline{F}$ central com $g_j$ e $g(n)$ graus de liberdade, $\overline{F}(.|g_j, g(n)), j = 1, ..., \tau$.

Considerando como dimensão mínima global para as amostras o valor $n^\bullet$, o que significa que $\ddot{N} \geq n^\bullet$, obtém-se a distribuição não condicional de $\Im_j, i = 1, \ldots, \tau$, quando $H_{0,j},\ i = 1, \ldots, \tau$, se verifica,

$$\begin{aligned}\overline{\overline{F}}_j(z) &= \sum_{n=n^\bullet}^{r} pr(\ddot{N}=n|\ddot{N}\geq n^\bullet)pr(\Im_j\leq z|\ddot{N}=n)\\ &= \sum_{n=n^\bullet}^{r}\ddot{p}_{n,n^\bullet}\overline{F}(z|g_j,g(n)),\ j=1,...,\tau,\end{aligned}$$

onde $\ddot{p}_{n,n^\bullet}=\ddot{p}_{n,m+1}\frac{1-\ddot{p}_m}{1-\sum_{u=m}^{n^\bullet-1}\ddot{p}_u},\ n=n^\bullet,\ldots,r$, como definido pela expressão (4.1.1).

#### 4.2.3.2 Uma aplicação a dados do cancro

Mais uma vez os dados que vamos usar nesta aplicação são da cidade de São Paulo, Brasil, 2010 referentes à idade de deteção da doença. Vamos considerar dois fatores, o *Tipo de cancro* e o *Género*. O primeiro fator com três níveis: *Cavidade nasal e ouvido médio; Meninges* e *Coração, mediastino e pleura*. O segundo fator com dois níveis: *Masculino* e *Feminino*. Isto resulta em $m=3\times 2=6$ diferentes tratamentos. As Tabelas do Anexo 6 mostram as frequências destes três tipos de cancro, para os dois géneros, agrupados por idade. A Tabela 4.5 ilustra o número de pacientes por tipo de cancro e género.

Tabela 4.5: Número de pacientes por tipo de cancro e género.

| | | Género (segundo fator) | |
|---|---|---|---|
| | | Masculino | Feminino |
| Tipos de cancro (primeiro fator) | | | |
| | Cavidade nasal e ouvido médio | 13 | 16 |
| | Meninges | 9 | 11 |
| | Coração, mediastino e pleura | 18 | 13 |

Vamos testar as hipóteses

$$H_{0,j}: \boldsymbol{A}_j\boldsymbol{\mu}=0,\ j=1,2,3.$$

Como já referimos no capítulo anterior, nestes casos a análise deve começar com o teste à interação e, se esta não for significativa, de seguida fazem-se os testes aos efeitos principais dos dois fatores. Se a interação for significativa, a inferência é realizada para cada nível de um dos fatores. Como neste caso o segundo fator tem apenas dois níveis, serão obtidos intervalos de confiança para a diferença das médias. Mais uma vez iremos seguir esta abordagem o que nos leva a ordenar as hipóteses $H_{0,j}$ e as matrizes $\boldsymbol{A}_j,\ j=1,2,3,$ da seguinte forma

- Índice $1$- interação;
- Índice $2$- primeiro fator;
- Índice $3$- segundo fator.

Como vimos, quando $\ddot{N} = n$ e $H_{0,j}$ se verifica, a distribuição condicional de $\Im_j,\ j = 1,2,3$, corresponde a $\overline{F}(.|g_j, n-6)$, com $g_j = car(\boldsymbol{A}_j),\ \ j = 1,2,3$. A distribuição não condicional, considerando $\ddot{N} > n^\bullet$, é dada por

$$\overline{\overline{F}}_j(z) = \sum_{n=n^\bullet}^{r} \ddot{p}_{n,n^\bullet} \overline{F}(z|g_j, n-6),\ j = 1,2,3,$$

com $\ddot{p}_{n,n^\bullet} = \ddot{p}_{u,7} \frac{1-\ddot{p}_6}{1-\sum_{u=6}^{n^\bullet-1} \ddot{p}_u},\ n = n^\bullet, \ldots, r.$

Devido às propriedades de monotonia da distribuição $\overline{F}$, que já foram referidas anteriormente, quando $n < n^o$, temos

$$\overline{F}(z|g_j, n-6) < \overline{F}(z|g_j, n^o-6),$$

portanto

$$\overline{F}(z|g_j, n^\bullet-6) \leq \overline{\overline{F}}_j(z) \leq 1,$$

o que significa que $\overline{F}(z|g_j, n^\bullet-6)$ nos dá um limite inferior para $\overline{\overline{F}}(z)$. Assim, a partir de $\overline{F}(z|g_j, n^\bullet-6)$, podemos obter limites superiores para os quantis da distribuição não condicional $\overline{\overline{F}}_j(z),\ j = 1,2,3.$

Em relação à **interação** entre os dois fatores tem-se

$$\boldsymbol{A}_1 = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & 0 & 0 & -\frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2\sqrt{3}} & \frac{1}{2\sqrt{3}} & \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} & -\frac{1}{2\sqrt{3}} & \frac{1}{2\sqrt{3}} \end{bmatrix},$$

com $g_1 = car(\boldsymbol{A}_1) = 2$. Obtemos

$$\left(\boldsymbol{A}_1 D\left(\frac{1}{13}, \frac{1}{16}, \frac{1}{9}, \frac{1}{11}, \frac{1}{18}, \frac{1}{13}\right) \boldsymbol{A}_1'\right)^{-1} = \begin{bmatrix} 14.7137 & 0.1639 \\ 0.1639 & 11.1131 \end{bmatrix},$$

e

$$\boldsymbol{A}_1 \boldsymbol{y}_\bullet = \begin{bmatrix} 0.3980 \\ 11.6077 \end{bmatrix},$$

em que o vetor $\boldsymbol{y}_\bullet$ tem como componentes

- $y_{1,\bullet} = 55.0769;\ y_{2,\bullet} = 63.8750;$
- $y_{3\bullet} = 47.0000;\ y_{4,\bullet} = 36.0909;$
- $y_{5\bullet} = 53.9444;\ y_{6,\bullet} = 63.5385.$

Obtém-se então

$$S_1 = (\boldsymbol{A}_1 \boldsymbol{y}_\bullet)' \left(\boldsymbol{A}_1 D\left(\frac{1}{13}, \frac{1}{16}, \frac{1}{9}, \frac{1}{11}, \frac{1}{18}, \frac{1}{13}\right) \boldsymbol{A}_1'\right)^{-1} (\boldsymbol{A}_1 \boldsymbol{y}_\bullet) = 1501.221$$

para o numerador da estatística $\Im_1$.

Para o denominador das estatísticas, temos

$$S = \sum_{i=1}^{6} \sum_{k=1}^{n_i} (Y_{i,k} - Y_{i,\bullet})^2 = 1501.221.$$

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

$$\Im_{1,Obs} = \frac{1501.221}{35037.7574} = 0.0428.$$

Os quantis, $z_{1-\alpha}$, da distribuição condicional $\Im_1$, que corresponde a $\overline{F}(z|2,74)$, são apresentados na Tabela 4.6.

Tabela 4.6: Os quantis da distribuição condicional e limites superiores para o quantis de $\Im_1$ e $\Im_2$.

| **Valores de** $\alpha$ | 0.10 | 0.05 | 0.01 |
|---|---|---|---|
| $z_{1-\alpha}$ | 0.0642 | 0.0843 | 0.1325 |
| $z^u_{1-\alpha}$ | 0.3594 | 0.4910 | 0.8478 |

Podemos concluir que não se rejeita $H_{0,1}$ para os níveis usuais de significância quando consideramos a distribuição condicional.

Vamos agora supor que se tem $n^\bullet \geq 21$, por exemplo que $n^\bullet = 21$, o que significa que se tem no mínimo pouco mais de três observações por tratamento. A Tabela 4.6 mostra os limites superiores para os quantis, para a probabilidade $1-\alpha$, $z^u_{1-\alpha}$, da distribuição não condicional $\overline{\overline{F}}_1(z)$. Como $\Im_{1,Obs} < z^u_{1-\alpha}$ podemos concluir, como era de esperar, que não se rejeita $H_{0,1}$ para os níveis usuais de significância. Os valores apresentados na Tabela 4.6 podem levar-nos a pensar que não temos dimensões suficientemente grandes para as amostras por forma a detetar a existência de interação. Assumindo que os valores da estatística permanecem inalterados, serão necessários os valores mínimos de $n^\bullet$ apresentados na Tabela 4.7 por forma a rejeitar $H_{0,1}$.

Tabela 4.7: Valor mínimo de $n^\bullet$ que leva à rejeição da hipótese $H_{0,1}$.

| **Valores de** $\alpha$ | 0.10 | 0.05 | 0.01 |
|---|---|---|---|
| $n^\bullet$ | 116 | 149 | 226 |

Uma vez que para valores maiores de $n^\bullet$ vamos obter valores menores para os quantis. Concluímos que para todo o $n^\bullet \geq 226$ se rejeita $H_{0,1}$, o que significa que nestas circunstâncias a interação é significativa. A Figura 4.1 parece apontar efetivamente para a existência de interação entre os dois fatores, uma vez que as linhas se cruzam.

Construímos também os intervalos de confiança para as diferenças das médias entre os géneros, para os três tipos de cancro. A Tabela 4.8 mostra os intervalos obtidos.

Para todos os graus de confiança considerados a origem pertence aos intervalos de confiança, o que significa que não existe diferença significativa quanto à média das idades entre os géneros. Os resultados podem ser igualmente observados pela análise da Figura 4.2, considerando o grau de confiança de $95\%$.

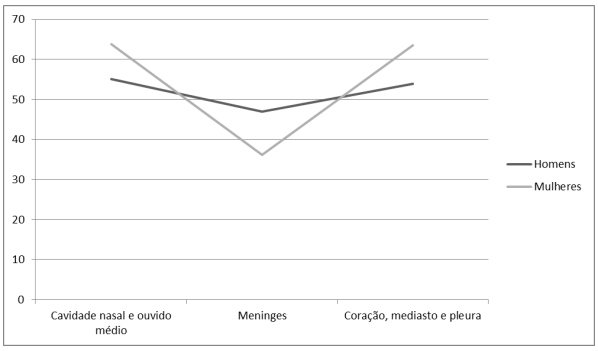


Figura 4.1: Interação entre os dois fatores.

Tabela 4.8: Intervalos de confiança para as diferenças das médias

| **Valores de** $1-\alpha$ | 0.90 | 0.95 | 0.99 |
|---|---|---|---|
| **Cavidade nasal e ouvido médio** | [-22.1056; 4.5094] | [-24.8286; 7.2325] | [-30.4449; 12.8487] |
| **Meninges** | [-8.7835; 30.6017] | [-12.9497; 34.7679] | [-21.7795; 43.5977] |
| **Coração, mediastino e pleura** | [-22.0300; 2.8418] | [-24.5632; 5.3750] | [-29.7681; 10.5799] |

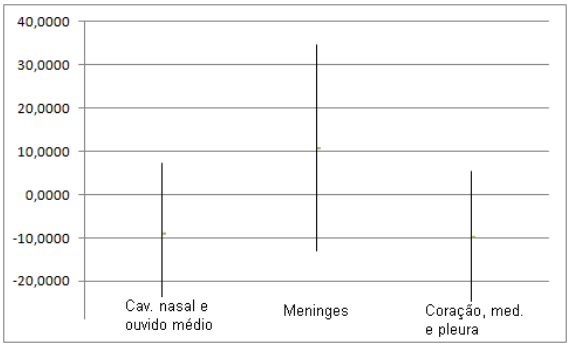


Figura 4.2: Intervalo de confiança a 95% para a diferenças das médias entre os dois géneros.

Apesar da existência de interação, iremos mesmo assim testar os efeitos principais dos dois fatores por forma a mostrar como estes testes se comportam descondicionando as distribuições.

Para o **primeiro fator** temos

$$\boldsymbol{A}_2 = \begin{bmatrix} -\frac{1}{2} & -\frac{1}{2} & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2\sqrt{3}} & \frac{1}{2\sqrt{3}} & -\frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} & \frac{1}{2\sqrt{3}} & \frac{1}{2\sqrt{3}} \end{bmatrix}$$

e $g_2 = car(\boldsymbol{A}_2) = 2$. Obtemos

$$\left(\boldsymbol{A}_2 D\left(\frac{1}{13}, \frac{1}{16}, \frac{1}{9}, \frac{1}{11}, \frac{1}{18}, \frac{1}{13}\right) \boldsymbol{A}_2'\right)^{-1} = \begin{bmatrix} 14.7137 & 0.1639 \\ 0.1639 & 11.1131 \end{bmatrix},$$

e

$$\boldsymbol{A}_2 \boldsymbol{y}_\bullet = \begin{bmatrix} -0.7345 \\ 20.2803 \end{bmatrix},$$

donde

$$S_2 = (\boldsymbol{A}_2\boldsymbol{y}_\bullet)' \left( \boldsymbol{A}_2 D\left(\frac{1}{13}, \frac{1}{16}, \frac{1}{9}, \frac{1}{11}, \frac{1}{18}, \frac{1}{13}\right) \boldsymbol{A}_2' \right)^{-1} (\boldsymbol{A}_2\boldsymbol{y}_\bullet) = 4573.7740.$$

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

$$\Im_{2,Obs} = \frac{4573.7740}{35037.7574} = 0.1305.$$

Os quantis, $z_{1-\alpha}$, da distribuição condicional de $\Im_2$, que corresponde a $\overline{F}(z|2,74)$, são apresentados mais uma vez na Tabela 4.6. Assim, usando a distribuição condicional podemos concluir que rejeitamos $H_{0,2}$ para $\alpha = 0.1$ e $0.05$ e não rejeitamos para a $\alpha = 0.01$.

Para a distribuição não condicional da estatística de teste $\overline{\overline{F}}_2(z)$, vamos considerar mais uma vez $n^\bullet = 21$ obtendo-se os limites superiores para os quantis da probabilidade $1-\alpha$, $z^u_{1-\alpha}$, apresentados na Tabela 4.6. Estes resultados podem contradizer a decisão que tomamos através da utilização da abordagem condicional, para $\alpha = 0.1$ e $0.05$.

Supondo que os valores da estatística de teste permanecem inalterados teremos que ter os tamanhos das amostras apresentados na Tabela 4.9 por forma a garantir a rejeição da hipótese.

Tabela 4.9: Valor mínimo de $n^\bullet$ que leva à rejeição da hipótese $H_{0,2}$.

| **Valores de** $\alpha$ | 0.10 | 0.05 | 0.01 |
|---|---|---|---|
| $n^\bullet$ | 44 | 55 | 82 |

Então, por exemplo para $\alpha = 0.05$, teremos que ter pelo menos $55$ observações para tomarmos a mesma decisão utilizando as duas abordagens.

Uma vez que para valores maiores de $n^\bullet$ obteríamos valores menores para os quantis, temos $\Im_{2,Obs} > z^u_{1-\alpha}$ para todo o $n^\bullet \geq 82$, o que significa que, neste caso, vamos rejeitar $H_{0,2}$ considerando os níveis usuais de significância o que significa que o primeiro fator é significativo.

Para o **segundo fator** temos

$$\boldsymbol{A}_3 = \left[ \begin{array}{cccccc} -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \end{array} \right],$$

e $g_3 = car(\boldsymbol{A}_3) = 1$.

Obtém-se

$$S_3 = \left( \boldsymbol{A}_3 D\left(\frac{1}{13}, \frac{1}{16}, \frac{1}{9}, \frac{1}{11}, \frac{1}{18}, \frac{1}{13}\right) \boldsymbol{A}_3' \right)^{-1} = \left[ \begin{array}{c} 12.6603 \end{array} \right],$$

e

$$\boldsymbol{A}_3\boldsymbol{y}_\bullet = \left[ \begin{array}{c} 3.0549 \end{array} \right],$$

e para o numerador da estatística $\Im_3$,

$$S_3 = (\boldsymbol{A}_3\boldsymbol{y}_\bullet)' \left( \boldsymbol{A}_3 D\left(\frac{1}{13}, \frac{1}{16}, \frac{1}{9}, \frac{1}{11}, \frac{1}{18}, \frac{1}{13}\right) \boldsymbol{A}_3' \right)^{-1} (\boldsymbol{A}_3\boldsymbol{y}_\bullet) = 118.1531$$

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

$$\Im_{3,Obs} = \frac{118.1531}{35037.7574} = 0.0034.$$

Considerando a distribuição condicional de $\Im_3$, que corresponde a $\overline{F}(z|1,74)$, obtêm-se os quantis apresentado na Tabela 4.10.

Tabela 4.10: Os quantis da distribuição condicional e limites superiores para o quantis de $\Im_3$.

| **Valores de** $\alpha$ | 0.10 | 0.05 | 0.01 |
|---|---|---|---|
| $z_{1-\alpha}$ | 0.0375 | 0.0537 | 0.0945 |
| $z^u_{1-\alpha}$ | 0.2049 | 0.3029 | 0.5789 |

Assim, podemos concluir que não se rejeita a hipótese nula para os níveis usuais de significância quando consideramos a distribuição condicional.

Para as distribuição não condicional da estatística de teste, $\overline{\overline{F}}_3(z)$, vamos considerar $n^\bullet = 21$ obtendo-se os limites superiores para os quantis da probabilidade $1-\alpha$, $z^u_{1-\alpha}$, apresentados também na Tabela 4.10.

Como esperado, esta abordagem leva-nos a tomar a mesma decisão que tínhamos tomado usando a abordagem condicional. No entanto, os resultados da Tabela 4.10 poderíam-nos levar a pensar que não temos observações suficientes para detectar as diferenças entre os géneros. Assumindo que os valores da estatística de teste permanecem inalterados, para garantir a rejeição da hipótese seria necessário ter os tamanhos mínimos das amostras apresentados na Tabela 4.11.

Tabela 4.11: Valor mínimo $n^\bullet$ que leva à rejeição da hipótese $H_{0,3}$.

| **Valores** $\alpha$ | 0.10 | 0.05 | 0.01 |
|---|---|---|---|
| $n^\bullet$ | 810 | 1148 | 1978 |

Na verdade, esses valores são elevados, logo somos levados a concluir que não existe diferença significativa entre os dois géneros. A mesma decisão já tinha sido tomada com base nos intervalos de confiança obtidos e apresentados na Tabela 4.8.

## 4.3 Modelos mistos

Na formulação dos modelos mistos, à semelhança do apresentado no Capítulo 3, vamos utilizar as extensões $L$.

Vamos supor que o vetor $\boldsymbol{Y}^o$ tem $m$ componentes que correspondem aos tratamentos de um modelo linear e que se tem

$$\boldsymbol{L} = D(\mathbf{1}_{n_1}, ..., \mathbf{1}_{n_m}) = \boldsymbol{L}(\boldsymbol{n}),$$

com $\boldsymbol{n} = (n_1, ..., n_m)'$. Temos então a extensão $L$

$$\boldsymbol{Y} = \boldsymbol{L}\boldsymbol{Y}^o + \boldsymbol{\varepsilon},$$

que corresponde a um modelo com amostras $n_1, ..., n_m$, onde $\varepsilon$ é o vetor dos erros com vetor médio nulo e matriz de covariância $\sigma^2\boldsymbol{I}_n$. Continuamos a considerar que

$$n = \sum_{i=1}^{m} n_i,$$

e que se tem o modelo misto

$$\boldsymbol{Y}^o = \sum_{i=0}^{w} \boldsymbol{X}_i\boldsymbol{\beta}_i, \quad (4.3.5)$$

com $\beta_0$ fixo e $\beta_1, ..., \boldsymbol{\beta}_w$ aleatórios e independentes, com vetores médios nulos e matrizes de covariância $\sigma_1^2\boldsymbol{I}_{c_1}, ..., \sigma_w^2\boldsymbol{I}_{c_w}$. O modelo $\boldsymbol{Y}$ também será um modelo misto.

Suponhamos mais uma vez que podem ocorrer falhas para os $m$ tratamentos e que $r_1, \ldots, r_m$ correspondem aos limites superiores para a dimensão das amostras. Portanto:

- $n_1, ..., n_m$ são as realizações das variáveis aleatórias independentes, $N_1, ..., N_m$;
- $N_i \sim B(r_i, 1-p),\ i = 1, \ldots, m$ e $N \sim B(r, 1-p)$, com $r = \sum_{i=1}^{m} r_i$.

Como vimos no Capítulo anterior os resultados sobre extensões $L$ podem ser aplicados no "contexto" dos tamanhos das amostras aleatórias. Vamos portanto passar a considerar a matriz

$$\boldsymbol{L}(\boldsymbol{N}) = D(\mathbf{1}_{N_1}, ..., \mathbf{1}_{N_m}),$$

com $\boldsymbol{N} = (N_1, \ldots, N_m)$.

### 4.3.1 Estatística de teste e suas distribuições

Iremos mais uma vez, por forma a obter as expressões das distribuições não condicionais das estatísticas, assumir que se tem uma dimensão mínima global para as amostras o que nos leva a considerar as variáveis $\ddot{N}_i,\ i = 1, \ldots, m$.

Como vimos na seção 3.4 estamos interessados em testas as hipóteses

$$H_{0,j,M} : \gamma_j = 0,\ j > z,$$

onde $\gamma_j,\ j = 1, \ldots, l$, correspondem às componentes de variância canónicas.

Teremos a estatística de teste definida em (2.3.20) dada por

$$\mathcal{T}_j = \frac{(\boldsymbol{^{oo}Y}_j)'(\boldsymbol{B}_j^{-1})\boldsymbol{^{oo}Y}_j}{S},\ j > z,$$

que, quando $H_{0,j}$ se verifica e $\ddot{N}_i = n_i,\ i = 1, \ldots, m$, tem distribuição $\overline{F}$ central com $f_j = car(\boldsymbol{Q}_j)$, $j > z$, e $g(n) = n - m$ graus de liberdade, $\overline{F}(.|f_j, g(n))$.

A distribuição não condicional de $\mathcal{T}_j,\ j > z$, quando a hipótese $H_{0,j,M}$ se verifica, e se tem como dimensão mínima global para as amostras $n^\bullet$, ou seja $\ddot{N} \geq n^\bullet$, será dada por

$$\begin{aligned} \overline{\overline{F}}_j(z) &= \sum_{n=n^\bullet}^{r} pr(\ddot{N} = n|\ddot{N} \geq n^\bullet) F(z|f_j, g(n)) \\ &= \sum_{n=n^\bullet}^{r} \ddot{p}_{n,n^\bullet} F(z|f_j, g(n)),\ j > z, \end{aligned} \tag{4.3.6}$$

com

$$\ddot{p}_{n,n^\bullet} = \ddot{p}_{n,m+1} \frac{1 - \ddot{p}_m}{1 - \sum_{u=m}^{n^\bullet - 1} \ddot{p}_u},\ n = n^\bullet, \ldots, r,$$

como definido em (4.1.1).

### 4.3.2 Uma aplicação a dados do cancro

Nesta seção, vamos considerar um modelo misto com um fator de efeitos fixos e outro aleatório. Mais uma vez os dados foram disponibilizados pelo INCA, e são referentes ao cancro do Cólon do ano de 2008, Brasil. O fator de efeitos fixos será o *género*, com dois níveis: *Masculino* e *Feminino* e o fator de efeitos aleatórios será o *Estado.* Foram selecionados três Estados recorrendo-se ao método de amostragem aleatória simples. A Tabela 4.12 ilustra o número de pacientes por género e pelos três Estados que foram selecionados.

Tabela 4.12: Estados selecionados e número de pacientes.

| | | género (segundo fator) | |
|---|---|---|---|
| | | Masculino | Feminino |
| Estados (primeiro fator) | Espírito Santo | 34 | 38 |
| | Goiás | 99 | 88 |
| | Paraná | 108 | 132 |

Temos portanto $m = 2 \times 3 = 6$ tratamentos diferentes. As Tabelas de frequências referentes aos três Estados, para homens e mulheres, são apresentadas no Anexo 7.

De (4.3.5), teremos o modelo

$$\boldsymbol{Y}^o = \boldsymbol{X}_0\boldsymbol{\beta}_0 + \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2, \quad (4.3.7)$$

onde $\boldsymbol{\beta}_0$ é fixo e $\boldsymbol{\beta}_1$ e $\boldsymbol{\beta}_2$ são aleatórios e independentes, e correspondem, respectivamente, ao fator de efeitos aleatórios e interação entre os fatores. Mais uma vez iremos utilizar os resultados apresentados na seção 2.3.5 do Capítulo 2 sobre extensões $L$. Teremos então, como vimos na seção 3.4.2,

$$\begin{cases} \boldsymbol{X}_0 = \boldsymbol{I}_2 \otimes \mathbf{1}_3 \\ \boldsymbol{X}_1 = \mathbf{1}_2 \otimes \boldsymbol{I}_3 \\ \boldsymbol{X}_2 = \boldsymbol{I}_2 \otimes \boldsymbol{I}_3 \end{cases}.$$

As matrizes $\boldsymbol{A}_j$, $j = 1, 2$ serão dada por

$$\begin{cases} \boldsymbol{A}_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{\sqrt{2}} \end{bmatrix} \\ \boldsymbol{A}_2 = \begin{bmatrix} -\frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & -\frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & -\frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{\sqrt{2}} \end{bmatrix} \end{cases}$$

e

$$\begin{cases} \boldsymbol{X}_1^0 = \boldsymbol{A}_1\boldsymbol{X}_0 = \frac{1}{\sqrt{2}}\mathbf{1}_2' \otimes \mathbf{1}_3 \\ \boldsymbol{X}_2^0 = \boldsymbol{A}_2\boldsymbol{X}_0 = \left[-\frac{1}{\sqrt{2}} \ \frac{1}{\sqrt{2}}\right] \otimes \mathbf{1}_3 \end{cases}.$$

As matrizes $\boldsymbol{Q}_j$, $j = 1, 2$, que correspondem às $MPO$ sobre $R(\boldsymbol{X}_j^0)^\perp$, $j = 1, 2$, serão neste caso

$$\begin{cases} \boldsymbol{Q}_1 = \boldsymbol{W}_1'\boldsymbol{W}_1 = \boldsymbol{I}_3 - \frac{1}{3}\boldsymbol{J}_3 \\ \boldsymbol{Q}_2 = \boldsymbol{W}_2'\boldsymbol{W}_2 = \boldsymbol{I}_3 - \frac{1}{3}\boldsymbol{J}_3 \end{cases},$$

com

$$\boldsymbol{W}_1 = \boldsymbol{W}_2 = \begin{bmatrix} -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \end{bmatrix},$$

e ter-se-á $f_1 = car(\boldsymbol{Q}_1) = 3$ e $f_2 = car(\boldsymbol{Q}_2) = 3$. Portanto, as $MPO$ sobre $R(\boldsymbol{X}_j^0)$, $j = 1, 2$, são

$$\begin{cases} \boldsymbol{P}_1 = \boldsymbol{X}_1^0(\boldsymbol{X}_1^0)^+ = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} \\ \\ \boldsymbol{P}_2 = \boldsymbol{X}_2^0(\boldsymbol{X}_2^0)^+ = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} \end{cases}.$$

Queremos testar as hipóteses

$$H_{0,j,M} : \gamma_j = 0,\ j = 1, 2,$$

que correspondem às hipóteses de ausências de efeitos do segundo fator e interação entre os dois fatores, respectivamente. Quando $\ddot{N} = n$ e $H_{0,j}$, $j = 1, 2$, se verifica, a distribuição condicional da estatística

$$\mathcal{T}_j = \frac{(^{\boldsymbol{oo}}\boldsymbol{Y}_j)'(\boldsymbol{B}_j^{-1})^{\boldsymbol{oo}}\boldsymbol{Y}_j}{S},\ j = 1, 2$$

será $\overline{F}(.|3, n-6)$, uma vez que $f_j = rank(\boldsymbol{Q}_j) = 3$, $j = 1, 2$ e $g(n) = n - 6$.

Assumindo que se tem $\ddot{N} \geq n^{\bullet}$, sendo portanto $n^{\bullet}$ a dimensão mínina global para as amostras, a distribuição não condicional da estatística será dada por

$$\overline{\overline{F}}_j(z) = \sum_{n=n^{\bullet}}^{r} \ddot{p}_{n,n^{\bullet}} F(z|3, n-6),\ j = 1, 2.$$

Pelas propriedades de monotonia da distribuição $\overline{F}$, já referidas e apresentadas na subseção 2.2.5.1,

$$\overline{F}(z|3, n^{\bullet} - 6) \leq \overline{\overline{F_j}}(z) \leq 1$$

e portanto, a partir de $\overline{F}(z|3, n^{\bullet}-6)$, podemos obter limites superiores para os quantis da distribuição não condicional, $\overline{\overline{F_j}}(z)$. Tal como na subseção 3.4.2, iremos manter os índices das matrizes $\boldsymbol{X}_j,\ j = 0, 1, 2$ e das hipóteses $H_{0,j}, j = 1, 2,$ apesar de começarmos com o teste à interação, logo

- Índice $0$- primeiro fator;
- Índice $1$- segundo fator;
- Índice $2$- interação.

Assim para a **interação**, obtemos

$$^{oo}\boldsymbol{Y}_2 = \boldsymbol{W}_2\boldsymbol{A}_2\boldsymbol{L}^+\boldsymbol{Y} = \begin{bmatrix} 5.4445 \\ 0.2594 \end{bmatrix},$$

onde

$$\boldsymbol{L}^{+} = D\left(\frac{1}{34}\mathbf{1}'_{34}, \frac{1}{38}\mathbf{1}'_{38}, \frac{1}{99}\mathbf{1}'_{99}, \frac{1}{88}\mathbf{1}'_{88}, \frac{1}{108}\mathbf{1}'_{108}, \frac{1}{132}\mathbf{1}'_{132}\right)$$

e $\boldsymbol{L}^{+}\boldsymbol{Y}$ corresponde ao vetor das médias amostrais com componentes

- $66.4118$; • $63.1842$;
- $58.9697$; • $61.2045$;
- $62.9722$; • $64.6515$.

Teremos

$$\boldsymbol{B}_2 = \boldsymbol{W}_2\boldsymbol{A}_2\boldsymbol{L}^{+}(\boldsymbol{L}^{+})'\boldsymbol{A}'_2\boldsymbol{W}'_2 = \begin{bmatrix} 0.0146 & -0.0033 \\ -0.0033 & 0.0167 \end{bmatrix}$$

e para o numerador da estatística $\mathcal{T}_2$,

$$({}^{oo}\boldsymbol{Y}_2)'(\boldsymbol{B}_2^{-1})^{oo}\boldsymbol{Y}_2 = 2169.713.$$

Quanto ao denominador obtém-se

$$S = 96590.0595$$

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

$$\mathcal{T}_{2,Obs} = \frac{2169.713}{96590.0595} = 0.0225.$$

Em relação à distribuição condicional de $\mathcal{T}_2$, que corresponde a $\overline{F}(z|3, 493)$ uma vez que $n = 499$, temos os quantis indicados na Tabela 4.13. Concluímos que se rejeita a hipótese para $\alpha = 0.1$ e $\alpha = 0.05$ e não se rejeita para $\alpha = 0.01$. Vamos assumir agora que $n^{\bullet} \geq 90$, digamos que $n^{\bullet} = 90$, o que significa que se tem $15$ observações por tratamento. Este valor não é muito elevado uma vez que os dados se referem a Estados do Brasil.

Tabela 4.13: Os quantis da distribuição condicional e limites superiores para os quantis de $\mathcal{T}_1$ e $\mathcal{T}_2$.

| **Valores de** $\alpha$ | 0.10 | 0.05 | 0.01 |
|---|---|---|---|
| $z_{1-\alpha}$ | 0.0127 | 0.0160 | 0.0232 |
| $z^{u}_{1-\alpha}$ | 0.0768 | 0.0969 | 0.1437 |

A Tabela 4.13 mostra os limites superiores dos quantis para a probabilidade $1-\alpha$, $z^{u}_{1-\alpha}$, da distribuição não condicional. Os resultados podem levar-nos a não rejeitar $H_{0,2,M}$ para os níveis usuais de significância, e consequentemente a tomar uma decisão contrária à que tínhamos tomado pela abordagem condicional, para $\alpha = 0.1$ e $\alpha = 0.05$. No entanto estes não são completamente conclusivos como já foi explicado anteriormente.

Assumindo que os valores da estatística de teste permanecem inalterados, teremos que ter as dimensões mínimas para as amostras apresentados na Tabela 4.14 para garantirmos a rejeição da hipótese nula.

Tabela 4.14: Valor minimo de $n^\bullet$ que leva à rejeição da hipótese $H_{0,2,M}$.

| **Valores de** $1-\alpha$ | 0.10 | 0.05 | 0.01 |
|---|---|---|---|
| $n^\bullet$ | 287 | 358 | 517 |

Estes valores podem parecer um pouco elevados como valores mínimos de $n^\bullet$, no entanto é importante referir que os dados se referem a diferentes Estados do Brasil e não apenas a uma cidade, como em exemplos considerados anteriormente. Assim, concluímos que para todo $n^\bullet \geq 517$ se rejeita $H_{0,2,M}$ para os níveis usuais de significância, o que significa que existe interação entre os dois fatores. A Figura 4.3 parece indicar efetivamente a existência de interação já que as linhas se cruzam.

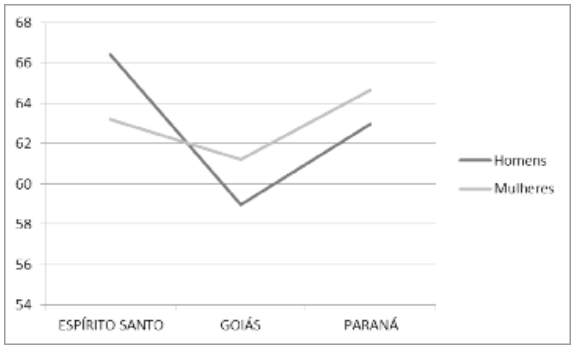


Figura 4.3: Interação entre os fatores.

Agora, para o **segundo fator**, teremos

$$^{oo}\boldsymbol{Y}_1 = \boldsymbol{W}_1\boldsymbol{A}_1\boldsymbol{L}^+\boldsymbol{Y} = \begin{bmatrix} -1.9975 \\ -0.3104 \end{bmatrix}$$

e

$$\boldsymbol{B}_1 = \boldsymbol{W}_1\boldsymbol{A}_1\boldsymbol{L}^+(\boldsymbol{L}^+)'\boldsymbol{A}_1'\boldsymbol{W}_1' = \begin{bmatrix} 0.0146 & -0.0033 \\ -0.0033 & 0.0167 \end{bmatrix}.$$

Assim, para o numerador da estatística $\mathcal{T}_1$, obtemos

$$(^{oo}\boldsymbol{Y}_1)'(\boldsymbol{B}_1^{-1})^{oo}\boldsymbol{Y}_1 = 309.8181.$$

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

$$\mathcal{T}_{1,Obs} = \frac{309.8181}{96590.0595} = 0.0032.$$

Se usarmos a distribuição condicional de $\mathcal{T}_1$, que corresponde também a $\overline{F}(z|3,493)$, obteremos os quantis indicados na Tabela 4.13 e portanto concluímos que pela abordagem condicional não rejeitamos $H_{0,1,M}$ para os níveis usuais de significância. A Tabela 4.13 mostra ainda os limites superiores para os quantis da probabilidade $1-\alpha$, $z^u_{1-\alpha}$, da distribuição não condicional $\overline{\overline{F}}_1(z)$. Como era de esperar não rejeitamos $H_{0,1,M}$ considerando estes limites superiores. Portanto a

idade de deteção da doença para este tipo de cancro não depende do Estado, ou seja, de região onde se vive.

## 4.4 Conclusões e discussão dos resultados obtidos

Nesta seção, tentámos abrir um "novo campo" com base no uso da distribuição Binomial, para modelos de fatores de efeitos fixos e modelos mistos, quando podem ocorrer falhas nas observações.

Verificámos, à semelhança do capítulo anterior em que a ocorrência das observações correspondia a processos de contagem, que também neste caso os quantis considerando a abordagem não condicional são superiores aos da abordagem clássica.

Podemos portanto concluir que ao considerarmos as dimensões das amostras como aleatórias, devido à possível ocorrência de falhas nas observações, teremos uma abordagem mais realista e mais robusta, uma vez que esta diminui a probabilidade de falsas rejeições.

# Capítulo 5

# Conclusões finais e trabalhos futuros

Com este trabalho propomos uma abordagem alternativa à ANOVA usual, baseada em situações em que as dimensões das amostras não são conhecidas à priori. Concluímos que nestes casos a ANOVA usual, usada rotineiramente por investigadores das mais diversas áreas da ciência, pode não ser a opção mais correta para a análise e tratamento dos dados.

A metodologia apresentada passa por, sempre que as dimensões das amostras são desconhecidas, considerar essas dimensões como realizações de variáveis aleatórias independentes. Mediante a situação prática em questão, assumimos duas distribuições para essas variáveis aleatórias:

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

- A distribuição Binomial, caso ocorram falhas nas observações.

Através dos Capítulos 3 e 4, mostrámos a relevância da abordagem não condicional, em evitar falsas rejeições. Podemos concluir portanto que a nossa abordagem, para além de ser mais realista quando as dimensões das amostras são desconhecidas, é igualmente mais robusta, uma vez que diminui a probabilidade de falsas rejeições.

Em termos de trabalho futuros, pensamos em estender este tratamento a outras situações reais onde se justifique considerar diferentes distribuições discretas para as dimensões das amostras. Por exemplo, assumir a distribuição Geométrica [Binomial Negativa] quando se consideram situações em que se desconhece o número de observações até à primeira falha [até à $s$-ésima falha]. Uma situação prática que poderá ser considerada neste caso é a de ambientes protegidos, em que cada vez que se recolhe uma observação há o risco de ter ocorrido alguma espécie de contaminação. Na área da investigação médica um exemplo que poderá justificar a utilização destas distribuições é a deslocação de pacientes com uma determinada patologia a um hospital e que correm o risco de sofrer algum tipo de infeção bacteriana.

Para o caso em que a distribuição Binomial é assumida como a distribuição adequada para as dimensões das amostras pretendemos realizar o estudo para modelos de efeitos aleatórios. Pretendemos ainda alargar a nossa abordagem ao caso em que a probabilidade da ocorrência de uma falha varia de tratamento para tratamento, ou seja, assumir que $N_i \sim B(r_i, 1-p_i)$, $i = 1, ..., m$.

É importante ainda referir que durante o nosso tratamento trabalhámos com as distribuições $\overline{F}$, uma vez que são mais "tratáveis" e estatisticamente equivalentes às distribuições $F$. Esta equivalência permite-nos considerar os nossos testes como testes $F$.

# Bibliografia

[1] Calinski, T. and Kageyama S. (2000). *Block Designs: A Randomization Approach: Vol. I, Analysis*. Lecture Note in Statistics, 150, New York : Springer-Verlag.

[2] Calinski, T. and Kageyama S. (2003). *Block Designs: A Randomization Approach: Vol. II, Design.* Lecture Note in Statistics, 170, New York : Springer-Verlag.

[3] 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.

[4] 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.

[5] 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. doi: 10.1007/s00184-014-0506-8.

[6] 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.

[7] Ferreira, D. (2006). *Variáveis pivot indutoras e componentes de variância em modelos normais ortogonais*. PhD Thesis, Universidade da Beira Interior, Covilhã, Portugal.

[8] Ferreira, S.S. (2006). *Inferência para modelos ortogonais com segregação*. PhD Thesis, Universidade da Beira Interior, Covilhã, Portugal.

[9] Ferreira, S. S. , Ferreira, D. , Fernandes, C. and Mexia, J.T. (2007). Orthogonal Mixed Models and Perfect Families of Symmetric Matrices. In proceedings of 56th session of the International Statistical Institute. Lisboa.

[10] 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.

[11] 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.

[12] Fonseca, M., Mexia, J.T. and Zmyślony, R. (2006). Binary Operations on Jordan algebras. *Linear Algebra and its Applications*, 117, 75-86.

[13] Fonseca, M. (2007). *Estimation and hypothesis testing in mixed linear models*. PhD Thesis. FCT - Universidade Nova de Lisboa, Lisboa, Portugal.

[14] 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.

[15] 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.

[16] Horn, R. , Johnson, C. (1985). *Matrix Analysis*. Cambridge University Press.

[17] Houtman, A.M. and Speed, T.P. (1983). Balance in designed experiments with orthogonal block structure. *Annals of Statistics*, 11(4), 1069-1983.

[18] Jacobson, N. (1953). *Lectures in Abstract Algebra*. Volume II-Linear Algebra. New York: D. Van Nostrand.

[19] 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.

[20] 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.

[21] Johnson, N.L. and Kotz, S. (1969). *Discrete distributions*. John Wiley $\&$ Sons, Inc., New York.

[22] 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.

[23] Khuri, A.I., Mathew, T. and Sinha, B. K. (1998). *Statistical Tests for Mixed Linear Models*. A Wiley-Interscience Publication. John Wiley $\&$ Sons, Inc., New York.

[24] Khuri, A.I. and Ghosh, M. (1990). *Minimal Sufficient Statistics for the Unbalanced Two- Fold Nested Model*. Statistics and Probability Letters, 10, 351-353.

[25] Meyer, P.L. (1970). *Introductory Probability and Statistical Applications*, Addison-Wesley Educational Publishers Inc, 2nd Revised Edition.

[26] Mejza, S. (1992). On some aspects of general balance in designed experiments. *Statistica*, 52, 263-278.

[27] Mexia, J.T. (1989). *Controlled Heteroscedasticity, Quocient Vector Spaces and $F$ Tests for Hypothesis on Mean Vectors*. Trabalhos de investigação, FCT/UNL, N$^o$ 1.

[28] 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.

[29] Mexia, J.T. (1995). *Introdução à Inferência Estatística Linear*. Centro de Estudos de Matemática Aplicada. Edições Lusófonas.

[30] 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.

[31] 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.

[32] Michalski, A. and Zmyślony, R. (1996). Testing hypothesis for variance components in mixed linear models. *Statistics*, 27, 297-310.

[33] Michalski, A. and Zmyślony, R. (1999). Testing hypothesis for linear functions of parameters in mixed linear models. *Tatra Mountain Mathematical Publications*, 17, 103- 110.

[34] Mood, M.M.; Graybill, F.A. and Boes, D.C. (1987). *Introdution to the Theory of Statistics*. McGraw-Hill.

[35] Moreira, E., Mexia J.T., Fonseca, M. and Zmyslony, R. (2009). $L$ models and multiple regressions designs. *Statistical Papers*, Springer, 50(4), 869-885.

[36] 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.

[37] Muller, K.E., and Stewart, P.W. (2006). *Linear Model Theory; Univariate, Multivariate, and Mixed Models*. John Wiley & Sons, Inc. Hoboken, New Jersey.

[38] 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.

[39] 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.

[40] Nunes, C. (2005). *Testes F e relacionados em modelos mistos com cross-nesting ortogonal*. PhD Thesis, Universidade da Beira Interior, Covilhã, Portugal.

[41] Nunes, C. and Mexia, J.T. (2006). Non-central generelized F distributions. *Discussiones Mathematicae -Probability and Statistics*, 26(I), 297-310.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] Pestana, D.D. e Velosa, S.F. (2006). *Introdução à Probabilidade e à Estatística*. Volune I, 2ª Edição. Lisboa: Fundação Calouste Gulbenkian.

[51] Pollock, D.S.G. (1979). *The Algebra of Econometrics.* John Wiley $\&$ Sons, Inc., New York.

[52] Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*. Second Edition; John Wiley $\&$ Sons, Inc., New York.

[53] Rao, C., Kleve, J. (1988). *Estimation of variance components and applications*. North Holland, Elsevier - Amsterdam.

[54] Robbins, H. (1948). Mixture of distribution. *The Annals of Mathematical Statistics*, 19, 360-369.

[55] 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.

[56] Rodrigues, P.C. and Mexia, J.T. (2006). ANOVA using commutative Jordan algebras. *Discussiones Mathematicae -Probability and Statistics*, 26, 179-191.

[57] 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.

[58] 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.

[59] 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.

[60] Santos, C. (2012). *Error orthogonal models: Structure, Operations and Inference*. PhD thesis. Universidade da Beira Interior, Covilhã, Portugal.

[61] Salvador, D. (2013). *Modelação de Matrizes Estocásticas Simétricas, Operadores do tipo vec*. PhD Thesis, FCT - Universidade Nova Lisboa, Lisboa, Portugal.

[62] Scheffé, H. (1959). *The analysis of variance*, Wiley series in Probability and Statistics, John Wiley $\&$ Sons, New York.

[63] Schott, J.R. (1997). *Matrix Analysis for Statistics*. Wiley - Interscience. John Wiley $\&$ Sons, Inc., New York.

[64] Searle, S.R., Casella, G. e McCulloch, C.E. (1992). *Variance Components.* Wiley series in Probability and statistics. John Wiley $\&$ Sons, Inc., New York.

[65] Seely, J. (1970a). Linear spaces and unbiased estimation. *The Annals of Mathematical Statistics*, (41), 1725- 1734.

[66] Seely, J. (1970b). Linear spaces and unbiased estimation. An aplication to the mixed linear model. *The Annals of Mathematical Statistics*, 41, 1735-1748.

[67] Seely, J. (1971). Quadratic subspaces and completeness. *The Annals of Mathematical Statistics*, 42, 710-721.

[68] Seely, J. e Zyskind, G. (1971). Linear spaces and minimum variance estimators. *The Annals of Mathematical Statistics*, 42(2), 691-703.

[69] Seely, J. (1977). Minimal sufficient statistics and completeness for multivariate normal families. *Sankhya*, 39, 170-185.

[70] Steeb, W.H. (1991). *Kronecker Products of Matrices and Applications*. BI-Wissenschaftsvlg.

[71] 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.

[72] 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.

# Apêndice A

# Anexos

## A.1 - Tabelas de frequência referentes às Subseções 3.2.2.2 e 4.2.2.2.

Tabela A.1: Cancro do tecidos moles do tórax

| **Idades** | $1-4$ | $5-9$ | $10-14$ | $15-19$ | $20-24$ | $25-29$ | $30-34$ | $35-39$ | $40-44$ |
|---|---|---|---|---|---|---|---|---|---|
| **Média das idades** | 2 | 7 | 12 | 17 | 22 | 27 | 32 | 37 | 42 |
| **Pacientes** | 0 | 0 | 0 | 2 | 1 | 2 | 1 | 1 | 2 |
| **Idades** | $45-49$ | $50-54$ | $55-59$ | $60-64$ | $65-69$ | $70-74$ | $75-79$ | $80-84$ | $85+$ |
| **Média das idades** | 47 | 52 | 57 | 62 | 67 | 72 | 77 | 82 | 87 |
| **Pacientes** | 2 | 0 | 0 | 0 | 2 | 1 | 1 | 2 | 1 |

Tabela A.2: Cancro do trato intestinal

| **Idades** | $1-4$ | $5-9$ | $10-14$ | $15-19$ | $20-24$ | $25-29$ | $30-34$ | $35-39$ | $40-44$ |
|---|---|---|---|---|---|---|---|---|---|
| **Média das idades** | 2 | 7 | 12 | 17 | 22 | 27 | 32 | 37 | 42 |
| **Pacientes** | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 |
| **Idades** | $45-49$ | $50-54$ | $55-59$ | $60-64$ | $65-69$ | $70-74$ | $75-79$ | $80-84$ | $85+$ |
| **Média das idades** | 47 | 52 | 57 | 62 | 67 | 72 | 77 | 82 | 87 |
| **Pacientes** | 3 | 3 | 2 | 1 | 3 | 0 | 1 | 2 | 4 |

Tabela A.3: Cancro da cavidade nasal

| **Idades** | $1-4$ | $5-9$ | $10-14$ | $15-19$ | $20-24$ | $25-29$ | $30-34$ | $35-39$ | $40-44$ |
|---|---|---|---|---|---|---|---|---|---|
| **Média das idades** | 2 | 7 | 12 | 17 | 22 | 27 | 32 | 37 | 42 |
| **Pacientes** | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 2 | 0 |
| **Idades** | $45-49$ | $50-54$ | $55-59$ | $60-64$ | $65-69$ | $70-74$ | $75-79$ | $80-84$ | $85+$ |
| **Média das idades** | 47 | 52 | 57 | 62 | 67 | 72 | 77 | 82 | 87 |
| **Pacientes** | 1 | 4 | 0 | 4 | 1 | 3 | 3 | 1 | 4 |

# A.2 - Tabelas de frequência referentes à Subseção 3.2.3.2.

Tabela A.4: Homens com cancro na amígdala

| **Idades** | $1-4$ | $5-9$ | $10-14$ | $15-19$ | $20-24$ | $25-29$ | $30-34$ | $35-39$ | $40-44$ |
|---|---|---|---|---|---|---|---|---|---|
| **Médias das idades** | 2 | 7 | 12 | 17 | 22 | 27 | 32 | 37 | 42 |
| **Pacientes** | 5 | 1 | 1 | 0 | 3 | 5 | 1 | 1 | 5 |
| **Idades** | $45-49$ | $50-54$ | $55-59$ | $60-64$ | $65-69$ | $70-74$ | $75-79$ | $80-84$ | $85+$ |
| **Médias das idades** | 47 | 52 | 57 | 62 | 67 | 72 | 77 | 82 | 87 |
| **Pacientes** | 3 | 5 | 6 | 3 | 1 | 4 | 0 | 1 | 6 |

Tabela A.5: Mulheres com cancro na amígdala

| **Idades** | $1-4$ | $5-9$ | $10-14$ | $15-19$ | $20-24$ | $25-29$ | $30-34$ | $35-39$ | $40-44$ |
|---|---|---|---|---|---|---|---|---|---|
| **Média das idades** | 2 | 7 | 12 | 17 | 22 | 27 | 32 | 37 | 42 |
| **Pacientes** | 3 | 3 | 0 | 1 | 1 | 2 | 2 | 0 | 2 |
| **Idade** | $45-49$ | $50-54$ | $55-59$ | $60-64$ | $65-69$ | $70-74$ | $75-79$ | $80-84$ | $85+$ |
| **Média das Idades** | 47 | 52 | 57 | 62 | 67 | 72 | 77 | 82 | 87 |
| **Pacientes** | 4 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 |

Tabela A.6: Homens com cancro na cavidade nasal e do ouvido médio

| **Idades** | $1-4$ | $5-9$ | $10-14$ | $15-19$ | $20-24$ | $25-29$ | $30-34$ | $35-39$ | $40-44$ |
|---|---|---|---|---|---|---|---|---|---|
| **Média das Idades** | 2 | 7 | 12 | 17 | 22 | 27 | 32 | 37 | 42 |
| **Pacientes** | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 3 | 0 |
| **Idades** | $45-49$ | $50-54$ | $55-59$ | $60-64$ | $65-69$ | $70-74$ | $75-79$ | $80-84$ | $85+$ |
| **Média das Idades** | 47 | 52 | 57 | 62 | 67 | 72 | 77 | 82 | 87 |
| **Pacientes** | 0 | 1 | 0 | 3 | 0 | 0 | 2 | 0 | 2 |

Tabela A.7: Mulheres com cancro na cavidade nasal e do ouvido médio

| **Idades** | $1-4$ | $5-9$ | $10-14$ | $15-19$ | $20-24$ | $25-29$ | $30-34$ | $35-39$ | $40-44$ |
|---|---|---|---|---|---|---|---|---|---|
| **Média das idades** | 2 | 7 | 12 | 17 | 22 | 27 | 32 | 37 | 42 |
| **Pacientes** | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| **Idades** | $45-49$ | $50-54$ | $55-59$ | $60-64$ | $65-69$ | $70-74$ | $75-79$ | $80-84$ | $85+$ |
| **Médias das idades** | 47 | 52 | 57 | 62 | 67 | 72 | 77 | 82 | 87 |
| **Pacientes** | 2 | 3 | 0 | 1 | 1 | 3 | 2 | 1 | 2 |

Tabela A.8: Homens com cancro no timo

| **Idades** | $1-4$ | $5-9$ | $10-14$ | $15-19$ | $20-24$ | $25-29$ | $30-34$ | $35-39$ | $40-44$ |
|---|---|---|---|---|---|---|---|---|---|
| **Média das idades** | 2 | 7 | 12 | 17 | 22 | 27 | 32 | 37 | 42 |
| **Pacientes** | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| **Idades** | $45-49$ | $50-54$ | $55-59$ | $60-64$ | $65-69$ | $70-74$ | $75-79$ | $80-84$ | $85+$ |
| **Média das idades** | 47 | 52 | 57 | 62 | 67 | 72 | 77 | 82 | 87 |
| **Pacientes** | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |

Tabela A.9: Mulheres com cancro no timo

| **Idades** | $1-4$ | $5-9$ | $10-14$ | $15-19$ | $20-24$ | $25-29$ | $30-34$ | $35-39$ | $40-44$ |
|---|---|---|---|---|---|---|---|---|---|
| **Média das idades** | 2 | 7 | 12 | 17 | 22 | 27 | 32 | 37 | 42 |
| **Pacientes** | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| **Idades** | $45-49$ | $50-54$ | $55-59$ | $60-64$ | $65-69$ | $70-74$ | $75-79$ | $80-84$ | $85+$ |
| **Média das idades** | 47 | 52 | 57 | 62 | 67 | 72 | 77 | 82 | 87 |
| **Pacientes** | 1 | 0 | 1 | 0 | 3 | 1 | 0 | 0 | 1 |

Tabela A.10: Homens com cancro no coração, mediasto e pleura

| **Idades** | $1-4$ | $5-9$ | $10-14$ | $15-19$ | $20-24$ | $25-29$ | $30-34$ | $35-39$ | $40-44$ |
|---|---|---|---|---|---|---|---|---|---|
| **Média das idades** | 2 | 7 | 12 | 17 | 22 | 27 | 32 | 37 | 42 |
| **Pacientes** | 0 | 0 | 0 | 0 | 1 | 1 | 2 | 1 | 3 |
| **Idades** | $45-49$ | $50-54$ | $55-59$ | $60-64$ | $65-69$ | $70-74$ | $75-79$ | $80-84$ | $85+$ |
| **Média das idades** | 47 | 52 | 57 | 62 | 67 | 72 | 77 | 82 | 87 |
| **Pacientes** | 0 | 3 | 0 | 1 | 1 | 1 | 0 | 2 | 2 |

Tabela A.11: Mulheres com cancro no coração, mediasto e pleura

| **Idades** | $1-4$ | $5-9$ | $10-14$ | $15-19$ | $20-24$ | $25-29$ | $30-34$ | $35-39$ | $40-44$ |
|---|---|---|---|---|---|---|---|---|---|
| **Média das idades** | 2 | 7 | 12 | 17 | 22 | 27 | 32 | 37 | 42 |
| **Pacientes** | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| **Idades** | $45-49$ | $50-54$ | $55-59$ | $60-64$ | $65-69$ | $70-74$ | $75-79$ | $80-84$ | $85+$ |
| **Média das idades** | 47 | 52 | 57 | 62 | 67 | 72 | 77 | 82 | 87 |
| **Pacientes** | 3 | 1 | 0 | 2 | 2 | 0 | 0 | 1 | 3 |

# A.3 - Tabelas de frequência referentes à Subseção 3.3.2.

Tabela A.12: Cancro do corpo do estômago

| **Idades** | $1-4$ | $5-9$ | $10-14$ | $15-19$ | $20-24$ | $25-29$ | $30-34$ | $35-39$ | $40-44$ |
|---|---|---|---|---|---|---|---|---|---|
| **Média das idades** | 2 | 7 | 12 | 17 | 22 | 27 | 32 | 37 | 42 |
| **Pacientes** | 0 | 1 | 0 | 1 | 0 | 0 | 3 | 4 | 2 |
| **Idades** | $45-49$ | $50-54$ | $55-59$ | $60-64$ | $65-69$ | $70-74$ | $75-79$ | $80-84$ | $85+$ |
| **Média das idades** | 47 | 52 | 57 | 62 | 67 | 72 | 77 | 82 | 87 |
| **Pacientes** | 13 | 4 | 15 | 4 | 11 | 8 | 9 | 5 | 11 |

Tabela A.13: Cancro do encéfalo

| **Idades** | $1-4$ | $5-9$ | $10-14$ | $15-19$ | $20-24$ | $25-29$ | $30-34$ | $35-39$ | $40-44$ |
|---|---|---|---|---|---|---|---|---|---|
| **Média das idades** | 2 | 7 | 12 | 17 | 22 | 27 | 32 | 37 | 42 |
| **Pacientes** | 2 | 5 | 1 | 2 | 6 | 1 | 3 | 7 | 7 |
| **Idades** | $45-49$ | $50-54$ | $55-59$ | $60-64$ | $65-69$ | $70-74$ | $75-79$ | $80-84$ | $85+$ |
| **Média das idades** | 47 | 52 | 57 | 62 | 67 | 72 | 77 | 82 | 87 |
| **Pacientes** | 3 | 11 | 10 | 8 | 11 | 7 | 5 | 1 | 3 |

Tabela A.14: Cancro da medula espinhal e outras partes S.N.C.

| **Idades** | $1-4$ | $5-9$ | $10-14$ | $15-19$ | $20-24$ | $25-29$ | $30-34$ | $35-39$ | $40-44$ |
|---|---|---|---|---|---|---|---|---|---|
| **Média das idades** | 2 | 7 | 12 | 17 | 22 | 27 | 32 | 37 | 42 |
| **Pacientes** | 4 | 2 | 0 | 1 | 1 | 2 | 2 | 4 | 3 |
| **Idades** | $45-49$ | $50-54$ | $55-59$ | $60-64$ | $65-69$ | $70-74$ | $75-79$ | $80-84$ | $85+$ |
| **Média das idades** | 47 | 52 | 57 | 62 | 67 | 72 | 77 | 82 | 87 |
| **Pacientes** | 1 | 5 | 4 | 2 | 0 | 5 | 5 | 1 | 0 |

Tabela A.15: Melanoma malígno do tronco

| **Idades** | $1-4$ | $5-9$ | $10-14$ | $15-19$ | $20-24$ | $25-29$ | $30-34$ | $35-39$ | $40-44$ |
|---|---|---|---|---|---|---|---|---|---|
| **Média das idades** | 2 | 7 | 12 | 17 | 22 | 27 | 32 | 37 | 42 |
| **Pacientes** | 0 | 0 | 1 | 0 | 1 | 1 | 7 | 5 | 6 |
| **Idades** | $45-49$ | $50-54$ | $55-59$ | $60-64$ | $65-69$ | $70-74$ | $75-79$ | $80-84$ | $85+$ |
| **Média das idades** | 47 | 52 | 57 | 62 | 67 | 72 | 77 | 82 | 87 |
| **Pacientes** | 5 | 7 | 6 | 2 | 7 | 4 | 5 | 3 | 47 |

Tabela A.16: Cancro do cólon ascendente

| **Idades** | $1-4$ | $5-9$ | $10-14$ | $15-19$ | $20-24$ | $25-29$ | $30-34$ | $35-39$ | $40-44$ |
|---|---|---|---|---|---|---|---|---|---|
| **Média das idades** | 2 | 7 | 12 | 17 | 22 | 27 | 32 | 37 | 42 |
| **Pacientes** | 0 | 0 | 0 | 0 | 0 | 2 | 1 | 4 | 6 |
| **Idades** | $45-49$ | $50-54$ | $55-59$ | $60-64$ | $65-69$ | $70-74$ | $75-79$ | $80-84$ | $85+$ |
| **Média das idades** | 47 | 52 | 57 | 62 | 67 | 72 | 77 | 82 | 87 |
| **Pacientes** | 7 | 9 | 14 | 20 | 20 | 23 | 21 | 16 | 58 |

Tabela A.17: Cancro do lobo superior, brônquios ou pulmão

| **Idades** | $1-4$ | $5-9$ | $10-14$ | $15-19$ | $20-24$ | $25-29$ | $30-34$ | $35-39$ | $40-44$ |
|---|---|---|---|---|---|---|---|---|---|
| **Média das idades** | 2 | 7 | 12 | 17 | 22 | 27 | 32 | 37 | 42 |
| **Pacientes** | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 2 | 3 |
| **Idades** | $45-49$ | $50-54$ | $55-59$ | $60-64$ | $65-69$ | $70-74$ | $75-79$ | $80-84$ | $85+$ |
| **Média das idades** | 47 | 52 | 57 | 62 | 67 | 72 | 77 | 82 | 87 |
| **Pacientes** | 7 | 14 | 17 | 15 | 28 | 24 | 24 | 10 | 9 |

# A.4 - Tabelas de frequência referentes à Subseção 3.4.2.

Tabela A.18: Homens com cancro nos ossos e articulações dos membros

| **Idades** | $1-4$ | $5-9$ | $10-14$ | $15-19$ | $20-24$ | $25-29$ | $30-34$ | $35-39$ | $40-44$ |
|---|---|---|---|---|---|---|---|---|---|
| **Média das idades** | 2 | 7 | 12 | 17 | 22 | 27 | 32 | 37 | 42 |
| **Pacientes** | 0 | 5 | 8 | 5 | 2 | 2 | 1 | 3 | 1 |
| **Idades** | $45-49$ | $50-54$ | $55-59$ | $60-64$ | $65-69$ | $70-74$ | $75-79$ | $80-84$ | $85+$ |
| **Média das idades** | 47 | 52 | 57 | 62 | 67 | 72 | 77 | 82 | 87 |
| **Pacientes** | 2 | 0 | 1 | 2 | 0 | 0 | 1 | 0 | 1 |

Tabela A.19: Mulheres com cancro nos ossos e articulações dos membros

| **Idades** | $1-4$ | $5-9$ | $10-14$ | $15-19$ | $20-24$ | $25-29$ | $30-34$ | $35-39$ | $40-44$ |
|---|---|---|---|---|---|---|---|---|---|
| **Média das idades** | 2 | 7 | 12 | 17 | 22 | 27 | 32 | 37 | 42 |
| **Pacientes** | 0 | 1 | 3 | 2 | 1 | 1 | 2 | 3 | 2 |
| **Idades** | $45-49$ | $50-54$ | $55-59$ | $60-64$ | $65-69$ | $70-74$ | $75-79$ | $80-84$ | $85+$ |
| **Média das idades** | 47 | 52 | 57 | 62 | 67 | 72 | 77 | 82 | 87 |
| **Pacientes** | 0 | 1 | 3 | 1 | 1 | 1 | 0 | 0 | 3 |

Tabela A.20: Homens com cancro na medula espinhal e outras partes S.N.C.

| **Idades** | $1-4$ | $5-9$ | $10-14$ | $15-19$ | $20-24$ | $25-29$ | $30-34$ | $35-39$ | $40-44$ |
|---|---|---|---|---|---|---|---|---|---|
| **Média das idades** | 2 | 7 | 12 | 17 | 22 | 27 | 32 | 37 | 42 |
| **Pacientes** | 3 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| **Idades** | $45-49$ | $50-54$ | $55-59$ | $60-64$ | $65-69$ | $70-74$ | $75-79$ | $80-84$ | $85+$ |
| **Média das idades** | 47 | 52 | 57 | 62 | 67 | 72 | 77 | 82 | 87 |
| **Pacientes** | 1 | 3 | 0 | 0 | 0 | 3 | 3 | 1 | 0 |

Tabela A.21: Mulheres com cancro na medula espinhal e outras partes S.N.C.

| **Idades** | $1-4$ | $5-9$ | $10-14$ | $15-19$ | $20-24$ | $25-29$ | $30-34$ | $35-39$ | $40-44$ |
|---|---|---|---|---|---|---|---|---|---|
| **Média das idades** | 2 | 7 | 12 | 17 | 22 | 27 | 32 | 37 | 42 |
| **Pacientes** | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 3 | 2 |
| **Idades** | $45-49$ | $50-54$ | $55-59$ | $60-64$ | $65-69$ | $70-74$ | $75-79$ | $80-84$ | $85+$ |
| **Média das idades** | 47 | 52 | 57 | 62 | 67 | 72 | 77 | 82 | 87 |
| **Pacientes** | 0 | 2 | 4 | 2 | 0 | 2 | 2 | 0 | 0 |

Tabela A.22: Homens com linfomas de células T cutâneas e periféricas

| **Idades** | $1-4$ | $5-9$ | $10-14$ | $15-19$ | $20-24$ | $25-29$ | $30-34$ | $35-39$ | $40-44$ |
|---|---|---|---|---|---|---|---|---|---|
| **Média das idades** | 2 | 7 | 12 | 17 | 22 | 27 | 32 | 37 | 42 |
| **Pacientes** | 0 | 0 | 0 | 1 | 3 | 2 | 0 | 1 | 3 |
| **Idades** | $45-49$ | $50-54$ | $55-59$ | $60-64$ | $65-69$ | $70-74$ | $75-79$ | $80-84$ | $85+$ |
| **Média das idades** | 47 | 52 | 57 | 62 | 67 | 72 | 77 | 82 | 87 |
| **Pacientes** | 2 | 1 | 1 | 5 | 2 | 6 | 2 | 1 | 4 |

Tabela A.23: Mulheres com linfomas de células T cutâneas e periféricas

| **Idades** | $1-4$ | $5-9$ | $10-14$ | $15-19$ | $20-24$ | $25-29$ | $30-34$ | $35-39$ | $40-44$ |
|---|---|---|---|---|---|---|---|---|---|
| **Média das idades** | 2 | 7 | 12 | 17 | 22 | 27 | 32 | 37 | 42 |
| **Pacientes** | 2 | 1 | 0 | 0 | 2 | 2 | 1 | 4 | 2 |
| **Idades** | $45-49$ | $50-54$ | $55-59$ | $60-64$ | $65-69$ | $70-74$ | $75-79$ | $80-84$ | $85+$ |
| **Média das idades** | 47 | 52 | 57 | 62 | 67 | 72 | 77 | 82 | 87 |
| **Pacientes** | 2 | 2 | 0 | 1 | 0 | 2 | 2 | 1 | 3 |

# A.5 - Tabelas de frequência referentes à Subseção 4.2.1.2.

Tabela A.24: Cancro dos ossos longos dos membros inferiores

| **Idades** | $1-4$ | $5-9$ | $10-14$ | $15-19$ | $20-24$ | $25-29$ | $30-34$ | $35-39$ | $40-44$ |
|---|---|---|---|---|---|---|---|---|---|
| **Média das idades** | 2 | 7 | 12 | 17 | 22 | 27 | 32 | 37 | 42 |
| **Pacientes** | 0 | 4 | 4 | 7 | 1 | 2 | 1 | 4 | 2 |
| **Idades** | $45-49$ | $50-54$ | $55-59$ | $60-64$ | $65-69$ | $70-74$ | $75-79$ | $80-84$ | $85+$ |
| **Média das idades** | 47 | 52 | 57 | 62 | 67 | 72 | 77 | 82 | 87 |
| **Pacientes** | 1 | 0 | 2 | 1 | 0 | 0 | 0 | 0 | 3 |

Tabela A.25: Cancro da parede lateral da bexiga urinária

| **Idades** | $1-4$ | $5-9$ | $10-14$ | $15-19$ | $20-24$ | $25-29$ | $30-34$ | $35-39$ | $40-44$ |
|---|---|---|---|---|---|---|---|---|---|
| **Média das idades** | 2 | 7 | 12 | 17 | 22 | 27 | 32 | 37 | 42 |
| **Pacientes** | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 1 |
| **Idades** | $45-49$ | $50-54$ | $55-59$ | $60-64$ | $65-69$ | $70-74$ | $75-79$ | $80-84$ | $85+$ |
| **Média das idades** | 47 | 52 | 57 | 62 | 67 | 72 | 77 | 82 | 87 |
| **Pacientes** | 0 | 1 | 6 | 8 | 1 | 5 | 3 | 2 | 3 |

Tabela A.26: Cancro do corpo do pâncreas

| **Idades** | $1-4$ | $5-9$ | $10-14$ | $15-19$ | $20-24$ | $25-29$ | $30-34$ | $35-39$ | $40-44$ |
|---|---|---|---|---|---|---|---|---|---|
| **Média das idades** | 2 | 7 | 12 | 17 | 22 | 27 | 32 | 37 | 42 |
| **Pacientes** | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| **Idades** | $45-49$ | $50-54$ | $55-59$ | $60-64$ | $65-69$ | $70-74$ | $75-79$ | $80-84$ | $85+$ |
| **Média das idades** | 47 | 52 | 57 | 62 | 67 | 72 | 77 | 82 | 87 |
| **Pacientes** | 2 | 2 | 4 | 3 | 6 | 4 | 4 | 1 | 2 |

# A.6 - Tabelas de frequência referentes à Subseção 4.2.3.2.

Tabela A.27: Homens com cancro na cavidade nasal e do ouvido médio

| **Idades** | $1-4$ | $5-9$ | $10-14$ | $15-19$ | $20-24$ | $25-29$ | $30-34$ | $35-39$ | $40-44$ |
|---|---|---|---|---|---|---|---|---|---|
| **Média das idades** | 2 | 7 | 12 | 17 | 22 | 27 | 32 | 37 | 42 |
| **Pacientes** | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 3 | 0 |
| **Idades** | $45-49$ | $50-54$ | $55-59$ | $60-64$ | $65-69$ | $70-74$ | $75-79$ | $80-84$ | $85+$ |
| **Média das idades** | 47 | 52 | 57 | 62 | 67 | 72 | 77 | 82 | 87 |
| **Pacientes** | 0 | 1 | 0 | 3 | 0 | 0 | 2 | 0 | 2 |

Tabela A.28: Mulheres com cancro na cavidade nasal e do ouvido médio

| **Idades** | $1-4$ | $5-9$ | $10-14$ | $15-19$ | $20-24$ | $25-29$ | $30-34$ | $35-39$ | $40-44$ |
|---|---|---|---|---|---|---|---|---|---|
| **Média das idades** | 2 | 7 | 12 | 17 | 22 | 27 | 32 | 37 | 42 |
| **Pacientes** | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| **Idades** | $45-49$ | $50-54$ | $55-59$ | $60-64$ | $65-69$ | $70-74$ | $75-79$ | $80-84$ | $85+$ |
| **Média das idades** | 47 | 52 | 57 | 62 | 67 | 72 | 77 | 82 | 87 |
| **Pacientes** | 2 | 3 | 0 | 1 | 1 | 3 | 2 | 1 | 2 |

Tabela A.29: Homens com cancro na meninge

| **Idades** | $1-4$ | $5-9$ | $10-14$ | $15-19$ | $20-24$ | $25-29$ | $30-34$ | $35-39$ | $40-44$ |
|---|---|---|---|---|---|---|---|---|---|
| **Média das idades** | 2 | 7 | 12 | 17 | 22 | 27 | 32 | 37 | 42 |
| **Pacientes** | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 |
| **Idades** | $45-49$ | $50-54$ | $55-59$ | $60-64$ | $65-69$ | $70-74$ | $75-79$ | $80-84$ | $85+$ |
| **Média das idades** | 47 | 52 | 57 | 62 | 67 | 72 | 77 | 82 | 87 |
| **Pacientes** | 0 | 1 | 1 | 1 | 0 | 0 | 2 | 0 | 0 |

Tabela A.30: Mulheres com cancro na meninge

| **Idades** | $1-4$ | $5-9$ | $10-14$ | $15-19$ | $20-24$ | $25-29$ | $30-34$ | $35-39$ | $40-44$ |
|---|---|---|---|---|---|---|---|---|---|
| **Média das idades** | 2 | 7 | 12 | 17 | 22 | 27 | 32 | 37 | 42 |
| **Pacientes** | 1 | 2 | 0 | 0 | 0 | 2 | 0 | 1 | 1 |
| **Idades** | $45-49$ | $50-54$ | $55-59$ | $60-64$ | $65-69$ | $70-74$ | $75-79$ | $80-84$ | $85+$ |
| **Média das idades** | 47 | 52 | 57 | 62 | 67 | 72 | 77 | 82 | 87 |
| **Pacientes** | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 1 | 0 |

Tabela A.31: Homens com cancro no coração, mediastino e pleura

| **Idades** | $1-4$ | $5-9$ | $10-14$ | $15-19$ | $20-24$ | $25-29$ | $30-34$ | $35-39$ | $40-44$ |
|---|---|---|---|---|---|---|---|---|---|
| **Média das idades** | 2 | 7 | 12 | 17 | 22 | 27 | 32 | 37 | 42 |
| **Pacientes** | 0 | 0 | 0 | 0 | 1 | 1 | 2 | 1 | 3 |
| **Idades** | $45-49$ | $50-54$ | $55-59$ | $60-64$ | $65-69$ | $70-74$ | $75-79$ | $80-84$ | $85+$ |
| **Média das idades** | 47 | 52 | 57 | 62 | 67 | 72 | 77 | 82 | 87 |
| **Pacientes** | 0 | 3 | 0 | 1 | 1 | 1 | 0 | 2 | 2 |

Tabela A.32: Mulheres com cancro no coração, mediastino e pleura

| **Idades** | $1-4$ | $5-9$ | $10-14$ | $15-19$ | $20-24$ | $25-29$ | $30-34$ | $35-39$ | $40-44$ |
|---|---|---|---|---|---|---|---|---|---|
| **Média das idades** | 2 | 7 | 12 | 17 | 22 | 27 | 32 | 37 | 42 |
| **Pacientes** | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| **Idades** | $45-49$ | $50-54$ | $55-59$ | $60-64$ | $65-69$ | $70-74$ | $75-79$ | $80-84$ | $85+$ |
| **Média das idades** | 47 | 52 | 57 | 62 | 67 | 72 | 77 | 82 | 87 |
| **Pacientes** | 3 | 1 | 0 | 2 | 2 | 0 | 0 | 1 | 3 |

# A.7 - Tabelas de frequência referentes à Subseção 4.3.2.

Tabela A.33: Espírito Santo-Homens com cancro no cólon

| **Idades** | $1-4$ | $5-9$ | $10-14$ | $15-19$ | $20-24$ | $25-29$ | $30-34$ | $35-39$ | $40-44$ |
|---|---|---|---|---|---|---|---|---|---|
| **Média das idades** | 2 | 7 | 12 | 17 | 22 | 27 | 32 | 37 | 42 |
| **Pacientes** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| **Idades** | $45-49$ | $50-54$ | $55-59$ | $60-64$ | $65-69$ | $70-74$ | $75-79$ | $80-84$ | $85+$ |
| **Média das idades** | 47 | 52 | 57 | 62 | 67 | 72 | 77 | 82 | 87 |
| **Pacientes** | 1 | 3 | 5 | 5 | 6 | 6 | 0 | 5 | 2 |

Tabela A.34: Espírito Santo-Mulheres com cancro no cólon

| **Idades** | $1-4$ | $5-9$ | $10-14$ | $15-19$ | $20-24$ | $25-29$ | $30-34$ | $35-39$ | $40-44$ |
|---|---|---|---|---|---|---|---|---|---|
| **Média das idades** | 2 | 7 | 12 | 17 | 22 | 27 | 32 | 37 | 42 |
| **Pacientes** | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 2 |
| **Idades** | $45-49$ | $50-54$ | $55-59$ | $60-64$ | $65-69$ | $70-74$ | $75-79$ | $80-84$ | $85+$ |
| **Média das idades** | 47 | 52 | 57 | 62 | 67 | 72 | 77 | 82 | 87 |
| **Pacientes** | 5 | 3 | 5 | 3 | 4 | 5 | 5 | 3 | 2 |

Tabela A.35: Goiás-Homens com cancro no cólon

| **Idades** | $1-4$ | $5-9$ | $10-14$ | $15-19$ | $20-24$ | $25-29$ | $30-34$ | $35-39$ | $40-44$ |
|---|---|---|---|---|---|---|---|---|---|
| **Média das idades** | 2 | 7 | 12 | 17 | 22 | 27 | 32 | 37 | 42 |
| **Pacientes** | 0 | 0 | 1 | 1 | 0 | 1 | 2 | 4 | 12 |
| **Idades** | $45-49$ | $50-54$ | $55-59$ | $60-64$ | $65-69$ | $70-74$ | $75-79$ | $80-84$ | $85+$ |
| **Média das idades** | 47 | 52 | 57 | 62 | 67 | 72 | 77 | 82 | 87 |
| **Pacientes** | 7 | 7 | 17 | 9 | 6 | 12 | 15 | 2 | 3 |

Tabela A.36: Goiás-Mulheres com cancro no cólon

| **Idades** | $1-4$ | $5-9$ | $10-14$ | $15-19$ | $20-24$ | $25-29$ | $30-34$ | $35-39$ | $40-44$ |
|---|---|---|---|---|---|---|---|---|---|
| **Média das idades** | 2 | 7 | 12 | 17 | 22 | 27 | 32 | 37 | 42 |
| **Pacientes** | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 5 | 7 |
| **Age** | $45-49$ | $50-54$ | $55-59$ | $60-64$ | $65-69$ | $70-74$ | $75-79$ | $80-84$ | $85+$ |
| **Média das idades** | 47 | 52 | 57 | 62 | 67 | 72 | 77 | 82 | 87 |
| **Pacientes** | 10 | 6 | 6 | 14 | 8 | 12 | 10 | 3 | 5 |

Tabela A.37: Paraná-Homens com cancro no cólon

| **Idades** | $1-4$ | $5-9$ | $10-14$ | $15-19$ | $20-24$ | $25-29$ | $30-34$ | $35-39$ | $40-44$ |
|---|---|---|---|---|---|---|---|---|---|
| **Média das idades** | 2 | 7 | 12 | 17 | 22 | 27 | 32 | 37 | 42 |
| **Pacientes** | 0 | 0 | 0 | 0 | 0 | 1 | 3 | 2 | 0 |
| **Idades** | $45-49$ | $50-54$ | $55-59$ | $60-64$ | $65-69$ | $70-74$ | $75-79$ | $80-84$ | $85+$ |
| **Média das idades** | 47 | 52 | 57 | 62 | 67 | 72 | 77 | 82 | 87 |
| **Pacientes** | 8 | 10 | 18 | 11 | 18 | 24 | 4 | 5 | 4 |

Tabela A.38: Paraná-Mulheres com cancro no cólon

| **Idades** | $1-4$ | $5-9$ | $10-14$ | $15-19$ | $20-24$ | $25-29$ | $30-34$ | $35-39$ | $40-44$ |
|---|---|---|---|---|---|---|---|---|---|
| **Média das idades** | 2 | 7 | 12 | 17 | 22 | 27 | 32 | 37 | 42 |
| **Pacientes** | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 7 |
| **Idades** | $45-49$ | $50-54$ | $55-59$ | $60-64$ | $65-69$ | $70-74$ | $75-79$ | $80-84$ | $85+$ |
| **Média das idades** | 47 | 52 | 57 | 62 | 67 | 72 | 77 | 82 | 87 |
| **Pacientes** | 9 | 14 | 17 | 14 | 14 | 17 | 15 | 14 | 8 |