

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
Departamento de Estatística e Investigação Operacional



Testes Conjuntos
Extensões da Teoria de Dorfman

Ricardo Augusto Domingues Gomes de Sá e Sousa

Doutoramento em Estatística e Investigação Operacional
(Especialidade de Probabilidades e Estatística)

2012

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
Departamento de Estatística e Investigação Operacional



Testes Conjuntos

Extensões da Teoria de Dorfman

Ricardo Augusto Domingues Gomes de Sá e Sousa

Tese orientada pelo Prof. Doutor Rui Sousa Santos e pelo Prof. Doutor Dinis Duarte Pestana, especialmente elaborada para a obtenção do grau de Doutor em Estatística e Investigação Operacional
(Especialidade de Probabilidades e Estatística)

2012

Dissertação apresentada à Faculdade de Ciências da Universidade de Lisboa, para a obtenção do grau de Doutor em Estatística e Investigação Operacional, especialidade de Probabilidades e Estatística

RESUMO

As análises conjuntas ao sangue, propostas por Dorfman durante a segunda grande Guerra, permitiram uma gestão mais eficiente de recursos na detecção dos infetados com sífilis no exército americano, e tornaram-se um paradigma, que se pode tornar mais realista e aplicável se considerarmos que os testes de diagnóstico são sujeitos a erros de classificação.

No âmbito deste trabalho estendemos os conceitos de sensibilidade e especificidade para a realização de testes conjuntos, adotando a proposta de Santos, Pestana e Martins (2012) para modelar a sensibilidade e a especificidade, que tem em linha de conta o problema da diluição e conseqüente rarefação.

Analizamos, via simulação, o comportamento de alguns estimadores para a taxa de prevalência baseados em testes conjuntos, quer na ausência quer na presença de erros de classificação.

Para os testes quantitativos discretos estendemos os cálculos da sensibilidade e da especificidade para o modelo de Poisson a populações mais dispersas, nomeadamente binomiais negativas, dando especial relevo ao caso mais tratável de população geométrica.

Por fim, no que toca a testes quantitativos contínuos, é investigada a informação da média sobre o máximo (ou sobre o mínimo) da amostra a fim de, com base num resultado conjunto, decidir os casos em que a amostra conjunta é classificada como suspeita de conter um ou mais infetados.

Palavras chave: Estimação, Classificação, Simulação, Testes Compostos.

AMS (2010) Subject Classification: 62D02.

ABSTRACT

The composite sampling proposed by Dorfman during the Second World War led to a more efficient management of resources in the detection of the infected with syphilis in the U.S. Army, and became a paradigm, which could be more realistic and enforceable considering that diagnostic tests are subjected to classification errors.

In this work we extended the concepts of sensitivity and specificity for the performance of compound tests, adopting the Santos, Pestana and Martins (2012) proposal to model the sensitivity and specificity, which takes into account the dilution and consequent rarefaction problem.

We analyse, through simulation, the behaviour of some estimators for the prevalence rate based on compound tests, let it be the absence or in the presence of errors of classification .

For quantitative discrete tests we extended the calculations of sensitivity and specificity for the Poisson model to more dispersed populations, namely negative binomial models, with more detailed analysis of geometric populations.

Finally, concerning the continuous quantitative tests, we investigated the information that the sample mean can provide about the maximum (or the minimum) of the sample, based on a joint result, and its bearing on identifying composite samples which eventually include infected individuals.

Keywords: Estimation, Classification, Simulation, Compound Tests

AMS (2010) Subject Classification: 62D02

Dedicatória

Num dia nasceu
Do nada começou
Cresceu, criou, lutou . . .
Hoje sou o que Sou

Às minhas filhas, Mafalda e Matilde
À minha mulher, Ana Luisa
Aos meus Pais

Deixo expresso os meus sinceros agradecimentos:

- Aos meus co-orientadores, Professores Doutores Rui Sousa Santos (Instituto Politécnico de Leiria) e Dinis Duarte Pestana (Universidade de Lisboa) pela sua disponibilidade, apoio, estímulo e principalmente pelos valiosos conhecimentos transmitidos durante a realização deste trabalho. Agradeço também aos seus colaboradores Miguel Felgueiras e João Paulo Martins, que também com eles têm investigado o tema central da minha tese, e que generosamente me deram acesso a resultados, permitindo-me citá-los e assim enriquecer esta dissertação com uma panorâmica mais completa e atual, nomeadamente no que se refere à informação que a média contém sobre os extremos
- À Escola Superior de Tecnologia da Saúde de Lisboa pela preciosa colaboração na impressão dos exemplares deste trabalho.
- Aos meus professores pelos ensinamentos prestados.
- À minha mulher pela enorme paciência que teve para me aturar nos períodos mais críticos.
- Ao amigo Eduardo Severino por tudo o que me ensinou desde o longínquo ano de 1994 em que fomos colegas de Mestrado.
- À amiga Elisabete Carolino pela total disponibilidade que demonstrou para me ajudar a trabalhar com o R.

A todos, bem hajam

Conteúdo

Resumo	i
Abstract	iii
Conteúdo	xi
Lista de Figuras	xiii
Lista de Tabelas	xvi
Lista de Notações	xvii
1 Introdução	1
2 Testes compostos qualitativos	3
2.1 Introdução	3
2.2 A metodologia de Dorfman	4
2.3 Uma estratégia alternativa para análises aos pares	9
2.4 A metodologia de Dorfman otimizada	12
2.5 A metodologia de Sterrett	14
2.6 A metodologia de Sterrett otimizada	15
2.7 A metodologia de Gill & Gottlieb	18
2.8 A metodologia de Gill & Gottlieb otimizada	19
2.9 Comparações entre as diversas metodologias	21
3 Testes de diagnóstico e erros de classificação	23
3.1 Introdução	23
3.2 Teoria estatística	25

3.3	Diagnóstico do problema	26
3.4	A curva ROC	28
3.5	Testes com erros de classificação	29
3.5.1	Introdução	29
3.5.2	Sensibilidade e especificidade num teste simples	30
3.5.3	Sensibilidade e especificidade num teste composto	30
3.5.4	Sensibilidade e especificidade na metodologia de Dorfman	33
3.5.5	Número esperado de análises em testes com erros de classificação	35
3.6	Subpopulações	37
4	Estimação da taxa de prevalência na ausência de erros de classificação	41
4.1	Introdução	41
4.2	Estimação por verosimilhança máxima	42
4.3	Estimação intervalar	50
4.3.1	Intervalo de confiança de Wald	51
4.3.2	Intervalo de confiança de score de Wilson	53
4.3.3	Intervalo de confiança de Agresti-Coull	54
4.3.4	Intervalo de confiança assintótico	55
4.3.5	Intervalo com correções de segunda ordem	57
4.3.6	Intervalo de confiança de Clopper-Pearson	59
4.3.7	Algoritmo de Blaker	60
4.3.8	Comparação de resultados via simulação	62
5	Estimação da taxa de prevalência na presença de erros de classificação	67
5.1	Estimação por verosimilhança máxima	67
5.2	Estimação intervalar	73
5.2.1	Intervalo de confiança assintótico	74
5.2.2	Comparação de intervalos de confiança via simulação	75
5.3	Robustez do modelo binomial	78
5.3.1	Primeiro modelo de diluição	79
5.3.2	Segundo modelo de diluição	80

6	Testes quantitativos discretos	81
6.1	Introdução	81
6.2	Classes de Panjer	82
6.3	O modelo de Poisson	85
6.4	Mistura Poisson-exponencial – o modelo geométrico	88
6.5	Mistura Poisson-gama – o modelo binomial negativo	91
6.6	O modelo logarítmico	93
6.7	O modelo binomial	94
 7	 Testes quantitativos contínuos	 97
7.1	Introdução	97
7.2	Testes de hipóteses para amostras individuais	97
7.2.1	Sensibilidade do teste simples para o modelo Gaussiano	98
7.2.2	Sensibilidade do teste simples para o modelo exponencial	103
7.3	Testes de hipóteses para amostras conjuntas	105
7.3.1	Sensibilidade do teste composto para o modelo gaussiano conside- rando grupos de dimensão n dos quais γ estão infetados	105
7.3.2	Sensibilidade do teste composto para o modelo exponencial conside- rando grupos de dimensão n dos quais γ estão infetados	108
7.3.3	Proposta de duas metodologias	111
7.4	Conclusões	119
 8	 Comentários finais	 121
 BIBLIOGRAFIA		 125

Lista de Figuras

2.1	Varição do CR para $p \in [0.01, 0.3]$	5
2.2	Varição do CR para $p \in [0.2, 0.3]$	7
2.3	Estratégia sequencial para grupos de dimensão 2.	10
2.4	Estratégia sequencial para grupos de dimensão 3	11
3.1	Funções densidade de duas populações	25
3.2	Sobreposição de duas populações hipotéticas	27
3.3	Curva ROC com a variação do critério de decisão	29
3.4	Varição do CR com $\varphi_s = 0.8$ e $\varphi_e = 0.9$	36
4.1	Variações do viés do estimador	45
4.2	Valores da variância ($m = 30$)	46
4.3	Valores do EQM ($m = 30$)	47
7.1	Densidade de \bar{Y} condicional a H_0 e H_1	99
7.2	Densidade de \bar{Y} condicional a H_0 e H_1 com $\mu_1 > \mu_0$ e $\sigma_1 > \sigma_0$	101
7.3	Densidade de \bar{Y} condicional a $\gamma = m$ com $m = \{0, 1, 2, 3\}$	109

Lista de Tabelas

2.1	Intervalos de Variação de Δ_n	7
2.2	Valores exatos de n^* e correspondente CR para a metodologia de Dorfman	8
2.3	Intervalos de variação de E_n^*	9
2.4	Valores aproximados de n^* e correspondente CR para a metodologia de Dorfman	9
2.5	Alterações à metodologia de Dorfman	12
2.6	Dimensão ótima e correspondente CR para a metodologia de Dorfman otimizada	13
2.7	Dimensão ótima e correspondente CR para a metodologia de Sterrett	16
2.8	Dimensão ótima e correspondente CR para a metodologia de Sterrett otimizada	17
2.9	Dimensão ótima e correspondente CR para a metodologia de Gill & Gottlieb	19
2.10	Dimensão ótima e correspondente CR para a metodologia de Gill & Gottlieb otimizada.	20
2.11	Tabela comparativa das três metodologias	21
3.1	Probabilidades num teste individual	30
3.2	Valores de λ_1 e λ_2 para algumas prevalências com $n = n^*$	32
3.3	Probabilidades num teste conjunto	33
3.4	Probabilidades num teste conjunto com $\varphi_s^{[\gamma, n]} = \varphi_s$	34
3.5	Valores de $1 + \frac{\epsilon}{n}$	37
3.6	Variação de $\varphi_{e_n}^{[S]}$ e do CR na presença de subpopulações	38
4.1	Valores do viés ($m = 30$)	44
4.2	Valores do EQM ($m = 30$)	45
4.3	Valores do viés ($m = 100$)	47
4.4	Valores do EQM ($m = 100$)	48
4.5	Valores do viés ($m = 1000$)	48

Lista de Tabelas

4.6	Valores do EQM ($m = 1000$)	49
4.7	Resultados de simulação de I. C. na ausência de erros de classificação ($m = 30$)	63
4.8	Resultados de simulação de I.C. na ausência de erros de classificação ($m = 100$)	64
4.9	Resultados de simulação de I.C. na ausência de erros de classificação ($m = 1000$)	65
5.1	Valores do viés ($m = 30, \varphi_s = \varphi_e = 0.95$)	70
5.2	Valores do EQM ($m = 30, \varphi_s = \varphi_e = 0.95$)	71
5.3	Valores do viés ($m = 100, \varphi_s = \varphi_e = 0.95$)	71
5.4	Valores do EQM ($m = 100, \varphi_s = \varphi_e = 0.95$)	72
5.5	Valores do viés ($m = 1000, \varphi_s = \varphi_e = 0.95$)	72
5.6	Valores do EQM ($m = 1000, \varphi_s = \varphi_e = 0.95$)	73
5.7	Resultados de simulação de I.C. utilizando $m = 30$ e $\varphi_s = \varphi_e = 0.95$	77
5.8	Resultados de simulação de I.C. utilizando $m = 100$ e $\varphi_s = \varphi_e = 0.95$. . .	77
5.9	Resultados de simulação de I.C. utilizando $m = 1000$ e $\varphi_s = \varphi_e = 0.95$. . .	78
5.10	Valores assumidos pelo primeiro modelo de diluição	79
7.1	Correlação entre a soma e o máximo	112
7.2	Modelo Gaussiano (100000 replicas)	114
7.3	Modelo exponencial (100000 replicas)	115
7.4	Modelo Pareto(5) (100000 replicas)	115
7.5	Modelo Pareto(3) (100000 replicas)	116
7.6	Modelo Pareto(1) (100000 replicas)	116
7.7	Modelo Pareto(5) com $1 < n < n^*$ (100000 replicas)	117
7.8	Modelo Pareto(3) com $1 < n < n^*$ (100000 replicas)	117
7.9	Modelo Pareto(I) com $1 < n < n^*$ (100000 replicas)	118

LISTA DE NOTAÇÕES

Notações	Significado
p	Taxa de prevalência
n	Dimensão dos grupos
m	Número de grupos
$N = m \times n$	Dimensão da população
X_i	$X_i = 1$ se infetado ; $X_i = 0$ se saudável
T_n	Número de testes necessários num grupo de n indivíduos
$E_n = CR$	Custo relativo
Δ_n	$E_{n+1} - E_n$
E_n^*	Custo relativo aproximado
φ_s	Sensibilidade de um teste individual
φ_e	Especificidade de um teste individual
X_i^+	Resultado positivo do teste individual
X_i^-	Resultado negativo do teste individual
$I^{[n]}$	Número de elementos infetados numa amostra de dimensão n
$X^{[+,n]}$	Ocorrência de um resultado positivo num teste composto
$X^{[-,n]}$	Ocorrência de um resultado negativo num teste composto
$\varphi_s^{[n]}$	Sensibilidade de um teste composto de dimensão n
$\varphi_e^{[n]}$	Especificidade de um teste composto de dimensão n
$\varphi_s^{[\gamma,n]}$	Sensibilidade de um teste composto com γ elementos infetados
λ_j	Peso associado a $\varphi_s^{[j,n]}$
$k_\gamma^{[n]}$	Pesos associados a $\varphi_s^{[\gamma,n]}$ e φ_s .
k	Número de subpopulações
w_i	Peso relativo à população i
$\varphi_{e_{n_i}}$	Especificidade para cada subpopulação
$\varphi_{e_n}^{[S]}$	Especificidade total
π_n	Probabilidade de um grupo de dimensão n estar contaminado
X_n^+	Número de testes positivos em m grupos de dimensão n
X_1^+	Número de testes positivos em m amostras individuais
Y_i^*	Número de bactérias presentes num ml de iogurte
$B_n = \sum_{i=1}^n Y_i^*$	Número de bactérias presentes em n ml numa amostra composta
B_1	Número de bactérias num ml retirado de uma mistura de n ml
$I^{[i,n]}$	$\mathbb{P}(I^{[n]} = i)$
γ	Número de infetados no grupo

Capítulo 1

Introdução

As análises conjuntas ou testes compostos são utilizadas em variadíssimas situações e têm como principal objetivo aceder à informação contida nas amostras individuais a custos reduzidos. Os testes compostos foram introduzidos na estatística por Dorfman (1943) e são efetuados usando um sangue combinado (mistura de sangues de n indivíduos) com o objetivo de determinar a dimensão ótima para cada grupo em função da taxa de prevalência p . Desta forma é possível minimizar o número esperado de testes necessários para a identificação de todos os elementos de uma determinada população

Na metodologia proposta por Dorfman é efetuado um teste conjunto a cada grupo. Caso o teste conjunto seja negativo, todos os indivíduos do grupo são classificados como negativos (saudáveis), caso contrário, pelo menos um dos elementos do grupo está infetado e consequentemente todos os elementos do grupo são testados individualmente por forma a identificar os elementos infetados. Para taxas de prevalência reduzidas consegue-se obter poupanças consideráveis. A metodologia de Dorfman é contudo muito limitada já que, por um lado, assume a inexistência de erros de classificação nos resultados dos testes e, por outro, só é aplicável a testes qualitativos (presença ou ausência de determinada característica nas unidades experimentais).

Novas metodologias foram posteriormente propostas por forma a minimizar o número esperado de testes necessários para classificar corretamente todos os elementos da população, tais como Sterret (1957), Gill & Gottlieb (1974), ou Kim *et al.*(2007).

A utilização de amostragem composta deve porem levar em linha de conta o problema de os resultados dos testes laboratoriais não serem completamente fiáveis. A presença de uma substância está sujeita a eventuais erros de deteção que são caracterizados pela sensibilidade (probabilidade de classificar corretamente os infetados) e pela especificidade (probabilidade de classificar corretamente os não infetados). Esta questão é analisada quer assumindo que a miscigenação da unidades não altera as características operacionais do teste, quer admitindo que a sensibilidade e/ou a especificidade são afetadas por fenómenos associados à ultra-diluição e consequente rarefação.

A metodologia de Dorfman pode ser estendida a testes quantitativos onde a quantidade de

qualquer substância em análise é descrita por uma variável aleatória X caracterizada por uma distribuição \mathbf{D} com vetor de parâmetros θ . Nestes casos a positividade é determinada caso uma determinada quantidade ultrapasse um determinado limiar l previamente definido.

O recurso a análises conjuntas não se esgota com a classificação de indivíduos sendo também aplicável na estimação da taxa de prevalência. Para este fim, perante um resultado positivo do teste conjunto, não é necessário proceder à realização de testes individuais uma vez que o objetivo da análise já não passa pela classificação de todos os indivíduos da população, mas apenas pela estimação da taxa de prevalência. Alguns autores, como por exemplo Chen & Swallow (1990), referem que sob determinadas condições, os estimadores obtidos pela aplicação de testes compostos têm melhor comportamento que os estimadores baseados em testes individuais. Assim, os testes compostos permitem não apenas poupar recursos monetários (minimizando o número de testes efetuados) como também obter estimativas mais precisas comparativamente às estimativas obtidas nos testes individuais.

O objetivo dos testes conjuntos quantitativos consiste em identificar se algum dos elementos do grupo está infetado. Assim, ao efetuar um teste conjunto estamos interessados em averiguar se o máximo (ou o mínimo) do grupo é superior (ou inferior) a um ponto de corte e, como tal, pretendemos realizar testes sobre o máximo do grupo tendo como única informação o valor da média desse grupo. A quantidade de informação que a média do grupo contém sobre o máximo do grupo desempenha um papel preponderante na avaliação da qualidade dos testes efetuados.

Capítulo 2

Testes compostos qualitativos

2.1 Introdução

Quando se pretende classificar qualitativamente os elementos de uma população em defeituosos ou não defeituosos, é possível, em certas situações, juntar um número de unidades num grupo e fazer um teste a esse grupo. É o que acontece, por exemplo, com análises ao sangue ou à urina, para despiste de doenças de prevalência baixa, onde a probabilidade de todos os indivíduos do grupo serem negativos é razoavelmente alta.

Se o teste for positivo então pelo menos um dos indivíduos do grupo é positivo, sendo por isso necessário proceder a uma segunda análise, essa nos moldes clássicos (análises individuais). Este método surgiu na literatura estatística durante a segunda guerra mundial com o intuito de identificar soldados infetados por sífilis (Dorfman, 1943; Feller, 1968). A ideia consiste em formar amostras misturando o sangue de grupos de soldados. Se o resultado do teste fosse negativo, todos os soldados daquele grupo seriam classificados como não tendo a infeção. No caso de um resultado positivo, um ou mais soldados do grupo estariam infetados e todos eles teriam que ser testados individualmente.

Se o grupo fosse constituído por n soldados, no primeiro caso, apenas um teste substituiria n testes. No segundo caso, como todos os n soldados tinham que ser testados individualmente, seriam necessários $n + 1$ testes. Consequentemente existiriam grupos onde se economizavam $n - 1$ testes e outros onde se tinha um gasto suplementar de um teste. Se a taxa de prevalência for alta, haverá predominância de grupos contaminados, caso contrário, haverá predominância de grupos não contaminados o que determina se o processo é ou não económico.

Uma economia considerável será obtida se a prevalência for baixa. Dorfman (1943) obteve na prática uma economia superior a 80%. O objetivo principal do método consiste em encontrar o tamanho ótimo para os grupos de observação em função da taxa de prevalência por forma a minimizar o número esperado de testes.

A metodologia inicial proposta por Dorfman é extremamente simples de formalizar e resolver já que considera a inexistência de erros nos resultados dos testes e unicamente a re-

alização de testes qualitativos (identificação da presença ou da ausência de uma qualquer substância no líquido composto analisado).

Posteriormente novos algoritmos foram propostos com o objetivo de minimizar o número de testes necessários para a correta classificação de todos os indivíduos da população, como por exemplo as propostas de Sterrett (1957), Sobel & Groll (1959), Finucan (1964), Gill & Gottlieb (1974), ou mais recentemente, Kim *et al.* (2007).

Sterrett (1957) sugeriu uma alternativa à metodologia de Dorfman onde, perante um resultado positivo da amostra combinada, as amostras individuais são testadas sequencialmente até à obtenção de um resultado positivo, sendo as remanescentes testadas conjuntamente. Caso se obtenha um resultado positivo o procedimento repete-se, caso contrário o procedimento termina.

Gill & Gottlieb (1974) propuseram que perante um resultado positivo da amostra combinada, esta fosse dividida em duas subamostras de dimensão tanto quanto possível idêntica. Cada uma das subamostras é testada e no caso de se obter um resultado positivo esta será novamente dividida. O procedimento termina quando todas as amostras individuais tiverem sido testadas.

Assim, neste capítulo são descritas três metodologias que visam minimizar o número esperado de testes necessários para a identificação de todos os elementos de uma determinada população. A metodologia inicial proposta por Dorfman (1943) e duas metodologias alternativas propostas por Sterret (1957) e por Gill & Gottlieb (1974). Para cada metodologia apresenta-se uma versão otimizada. Por fim procede-se à comparação das três metodologias com o intuito de identificar a mais eficiente.

2.2 A metodologia de Dorfman

Consideremos que o valor p , probabilidade de infeção, é igual para todos os elementos da população e que não há interação entre os indivíduos (observações independentes). Neste caso podemos caracterizar os membros da população através de variáveis aleatórias independentes $X_i, i = 1, 2, \dots, N$, com distribuição de Bernoulli de parâmetro p (taxa de prevalência da doença), onde $X_i = 1$ representa a presença e $X_i = 0$ a ausência de infeção no i -ésimo indivíduo da população.

Ao aplicar o teste a grupos de n indivíduos duas situações podem ocorrer: ou o teste é negativo (*i.e.* $\sum_{i=1}^n X_i = 0$) e por conseguinte será necessário efetuar apenas um teste para catalogar todo o grupo, o que ocorre com probabilidade $(1 - p)^n$, ou o teste é positivo e pelo menos um dos membros do grupo está infetado (*i.e.* $\sum_{i=1}^n X_i \geq 1$) sendo necessário efetuar testes individuais. Este acontecimento ocorre com probabilidade $1 - (1 - p)^n$ e serão efetuados $n + 1$ testes para identificar todos os indivíduos infetados.

Denotando por T_n o número de análises que é necessário efetuar em cada grupo de n

indivíduos tem-se

$$T_n = \begin{cases} 1 & n+1 \\ (1-p)^n & 1 - (1-p)^n \end{cases},$$

sendo o número esperado de análises em cada grupo de n rastreados simultaneamente dado por

$$\begin{aligned} \mathbb{E}(T_n) &= (1-p)^n + (n+1)[1 - (1-p)^n] \\ &= 1 + n[1 - (1-p)^n], \quad \forall n \geq 2. \end{aligned}$$

Através da comparação com o número de testes necessários no caso de análises individuais, n , podemos definir o custo relativo, ou seja, o número esperado de análises por indivíduo, através de

$$CR = E_n = \frac{\mathbb{E}(T_n)}{n} = \frac{1}{n} + 1 - (1-p)^n, \quad n \geq 2. \quad (2.1)$$

A economia alcançada pelo método depende da dimensão dos grupos e da prevalência da doença. Na Figura 2.1 é visível o comportamento de E_n para diversos valores de n e de p .

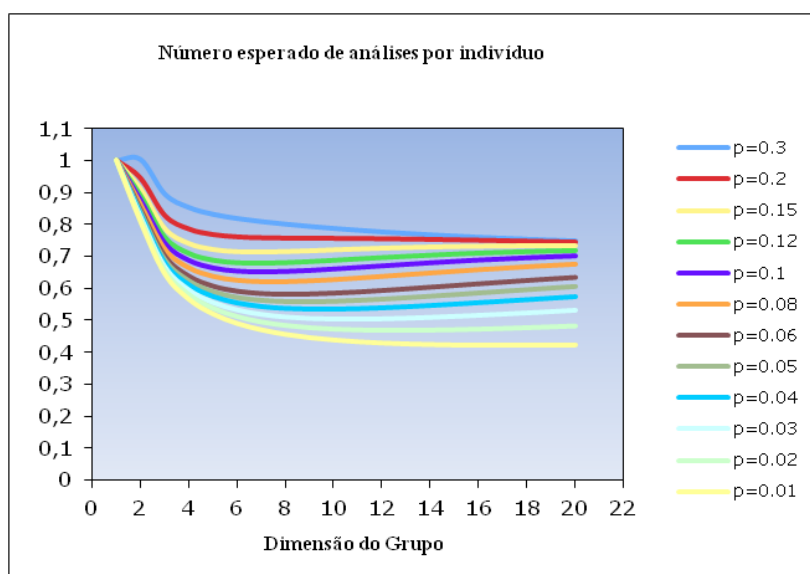


Figura 2.1: Variação do CR para $p \in [0.01, 0.3]$

Considerem-se as diferenças

$$\begin{aligned} \Delta_n(p) &= E_{n+1} - E_n \\ &= \left[\frac{1}{n+1} + 1 - (1-p)^{n+1} \right] - \left[\frac{1}{n} + 1 - (1-p)^n \right] \\ &= p(1-p)^n - \frac{1}{n(n+1)}. \end{aligned}$$

Derivando $\Delta_n(p)$ relativamente a p e igualando a zero obtém-se:

$$\begin{aligned}\Delta'_n(p) = 0 &\Leftrightarrow (1-p)^n - np(1-p)^{n-1} = 0 \\ &\Leftrightarrow (1-p)^{n-1} [1-p-np] = 0 \\ &\Leftrightarrow (1-p)^{n-1} = 0 \vee p(n+1) = 1 \\ &\Leftrightarrow p = \frac{1}{n+1} \vee p = 1,\end{aligned}$$

onde $p = 1$ é uma raiz de multiplicidade $n - 1$. Assim obtém-se

	0		$\frac{1}{n+1}$		1
$\Delta'_n(p)$	+	+	0	-	0
$\Delta_n(p)$		↗	Max	↘	

ou seja, para um valor fixo de n , $\Delta_n(p)$ é máximo em $p = \frac{1}{n+1}$.

Para $n = 2$ tem-se

$$\Delta_2\left(\frac{1}{3}\right) = \frac{1}{3} \times \left(\frac{2}{3}\right)^2 - \frac{1}{6} = \frac{-1}{54} < 0 \Rightarrow E_3 < E_2.$$

Dado que $\Delta_2(p)$ é máximo em $p = \frac{1}{3}$ e se obteve $\Delta_2\left(\frac{1}{3}\right) < 0$, podemos concluir que $E_3 < E_2, \forall p$. Para $n \geq 3$ tem-se

$$\begin{aligned}\Delta_n\left(\frac{1}{n+1}\right) &= \frac{1}{n+1} \left(1 - \frac{1}{n+1}\right)^n - \frac{1}{n(n+1)} \\ &= \frac{1}{n+1} \left[\left(1 - \frac{1}{n+1}\right)^n - \frac{1}{n}\right] \\ &= \frac{1}{n+1} \left[\left(\frac{n+1}{n}\right) \left(1 - \frac{1}{n+1}\right)^{n+1} - \frac{1}{n}\right] \\ &> \frac{1}{n+1} \left(\frac{1}{e} - \frac{1}{n}\right) > 0.\end{aligned}$$

Pelo teorema de Rolle sabemos que:

1. Se para um determinado valor de n , R_1 e R_2 forem zeros da função $\Delta_n(p)$, contínua no intervalo $[R_1, R_2]$ e diferenciável no seu interior, existe pelo menos um zero de $\Delta'_n(p)$ em $]R_1, R_2[$, ou seja, $R_1 < \frac{1}{n+1} < R_2$.
2. Sendo $\frac{1}{n+1}$ e 1, zeros consecutivos de $\Delta'_n(p)$, então $\Delta_n(p)$ não pode ter mais do que um zero no intervalo $] \frac{1}{n+1}, 1[$.

	0		R_1		$\frac{1}{n+1}$		R_2		1
$\Delta'_n(p)$	+	+	+	+	0	-	-	-	0
Sinal de $\Delta_n(p)$	-	-	0	+	+	+	0	-	-
$\Delta_n(p)$	\nearrow	\nearrow	\nearrow	\nearrow	Max	\searrow	\searrow	\searrow	\searrow

Tabela 2.1: Intervalos de Variação de Δ_n

3. Conjugando este resultado com o teorema de Bolzano sobre funções contínuas podemos afirmar que $\Delta_n(p)$ admite apenas um zero (designado por R_2) no intervalo $]\frac{1}{n+1}, 1[$ já que $\Delta_n(\frac{1}{n+1}) \times \Delta_n(1) < 0$.
4. O teorema de Bolzano garante a existência de pelo menos um zero de $\Delta_n(p)$ no intervalo $]0, \frac{1}{n+1}[$ já que $\Delta_n(0) \times \Delta_n(\frac{1}{n+1}) < 0$. Por outro lado, como $\frac{1}{n+1}$ é o menor zero de $\Delta'_n(p)$ não pode existir mais do que um zero de $\Delta_n(p)$ menor que $\frac{1}{n+1}$ e por conseguinte R_1 é o único zero de $\Delta_n(p)$ no intervalo $]0, \frac{1}{n+1}[$ (ver Tabela 2.1).

Assim, para cada valor de n , determinam-se numericamente os zeros de $\Delta_n(p)$ e posteriormente obtém-se o intervalo de valores de p que minimizam o custo relativo (ver Samuels (1978)). As dimensões ótimas (n^*) e correspondentes custos relativos (CR) são apresentados na Tabela 2.2.

Na Figura 2.2, é visível que não existem valores de prevalência para os quais seja mais favorável juntar os sangues 2 a 2 do que 3 a 3. Assiste-se a uma variação brusca de isoladas ($n = 1$) para trios ($n = 3$).

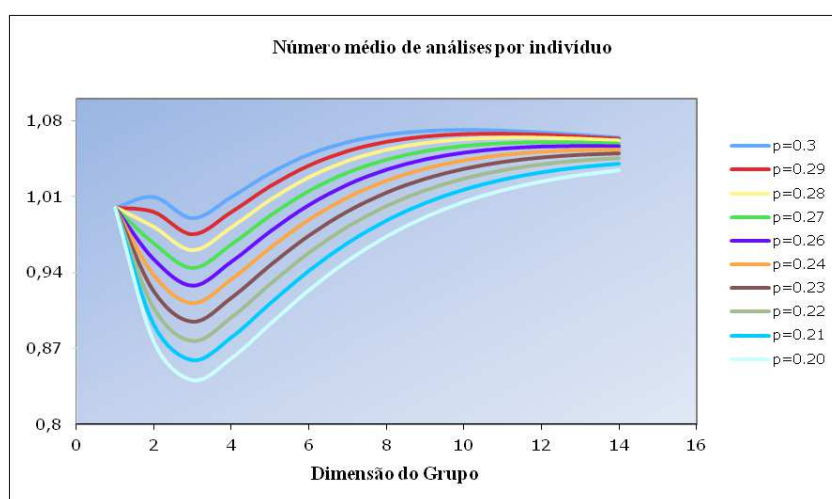


Figura 2.2: Variação do CR para $p \in [0.2, 0.3]$

Na Tabela 2.2 é visível um salto inesperado, ao passar de amostras de sangue individuais ($p \geq 0.30663$), para amostras de 3 sangues ($p \in (0.12394, 0.30663)$). Claro que, do ponto de vista prático, pareceria indicado e expedito analisar pares, para uma prevalência intermédia.

Recorrendo ao binómio de Newton e utilizando uma aproximação linear (razoável para prevalências p próximas de 0) obtém-se:

$$\begin{aligned} (1-p)^n &= \sum_{k=0}^n (-1)^k \binom{n}{k} p^k 1^{n-k} \\ &= \binom{n}{0} p^0 1^n - \binom{n}{1} p^1 1^{n-1} + \sum_{k=2}^n (-1)^k \binom{n}{k} p^k 1^{n-k} \\ &\approx 1 - np. \end{aligned}$$

O número esperado de análises por indivíduo, E_n , (ver equação (2.1) na página 5) pode ser aproximado por

$$E_n^* = \frac{1}{n} + 1 - (1 - np) = \frac{1}{n} + np.$$

Derivando E_n^* em ordem a n e igualando a zero obtém-se:

$$\frac{\partial E_n^*}{\partial n} = 0 \Leftrightarrow \frac{-1}{n^2} + p = 0 \Leftrightarrow n = \frac{1}{\sqrt{p}}.$$

Assim, podemos concluir (ver Tabela 2.3) que E_n^* tem um mínimo aproximadamente igual a $2\sqrt{p}$ em $n = \mathbf{I}\left(\frac{1}{\sqrt{p}}\right)$ onde, segundo Finucan (1964), $\mathbf{I}(x)$ representa o menor inteiro maior ou igual $x+0.5$. Na Tabela 2.4 são apresentados os valores aproximados de n^* bem como o correspondente custo relativo

n^*	p	CR
1	(0.30663, 1]	1
3	(0.12394, 0.30663]	(0.660973, 0.999987]
4	(0.06558, 0.12394]	(0.487622, 0.660973]
5	(0.04112, 0.06558]	(0.389372, 0.487622]
6	(0.02828, 0.04112]	(0.324792, 0.389372]
7	(0.02066, 0.02828]	(0.278810, 0.324792]
8	(0.01577, 0.02066]	(0.244410, 0.278810]
9	(0.01243, 0.01577]	(0.217573, 0.244410]
10	(0.01005, 0.01243]	(0.196068, 0.217573]
11	(0.00830, 0.01005]	(0.178510, 0.196068]
12	(0.00697, 0.00830]	(0.163839, 0.178510]
13	(0.00593, 0.00697]	(0.151323, 0.163839]
14	(0.00511, 0.00593]	(0.140635, 0.151323]
15	(0.00445, 0.00511]	(0.131373, 0.140635]

Tabela 2.2: Valores exatos de n^* e correspondente CR para a metodologia de Dorfman

		$\frac{1}{\sqrt{p}}$	
$\frac{\partial E_n^*}{\partial n}$	-	0	+
E_n^*	\searrow	Min	\nearrow

Tabela 2.3: Intervalos de variação de E_n^*

n^*	p	CR
4	(0.0816, 0.16]	(0.5781, 0.82]
5	(0.0494, 0.0816]	(0.4476, 0.5781]
6	(0.0331, 0.0494]	(0.3655, 0.4476]
7	(0.0237, 0.0331]	(0.3089, 0.3655]
8	(0.0178, 0.02237]	(0.2675, 0.3089]
9	(0.0138, 0.0178]	(0.2354, 0.2675]
10	(0.0111, 0.0138]	(0.2110, 0.2354]
11	(0.0091, 0.0111]	(0.191, 0.2110]
12	(0.0076, 0.0091]	(0.1745, 0.191]
13	(0.0064, 0.0076]	(0.1601, 0.1745]
14	(0.0055, 0.0064]	(0.1484, 0.1601]
15	(0.0048, 0.0055]	(0.1386, 0.1484]
16	(0.0042, 0.0048]	(0.1297, 0.1386]

Tabela 2.4: Valores aproximados de n^* e correspondente CR para a metodologia de Dorfman

2.3 Uma estratégia alternativa para análises aos pares

Vamos apresentar uma estratégia de análises sequenciais que nos fornece resultados ótimos para análises aos pares (ver Sousa (2005)) e constatar que tal acontece quando $p \in (0.203, 0.382)$.

A estratégia está descrita na Figura 2.3 e advém da constatação de que, caso a análise ao sangue combinado dê positiva e a análise ao primeiro sangue individual dê negativa então já não é necessário fazer a análise ao segundo sangue, pois será necessariamente positiva. Assim:

- Se a análise combinada der negativa, essa é suficiente (número de análises igual a 1);
- Se a análise combinada der positiva e a primeira análise individual der negativa, essas são suficientes (número de análises de análises igual a 2);
- Na pior das hipóteses, a sequência positivo, positivo obriga a uma terceira análise, para determinar se ambos os indivíduos serão positivos (número de análises igual a 3).

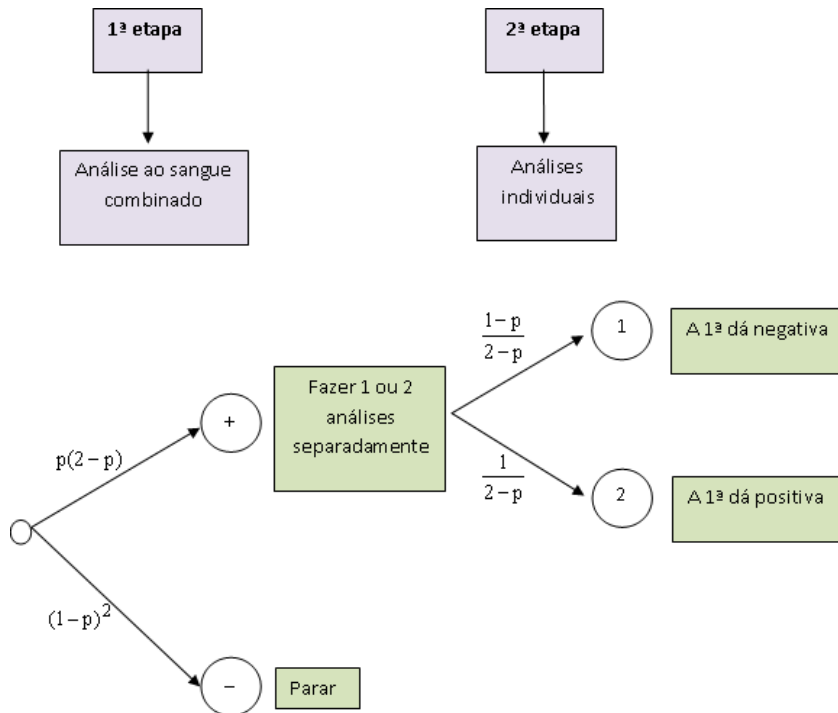


Figura 2.3: Estratégia sequencial para grupos de dimensão 2.

A função massa de probabilidade de T_2 , variável aleatória que conta o número de análises necessárias para determinar a positividade/negatividade dos dois indivíduos do grupo, é dada por

t	1	2	3
$\mathbb{P}(T_2 = t)$	$(1 - p)^2$	$p(1 - p)$	p

e, por conseguinte, o número médio de análises por indivíduo é dado por

$$E_2 = \frac{\mathbb{E}(T_2)}{2} = \frac{1}{2} [(1 - p)^2 + 2p(1 - p) + 3p] = \frac{-p^2 + 3p + 1}{2}.$$

Dado que $E_1 = 1$ tem-se

$$\begin{aligned} \Delta_1(p) < 0 &\Leftrightarrow E_2 - E_1 < 0 \Leftrightarrow E_2 < 1 \Leftrightarrow \frac{-p^2 + 3p + 1}{2} < 1 \Leftrightarrow \\ &\Leftrightarrow 0 \leq p < 0.382. \end{aligned}$$

Considere-se a estratégia descrita na Figura 2.4 que advém da constatação de que, caso a análise ao sangue combinado dê positiva e caso as análises individuais aos dois primeiros sangues sejam negativas, já não é necessário fazer a análise ao terceiro sangue, que será necessariamente positiva.

Assim:

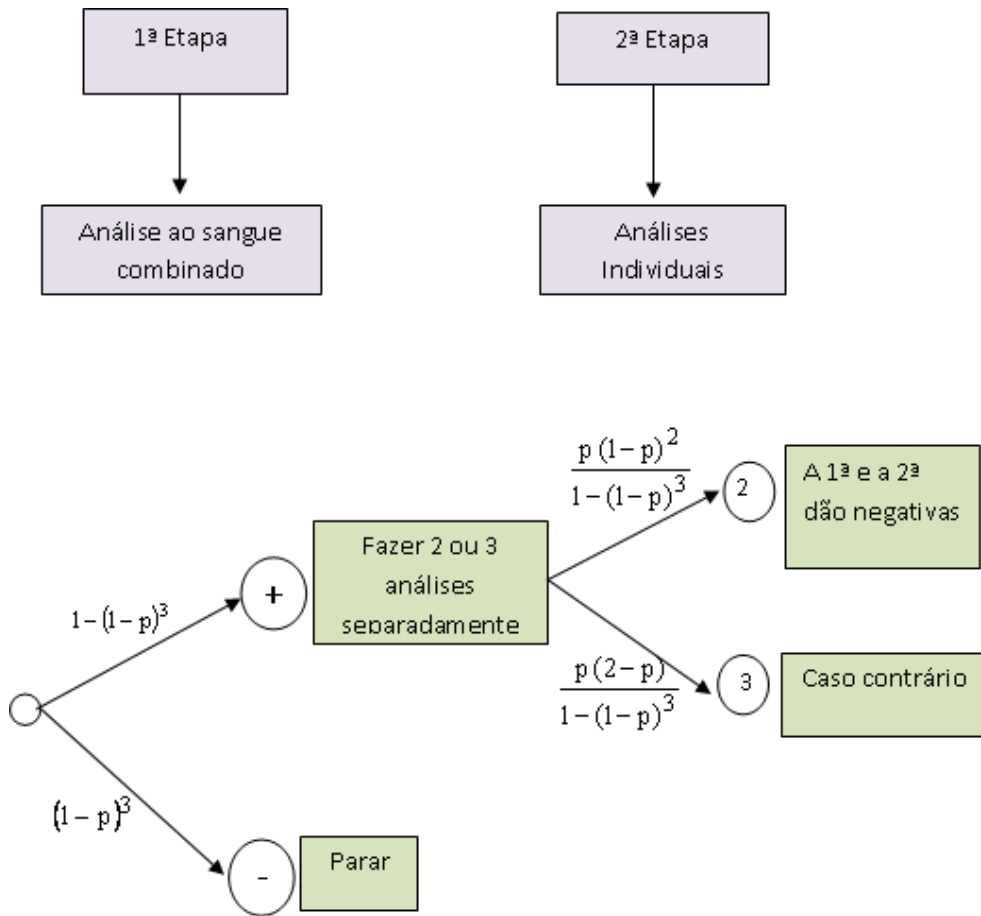


Figura 2.4: Estratégia sequencial para grupos de dimensão 3

- Se a análise combinada der negativa, essa é suficiente (número de análises igual a 1)
- Se a análise combinada der positiva e se as duas primeiras análises individuais derem negativas, essas são suficientes (número de análises igual 3)
- Caso contrário será necessário fazer a análise a todos os sangues (número de análises igual a 4)

A função massa de probabilidade de T_3 , variável aleatória que conta o número de análises necessárias para determinar a positividade/negatividade dos três indivíduos do grupo é dada por

t	1	3	4
$\mathbb{P}(T_3 = t)$	$(1 - p)^3$	$p(1 - p)^2$	$p(2 - p)$

e, por conseguinte, o número médio de análises por indivíduo é dado por

$$\begin{aligned} E_3 = \frac{\mathbb{E}(T_3)}{3} &= \frac{1}{3} [(1-p)^3 + 3p(1-p)^2 + 4p(2-p)] \\ &= \frac{1}{3} [2p^3 - 7p^2 + 8p + 1] \\ &= \frac{2}{3}p^3 - \frac{7}{3}p^2 + \frac{8}{3}p + \frac{1}{3}. \end{aligned}$$

Assim tem-se

$$\begin{aligned} \Delta_2(p) = E_3 - E_2 &= \left(\frac{2}{3}p^3 - \frac{7}{3}p^2 + \frac{8}{3}p + \frac{1}{3} \right) - \left(\frac{-p^2 + 3p + 1}{2} \right) \\ &= \frac{2}{3}p^3 - \frac{11}{6}p^2 + \frac{7}{6}p - \frac{1}{6}, \end{aligned}$$

e portanto

$$\Delta_2(p) > 0 \Leftrightarrow \frac{2}{3}p^3 - \frac{11}{6}p^2 + \frac{7}{6}p - \frac{1}{6} > 0 \Leftrightarrow 0.203 < p < 0.650,$$

o que permite concluir que para valores de prevalência $p \in (0.203, 0.382)$ é mais económico juntar os sangues 2 a 2.

As alterações relativamente à metodologia de Dorfman para $n \leq 2$ são apresentadas na Tabela 2.5. Para valores de $n \geq 3$ consultar a Tabela 2.6 na página 13.

n^*	p
1	(0.381966, 1]
2	(0.202616, 0.381966]

Tabela 2.5: Alterações à metodologia de Dorfman

Consegue-se assim melhorar a estratégia de Dorfman, para doenças de prevalência moderada (na prática de 20% a 38%), a qual, foi em si mesma, um salto qualitativo importante na gestão da saúde pública. Note-se ainda que este procedimento é, do ponto de vista prático, facilmente implementável.

2.4 A metodologia de Dorfman otimizada

A estratégia sequencial apresentada na secção anterior pode ser generalizada para grupos de dimensão n . Assim, ao aplicar o teste a grupos de n indivíduos três situações podem ocorrer:

1. O teste é negativo e por conseguinte será necessário efetuar apenas um teste para catalogar todo o grupo, o que ocorre com probabilidade q^n ;

2. O teste é positivo e os primeiros $n-1$ membros do grupo têm um resultado negativo, não sendo por conseguinte necessário efetuar o teste ao último elemento do grupo. Serão necessários n testes para identificar todos os elementos do grupo, o que ocorre com probabilidade pq^{n-1} ;
3. Caso contrário serão necessários $n+1$ testes para identificar todos os elementos contaminados, o que ocorre com probabilidade $1 - q^{n-1}$.

Assim, a função massa de probabilidade de T_n é dada por

t	1	n	$n+1$
$\mathbb{P}(T_n = t)$	q^n	pq^{n-1}	$1 - q^{n-1}$

donde

$$\begin{aligned} \mathbb{E}[T_n] &= q^n + npq^{n-1} + (n+1)(1 - q^{n-1}) \\ &= n+1 - nq^n - q^{n-1}p, \end{aligned}$$

e, por conseguinte, o custo relativo é dado por

$$\begin{aligned} \text{CR} &= \frac{1}{n} [n+1 - nq^n - q^{n-1}p] \\ &= 1 - q^n + \frac{1}{n} [1 - q^{n-1}p]. \end{aligned}$$

A dimensão ótima dos grupos para uma determinada prevalência e o seu CR constam na Tabela 2.6.

n^*	p	CR
1	(0.38196, 1]	1
2	(0.20261, 0.38196]	(0.783386, 0.999993]
3	(0.10291, 0.20261]	(0.583770, 0.783386]
4	(0.06010, 0.10291]	(0.457106, 0.583770]
5	(0.03906, 0.06010]	(0.373965, 0.457106]
6	(0.02733, 0.03906]	(0.315871, 0.373965]
7	(0.02017, 0.02733]	(0.273231, 0.315871]
8	(0.01549, 0.02017]	(0.240669, 0.273231]
9	(0.01226, 0.01549]	(0.214956, 0.240669]
10	(0.00994, 0.01226]	(0.194156, 0.214956]

Tabela 2.6: Dimensão ótima e correspondente CR para a metodologia de Dorfman otimizada

2.5 A metodologia de Sterrett

Perante um resultado positivo do teste conjunto, Sterrett (1957) sugere que todos os elementos do grupo sejam sequencialmente testados até à identificação do primeiro elemento com resultado positivo, sendo os remanescentes testados conjuntamente. Se o resultado do teste conjunto for positivo o procedimento é repetido até serem identificados todos os elementos infetados, caso contrário o procedimento termina. Seja T_n o número de testes necessários para a classificação dos n indivíduos que compõem o grupo. Para valores reduzidos de n , o número esperado de testes pode ser calculado diretamente, obtendo-se:

$$\mathbb{E}[T_1] = 1, \quad \mathbb{E}[T_2] = 3 - 2q^2, \quad \mathbb{E}[T_3] = 5 - q - 2q^2 - q^3.$$

No cálculo de $\mathbb{E}[T_2] = 3 - 2q^2$, estamos a considerar que quando o teste ao grupo dá positivo e o primeiro teste individual dá negativo, se testa o segundo indivíduo.

Vamos determinar uma fórmula de recorrência condicionando T_n ao número de testes individuais J efetuados até encontrar o primeiro elemento do grupo com resultado positivo. Assim, $J = 0$ se o teste ao grupo der negativo, $J = 1$ se o teste ao grupo der positivo e o teste ao primeiro elemento do grupo der positivo e assim sucessivamente. Então

$$\begin{aligned} \mathbb{E}[T_n] &= \mathbb{E}[\mathbb{E}(T_n|J)] \\ &= \sum_{j=0}^n \mathbb{E}[T_n|J=j] \mathbb{P}[J=j] \\ &= \mathbb{E}[T_n|J=0] q^n + \sum_{j=1}^n \mathbb{E}[T_n|J=j] q^{j-1} p, \quad n = 2, 3, \dots \end{aligned}$$

Para que o primeiro teste individual com resultado positivo seja o j -ésimo é necessário que o teste ao grupo dê positivo e que os primeiros $j - 1$ testes individuais deem negativos. Dado que os remanescentes $n - j$ elementos do grupo são variáveis aleatórias com distribuição de Bernoulli de parâmetro p tem-se

$$\mathbb{E}[T_n|J=j] = j + 1 + \mathbb{E}[T_{n-j}], \quad j > 0,$$

donde se obtém

$$\begin{aligned} \mathbb{E}[T_n] &= q^n + \sum_{j=1}^n \{j + 1 + \mathbb{E}[T_{n-j}]\} q^{j-1} p \\ &= q^n + \sum_{j=1}^n (j + 1) q^{j-1} p + \sum_{j=1}^n \mathbb{E}[T_{n-j}] q^{j-1} p \\ &= q^n + \sum_{j=0}^{n-1} (j + 2) q^j p + \sum_{j=0}^{n-1} \mathbb{E}[T_j] q^{n-j-1} p. \end{aligned}$$

por outro lado

$$q\mathbb{E}[T_{n-1}] = q^n + \sum_{j=1}^{n-1} (j+1)q^j p + \sum_{j=0}^{n-2} \mathbb{E}[T_j] q^{n-j-1} p, \quad n = 3, 4, \dots$$

e portanto

$$\mathbb{E}[T_n] - q\mathbb{E}[T_{n-1}] = \left[2p + \sum_{j=1}^{n-1} q^j p \right] + \mathbb{E}[T_{n-1}] p,$$

donde

$$\mathbb{E}[T_n] - \mathbb{E}[T_{n-1}] = 2p + \frac{qp(1 - q^{n-1})}{1 - q} = 2p + q - q^n.$$

No entanto

$$\begin{aligned} \mathbb{E}[T_n] - \mathbb{E}[T_2] &= \sum_{j=3}^n \{\mathbb{E}[T_j] - \mathbb{E}[T_{j-1}]\} \\ &= (n-2)(2p+q) - \sum_{j=3}^n q^j \\ &= (n-2)(2p+q) - \frac{q^3(1 - q^{n-2})}{1 - q}, \quad n = 2, 3, \dots \end{aligned}$$

Somando a ambos os membros o valor de $\mathbb{E}[T_2]$ obtém-se

$$\begin{aligned} \mathbb{E}[T_n] &= 3 - 2q^2 + (n-2)(2p+q) + \frac{1 - q^3}{p} - \frac{1 - q^{n+1}}{p} \\ &= 2n - (n-3)q - q^2 - \frac{1 - q^{n+1}}{1 - q}, \quad n = 2, 3, \dots \end{aligned}$$

O custo relativo é dado pela expressão

$$\text{CR} = \frac{\mathbb{E}[T_n]}{n} = 2 - q + \frac{1}{n} \left[3q - q^2 - \frac{1 - q^{n+1}}{p} \right].$$

Para uma determinada taxa de prevalência, podemos visualizar a dimensão ótima do grupo e o correspondente custo relativo CR na Tabela 2.7 da página 16.

2.6 A metodologia de Sterrett otimizada

A metodologia de Sterrett pode ser otimizada quando o teste ao grupo der positivo e os testes individuais aos primeiros $n - 1$ elementos que o compõem derem negativos. Nestas circunstâncias não é necessário efetuar o teste ao último elemento do grupo pois sabe-se *a priori* que é positivo. Para valores reduzidos de n , o número esperado de testes pode ser calculado

n^*	p	CR
1	(0.30437, 1]	1
3	(0.21253, 0.30437]	(0.827997, 0.999984]
4	(0.12774, 0.21253]	(0.622819, 0.827997]
5	(0.08283, 0.12774]	(0.487586, 0.622819]
6	(0.05759, 0.08283]	(0.397398, 0.487586]
7	(0.04224, 0.05759]	(0.334221, 0.397398]
8	(0.03226, 0.04224]	(0.287849, 0.334221]
9	(0.02543, 0.03226]	(0.252541, 0.287849]
10	(0.02005, 0.02543]	(0.224784, 0.252541]
11	(0.01695, 0.02005]	(0.202456, 0.224784]
12	(0.01422, 0.01695]	(0.184127, 0.202456]
13	(0.01210, 0.01422]	(0.168814, 0.184127]
14	(0.01042, 0.01210]	(0.155825, 0.168814]
15	(0.00906, 0.01042]	(0.144618, 0.155825]

Tabela 2.7: Dimensão ótima e correspondente CR para a metodologia de Sterrett

diretamente, obtendo-se: $\mathbb{E}[T_1] = 1$, $\mathbb{E}[T_2] = 3 - q - q^2$, $\mathbb{E}[T_3] = 5 - 2q - q^2 - q^3$. Designando por J o número de testes individuais que é necessário efetuar até encontrar o primeiro elemento do grupo com resultado positivo tem-se

$$\begin{aligned}
 \mathbb{E}[T_n] &= \sum_{j=0}^n \mathbb{E}[T_n | J = j] \mathbb{P}(J = j) \\
 &= q^n + \sum_{j=1}^{n-1} \mathbb{E}[T_n | J = j] q^{j-1} p + n q^{n-1} p \\
 &= q^n + n q^{n-1} p + \sum_{j=1}^{n-1} \{j + 1 + \mathbb{E}[T_{n-j}]\} q^{j-1} p \\
 &= q^n + n q^{n-1} p + \sum_{j=0}^{n-2} (j + 2) q^j p + \sum_{j=1}^{n-1} \mathbb{E}[T_j] q^{n-j-1} p,
 \end{aligned}$$

por outro lado,

$$\begin{aligned}
 q \mathbb{E}[T_{n-1}] &= q^n + (n-1) q^{n-1} p + \sum_{j=0}^{n-3} (j+2) q^{j+1} p + \sum_{j=1}^{n-2} \mathbb{E}[T_j] q^{n-j-1} p \\
 &= q^n + (n-1) q^{n-1} p + \sum_{j=1}^{n-2} (j+1) q^j p + \sum_{j=1}^{n-2} \mathbb{E}[T_j] q^{n-j-1} p, \quad n = 3, 4, \dots
 \end{aligned}$$

e portanto,

$$\mathbb{E}[T_n] - q\mathbb{E}[T_{n-1}] = q^{n-1}p + 2p + \sum_{j=1}^{n-2} q^j p + \mathbb{E}[T_{n-1}]p$$

ou

$$\begin{aligned} \mathbb{E}[T_n] - \mathbb{E}[T_{n-1}] &= q^{n-1}p + 2p + \frac{pq(1 - q^{n-2})}{1 - q} \\ &= q^{n-1}p + 2p + q(1 - q^{n-2}), \quad n = 3, 4, \dots \end{aligned}$$

O número esperado de testes para classificar os n elementos do grupo assume todos os valores de 1 a $2n - 1$ à exceção do 2 e é dado por

$$\mathbb{E}[T_n] = n(2 - q) + 2q - \frac{1 - q^{n+1}}{p},$$

sendo o custo relativo dado por

$$\text{CR} = \frac{\mathbb{E}[T_n]}{n} = 2 - q + \frac{\left[2q - \frac{1 - q^{n+1}}{p}\right]}{n}.$$

Para uma determinada taxa de prevalência, podemos visualizar a dimensão ótima do grupo e o correspondente custo relativo CR na Tabela 2.8.

n^*	p	CR
1	(0.38196, 1]	1
2	(0.26101, 0.38196]	(0.857449, 0.999993]
3	(0.16832, 0.26101]	(0.689891, 0.857449]
4	(0.10986, 0.16832]	(0.551025, 0.689891]
5	(0.07506, 0.10986]	(0.448910, 0.551025]
6	(0.05381, 0.07506]	(0.374861, 0.448910]
7	(0.04022, 0.05381]	(0.320154, 0.374861]
8	(0.03109, 0.04022]	(0.278530, 0.320154]
9	(0.02471, 0.03109]	(0.246079, 0.278530]
10	(0.02008, 0.02471]	(0.220107, 0.246079]
11	(0.01664, 0.02008]	(0.199027, 0.220107]
12	(0.01400, 0.01664]	(0.181486, 0.199027]
13	(0.01194, 0.01400]	(0.166741, 0.181486]
14	(0.01030, 0.01194]	(0.154162, 0.166741]
15	(0.00897, 0.01030]	(0.143279, 0.154162]

Tabela 2.8: Dimensão ótima e correspondente CR para a metodologia de Sterrett otimizada

2.7 A metodologia de Gill & Gottlieb

Gill & Gottlieb (1974) argumentam que o procedimento de Dorfman e o procedimento de Sterrett são limitados no que respeita ao binómio custo-eficiência já que perante um resultado positivo do teste conjunto ambos recomendam (pelo menos inicialmente) que se teste separadamente cada um dos elementos do grupo. Os autores argumentam que condicionalmente à positividade do grupo, a prevalência dos elementos que formam o grupo é dada por $p^+ = \frac{p}{1-(1-p)^n} > p$ e por conseguinte a dimensão ótima correspondente a p^+ não será superior a n^* (dimensão ótima correspondente a p). Assim, caso o teste ao grupo dê positivo, é mais razoável formar subamostras do que testar isoladamente cada um dos elementos que o compõem.

Com esta modificação e assumindo o modelo binomial, Gill & Gottlieb (1974) propõem que na presença de um grupo positivo, este seja dividido em dois subgrupos de dimensão idêntica (ou próxima) e testados posteriormente. Os subgrupos testados positivamente são novamente divididos até todos os elementos que o compõem serem testados.

Para valores reduzidos de n , o número esperado de testes pode ser calculado diretamente, obtendo-se:

$$\mathbb{E}[T_1] = 1, \mathbb{E}[T_2] = 3 - 2q^2, \mathbb{E}[T_3] = 5 - 2q^2 - q^3.$$

No cálculo de $\mathbb{E}[T_2] = 3 - 2q^2$, estamos a considerar que quando o teste conjunto dá positivo e o teste à primeira subamostra dá negativo, se testa a segunda subamostra.

Quando o teste ao grupo dá positivo, os n elementos que o compõem são divididos em dois subgrupos de dimensão n_1 e $n_2 = n - n_1$, onde $n_1 = n_2 = \frac{n}{2}$ se n for par e onde $n_1 = \frac{n-1}{2}$ e $n_2 = \frac{n+1}{2}$ se n for ímpar. Denotando por B o acontecimento “Pelo menos um dos n elementos que compõem o grupo está contaminado” então

$$\begin{aligned} \mathbb{E}[T_n] &= \mathbb{E}[T_n|B] + \mathbb{E}[T_n|\overline{B}] \\ &= (1 + \mathbb{E}[T_{n_1} + T_{n_2}|B]) \mathbb{P}(B) + 1 \times \mathbb{P}[\overline{B}] \\ &= \mathbb{E}[T_{n_1} + T_{n_2}|B] \underbrace{\mathbb{P}(B)}_{1-q^n} + 1. \end{aligned}$$

Considerando agora os subgrupos de dimensão n_1 e n_2 tem-se

$$\mathbb{E}[T_{n_1} + T_{n_2}] = \mathbb{E}[T_{n_1} + T_{n_2}|B] \mathbb{P}(B) + \mathbb{E}[T_{n_1} + T_{n_2}|\overline{B}] \mathbb{P}(\overline{B}),$$

ou seja,

$$\mathbb{E}[T_{n_1}] + \mathbb{E}[T_{n_2}] = \underbrace{\mathbb{E}[T_{n_1} + T_{n_2}|B] \mathbb{P}(B)}_{=\mathbb{E}(T_n)-1} + 2q^n$$

e portanto

$$\mathbb{E}[T_n] = \mathbb{E}[T_{n_1}] + \mathbb{E}[T_{n_2}] + 1 - 2q^n.$$

Esta fórmula pode ser utilizada recursivamente para o cálculo de $\mathbb{E}[T_n]$ para qualquer valor de n , obtendo-se, por exemplo, $\mathbb{E}[T_2] = 2\mathbb{E}[T_1] + 1 - 2q^2$ e $\mathbb{E}[T_3] = \mathbb{E}[T_1] + \mathbb{E}[T_2] + 1 - 2q^3$. Desta forma vamos obter:

$$\begin{aligned} \mathbb{E}[T_2] &= 3 - 2q^2; \\ \mathbb{E}[T_3] &= 5 - 2q^2 - 2q^3; \\ \mathbb{E}[T_4] &= 7 - 4q^2 - 2q^4; \\ \mathbb{E}[T_5] &= 9 - 4q^2 - 2q^3 - 2q^5; \\ \mathbb{E}[T_6] &= 11 - 4q^2 - 4q^3 - 2q^6; \\ \mathbb{E}[T_7] &= 13 - 6q^2 - 2q^3 - 2q^4 - 2q^7; \\ \mathbb{E}[T_8] &= 15 - 8q^2 - 4q^4 - 2q^8; \\ \mathbb{E}[T_9] &= 17 - 8q^2 - 2q^3 - 2q^4 - 2q^5 - 2q^9; \\ \mathbb{E}[T_{10}] &= 19 - 8q^2 - 4q^3 - 4q^5 - 2q^{10}; \\ \mathbb{E}[T_{11}] &= 21 - 8q^2 - 6q^3 - 2q^5 - 2q^6 - 2q^{11}; \\ \mathbb{E}[T_{12}] &= 23 - 8q^2 - 8q^3 - 4q^6 - 2q^{12}. \end{aligned}$$

O custo relativo é calculado dividindo o número esperado de testes pela dimensão do grupo. Na Tabela 2.9 podemos visualizar a dimensão ótima e o correspondente custo relativo para uma determinada taxa de prevalência.

n^*	p	CR
1	(0.29289, 1]	1
2	(0.15910, 0.29289]	(0.792883, 0.999995]
4	(0.08299, 0.15910]	(0.555524, 0.792883]
8	(0.04239, 0.08299]	(0.360729, 0.555524]
16	(0.02142, 0.04239]	(0.222719, 0.360729]
32	(0.01077, 0.02142]	(0.132800, 0.222719]
64	(0.00540, 0.01077]	(0.077175, 0.132800]
128	(0.00267, 0.00540]	(0.043552, 0.077175]

Tabela 2.9: Dimensão ótima e correspondente CR para a metodologia de Gill & Gottlieb

2.8 A metodologia de Gill & Gottlieb otimizada

A metodologia de Gill & Gottlieb pode ser otimizada quando o teste ao grupo der positivo e o teste ao primeiro subgrupo der negativo. Nestas circunstâncias não é obviamente necessário testar o segundo subgrupo. Assim, o segundo subgrupo só é testado quando o primeiro der

positivo. De forma análoga ao apresentado na metodologia de Gill & Gottlieb, obtém-se a seguinte fórmula recursiva:

$$\mathbb{E}[T_n] = \mathbb{E}[T_{n_1}] + \mathbb{E}[T_{n_2}] + 1 - q^{n-1} - q^n.$$

Quando n é par, então $n_1 = n_2 = \frac{n}{2}$. Caso n seja ímpar, o número esperado de testes é menor com $n_1 = \frac{n-1}{2}$ do que com $n_1 = \frac{n+1}{2}$. Com esta escolha para n_1 , a fórmula recursiva anterior pode ser utilizada iterativamente para obter o valor médio de T_n . Assim tem-se:

$$\begin{aligned} \mathbb{E}[T_2] &= 3 - q - q^2; \\ \mathbb{E}[T_3] &= 5 - q - 2q^2 - q^3; \\ \mathbb{E}[T_4] &= 7 - 2q - 2q^2 - q^3 - q^4; \\ \mathbb{E}[T_5] &= 9 - 2q - 3q^2 - q^3 - q^4 - q^5; \\ \mathbb{E}[T_6] &= 11 - 2q - 4q^2 - 2q^3 - q^5 - q^6; \\ \mathbb{E}[T_7] &= 13 - 3q - 4q^2 - 2q^3 - q^4 - q^6 - q^7; \\ \mathbb{E}[T_8] &= 15 - 4q - 4q^2 - 2q^3 - 2q^4 - q^7 - q^8; \\ \mathbb{E}[T_9] &= 17 - 4q - 5q^2 - 2q^3 - 2q^4 - q^5 - q^8 - q^9; \\ \mathbb{E}[T_{10}] &= 19 - 4q - 6q^2 - 2q^3 - 2q^4 - 2q^5 - q^9 - q^{10}; \\ \mathbb{E}[T_{11}] &= 21 - 4q - 7q^2 - 3q^3 - q^4 - 2q^5 - q^6 - q^{10} - q^{11}; \\ \mathbb{E}[T_{12}] &= 23 - 4q - 8q^2 - 4q^3 - 2q^5 - 2q^6 - q^{11} - q^{12}. \end{aligned}$$

O custo relativo é calculado dividindo o número esperado de testes pela dimensão do grupo. Para uma determinada prevalência, podemos visualizar a dimensão ótima e o custo relativo na Tabela 2.10.

n^*	p	CR
1	(0.38196, 1]	1
2	(0.26101, 0.38196]	(0.857449, 0.999993]
3	(0.16582, 0.26101]	(0.685100, 0.857449]
5	(0.10106, 0.16582]	(0.513090, 0.685100]
7	(0.08439, 0.10106]	(0.458100, 0.513090]
9	(0.06817, 0.08439]	(0.397968, 0.458100]
10	(0.06640, 0.06817]	(0.391045, 0.397968]
11	(0.05054, 0.06640]	(0.325611, 0.391045]

Tabela 2.10: Dimensão ótima e correspondente CR para a metodologia de Gill & Gottlieb otimizada.

2.9 Comparações entre as diversas metodologias

		Dorfman		Sterrett		Gill & Gottlieb	
p	q	n^*	CR	n^*	CR	n^*	CR
0.2	0.8	3	0.7787	3	0.7493	3	0.7493
0.19	0.81	3	0.7603	3	0.7308	3	0.7308
0.18	0.82	3	0.7416	3	0.7121	3	0.7121
0.17	0.83	3	0.7225	3	0.6931	3	0.6931
0.16	0.84	3	0.7030	4	0.6710	5	0.6705
0.15	0.85	3	0.6831	4	0.6478	5	0.6449
0.14	0.86	3	0.6628	4	0.6243	5	0.6189
0.13	0.87	3	0.6420	4	0.6004	5	0.5924
0.12	0.88	3	0.6209	4	0.5761	5	0.5655
0.11	0.89	3	0.5993	4	0.5514	5	0.5381
0.10	0.90	4	0.5757	5	0.5229	7	0.5097
0.09	0.91	4	0.5473	5	0.4937	7	0.4769
0.08	0.92	4	0.5180	5	0.4639	9	0.4421
0.07	0.93	4	0.4879	6	0.4317	9	0.4049
0.06	0.94	5	0.4567	6	0.3969	11	0.3611
0.05	0.95	5	0.4181	7	0.3798	13	0.3254
0.04	0.96	5	0.3738	8	0.3192	19	0.2907
0.01	0.99	10	0.1947	15	0.1517	75	0.1049
0.005	0.995	15	0.1388	21	0.1054	149	0.0602

Tabela 2.11: Tabela comparativa das três metodologias

Na Tabela 2.11 procede-se à comparação entre as três metodologias. Para cada metodologia, determinou-se o custo relativo para o valor de n ótimo, para um conjunto de valores de prevalência. Os valores a negrito identificam para cada valor de p a metodologia mais eficiente, ou seja, a que tem um custo relativo menor.

Assim, concluímos que a metodologia de Gill & Gottlieb é a mais eficiente para a generalidade dos valores de p sendo igualmente eficiente à metodologia de Sterrett apenas para grupos de dimensão 3.

Capítulo 3

Testes de diagnóstico e erros de classificação

3.1 Introdução

Os testes de diagnóstico permitem a previsão ou deteção de determinados acontecimentos numa fase incipiente de desenvolvimento. Variadas questões problemáticas se podem colocar na utilização de um teste de diagnóstico. Desde logo o problema da discriminação, que consiste em conseguir classificar, de uma forma precisa, os casos considerados normais e os casos anormais.

Outra questão que se torna problemática num teste de diagnóstico prende-se com as definições de exatidão e precisão. A precisão está relacionada com a dispersão dos valores em sucessivas observações, enquanto que a exatidão se refere à proximidade da estimativa do verdadeiro valor que se pretende representar. As limitações da exatidão e da precisão no diagnóstico, originaram a introdução dos conceitos de *sensibilidade* (probabilidade de um teste positivo num indivíduo infetado) e *especificidade* (probabilidade de um teste negativo num indivíduo saudável) dum teste de diagnóstico. Estas medidas e os índices a elas associados, como a *proporção de verdadeiros positivos*, a *proporção de falsos positivos*, o *valor preditivo positivo* e o *valor preditivo negativo* são mais significantes do que a exatidão, embora não forneçam uma descrição única do desempenho do diagnóstico.

A *sensibilidade* e a *especificidade* pecam por dependerem do critério de diagnóstico ou dum valor de *corte*, o qual é por vezes selecionado arbitrariamente. Assim, mudando o critério pode-se aumentar a *sensibilidade* com o conseqüente detrimento da *especificidade* e vice versa. Conseqüentemente, estas medidas representam um quadro incompleto do desempenho de um teste de diagnóstico.

É contudo necessário ter em consideração que um determinado critério de decisão depende dos benefícios associados aos resultados corretos bem como dos custos associados aos resultados incorretos. A previsão de um ciclone que acaba por não ocorrer (*falso positivo*) é tipicamente vista como tendo um custo menor comparativamente à falha na previsão de um

ciclone que ocorre (*falso negativo*), assim o critério a adotar para um diagnóstico positivo deverá estar do lado mais brando.

Num teste de diagnóstico existem dois tipos de erro que podem ocorrer na decisão, a escolha de uma *falha* (no sentido de declarar um doente como saudável) ou a escolha de um *falso alarme* (ao declarar uma pessoa saudável como doente). Qualquer pessoa perante um diagnóstico de doença contagiosa prefere um *falso alarme* a uma *falha* pois este tipo de erro conduzirá ao que se pode designar por “*um mal menor*” e, por conseguinte, irá optar sempre por um teste mais sensível. No entanto, ele deverá estar consciente que uma terapia disponível para este tipo de doença poderá ser, efectivamente, cara e deficiente, o que torna o teste pouco específico.

Observe-se por outro lado que em saúde pública, nomeadamente rastreios de doenças raras (em particular se não forem contagiosas), interessa mais o desempenho coletivo do que o individual, e nesse caso torna-se imperativo escolher um ponto de corte correspondente a uma especificidade elevada

Para contornar este tipo de situações, foi necessário desenvolver medidas alternativas de diagnóstico com propriedades mais robustas do que a *sensibilidade* e a *especificidade*. A análise ROC (*Receiver Operating Characteristic*) foi a técnica desenvolvida para tornar este tipo de problema.

A análise ROC é um método gráfico ⁽¹⁾ para avaliação, organização e seleção de sistemas de diagnóstico e/ou predição. Os gráficos ROC foram originalmente utilizados em deteção de sinais, para se avaliar a qualidade de transmissão de um sinal num canal com ruído. Os gráficos ROC também são muito utilizados em psicologia para se avaliar a capacidade de indivíduos distinguirem entre estímulo e não estímulo; em medicina, para analisar a qualidade de um determinado teste clínico; em economia (onde é conhecido como gráfico de Lorenz), para a avaliação de desigualdade de rendimento; e em previsão do tempo, para se avaliar a qualidade das predições de eventos raros.

Deste modo, neste capítulo começamos por fazer um paralelo entre a teoria dos testes de hipóteses e os testes de diagnóstico referindo o importante papel desempenhado pela curva ROC como uma medida alternativa de diagnóstico com propriedades robustas. A introdução de erros de classificação nos teste diagnóstico conduz-nos à redefinição dos conceitos de sensibilidade e de especificidade num teste composto. O cálculo da sensibilidade e da especificidade na metodologia de Dorfman é efetuado tendo em consideração o efeito de diluição e consequente rarefação. Por fim, analisam-se as alterações que os erros de classificação provocam no número esperado de análises efetuadas bem como as alterações provocadas pela existência de subpopulações na sensibilidade e na especificidade do teste.

⁽¹⁾Raramente são considerados procedimentos de inferência estatística que os tornem mais sofisticados e adequados à realidade, por exemplo na determinação do ponto de corte raramente se usa funções de perda que modelassem os riscos associados a erros de diagnóstico.

3.2 Teoria estatística

Suponha-se que se pretende estudar o comportamento probabilístico de um atributo de certa população – representado pela variável aleatória X – e admita-se que a respetiva função de distribuição depende de um parâmetro desconhecido, μ , sobre o qual se vão estabelecer conjeturas. O objetivo do teste de hipóteses vai ser o de decidir, com base na informação fornecida pelos dados, sobre a rejeição ou não rejeição de determinada hipótese paramétrica, que estabelece no espaço-parâmetro Θ uma partição,

$$\Theta_0 \cup \Theta_1 = \Theta \quad , \quad \Theta_0 \cap \Theta_1 = \emptyset,$$

onde $H_0 : \mu \in \Theta_0$ é a hipótese a testar, e $H_1 : \mu \in \Theta_1$ é a hipótese que corresponde ao conjunto das alternativas. À hipótese H_0 dá-se o nome de hipótese nula, designação tradicional que geralmente corresponde ao *statu quo* ou a algo que se pretende manter; a hipótese H_1 é designada por hipótese alternativa. Considerando o caso mais simples em que $\Theta_0 = \mu_0$ e $\Theta_1 = \mu_1$ tem-se:

$$H_0 : \text{A população tem média } \mu = \mu_0,$$

$$H_1 : \text{A população tem média } \mu = \mu_1.$$

Com base numa observação x , uma das hipóteses é aceite.

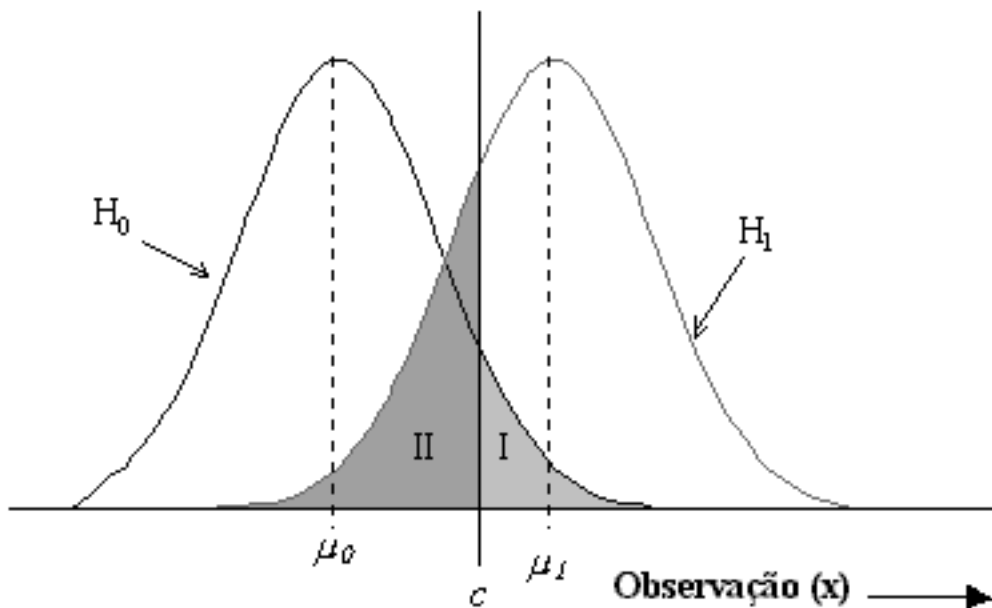


Figura 3.1: Funções densidade de duas populações

Como se pode verificar na Figura 3.1, H_0 é a hipótese nula que considera que a população tem valor médio $\mu = \mu_0$ e H_1 é a hipótese alternativa que considera que a população tem valor médio $\mu = \mu_1$. A área sombreada à direita do critério de decisão ($x > c$) representa a probabilidade de cometer um *erro do tipo I* e corresponde à probabilidade de rejeitar H_0 quando H_0 é verdadeira; a área sombreada à esquerda do critério de decisão ($x < c$) representa a probabilidade de cometer um *erro de tipo II* e corresponde à probabilidade de não rejeitar H_0 quando H_0 é falsa. A potência do teste é definida como a probabilidade de rejeitar H_0 quando H_0 é falsa.

O teste estatístico baseia-se na divisão do eixo das abcissas x em duas regiões separadas pelo ponto de corte c . Valores de x menores que c conduzirão à não rejeição da hipótese nula, H_0 , e valores de x maiores que c conduzirão à aceitação da hipótese alternativa, H_1 . Consoante o critério de decisão escolhido, pode-se determinar a probabilidade de cometer um *erro de tipo I ou tipo II* (Figura 3.1).

Claro que ao deslocarmos o ponto de corte para a direita (Figura 3.1) estamos a diminuir a probabilidade de cometer um *erro de tipo I* e necessariamente a aumentar a probabilidade de cometer um *erro de tipo II*. A redução simultânea das duas probabilidades (ou de uma delas, supondo a outra fixa) só se consegue à custa do aumento da dimensão da amostra suportando os custos que isso implica. Caso não se pretenda seguir essa via, resta-nos uma solução de compromisso, após uma avaliação das consequências de cada tipo de erro.

Na impossibilidade de minimizar simultaneamente os dois erros, torna-se necessário definir uma abordagem que permita considerá-los de alguma forma. Das várias alternativas possíveis, assume particular relevância a abordagem de Neyman-Pearson, que consiste em fixar a probabilidade associada ao *erro de tipo I* e minimizar a *probabilidade do erro de tipo II*, ou, dito de outra forma, fixar o tamanho do teste e maximizar a sua potência.

Assinale-se que esta forma de proceder atribui maior importância ao *erro de tipo I*, uma vez que é fixado num valor conveniente, ao passo que a potência é maximizada dentro das condicionantes existentes. Consequentemente, quando se rejeita H_0 tem-se sempre presente a probabilidade associada ao erro que se pode estar a cometer, situação que nem sempre acontece quando não se rejeita H_0 .

3.3 Diagnóstico do problema

Representando por x a variável em estudo, vamos supor que valores reduzidos de x favorecem a decisão “normal” (T^-) e que valores elevados de x favorecem a decisão “anormal” (T^+). Denote-se por $f(x|A)$ a distribuição dos valores de x para os casos designados anormais, x_A , e por $f(x|N)$ a distribuição dos valores de x para os casos designados normais, x_N . Graficamente, a situação descrita é ilustrada na Figura 3.2. Em termos de diagnóstico, a *fração de verdadeiros positivos* (FVP) corresponde à probabilidade de decidir que a característica em questão está presente, quando na realidade assim acontece. Por outro lado, a *fração de*

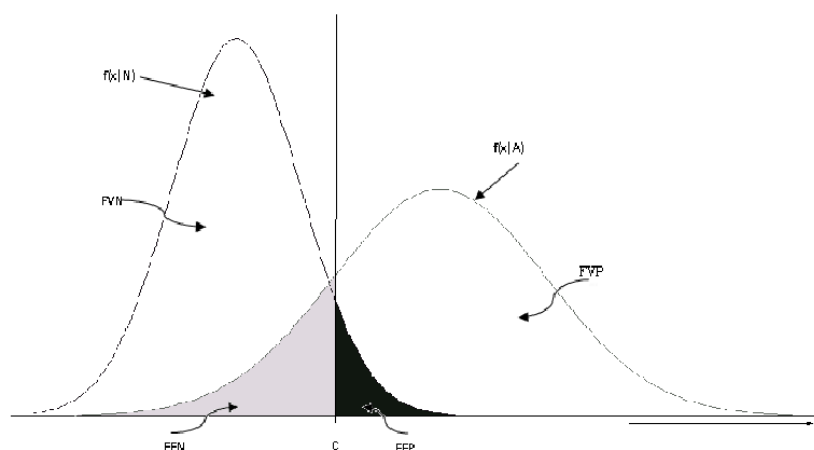


Figura 3.2: Sobreposição de duas populações hipotéticas

verdadeiros negativos (FVN) corresponde à probabilidade de decidir que a característica está ausente, quando de facto está ausente. Estas duas definições conduzem a outras duas diretamente relacionadas, a *fração de falsos positivos* (FFP) e a *fração de falsos negativos* (FFN), dadas por:

$$FFP = \frac{\text{número de decisões falsas positivas}}{\text{número de casos realmente negativos}}$$

e

$$FFN = \frac{\text{número de decisões falsas negativas}}{\text{número de casos realmente positivos}}$$

Assumindo-se que todos os casos podem ser diagnosticados como positivos ou negativos (no que diz respeito a uma determinada doença), então o número de decisões corretas adicionado ao número de decisões incorretas deverá ser igual ao número de casos com esse estado atual.

Assim, verifica-se que

$$FVP + FFN = 1$$

e

$$FVN + FFP = 1.$$

Geralmente um teste de diagnóstico é avaliado por duas destas medidas, FVP (*sensibilidade*) e FVN (*especificidade*). Em termos de diagnóstico, pode definir-se *sensibilidade* como a capacidade que um teste tem para detetar a doença no indivíduo, e a *especificidade* como a capacidade que o teste tem para excluir os indivíduos isentos de doença. Assim, caso a positividade ocorra quando o resultado da análise é superior ao ponto de corte tem-se que *valores de corte* elevados conduzem a testes pouco sensíveis e muito específicos, por outro lado, valores de corte baixos, conduzem a testes muito sensíveis e pouco específicos.

Num teste de diagnóstico as hipóteses podem ser definidas como:

H_0 : o indivíduo é anormal, X_A

H_1 : o indivíduo é normal, X_N ,

e conseqüentemente,

$$\begin{aligned}\alpha &= \mathbb{P}(\text{erro de tipo I}) = \mathbb{P}(\text{Rejeitar } H_0 | H_0 \text{ Verdadeira}) = \mathbb{P}(T^- | X_A) \\ &= 1 - \mathbb{P}(T^+ | X_A) = 1 - \text{sensibilidade},\end{aligned}$$

$$\begin{aligned}\beta &= \mathbb{P}(\text{erro de tipo II}) = \mathbb{P}(\text{Não Rejeitar } H_0 | H_1 \text{ Verdadeira}) = \mathbb{P}(T^+ | X_N) = \\ &= 1 - \mathbb{P}(T^- | X_N) = 1 - \text{especificidade}.\end{aligned}$$

Atendendo a que o *valor de corte* define a região de rejeição, isto é, define a dimensão dos erros de *tipo I* e de *tipo II*, à medida que se varia o *valor de corte*, estes erros vão variando em sentidos opostos, isto é, à medida que α aumenta, β diminui, e vice-versa.

Equilibrando a escolha do binómio sensibilidade/especificidade de uma forma a otimizar resultados, o que em geral é feito “graficamente”, através da distância dos pontos da curva ROC ao vértice superior esquerdo do quadrado $0 < x, y < 1$. Melhor seria – mas muito mais complexo – otimizar com base em funções de perda quantificando as conseqüências de erros de diagnóstico num e noutro sentido.

3.4 A curva ROC

A curva ROC é baseada na probabilidade de deteção, ou *fração de verdadeiros positivos* (FVP) e na probabilidade de falso alarme, ou *fração de falsos positivos* (FFP). Por definição, uma curva ROC é a representação gráfica dos pares *sensibilidade* ou FVP (ordenadas) e $1 - \text{especificidade}$ ou FFP (abscissas), resultante da variação do *valor de corte* ao longo de um eixo de decisão, x . Com efeito, uma curva ROC é uma descrição empírica da capacidade do sistema de diagnóstico poder discriminar entre dois estados num universo, onde cada ponto da curva representa um compromisso diferente entre a FVP e a FFP que pode ser adquirido pela adoção de um diferente *valor de corte*.

Sob o ponto de vista da teoria dos testes de hipóteses estatísticas, uma curva ROC é conceptualmente equivalente a uma curva que mostra a relação entre a potência do teste e a probabilidade de cometer um *erro de tipo I* com a variação do “valor crítico” (*valor de corte*) do teste estatístico.

Consoante os critérios adotados poder-se-á fazer corresponder um ponto na curva ROC. Assim, pode-se definir, um critério “estrito” (por exemplo, apenas se designa o paciente positivo quando a evidência da doença é muito forte) como sendo aquele que conduz a uma

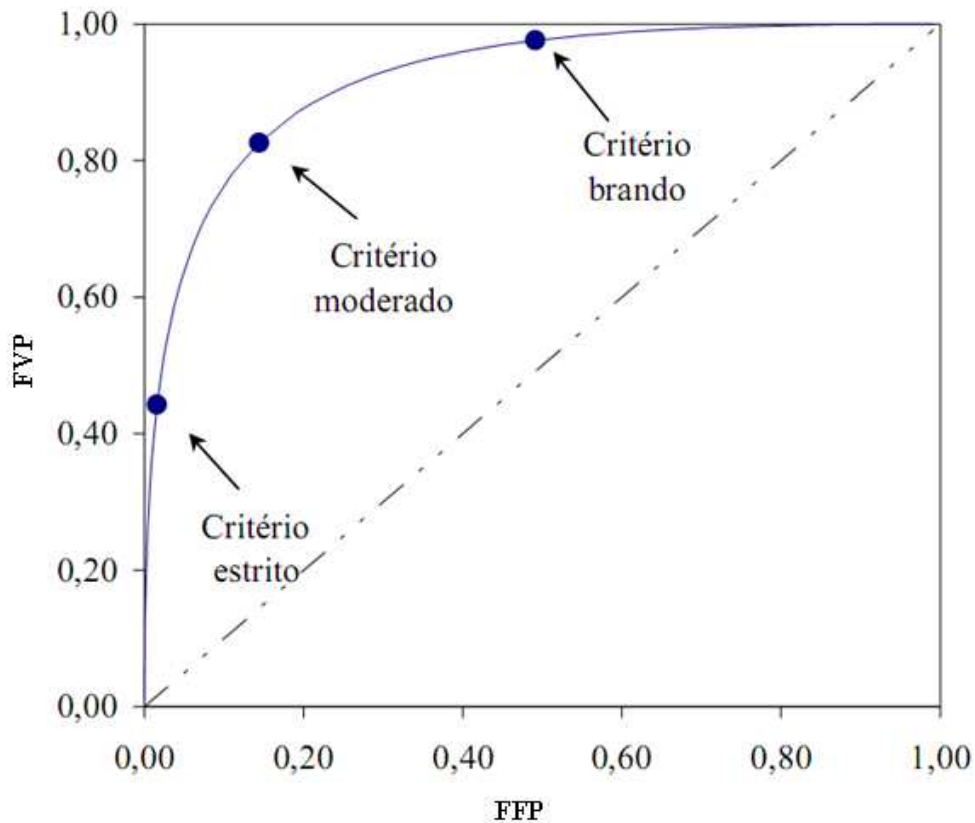


Figura 3.3: Curva ROC com a variação do critério de decisão

pequena fração de falsos positivos e também a uma relativamente pequena fração de verdadeiros positivos, isto é, gera um ponto na curva ROC que se situa no canto inferior esquerdo do espaço ROC. Progressivamente critérios menos estritos conduzem a maiores frações de ambos os tipos, isto é, pontos situados no canto superior direito da curva no espaço ROC. Esta situação pode ser descrita graficamente pela curva ROC apresentada na Figura 3.3. A localização ideal do ponto de corte é assim um processo difícil que depende de um equilíbrio adequado entre a sensibilidade e a especificidade, dado que um aumento da sensibilidade resulta no sacrifício da especificidade e vice-versa.

3.5 Testes com erros de classificação

3.5.1 Introdução

Ao relaxar a hipótese de ausência de erros de classificação, surge a necessidade de estender os conceitos de sensibilidade e especificidade para a realização de testes conjuntos. A maioria dos trabalhos em análises conjuntas considera que as probabilidades associadas aos erros de classificação são iguais nos testes individuais e nos testes conjuntos, como por exemplo Tu *et*

al. (1994, 1995), Kim *et al.* (2007) e Liu *et al.* (2011). Alguns autores consideram que estas probabilidades são distintas nos testes conjuntos, mas não dependem do número de indivíduos infetados dentro do grupo (Hwang, 1976); outros consideram modelos bastantes simplificados de forma a incluir o número de indivíduos infetados em cada grupo na modelação destas medidas (Hung & Swallow, 1999). Neste trabalho adotou-se a proposta de Santos, Pestana e Martins (2012) para modelar a sensibilidade e especificidade que tem em linha de conta o problema da diluição e conseqüente rarefação, uma vez que o valor da sensibilidade será condicionado ao número de elementos infetados no grupo. De facto, ao misturar o sangue de um indivíduo infetado com o de muitos indivíduos não infetados, com grande probabilidade a análise conjunta não irá identificar o sangue infetado devido à diluição e conseqüente rarefação das características que distinguem os dois tipos de sangue, o que provocará uma perda da sensibilidade do teste.

3.5.2 Sensibilidade e especificidade num teste simples

Como já repetidamente referimos, em geral um teste de diagnóstico é avaliado por duas características designadas por *sensibilidade* e *especificidade* que conjuntamente com a estimativa da prevalência permitem estimar o valor preditivo positivo e o valor preditivo negativo do teste. Denote-se por $\varphi_s \in (0, 1]$ a sensibilidade de um teste individual – probabilidade de obter um teste positivo (X_i^+) numa amostra infetada ($X_i = 1$) – e $\varphi_e \in (0, 1]$ a especificidade do teste – probabilidade de obter um teste negativo (X_i^-) numa amostra limpa ($X_i = 0$). Conseqüentemente $1 - \varphi_s$ representará a probabilidade de um falso negativo e $1 - \varphi_e$ a probabilidade de um falso positivo. As probabilidades dos possíveis acontecimentos num teste individual figuram na Tabela 3.1 (onde $q = 1 - p$).

		Resultado do teste		
		X_i^+	X_i^-	
$X_i = 1$		$\varphi_s p$	$(1 - \varphi_s) p$	p
$X_i = 0$		$(1 - \varphi_e) q$	$\varphi_e q$	q
		$\varphi_s p + (1 - \varphi_e) q$	$(1 - \varphi_s) p + \varphi_e q$	1

Tabela 3.1: Probabilidades num teste individual

3.5.3 Sensibilidade e especificidade num teste composto

Seja $I^{[n]} = \sum_{i=1}^n X_i$ o número de elementos infetados numa amostra de dimensão n e $I^{[i,n]} = \mathbb{P}(I^{[n]} = i) = \binom{n}{i} p^i (1 - p)^{n-i}$, $i = 0, 1, \dots, n$. Denote-se por $X^{[+,n]}$ ($X^{[-,n]}$) a ocorrência de um resultado positivo (negativo) num teste composto. Santos, Pestana e Martins (2012) compararam a sensibilidade φ_s e a especificidade φ_e de um teste simples ($n = 1$) com a

sensibilidade $\varphi_s^{[n]}$ e especificidade $\varphi_e^{[n]}$ de um teste composto por amostras de n indivíduos. A especificidade de um teste composto é definida como sendo a probabilidade de obter um teste negativo quando nenhuma das amostras está infectada, ou seja,

$$\varphi_e^{[n]} = \mathbb{P} (X^{[-,n]} | I^{[n]} = 0) .$$

Suponhamos que se pretende detetar a presença de uma bactéria para testar a contaminação de iogurtes numa unidade fabril. Caso se testem n iogurtes simultaneamente, começa-se por misturar as n amostras para posteriormente se retirar um mililitro da mistura para teste. Caso as n amostras estejam limpas, o teste conjunto dará negativo e, por conseguinte, este quadro será equivalente a retirar um mililitro de um iogurte não contaminado. Assim, podemos concluir que o valor de n não afeta a especificidade do teste ($\varphi_e^{[n]} = \varphi_e$).

Por outro lado, a sensibilidade de um teste composto $\varphi_s^{[n]}$ é definida como sendo a probabilidade de obter um teste positivo quando pelo menos uma amostra está infectada

$$\varphi_s^{[n]} = \mathbb{P} (X^{[+,n]} | I^{[n]} \geq 1) .$$

A sensibilidade do teste composto, $\varphi_s^{[n]}$, irá assumir valores distintos consoante o número de indivíduos infectados entre os n testados simultaneamente. Assim, denote-se por

$$\varphi_s^{[\gamma,n]} = \mathbb{P} (X^{[+,n]} | I^{[n]} = \gamma)$$

a sensibilidade do teste composto quando γ dos n elementos do grupo estão infetados. Nestas condições, (as diversas sensibilidades para o modelo Gaussiano podem ser consultadas na fórmula 7.8 da página 107) devido ao efeito de diluição e consequente rarefação é espetável que

$$0 \leq \varphi_s^{[1,n]} \leq \varphi_s^{[2,n]} \leq \dots \leq \varphi_s^{[n,n]} .$$

Recorrendo ao teorema da probabilidade total obtém-se:

$$\begin{aligned} \varphi_s^{[n]} &= \mathbb{P} (X^{[+,n]} | I^{[n]} \geq 1) \\ &= \frac{\sum_{j=1}^n \mathbb{P} (X^{[+,n]} | I^{[n]} = j) \mathbb{P} (I^{[n]} = j)}{\mathbb{P} (I^{[n]} \geq 1)} \\ &= \sum_{j=1}^n \varphi_s^{[j,n]} \frac{I^{[j,n]}}{1 - q^n} \\ &= \sum_{j=1}^n \varphi_s^{[j,n]} \lambda_j, \end{aligned}$$

onde a sucessão de números $\{\lambda_j\}_{j=1}^n \in (0, 1)$ com $\sum_{j=1}^n \lambda_j = 1$ representa a função massa de probabilidade de uma variável aleatória com distribuição binomial(n, p) truncada na origem. Assim, $\varphi_s^{[n]}$ é uma média ponderada dos $\varphi_s^{[j,n]}$.

p	0.15	0.1	0.05	0.03	0.025	0.02	0.015	0.01	0.005
n^*	3	4	5	6	7	8	9	11	15
λ_1	0.843	0.8478	0.9	0.925	0.926	0.931	0.941	0.951	0.965
λ_2	0.149	0.141	0.095	0.072	0.071	0.066	0.057	0.048	0.034
$\lambda_1 + \lambda_2$	0.992	0.989	0.995	0.997	0.997	0.997	0.998	0.999	0.999

Tabela 3.2: Valores de λ_1 e λ_2 para algumas prevalências com $n = n^*$

Na Tabela 3.2 figuram os valores de λ_1 e λ_2 para um conjunto de valores de p próximos de zero considerando-se grupos de dimensão ótima, ou seja $n = n^*$, para a metodologia de Dorfman. Podemos constatar que $\lambda_i > \lambda_j$ para $i < j$, sendo o primeiro peso decrescente com p e os restantes crescentes com p . Para $p = 0.15$, considerando $n^* = 3$ tem-se $\lambda_1 \approx 0.843$ e $\lambda_2 \approx 0.149$. Para uma prevalência $p = 0.01$, considerando $n^* = 11$, tem-se $\lambda_1 = 0.951$ e $\lambda_2 = 0.048$ e, por conseguinte, os restantes $\varphi_s^{[j,n]}$ com $j > 2$ podem ser negligenciados. Deste modo o valor de $\varphi_s^{[1,n]}$, sensibilidade de um teste composto numa amostra de n indivíduos dos quais apenas um está infectado, é fundamental para a determinação da sensibilidade de um teste composto e, conseqüentemente, caso a rarefação tenha grande influência na sensibilidade, ($\varphi_s^{[1,n]}$ tenha um valor reduzido comparativamente a φ_s) o teste composto não é recomendado. Considerando que a relação entre as diversas sensibilidades pode ser modelada através de

$$\varphi_s^{[\gamma,n]} = (1 - k_\gamma^{[n]}) \varphi_s \quad (3.1)$$

com $k_n^{[n]} \leq \dots \leq k_2^{[n]} \leq k_1^{[n]}$ tem-se

$$\begin{aligned} \varphi_s^{[n]} &= \sum_{j=1}^n \lambda_j [1 - k_j^{[n]}] \varphi_s \\ &= \sum_{j=1}^n \lambda_j \varphi_s - \sum_{j=1}^n \lambda_j k_j^{[n]} \varphi_s \\ &= \varphi_s - \sum_{j=1}^n \lambda_j k_j^{[n]} \varphi_s \end{aligned}$$

e, por conseguinte, a diferença entre a sensibilidade do teste simples e a sensibilidade do teste composto é dada por

$$\varphi_s - \varphi_s^{[n]} = \sum_{j=1}^n \lambda_j k_j^{[n]} \varphi_s.$$

Caso $p \approx 0$ tem-se

$$\varphi_s - \varphi_s^{[n]} \approx (\lambda_1 k_1^{[n]} + \lambda_2 k_2^{[n]}) \varphi_s. \quad (3.2)$$

As probabilidades dos possíveis acontecimentos num teste conjunto, considerando grupos de n indivíduos, figuram na Tabela 3.3

A utilização destes pesos será ilustrada no capítulo 6 com o modelo de Poisson e no capítulo 7 com os modelos exponencial e Gaussiano.

Resultado do teste

	$X^{[+,n]}$	$X^{[-,n]}$	
$I^{[n]} \geq 1$	$\varphi_s^{[n]} (1 - q^n)$	$(1 - \varphi_s^{[n]}) (1 - q^n)$	$1 - q^n$
$I^{[n]} = 0$	$(1 - \varphi_e) q^n$	$\varphi_e q^n$	q^n
	$\varphi_s^{[n]} - q^n (\varphi_s^{[n]} + \varphi_e - 1)$	$1 - \varphi_s^{[n]} + q^n (\varphi_s^{[n]} + \varphi_e - 1)$	1

Tabela 3.3: Probabilidades num teste conjunto

3.5.4 Sensibilidade e especificidade na metodologia de Dorfman

Para um elemento infetado ser corretamente identificado utilizando a metodologia de Dorfman, o grupo a que pertence terá que ser classificado como positivo (o que ocorre com probabilidade $\varphi_s^{[n]}$ que depende de $I^{[n]}$) e no teste individual terá novamente que obter um teste positivo (o que ocorre com probabilidade φ_s). Assim sendo, supondo que os resultados dos testes são independentes, a sensibilidade será dada por

$$\begin{aligned}
 \varphi_{s_n} &= \mathbb{P}(X_1^+ | X_1 = 1) = \sum_{i=0}^{n-1} \mathbb{P}(X_1^+ | X_1 = 1, I^{[n-1]} = i) \mathbb{P}(I^{[n-1]} = i) \\
 &= \sum_{i=0}^{n-1} \mathbb{P}(X_1^+ | X_1 = 1) \mathbb{P}(X^{[+,n]} | I^{[n]} = i + 1) I[i, n - 1] \\
 &= \varphi_s \sum_{i=0}^{n-1} \varphi_s^{[i+1, n]} I[i, n - 1],
 \end{aligned}$$

e $\varphi_{s_n} \leq \varphi_s$, ou seja, a sensibilidade utilizando a metodologia de Dorfman é inferior à dos testes individuais.

Existem duas possibilidades de um indivíduo não infetado ser corretamente identificado: ou o grupo a que pertence tem um resultado negativo ou o teste composto tem resultado positivo e o resultado do teste individual dá negativo. Assim, a especificidade φ_{e_n} será dada por

$$\begin{aligned}
 \varphi_{e_n} &= \mathbb{P}(X_1^- | X_1 = 0) = \sum_{i=0}^{n-1} \mathbb{P}(X_1^- | X_1 = 0, I^{[n-1]} = i) \mathbb{P}(I^{[n-1]} = i) \\
 &= \sum_{i=0}^{n-1} [\mathbb{P}(X_1^- | X_1 = 0) \mathbb{P}(X^{[+,n]} | I^{[n]} = i) + \mathbb{P}(X^{[-,n]} | I^{[n]} = i)] I[i, n - 1] \\
 &= [\varphi_e + \varphi_e (1 - \varphi_e)] q^{n-1} + \sum_{i=1}^{n-1} [\varphi_e \varphi_s^{[i, n]} + (1 - \varphi_s^{[i, n]})] I[i, n - 1] \\
 &= 1 - (1 - \varphi_e) \left[(1 - \varphi_e) q^{n-1} + \sum_{i=1}^{n-1} \varphi_s^{[i, n]} I[i, n - 1] \right] = 1 - (1 - \varphi_e) \xi,
 \end{aligned}$$

onde $\xi = \xi(\varphi_s^{[i,n]}, \varphi_e, p, n)$ é uma média ponderada de $(1 - \varphi_e)$ e $\varphi_s^{[i,n]}$, $i = 1, \dots, n - 1$, o que implica que $\xi \leq 1$ e consequentemente $\varphi_{e_n} \geq \varphi_e$. Assim se constata que a especificidade na metodologia de Dorfman é superior à especificidade dum teste individual.

Vamos admitir que a sensibilidade e a especificidade de um teste não sofrem alterações pelo facto de ao compor uma amostra combinada com frações de cada uma das unidades amostrais individuais haver um efeito de diluição e consequente rarefação, o que decerto nas situações mais usuais deve alterar a probabilidade de deteção de infetados ($\varphi_s^{[\gamma,n]} = \varphi_s$, $1 \leq \gamma \leq n$), ou seja, vamos excluir o efeito da diluição e consequente rarefação. Neste caso particular, as probabilidades dos possíveis acontecimentos num teste conjunto, considerando grupos de n indivíduos, são apresentadas na Tabela 3.4

Resultado do teste

	$X^{[+,n]}$	$X^{[-,n]}$	
$I^{[n]} \geq 1$	$\varphi_s (1 - q^n)$	$(1 - \varphi_s) (1 - q^n)$	$1 - q^n$
$I^{[n]} = 0$	$(1 - \varphi_e) q^n$	$\varphi_e q^n$	q^n
	$\varphi_s (1 - q^n) + (1 - \varphi_e) q^n$	$(1 - \varphi_s) (1 - q^n) + \varphi_e q^n$	1

Tabela 3.4: Probabilidades num teste conjunto com $\varphi_s^{[\gamma,n]} = \varphi_s$

Excluindo o fator de rarefação, ou seja, assumindo que $\varphi_s^{[\gamma,n]} = \varphi_s$ tem-se $\varphi_{s_n} = \varphi_s^2 \leq \varphi_s$ que não depende do valor de n utilizado.

Em relação à especificidade teremos:

$$\begin{aligned} \varphi_{e_n} &= \varphi_e q^{n-1} + \varphi_e (1 - \varphi_e) q^{n-1} + \varphi_e \varphi_s (1 - q^{n-1}) + (1 - \varphi_s) (1 - q^{n-1}) \\ &= 1 - (1 - \varphi_e) [(1 - \varphi_e) q^{n-1} + \varphi_s (1 - q^{n-1})] = 1 - (1 - \varphi_e) \xi \end{aligned}$$

onde $\xi = \xi(\varphi_s, \varphi_e, p, n)$ corresponde à probabilidade de se obter um teste positivo num grupo com $n - 1$ indivíduos. Assim, tem-se $\xi(\varphi_s, \varphi_e, p, n) \leq 1$ e consequentemente obtém-se $\varphi_{e_n} \geq \varphi_e$.

Conclui-se assim que a especificidade na metodologia de Dorfman é superior à dos testes individuais e, por conseguinte, a taxa global de falsos positivos é inferior (ver Sousa (2006)). Note-se que quando se analisa a mistura e se obtém um resultado positivo efetuam-se análises individuais com o intuito de identificar qual ou quais dessas análises dão positivas.

A especificidade pode ainda escrever-se da seguinte forma

$$\varphi_{e_n} = 1 - (1 - \varphi_e) \varphi_s - (1 - \varphi_e) (1 - \varphi_e - \varphi_s) q^{n-1}, \quad (3.3)$$

donde ressalta que para valores razoáveis de φ_s e de φ_e , isto é, para $\varphi_s + \varphi_e > 1$, a especificidade na metodologia de Dorfman é crescente com q^{n-1} . Caso se recorra à aproximação $n \approx \frac{1}{\sqrt{p}}$ teremos $q^{n-1} \approx (1 - p)^{\frac{1}{\sqrt{p}} - 1}$ que é decrescente com p . Assim, quanto menor for

o valor de p maior será o ganho na utilização de testes compostos no que respeita à especificidade. Refira-se ainda que nos casos em que a especificidade φ_e é elevada, a expressão $(1 - \varphi_e)(1 - \varphi_e - \varphi_s)$ assumirá valores reduzidos e, por conseguinte, a variação de p não implicará alterações significativas na especificidade φ_{e_n} .

3.5.5 Número esperado de análises em testes com erros de classificação

A existência de erros de classificação provoca alterações no número esperado de análises em cada grupo de n indivíduos. Assim,

$$\mathbb{P}[T_n = 1] = (1 - \varphi_s)(1 - q^n) + \varphi_e q^n,$$

e

$$\begin{aligned} \mathbb{P}[T_n = n + 1] &= 1 - P[T_n = 1] \\ &= 1 - (1 - \varphi_s)(1 - q^n) - \varphi_e q^n \\ &= \varphi_s(1 - q^n) + (1 - \varphi_e)q^n. \end{aligned}$$

O número esperado de análises em cada grupo de n indivíduos é então

$$\begin{aligned} \mathbb{E}(T_n) &= (1 - \varphi_s)(1 - q^n) + \varphi_e q^n + (n + 1) \times \\ &\quad \times [\varphi_s(1 - q^n) + (1 - \varphi_e)q^n] \\ &= 1 + n[\varphi_s(1 - q^n) + (1 - \varphi_e)q^n], \quad n \geq 2, \end{aligned}$$

e consequentemente o custo relativo é dado por

$$E_n = \frac{1}{n} + \varphi_s(1 - q^n) + (1 - \varphi_e)q^n, \quad n \geq 2.$$

Note-se que o custo relativo, E_n , é uma função crescente com φ_s e decrescente com φ_e . Na Figura 3.4 pode-se observar a variação do custo relativo para valores de prevalência $p \in [0.2, 0.3]$, assumindo valores para a sensibilidade e para a especificidade iguais a 0.8 e 0.9 respetivamente. É visível que não existem valores de prevalência $p \in [0.2, 0.3]$ onde não compense juntar sangues em virtude de o número médio de análises por indivíduo ser sempre inferior a 1.

Considerem-se as diferenças:

$$\begin{aligned} \Delta_n(p) &= E_{n+1} - E_n \\ &= \left[\frac{1}{n+1} + \varphi_s(1 - q^{n+1}) + (1 - \varphi_e)q^{n+1} \right] - \\ &\quad - \left[\frac{1}{n} + \varphi_s(1 - q^n) + (1 - \varphi_e)q^n \right] \\ &= \frac{-1}{n(n+1)} + p(1-p)^n[\varphi_s + \varphi_e - 1]. \end{aligned}$$

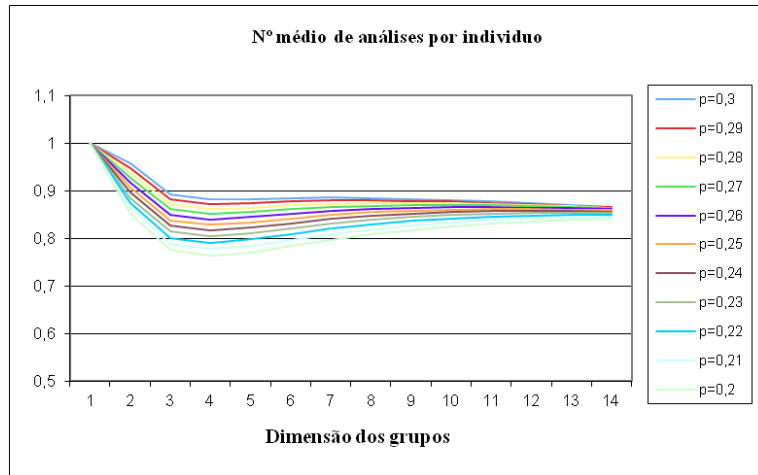


Figura 3.4: Variação do CR com $\varphi_s = 0.8$ e $\varphi_e = 0.9$

Derivando $\Delta_n(p)$ relativamente a p e igualando a zero obtém-se:

$$\begin{aligned} \Delta'_n(p) = 0 &\Leftrightarrow [\varphi_s + \varphi_e - 1] [(1 - p)^n - np(1 - p)^{n-1}] = 0 \\ &\Leftrightarrow (1 - p)^{n-1} [1 - p - np] = 0 \\ &\Leftrightarrow (1 - p)^{n-1} = 0 \vee p(n + 1) = 1 \\ &\Leftrightarrow p = \frac{1}{n + 1} \vee p = 1, \end{aligned}$$

onde $p = 1$ é uma raiz de multiplicidade $n - 1$. Assim obtém-se

	0		$\frac{1}{n+1}$		1
$\Delta'_n(p)$	+	+	0	-	0
$\Delta_n(p)$		↗	Max	↘	

ou seja, fixado um valor de n , $\Delta_n(p)$ é máximo em $p = \frac{1}{n+1}$.

Para $n = 2$ tem-se

$$\Delta_2\left(\frac{1}{3}\right) = -\frac{1}{6} + \underbrace{(\varphi_s + \varphi_e - 1)}_{<1} \underbrace{\left[\frac{1}{3} \times \left(\frac{2}{3}\right)^2\right]}_{<\frac{1}{6}} < 0 \Rightarrow E_3 < E_2.$$

Dado que $\Delta_2(p)$ é máximo em $p = \frac{1}{3}$ e se obteve $\Delta_2\left(\frac{1}{3}\right) < 0$, podemos concluir que $E_3 < E_2, \forall p$. Ou seja, não existem valores de prevalência para os quais seja mais favorável juntar análises 2 a 2 do que 3 a 3. Assiste-se, também neste caso, a uma variação brusca de isoladas $n = 1$ para análises de trios $n = 3$.

Para $n \geq 3$ tem-se

$$\begin{aligned} \Delta_n \left(\frac{1}{n+1} \right) &= \frac{-1}{n(n+1)} + \frac{1}{n+1} \left(1 - \frac{1}{n+1} \right)^n (\varphi_s + \varphi_e - 1) \\ &= \frac{1}{n+1} \left[\left(1 + \frac{1}{n} \right)^{-n} (\varphi_s + \varphi_e - 1) - \frac{1}{n} \right] \\ &> \frac{1}{n+1} \underbrace{\left[\frac{(\varphi_s + \varphi_e - 1)}{e} - \frac{1}{n} \right]}_{<0 \text{ se } \varphi_s + \varphi_e < 1 + \frac{e}{n}} \end{aligned}$$

Podemos concluir que $\forall n \geq 3$, se $\varphi_s + \varphi_e < 1 + \frac{e}{n}$, não existem valores de prevalência para os quais seja mais favorável juntar os indivíduos n a n do que $n+1$ a $n+1$ para $\forall p$.

Concretizando para alguns valores de n tem-se:

n	3	4	5	6	7	8	9	10
$1 + \frac{e}{n}$	1.9061	1.6796	1.5437	1.4530	1.3883	1.3798	1.3020	1.2718

Tabela 3.5: Valores de $1 + \frac{e}{n}$

- Se $\varphi_e + \varphi_s < 1.9061$ não existem valores de prevalência para os quais seja mais favorável juntar os indivíduos 3 a 3 do que 4 a 4.
- Se $\varphi_e + \varphi_s < 1.6796$ não existem valores de prevalência para os quais seja mais favorável juntar os indivíduos 4 a 4 do que 5 a 5.
- ⋮
- Se $\varphi_e + \varphi_s < 1.2718$ não existem valores de prevalência para os quais seja mais favorável juntar os indivíduos 10 a 10 do que 11 a 11.

3.6 Subpopulações

Santos, Pestana e Martins (2012) consideram a divisão da população em k subpopulações, com pesos w_1, w_2, \dots, w_k tais que $\sum_{i=1}^k w_i = 1$, e taxas de prevalência p_1, p_2, \dots, p_k com $\sum_{i=1}^k w_i p_i = p$. Nestas condições, ao efetuar testes conjuntos, coloca-se a questão de saber se a separação das k subpopulações se traduz, ou não, num ganho de eficiência. Outra questão relevante consiste em saber se a existência de subpopulações provoca alterações na sensibilidade e na especificidade.

No caso de separarmos as k subpopulações e considerando que cada população é composta por N indivíduos, o número esperado de testes é dado por

$$\mathbb{E}[T_S] = \sum_{i=1}^k w_i N \min \left\{ \frac{n_i + 1}{n_i} - (1 - p_i)^{n_i}, 1 \right\}, n_i \geq 2,$$

onde n_i é determinado em função de p_i de forma análoga à apresentada para a determinação de n . O custo relativo, $CR = \frac{\mathbb{E}[T_S]}{N}$, corresponde à média ponderada dos custos relativos de cada população. A sensibilidade dos testes conjuntos não sofre alterações, mas a especificidade para cada subpopulação será dada por

$$\varphi_{e_{n_i}} = 1 - (1 - \varphi_e) \varphi_s - (1 - \varphi_e) (1 - \varphi_e - \varphi_s) q_i^{n_i - 1}.$$

p_1	.05	.025	.005	.0005	.05	.05	.05	.05	.005	.005	.005
w_1	.05	.05	.05	.05	.25	.1	.75	.9	.25	.1	.75
p_2	.15	.175	.195	.1995	.11(6)	.10(5)	.25	.55	.131(6)	.110(5)	.385
w_2	.5	.5	.5	.5	.75	.9	.25	.1	.75	.9	.25
n_1	4	6	14	45	4	4	4	4	14	14	14
n_2	3	2	2	2	3	3	2	1	3	3	2
$\varphi_{e_n}^{[S]}$.9939	.9936	.9934	.9933	.994	.994	.9936	.9931	.9938	.9939	.9926
CR	.5004	.4686	.408	.3658	.5124	.5173	.4776	.4465	.4704	.5013	.3203

Tabela 3.6: Variação de $\varphi_{e_n}^{[S]}$ e do CR na presença de subpopulações

Quanto à especificidade total tem-se

$$\begin{aligned} \varphi_{e_n}^{[S]} &= \sum_{i=1}^k w_i [1 - (1 - \varphi_e) \varphi_s - (1 - \varphi_e) (1 - \varphi_e - \varphi_s) q_i^{n_i - 1}] \\ &= 1 - (1 - \varphi_e) \varphi_s - (1 - \varphi_e) (1 - \varphi_e - \varphi_s) \sum_{i=1}^k w_i q_i^{n_i - 1}, \end{aligned}$$

que pode ser aproximada, para prevalências próximas de zero, por

$$\varphi_{e_n}^{[S]} \approx 1 - (1 - \varphi_e) \varphi_s - (1 - \varphi_e) (1 - \varphi_e - \varphi_s) \sum_{i=1}^k w_i (1 - p_i)^{\frac{1}{\sqrt{p_i}} - 1}.$$

Dado que $f(p) = (1 - p)^{\frac{1}{\sqrt{p}} - 1}$ é uma função convexa para valores de $p \in (0, 1)$, isto é, $f''(p) \geq 0 \forall p \in (0, 1)$ tem-se pela desigualdade de Jensen $\sum_{i=1}^k w_i f(p_i) \geq f\left(\sum_{i=1}^k w_i p_i\right)$ e, por conseguinte,

$$\sum_{i=1}^k w_i (1 - p_i)^{\frac{1}{\sqrt{p_i}} - 1} \geq \left(1 - \sum_{i=1}^k w_i p_i\right) \sqrt{\sum_{i=1}^k w_i p_i}^{-1} = (1 - p)^{\frac{1}{\sqrt{p}} - 1}.$$

Conclui-se assim que a especificidade do teste aumenta no caso de separarmos as subpopulações.

Considerando, por exemplo, $k = 2$, $p = 0.1$, $\varphi_s = 0.95$ e $\varphi_e = 0.90$ onde, caso não se considerasse as subpopulações se obteria $\varphi_{e_n} = 0.9727$ e $CR = 0.5939$, obtêm-se os resultados apresentados na Tabela 3.6.

A redução do CR (ver última linha da Tabela 3.6) permite-nos concluir que é mais eficiente testar as subpopulações separadamente (caso pelo menos uma das subpopulações tenha taxa de prevalência inferior a 30,7%). Por outro lado, a sensibilidade do teste não sofre alteração, no entanto a especificidade tem um ligeiro aumento. A redução do custo relativo é tanto maior quanto maior for a distância entre p_1 e p_2 . Refira-se ainda que o ganho em eficiência é tanto maior quanto maior for a dimensão da subpopulação com menor taxa de prevalência.

Capítulo 4

Estimação da taxa de prevalência na ausência de erros de classificação

4.1 Introdução

Nos capítulos anteriores realçamos a capacidade da teoria de Dorfman e seus aperfeiçoamentos permitirem determinar a dimensão n ótima para cada grupo em função da taxa de prevalência p , por forma a minimizar o número esperado de testes necessários para a identificação de todos os indivíduos de uma determinada população.

Porém, o recurso a análises conjuntas não se esgota com a classificação de indivíduos, podendo adotar-se a perspectiva direcionada ao contrário: com base em classificações de análises combinadas baseadas na combinação de diversos números n de unidades amostrais, e concomitante determinação de quantos são detetados como infetados nesses diversos grupos, estimar a taxa de prevalência da doença.

Os estimadores obtidos para testes compostos, têm, sob determinadas condições, melhor comportamento que os estimadores tradicionais baseados em testes individuais, cf. Sobel e Elashoff (1975), Loyer (1983) e Garner *et al.* (1989).

Chen & Swallow (1990), Lancaster & Keller-McNulty (1998) e Hung & Swallow (1999) analisaram o enviesamento, eficiência e a robustez destes estimadores e concluíram que os estimadores baseados em testes conjuntos permitem não apenas a obtenção de ganhos monetários (minimizando o número de testes efetuados) mas também a obtenção de estimativas mais precisas (minimizando o erro quadrático médio do estimador) comparativamente às utilizadas nos testes individuais.

Refira-se ainda que as suas aplicações não se restringem a análises ao sangue ou à urina, podendo ser aplicadas a qualquer análise em que a mistura homogénea das amostras individuais se possa efetuar. Um grande número de aplicações e referências bibliográficas referentes aos testes conjuntos foi proposta por Boswell *et al.* (1996). Entre estas aplicações sublinhamos a potencial utilidade desta metodologia na área do controlo de qualidade, nomeadamente

na amostragem para aceitação de, por exemplo, produtos alimentares.

Neste capítulo começamos por analisar a qualidade do estimador pontual da taxa de prevalência. Para tal, determinam-se os valores da variância, do viés e do erro quadrático médio para um conjunto de valores da prevalência e para grupos de diversas dimensões. No que respeita à estimação intervalar da taxa de prevalência, analisamos, via simulação, a qualidade dos estimadores disponíveis no package *binGroup* para o R (Bilder *et al.*, 2010).

4.2 Estimação por verosimilhança máxima

Em grupos de dimensão n , numa primeira fase apenas nos interessa saber se o grupo está limpo ($\sum_{i=1}^n X_i = 0$) ou contaminado ($\sum_{i=1}^n X_i \geq 1$), sendo a probabilidade de o grupo estar contaminado dada por $\pi_n = \mathbb{P}(\sum_{i=1}^n X_i \geq 1) = 1 - (1 - p)^n$. Resolvendo a equação em ordem a p obtém-se

$$p = 1 - (1 - \pi_n)^{1/n}.$$

Denotando por X_n^+ o número de testes positivos quando m grupos (com n indivíduos cada) são analisados, tem-se que $X_n^+ \sim \text{binomial}(m, \pi_n)$ e, por conseguinte, a probabilidade de observar x grupos positivos é dada por

$$\mathbb{P}(X_n^+ = x | m, \pi_n) = \binom{m}{x} \pi_n^x (1 - \pi_n)^{m-x}.$$

Dado que π_n depende da taxa de prevalência, p , e da dimensão do grupo, n , a probabilidade referida pode ser calculada como

$$\mathbb{P}(X_n^+ = x | m, n, \pi_n) = \binom{m}{x} [1 - (1 - p)^n]^x [(1 - p)^n]^{m-x}. \quad (4.1)$$

Atendendo a que o estimador de máxima verosimilhança de $[1 - (1 - p)^n]$ é $\frac{X_n^+}{m}$ e recorrendo à propriedade da invariância dos estimadores de máxima verosimilhança obtém-se

$$\hat{p} = 1 - \left(1 - \frac{X_n^+}{m}\right)^{1/n}.$$

Seja $\hat{\theta} = \hat{\theta}(\mathbf{X})$ um estimador pontual do parâmetro θ . O erro quadrático médio de $\hat{\theta}$ é dado por

$$EQM(\hat{\theta}) = \mathbb{E}(\hat{\theta} - \theta)^2. \quad (4.2)$$

Mostra-se, sem dificuldade, que

$$EQM(\hat{\theta}) = \text{var}(\hat{\theta}) + [B(\hat{\theta})]^2, \quad (4.3)$$

onde $B(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$ representa o viés do estimador. Assim, o erro quadrático médio de um estimador é igual à soma da sua variância com o quadrado do seu enviesamento. Com efeito,

$$\begin{aligned} EQM(\hat{\theta}) &= \mathbb{E} \left\{ \left[(\hat{\theta} - \mathbb{E}(\hat{\theta})) + (\mathbb{E}(\hat{\theta}) - \theta) \right]^2 \right\} \\ &= \mathbb{E} \left[(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2 + (\mathbb{E}(\hat{\theta}) - \theta)^2 + 2(\hat{\theta} - \mathbb{E}(\hat{\theta}))(\mathbb{E}(\hat{\theta}) - \theta) \right] \\ &= \mathbb{E} \left[(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2 \right] + (\mathbb{E}(\hat{\theta}) - \theta)^2 \\ &= var(\hat{\theta}) + [B(\hat{\theta})]^2. \end{aligned}$$

No caso de amostras individuais ($n = 1$), considere-se a variável aleatória X_1^+ que representa o número de testes positivos em m testes individuais. Dado que a probabilidade de sucesso é dada por $\pi_1 = 1 - (1 - p)^1 = p$ tem-se que $X_1^+ \sim$ binomial (m, p) . Atendendo a que o estimador de máxima verosimilhança de p é dado por

$$\hat{p} = \frac{X_1^+}{m},$$

para o qual se tem:

$$\mathbb{E}(\hat{p}) = \mathbb{E} \left(\frac{X_1^+}{m} \right) = \frac{1}{m} \mathbb{E}(X_1^+) = \frac{1}{m} mp = p.$$

Constata-se assim que \hat{p} é um estimador centrado de p e, por conseguinte, tem-se

$$\begin{aligned} EQM(\hat{p}) &= var(\hat{p}) = var \left(\frac{X_1^+}{m} \right) = \frac{var(X_1^+)}{m^2} = \\ &= \frac{mp(1-p)}{m^2} = \frac{p(1-p)}{m}. \end{aligned}$$

Refira-se que a variância do estimador é igual ao limite inferior de Cramér-Rao e por conseguinte $\hat{p} = \frac{X_1^+}{m}$ é um estimador centrado de variância uniformemente mínima de p . Contudo, recorrendo à desigualdade de Jensen, demonstra-se que para $n > 1$ se tem $\mathbb{E}(\hat{p}) > p$, ou seja, para testes conjuntos, \hat{p} tem um viés positivo.

O valor médio, a variância e o erro quadrático médio de \hat{p} podem ser calculados através das seguintes expressões

$$\mathbb{E}(\hat{p}) = \sum_{x=0}^m \left[1 - \left(1 - \frac{x}{m} \right)^{\frac{1}{n}} \right] \binom{m}{x} \pi_n^x (1 - \pi_n)^{m-x}, \quad (4.4)$$

$$var(\hat{p}) = \sum_{x=0}^m \left[1 - \left(1 - \frac{x}{m} \right)^{\frac{1}{n}} \right]^2 \binom{m}{x} \pi_n^x (1 - \pi_n)^{m-x} - \mathbb{E}^2(\hat{p}) \quad (4.5)$$

e

$$EQM(\hat{p}) = var(\hat{p}) + [B(\hat{p})]^2 \tag{4.6}$$

Nas Tabelas 4.1 e 4.2 figuram respectivamente os valores do viés e do erro quadrático médio para $m = 30$. Note-se que no caso de amostras individuais ($n = 1$) o estimador tem viés nulo. Para $n > 1$ verifica-se que o viés aumenta com a dimensão do grupo. Para $n > 1$ o enviesamento do estimador de p é uma consequência do desconhecimento do número de indivíduos infetados que estão na origem da positividade do teste ao grupo. Nas tabelas apresentadas seguidamente recorreu-se à notação $e^{-a} = 10^{-a}$.

p	.50	.25	.15	.1	.05	.025	.01	.005	.001
$n = 1$	0	0	0	0	0	0	0	0	0
$n = 2$.0067	.0025	.0014	.0009	.0004	.0002	$9e^{-5}$	$4e^{-5}$	$8e^{-6}$
$n = 3$.0187	.0040	.0020	.0013	.0006	.0003	.0001	$6e^{-5}$	e^{-5}
$n = 4$.0664	.0055	.0025	.0015	.0007	.0003	.0001	$6e^{-5}$	e^{-5}
$n = 5$.1669	.0073	.0030	.0017	.0008	.0004	.0001	$7e^{-5}$	e^{-5}
$n = 6$.2803	.0107	.0035	.0019	.0008	.0004	.0001	$7e^{-5}$	e^{-5}
$n = 7$.3695	.0190	.0040	.0021	.0009	.0004	.0002	$7e^{-5}$	e^{-5}
$n = 8$.4272	.0383	.0046	.0023	.0009	.0004	.0002	$8e^{-5}$	e^{-5}
$n = 9$.4609	.0739	.0055	.0025	.0010	.0004	.0002	$8e^{-5}$	$2e^{-5}$
$n = 10$.4794	.1273	.0070	.0027	.0010	.0004	.0002	$8e^{-5}$	$2e^{-5}$
$n = 15$.4993	.4828	.0581	.0049	.0012	.0005	.0002	$8e^{-5}$	$2e^{-5}$
$n = 20$.5000	.6732	.2520	.0227	.0015	.0005	.0002	$9e^{-5}$	$2e^{-5}$
$n = 25$.5000	.7305	.4933	.0970	.0018	.0006	.0002	$9e^{-5}$	$2e^{-5}$
$n = 50$.5000	.7500	.8417	.7657	.0868	.0009	.0002	.0001	$2e^{-5}$
$n = 100$.5000	.7500	.8500	.8992	.7922	.0817	.0003	.0001	$2e^{-5}$
$n = 1000$.5000	.7500	.8500	.9000	.9500	.9750	.9887	.8141	$3e^{-5}$

Tabela 4.1: Valores do viés ($m = 30$)

Recorde-se que o facto de o resultado do teste conjunto ser positivo significa apenas que pelo menos uma das amostras individuais está infetada, desconhecendo-se portanto se é apenas uma, duas, ou se inclusivamente estão todas infetadas. Esta ausência de informação é que conduz à existência de enviesamento.

Se a dimensão do grupo for demasiadamente elevada para uma determinada prevalência, a probabilidade do grupo ser positivo $\pi_n = 1 - (1 - p)^n$ aumenta e, por conseguinte, o valor médio do estimador vai ser superior ao verdadeiro valor do parâmetro (ver equação 4.4 da página 43) originando um aumento do viés. Verifica-se assim que quanto maior for a dimensão do grupo maior será o viés do estimador. Para valores de n muito elevados começamos a ter todos os grupos infetados (quase certamente) e, consequentemente, a estimativa será igual a 1 (logo o viés converge para $1 - p$ quando $n \rightarrow \infty$). Refira-se que o viés também aumenta com a taxa de prevalência já que π_n é uma função crescente com p .

p	.50	.25	.15	.1	.05	.025	.01	.005	.001
$n = 1$.0083	.0063	.0043	.0030	.0016	.0008	.0003	.0002	$3e^{-5}$
$n = 2$.0067	.0038	.0024	.0016	.0008	.0004	.0002	$8e^{-5}$	$2e^{-5}$
$n = 3$.0111	.0030	.0018	.0012	.0006	.0003	.0001	$6e^{-5}$	e^{-5}
$n = 4$.0400	.0028	.0015	.0009	.0004	.0002	$9e^{-5}$	$4e^{-5}$	$9e^{-6}$
$n = 5$.0989	.0030	.0013	.0008	.0004	.0002	$7e^{-5}$	$3e^{-5}$	$7e^{-6}$
$n = 6$.1589	.0045	.0012	.0007	.0003	.0001	$6e^{-5}$	$3e^{-5}$	$6e^{-6}$
$n = 7$.2008	.0105	.0012	.0007	.0003	.0001	$5e^{-5}$	$2e^{-5}$	$5e^{-6}$
$n = 8$.2251	.0262	.0012	.0006	.0003	.0001	$4e^{-5}$	$2e^{-5}$	$4e^{-6}$
$n = 9$.2378	.0559	.0015	.0006	.0002	.0001	$4e^{-5}$	$2e^{-5}$	$4e^{-6}$
$n = 10$.2441	.1001	.0023	.0006	.0002	.0001	$4e^{-5}$	$2e^{-5}$	$3e^{-6}$
$n = 15$.2499	.3768	.0474	.0014	.0002	$7e^{-5}$	$2e^{-5}$	e^{-5}	$2e^{-6}$
$n = 20$.2500	.5122	.2211	.0171	.0002	$6e^{-5}$	$2e^{-5}$	$9e^{-6}$	$2e^{-6}$
$n = 25$.2500	.5503	.4298	.0870	.0002	$5e^{-5}$	$2e^{-5}$	$7e^{-6}$	e^{-6}
$n = 50$.2500	.5625	.7162	.6939	.0818	$9e^{-5}$	$9e^{-6}$	$4e^{-5}$	$7e^{-7}$
$n = 100$.2500	.5625	.7225	.8094	.7553	.0792	$8e^{-6}$	$2e^{-6}$	$4e^{-7}$
$n = 1000$.2500	.5625	.7225	.8100	.9025	.9506	.9788	.8103	e^{-6}

Tabela 4.2: Valores do EQM ($m = 30$)

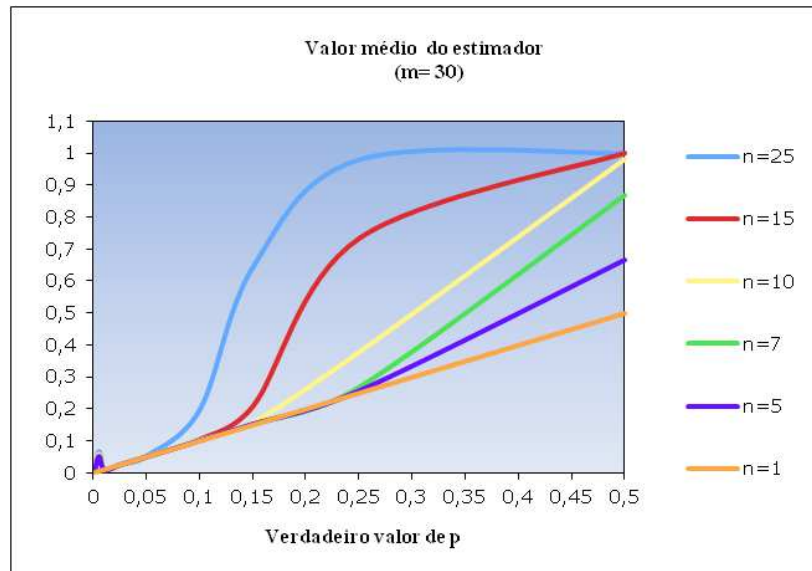


Figura 4.1: Variações do viés do estimador

Na Figura 4.1 são visíveis as oscilações do viés para grupos de diferentes dimensões. No caso de amostras individuais ($n = 1$), o valor esperado de \hat{p} é igual a p , ou seja, o estimador é centrado e, por conseguinte, tem viés nulo. Para $n > 1$ o estimador tem um viés positivo, ou seja, \hat{p} sobre-estima o verdadeiro valor de p . Facilmente se constata que o viés é fortemente

dependente da dimensão do grupo. Note-se que quando $n = 5$ o viés é reduzido para $p \leq 0.25$ sendo negligível para $p < 0.05$. No caso de $n = 25$ o viés é negligível para $p < 0.06$ sendo colossal para valores de p mais elevados. Notemos ainda que para $p = 0.3$ são necessários grupos de dimensão reduzida ($n \leq 5$) para manter o viés reduzido.

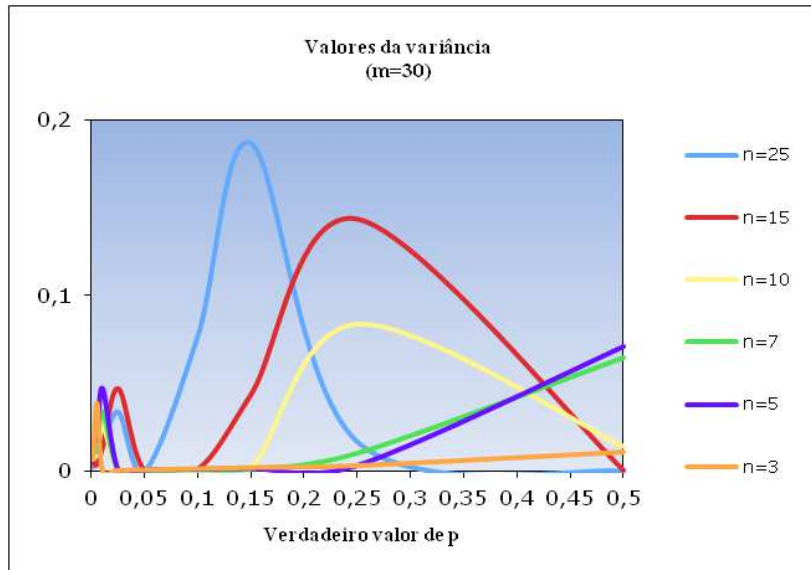


Figura 4.2: Valores da variância ($m = 30$)

Na Figura 4.2 podemos constatar que a dimensão do grupo não se baseia no critério de variância mínima. Note-se, por exemplo, que para $p = 0.3$ a variância é mínima quando $n = 25$, contudo (ver Figura 4.1) o seu viés é extremamente elevado (aproximadamente igual a 0.7).

A Figura 4.3 condensa a informação das duas figuras anteriores. O erro quadrático médio é igual à soma da variância com o quadrado do viés, e por conseguinte o erro quadrático médio será elevado não apenas quando o viés for elevado (estimador inadequado) mas também quando a variância for elevada (estimador pouco preciso).

Na Tabela 4.2 é visível que a utilização de testes conjuntos permite, para prevalências reduzidas, obter uma redução significativa no erro quadrático médio. Refira-se que comparativamente aos testes individuais, os testes conjuntos permitem a obtenção de um erro quadrático médio inferior para prevalências $p \in (0, 0.58)$. Os valores a negrito identificam os grupos de dimensão ótima (de entre as dimensões apresentadas na Tabela).

Quanto maior for o número de testes efetuados menores serão os valores do viés do estimador. Note-se que a probabilidade de todos os testes serem positivos, dada por $[1 - (1 - p)^n]^m$, é uma função decrescente com m o que provoca uma redução no viés do estimador. Os valores do viés para $m = 100$ e para $m = 1000$ figuram nas Tabelas 4.3 e 4.5 respetivamente.

O aumento do número de testes provoca igualmente uma redução no erro quadrático médio

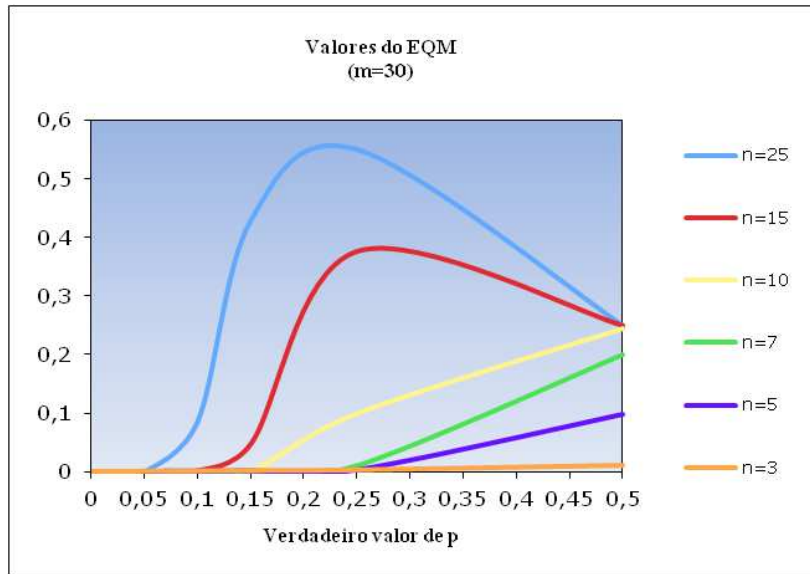


Figura 4.3: Valores do EQM ($m = 30$)

p	.50	.25	.15	.1	.05	.025	.01	.005	.001
$n = 1$	0	0	0	0	0	0	0	0	0
$n = 2$.0019	.0007	.0004	.0003	.0001	$6e^{-5}$	$3e^{-5}$	e^{-5}	$3e^{-6}$
$n = 3$.0041	.0012	.0006	.0004	.0002	$8e^{-5}$	$3e^{-5}$	$2e^{-5}$	$3e^{-6}$
$n = 4$.0083	.0015	.0007	.0004	.0002	.0001	$4e^{-5}$	$2e^{-5}$	$4e^{-6}$
$n = 5$.0273	.0020	.0009	.0005	.0002	.0001	$4e^{-5}$	$2e^{-5}$	$4e^{-6}$
$n = 6$.0955	.0025	.0010	.0006	.0002	.0001	$4e^{-5}$	$2e^{-5}$	$4e^{-6}$
$n = 7$.2068	.0032	.0011	.0006	.0003	.0001	$4e^{-5}$	$2e^{-5}$	$4e^{-6}$
$n = 8$.3146	.0040	.0013	.0007	.0003	.0001	$5e^{-5}$	$2e^{-5}$	$4e^{-6}$
$n = 9$.3927	.0054	.0014	.0007	.0003	.0001	$5e^{-5}$	$2e^{-5}$	$4e^{-6}$
$n = 10$.4411	.0084	.0016	.0008	.0003	.0001	$5e^{-5}$	$2e^{-5}$	$5e^{-6}$
$n = 15$.4978	.1885	.0031	.0011	.0003	.0001	$5e^{-5}$	$2e^{-5}$	$5e^{-6}$
$n = 20$.4999	.5326	.0206	.0017	.0004	.0002	$5e^{-5}$	$3e^{-5}$	$5e^{-6}$
$n = 25$.5000	.6896	.1483	.0030	.0005	.0002	$5e^{-5}$	$3e^{-5}$	$5e^{-6}$
$n = 50$.5000	.7499	.8234	.5307	.0016	.0003	$6e^{-5}$	$3e^{-5}$	$5e^{-6}$
$n = 100$.5000	.7500	.8500	.8975	.5215	.0009	$9e^{-5}$	$3e^{-5}$	$5e^{-6}$
$n = 1000$.5000	.7500	.8500	.9000	.9500	.9750	.9857	.5101	$9e^{-6}$

Tabela 4.3: Valores do viés ($m = 100$)

do estimador (ver Tabelas 4.4 e 4.6). Refira-se ainda que o aumento do número de testes não provoca alterações significativas na dimensão ótima dos grupos.

Estas tabelas fixam o número de testes (valor de m). A ideia é considerar custos fixos por teste uma vez que, em muitas aplicações, o custo associado à mistura dos sangues é usual-

p	.50	.25	.15	.1	.05	.025	.01	.005	.001
$n = 1$.0025	.0019	.0013	.0009	.0005	.0002	.0001	$5e^{-5}$	e^{-5}
$n = 2$.0019	.0011	.0007	.0005	.0002	.0001	$5e^{-5}$	$3e^{-5}$	$5e^{-6}$
$n = 3$.0021	.0009	.0005	.0003	.0002	$8e^{-5}$	$3e^{-5}$	$2e^{-5}$	$3e^{-6}$
$n = 4$.0031	.0007	.0004	.0003	.0002	$6e^{-5}$	$3e^{-5}$	e^{-5}	$3e^{-6}$
$n = 5$.0132	.0008	.0004	.0002	.0001	$5e^{-5}$	$2e^{-5}$	e^{-5}	$2e^{-6}$
$n = 6$.0534	.0008	.0003	.0002	$9e^{-5}$	$4e^{-5}$	$2e^{-5}$	$8e^{-6}$	$2e^{-6}$
$n = 7$.1155	.0008	.0003	.0002	$8e^{-5}$	$4e^{-5}$	e^{-5}	$7e^{-6}$	e^{-6}
$n = 8$.1788	.0009	.0003	.0002	$7e^{-5}$	$3e^{-5}$	e^{-5}	$6e^{-6}$	e^{-6}
$n = 9$.2076	.0013	.0003	.0002	$7e^{-5}$	$3e^{-5}$	e^{-5}	$6e^{-6}$	e^{-6}
$n = 10$.2284	.0029	.0003	.0002	$6e^{-5}$	$3e^{-5}$	e^{-5}	$5e^{-6}$	e^{-6}
$n = 15$.2494	.1470	.0005	.0001	$5e^{-5}$	$2e^{-5}$	$7e^{-6}$	$3e^{-6}$	$7e^{-7}$
$n = 20$.2500	.4101	.0144	.0002	$4e^{-5}$	$2e^{-5}$	$5e^{-6}$	$3e^{-6}$	$5e^{-7}$
$n = 25$.2500	.5222	.1277	.0007	$4e^{-5}$	e^{-5}	$5e^{-6}$	$2e^{-6}$	$4e^{-7}$
$n = 50$.2500	.5625	.7016	.4833	.0004	e^{-5}	$3e^{-6}$	e^{-6}	$2e^{-7}$
$n = 100$.2500	.5625	.7225	.8079	.4984	.0003	$2e^{-6}$	$7e^{-7}$	e^{-7}
$n = 1000$.2500	.5625	.7225	.8100	.9025	.9506	.9759	.5078	$2e^{-8}$

Tabela 4.4: Valores do EQM ($m = 100$)

p	.50	.25	.15	.1	.05	.025	.01	.005	.001
$n = 1$	0	0	0	0	0	0	0	0	0
$n = 2$.0002	$7e^{-5}$	$4e^{-5}$	$3e^{-5}$	e^{-5}	$6e^{-6}$	$3e^{-6}$	e^{-6}	$3e^{-7}$
$n = 3$.0004	.0001	$6e^{-5}$	$4e^{-5}$	$2e^{-5}$	$9e^{-6}$	$3e^{-6}$	$2e^{-6}$	$3e^{-7}$
$n = 4$.0007	.0002	$7e^{-5}$	$4e^{-5}$	$2e^{-5}$	e^{-5}	$4e^{-6}$	$2e^{-6}$	$4e^{-7}$
$n = 5$.0013	.0002	$9e^{-5}$	$5e^{-5}$	$2e^{-5}$	e^{-5}	$4e^{-6}$	$2e^{-6}$	$4e^{-7}$
$n = 6$.0023	.0002	.0001	$6e^{-5}$	$2e^{-5}$	e^{-5}	$4e^{-6}$	$2e^{-6}$	$4e^{-7}$
$n = 7$.0045	.0003	.0001	$6e^{-5}$	$2e^{-5}$	e^{-5}	$4e^{-6}$	$2e^{-6}$	$4e^{-7}$
$n = 8$.0156	.0004	.0001	$7e^{-5}$	$3e^{-5}$	e^{-5}	$5e^{-6}$	$2e^{-6}$	$4e^{-7}$
$n = 9$.0696	.0005	.0001	$7e^{-5}$	$3e^{-5}$	e^{-5}	$5e^{-6}$	$2e^{-6}$	$4e^{-7}$
$n = 10$.1763	.0006	.0002	$8e^{-5}$	$3e^{-5}$	e^{-5}	$5e^{-6}$	$2e^{-6}$	$5e^{-7}$
$n = 15$.4810	.0018	.0003	.0001	$3e^{-5}$	e^{-5}	$5e^{-6}$	$2e^{-6}$	$5e^{-7}$
$n = 20$.4993	.0346	.0005	.0002	$4e^{-5}$	$2e^{-5}$	$5e^{-6}$	$2e^{-6}$	$5e^{-7}$
$n = 25$.5000	.3444	.0010	.0002	$5e^{-5}$	$2e^{-5}$	$5e^{-6}$	$3e^{-6}$	$5e^{-7}$
$n = 50$.5000	.7495	.6265	.6926	.0001	$2e^{-5}$	$6e^{-6}$	$3e^{-6}$	$5e^{-7}$
$n = 100$.5000	.7500	.8499	.8755	.0034	$6e^{-5}$	$9e^{-6}$	$3e^{-6}$	$5e^{-7}$
$n = 1000$.5000	.7500	.8500	.9000	.9500	.9750	.9480	.0013	$9e^{-7}$

Tabela 4.5: Valores do viés ($m = 1000$)

Estimação da taxa de prevalência na ausência de erros de classificação

p	.50	.25	.15	.1	.05	.025	.01	.005	.001
$n = 1$.0003	.0002	.0001	$9e^{-5}$	$5e^{-5}$	$2e^{-5}$	e^{-5}	$5e^{-6}$	e^{-6}
$n = 2$.0002	.0001	$7e^{-5}$	$5e^{-5}$	$2e^{-5}$	e^{-5}	$5e^{-6}$	$2e^{-6}$	$5e^{-7}$
$n = 3$.0002	$9e^{-5}$	$5e^{-5}$	$3e^{-5}$	$2e^{-5}$	$8e^{-6}$	$3e^{-6}$	$2e^{-6}$	$3e^{-7}$
$n = 4$.0002	$8e^{-5}$	$4e^{-5}$	$3e^{-5}$	e^{-5}	$6e^{-6}$	$3e^{-6}$	e^{-6}	$3e^{-7}$
$n = 5$.0003	$7e^{-5}$	$4e^{-5}$	$2e^{-5}$	e^{-5}	$5e^{-6}$	$2e^{-6}$	e^{-6}	$2e^{-7}$
$n = 6$.0005	$7e^{-5}$	$3e^{-5}$	$2e^{-5}$	$9e^{-6}$	$4e^{-6}$	$2e^{-6}$	$8e^{-7}$	$2e^{-7}$
$n = 7$.0009	$8e^{-5}$	$3e^{-5}$	$2e^{-5}$	$8e^{-6}$	$4e^{-6}$	e^{-6}	$7e^{-7}$	e^{-7}
$n = 8$.0061	$8e^{-5}$	$3e^{-5}$	$2e^{-5}$	$7e^{-6}$	$3e^{-6}$	e^{-6}	$6e^{-7}$	e^{-7}
$n = 9$.0362	$9e^{-5}$	$3e^{-5}$	$2e^{-5}$	$7e^{-6}$	$3e^{-6}$	e^{-6}	$6e^{-7}$	e^{-7}
$n = 10$.0947	.0001	$3e^{-5}$	$2e^{-5}$	$6e^{-6}$	$3e^{-6}$	e^{-6}	$5e^{-7}$	e^{-7}
$n = 15$.2430	.0002	$3e^{-5}$	e^{-5}	$5e^{-6}$	$2e^{-6}$	$7e^{-7}$	$3e^{-7}$	$7e^{-8}$
$n = 20$.2498	.0239	$5e^{-5}$	e^{-5}	$4e^{-6}$	$2e^{-6}$	$5e^{-7}$	$3e^{-7}$	$5e^{-8}$
$n = 25$.2500	.2652	$7e^{-5}$	$2e^{-5}$	$4e^{-6}$	e^{-6}	$4e^{-7}$	$2e^{-7}$	$4e^{-8}$
$n = 50$.2500	.5622	.5376	.0047	$4e^{-6}$	e^{-6}	$3e^{-7}$	e^{-7}	$2e^{-8}$
$n = 100$.2500	.5625	.7224	.7888	.0024	e^{-6}	$2e^{-7}$	$6e^{-8}$	e^{-8}
$n = 1000$.2500	.5625	.7225	.8100	.9025	.9506	.9387	.0012	$2e^{-9}$

Tabela 4.6: Valores do EQM ($m = 1000$)

mente diminuto comparativamente ao custo de cada teste. Contudo o número de indivíduos utilizados em m testes com grupos de n indivíduos será $m \times n$ e, conseqüentemente, os testes conjuntos poderão ter mais informação.

4.3 Estimação intervalar

A estimação pontual permite aproximar o verdadeiro valor de um parâmetro, mas nada diz sobre a precisão ou fiabilidade da estimativa usada. Uma indicação sobre a precisão do processo de estimação é dada pelo erro-padrão da estimativa. Em alternativa à obtenção de um estimador $\widehat{\pi}_n$ vamos associar um intervalo de valores para π_n por forma a que com uma confiança $1 - \alpha$, o intervalo contenha o verdadeiro valor do parâmetro desconhecido. A amplitude do intervalo depende, entre outras coisas, da dimensão da amostra e do coeficiente de confiança $(1 - \alpha)$.

Genericamente, se \mathbf{X} for uma amostra aleatória proveniente de uma população com função de distribuição $F_{\pi_n}(x)$ e se $(l(\mathbf{X}), U(\mathbf{X}))$ forem duas estatísticas tais que $\mathbb{P}[l(\mathbf{X}) < \pi_n < U(\mathbf{X})] = 1 - \alpha$, dizemos que $(l(\mathbf{X}), U(\mathbf{X}))$ é um estimador intervalar de π_n e chamamos probabilidade de cobertura à probabilidade de o intervalo aleatório cobrir o verdadeiro valor do parâmetro, $\mathbb{P}[\pi_n \in (l(\mathbf{X}), U(\mathbf{X})) | \pi_n]$. O coeficiente de confiança de $(l(\mathbf{X}), U(\mathbf{X}))$ é o ínfimo das probabilidades de cobertura, $\inf_{\pi_n} \mathbb{P}[\pi_n \in (l(\mathbf{X}), U(\mathbf{X})) | \pi_n]$.

Dada uma amostra \mathbf{x} , a concretização $(l(\mathbf{x}), U(\mathbf{x}))$ é um intervalo de confiança (determinista) a $(1 - \alpha) \times 100\%$ para π_n . O valor $(1 - \alpha)$ é o grau ou coeficiente de confiança do intervalo.

Os adeptos da escola frequentista interpretam o intervalo de confiança numa perspetiva de experiências repetidas – procedendo à estimação de π_n a partir de inúmeras realizações da amostra aleatória, a proporção de casos em que o parâmetro se encontra de facto naquele intervalo estará a convergir para $1 - \alpha$.

Por definição, um intervalo de confiança a $(1 - \alpha) \times 100\%$ de confiança para um parâmetro π_n engloba todos os valores π_0 que conduzam à não rejeição da hipótese nula $H_0 : \pi_n = \pi_0$ de um teste bilateral com nível de confiança $1 - \alpha$. Na presença de um intervalo de confiança é desejável que a sua probabilidade de cobertura esteja próxima do coeficiente de confiança do mesmo. O facto de o modelo binomial ser do tipo discreto impede que o coeficiente de confiança seja sempre alcançado. Assim, o nosso objetivo consiste em construir um intervalo de confiança para π_n tal que a sua probabilidade de cobertura seja aproximadamente igual a $1 - \alpha$.

Denotando por X_n^+ o número de testes positivos quando m grupos de dimensão n são analisados, vimos que $X_n^+ \sim \text{binomial}(m, \pi_n)$ onde $\pi_n = 1 - (1 - p)^n$ representa a proporção de testes positivos. Com base em $\widehat{\pi}_n = \frac{X_n^+}{m}$ serão deduzidos vários intervalos de confiança para π_n cujos limites (inferior e superior) denotaremos por $\pi_n(l)$ e $\pi_n(u)$ respetivamente. Tendo em conta que $\hat{p} = 1 - (1 - \frac{X_n^+}{m})^{\frac{1}{n}}$, os limites de confiança para p serão obtidos através das expressões $p_l = 1 - (1 - \pi_n(l))^{\frac{1}{n}}$ e $p_u = 1 - (1 - \pi_n(u))^{\frac{1}{n}}$. Refira-se que este procedimento só é possível pelo facto de π_n ser uma função monótona crescente de p .

A convergência da distribuição de somas para uma distribuição Gaussiana foi descoberta por Abraham de Moivre, no caso especial de parcelas binomiais. O resultado geral foi estabelecido por Bienaymé, Laplace, Gauss, e a demonstração, com hipóteses cada vez

menos exigentes, retomada por Chebycheff, Liapunov, Markov, Lindeberg, Lévy e Feller. Em 1920 Pólya referiu os resultados sobre a convergência em distribuição de somas para um limite não degenerado como o teorema central da teoria das convergências estocásticas. O Teorema Limite Central, que seguidamente se enuncia, é desde então usado para referir, em sentido lato, a convergência de somas para uma variável aleatória não degenerada.

Teorema 4.1. (Teorema limite central)

Seja $\{X_k\}_{k=1}^{\infty}$ uma sucessão de variáveis i.i.d., $X_k \stackrel{d}{=} X$, $S_n = \sum_{k=1}^n X_k$. Então, se existir $\text{var}(X) = \sigma^2$ e denotarmos $\mathbb{E}(X) = \mu$,

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} Z \sim \text{Gaussiana}(0, 1).$$

A demonstração pode ser consultada, por exemplo, em Pestana e Velosa (2006), 2ª edição, página 1006.

4.3.1 Intervalo de confiança de Wald

Dado que o número de grupos positivos de dimensão n , é modelado por uma variável aleatória $X_n^+ \sim \text{binomial}(m, \pi_n)$ tem-se que $m\widehat{\pi}_n$ é uma soma de variáveis indicatrizes e, por conseguinte, para valores elevados de m ,

$$\frac{m\widehat{\pi}_n - m\pi_n}{\sqrt{m\pi_n(1 - \pi_n)}} \overset{o}{\sim} \text{Gaussiana}(0, 1),$$

e conseqüentemente

$$\frac{\widehat{\pi}_n - \pi_n}{\sqrt{\frac{\pi_n(1 - \pi_n)}{m}}} \overset{o}{\sim} \text{Gaussiana}(0, 1),$$

ou seja

$$\widehat{\pi}_n \overset{o}{\sim} \text{Gaussiana}\left(\pi_n, \sqrt{\frac{\pi_n(1 - \pi_n)}{m}}\right),$$

aproximação que é aceitável quando π_n não for um valor nem muito próximo de 0 nem muito próximo de 1. Deste modo o estimador intervalar de π_n será dado por

$$\left(\widehat{\pi}_n - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\pi_n(1 - \pi_n)}{m}}, \widehat{\pi}_n + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\pi_n(1 - \pi_n)}{m}}\right),$$

o que é pouco interessante porque na fórmula aparece o valor de π_n que estamos a estimar. Uma forma de contornar este obstáculo consiste em substituir π_n pela sua estimativa de máxima verosimilhança, $\widehat{\pi}_n = \frac{x_n^+}{m}$, e desprezar o erro cometido. Note-se que a substituição é

feita numa expressão que está dividida por um m elevado. Obtém-se assim, para o intervalo de confiança, a solução habitual

$$\left(\widehat{\pi}_n - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\widehat{\pi}_n(1-\widehat{\pi}_n)}{m}}, \widehat{\pi}_n + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\widehat{\pi}_n(1-\widehat{\pi}_n)}{m}} \right). \quad (4.7)$$

Este intervalo é conhecido por **intervalo de confiança de Wald** por poder ser obtido através da inversão do teste de Wald para π_n . Assim, se para testar as hipóteses $H_0 : \pi_n = \pi_0$ versus $H_1 : \pi_n \neq \pi_0$ utilizarmos a estatística de Wald

$$z = \frac{\widehat{\pi}_n - \pi_0}{\sqrt{\frac{\widehat{\pi}_n(1-\widehat{\pi}_n)}{m}}} \quad (4.8)$$

baseada na normalidade assintótica do estimador de máxima verosimilhança $\widehat{\pi}_n$, o intervalo de confiança será constituído pelo conjunto de valores π_0 que conduzam à não rejeição da hipótese nula $H_0 : \pi_n = \pi_0$ ao nível de significância α .

O facto de no erro padrão se proceder à substituição do parâmetro pela sua estimativa de máxima verosimilhança vai violar uma das premissas na construção do intervalo já que, mesmo para valores elevados de m , a estatística (4.8) não segue uma distribuição normal.

Este intervalo de confiança continua a ser usado apenas porque é muito simples, porque o seu desempenho é de facto muito mau. Na literatura da especialidade (ver, por exemplo, Agresti & Coull (1998) e Pires & Amado (2008)) reporta-se que um intervalo de confiança com coeficiente de confiança nominal de 95% chega a ter uma probabilidade de cobertura de π_n de apenas cerca de 80%.

O mau desempenho deste intervalo persiste perante amostras de dimensão muito elevada. As probabilidades de cobertura deste intervalo apresentam grandes oscilações em torno do nível de confiança nominal. Para além deste facto é visível a presença sistemática de um viés negativo. As oscilações estão relacionadas com o facto de o modelo binomial ser do tipo discreto ao passo que a existência do viés está relacionada com a má escolha para o centro do intervalo. De facto, apesar de $\widehat{\pi}_n$ ser o estimador centrado de π_n , deslocando o centro do intervalo para

$$\widetilde{\pi}_n = \frac{\widehat{\pi}_n + \frac{z_{1-\frac{\alpha}{2}}^2}{2m}}{1 + \frac{z_{1-\frac{\alpha}{2}}^2}{m}} \quad (4.9)$$

obtém-se um aumento significativo nas probabilidades de cobertura bem como a eliminação do viés sistemático (ver Brown, Cai & DasGupta (1999)).

A utilização de uma correção de continuidade ($\frac{1}{2m}$) se por um lado melhora a probabilidade de cobertura, por outro lado aumenta a probabilidade de o intervalo conter valores fora do intervalo (0,1). A inclusão da correção de continuidade dá origem ao seguinte estimador intervalar

$$\left(\widehat{\pi}_n - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\widehat{\pi}_n(1-\widehat{\pi}_n)}{m}} - \frac{1}{2m}, \widehat{\pi}_n + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\widehat{\pi}_n(1-\widehat{\pi}_n)}{m}} + \frac{1}{2m} \right).$$

4.3.2 Intervalo de confiança de score de Wilson

Este intervalo foi introduzido por Wilson (1927) e é deduzido através da inversão do teste score para π_n . Este teste baseia-se nas propriedades assintóticas da função score para comparar uma proporção observada com uma proporção hipotética. Assim, para testar as hipóteses $H_0 : \pi_n = \pi_0$ versus $H_1 : \pi_n \neq \pi_0$ utilizamos a estatística de score

$$z = \frac{\widehat{\pi}_n - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{m}}},$$

e o intervalo de confiança será constituído pelo conjunto de valores π_0 que conduzam à não rejeição da hipote nula $H_0 : \pi_n = \pi_0$ ao nível de significância α .

Pretende-se assim determinar os valores de π_0 que satisfaçam a desigualdade

$$\frac{|\widehat{\pi}_n - \pi_0|}{\sqrt{\frac{\pi_0(1-\pi_0)}{m}}} < z_{1-\frac{\alpha}{2}}, \quad (4.10)$$

o que é equivalente a

$$H(\pi_0) = (\widehat{\pi}_n - \pi_0)^2 - \frac{z_{1-\frac{\alpha}{2}}^2 \pi_0 (1 - \pi_0)}{m} \leq 0.$$

A expressão anterior pode ser escrita na forma

$$H(\pi_0) = \left(1 + \frac{z_{1-\frac{\alpha}{2}}^2}{m}\right) (\pi_0)^2 - \left(2\widehat{\pi}_n + \frac{z_{1-\frac{\alpha}{2}}^2}{m}\right) \pi_0 + (\widehat{\pi}_n)^2.$$

Atendendo a que $H(\pi_0)$ é uma expressão quadrática em π_0 a determinação dos seus zeros conduz-nos aos limites do intervalo de confiança

$$\pi_n(l) = \frac{\widehat{\pi}_n + \frac{z_{1-\frac{\alpha}{2}}^2}{2m} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\widehat{\pi}_n(1-\widehat{\pi}_n)}{m} + \frac{z_{1-\frac{\alpha}{2}}^2}{4m^2}}}{1 + \frac{z_{1-\frac{\alpha}{2}}^2}{m}}$$

$$\pi_n(u) = \frac{\widehat{\pi}_n + \frac{z_{1-\frac{\alpha}{2}}^2}{2m} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\widehat{\pi}_n(1-\widehat{\pi}_n)}{m} + \frac{z_{1-\frac{\alpha}{2}}^2}{4m^2}}}{1 + \frac{z_{1-\frac{\alpha}{2}}^2}{m}}.$$

Optando-se pela inclusão da correção de continuidade terá que se determinar os valores de π_0 que satisfaçam a desigualdade

$$\frac{|\widehat{\pi}_n - \pi_0| - \frac{1}{2m}}{\sqrt{\frac{\pi_0(1-\pi_0)}{m}}} < z_{1-\frac{\alpha}{2}},$$

obtendo de forma análoga os seguintes limites

$$\pi_n(l) = \frac{2m\widehat{\pi}_n + z_{1-\frac{\alpha}{2}}^2 - 1 - z_{1-\frac{\alpha}{2}} \sqrt{z_{1-\frac{\alpha}{2}}^2 - \left(2 + \frac{1}{m}\right) + 4\widehat{\pi}_n(m - m\widehat{\pi}_n + 1)}}{2\left(m + z_{1-\frac{\alpha}{2}}^2\right)}$$

$$\pi_n(l) = \frac{2m\widehat{\pi}_n + z_{1-\frac{\alpha}{2}}^2 - 1 + z_{1-\frac{\alpha}{2}} \sqrt{z_{1-\frac{\alpha}{2}}^2 - \left(2 + \frac{1}{m}\right) + 4\widehat{\pi}_n(m - m\widehat{\pi}_n + 1)}}{2\left(m + z_{1-\frac{\alpha}{2}}^2\right)}$$

Este intervalo tem um desempenho muito superior ao intervalo de confiança de Wald já que tem probabilidades de cobertura muito próximas do coeficiente de confiança nominal. No entanto, como tem uma expressão complicada, é preferível usar o intervalo de confiança de Agresti-Coull.

4.3.3 Intervalo de confiança de Agresti-Coull

À primeira vista, a fórmula do intervalo de confiança de score de Wilson parece ser de difícil interpretação comparativamente à expressão (4.7) da página 52. Agresti & Coull (1998) propuseram um ajustamento ao intervalo de confiança de Wald por forma a melhorar o seu desempenho. Começemos por notar que o centro do intervalo de confiança de score de Wilson pode ser escrito na forma

$$\widetilde{\pi}_n = \frac{\widehat{\pi}_n + \frac{z_{1-\frac{\alpha}{2}}^2}{2m}}{1 + \frac{z_{1-\frac{\alpha}{2}}^2}{m}} = \widehat{\pi}_n \left(\frac{m}{m + z_{1-\frac{\alpha}{2}}^2} \right) + \frac{1}{2} \left(\frac{z_{1-\frac{\alpha}{2}}^2}{m + z_{1-\frac{\alpha}{2}}^2} \right). \quad (4.11)$$

Note-se que $\widetilde{\pi}_n$ é uma combinação convexa de $\widehat{\pi}_n$ e $\frac{1}{2}$. Substituindo em (4.11) $\widehat{\pi}_n$ por $\frac{x_n^+}{m}$ obtém-se

$$\widetilde{\pi}_n = \left(\frac{x_n^+}{m + z_{1-\frac{\alpha}{2}}^2} \right) + \frac{1}{2} \left(\frac{z_{1-\frac{\alpha}{2}}^2}{m + z_{1-\frac{\alpha}{2}}^2} \right),$$

e, por conseguinte, verifica-se que o ponto médio do intervalo coincide com a proporção amostral após se terem adicionado $\frac{z_{1-\frac{\alpha}{2}}^2}{2}$ pseudo observações limpas e $\frac{z_{1-\frac{\alpha}{2}}^2}{2}$ observações contaminadas.

O erro padrão do referido intervalo pode ser reescrito na forma

$$\sqrt{\frac{1}{m + z_{1-\frac{\alpha}{2}}^2} \left[\widehat{\pi}_n (1 - \widehat{\pi}_n) \left(\frac{m}{m + z_{1-\frac{\alpha}{2}}^2} \right) + \left(\frac{1}{2} \right) \left(\frac{1}{2} \right) \left(\frac{z_{1-\frac{\alpha}{2}}^2}{m + z_{1-\frac{\alpha}{2}}^2} \right) \right]}.$$

O radicando da expressão anterior é a média ponderada da variância de uma proporção amostral quando $\pi_n = \widehat{\pi}_n$ com a variância de uma proporção amostral quando $\pi_n = \frac{1}{2}$, utilizando $m + z_{1-\frac{\alpha}{2}}^2$ em substituição da dimensão amostral usual m .

A variância de $\widetilde{\pi}_n$ pode ser estimada por

$$\text{var}(\widetilde{\pi}_n) = \frac{m\pi_n(1-\pi_n)}{\left(m + z_{1-\frac{\alpha}{2}}^2\right)^2} \approx \frac{\pi_n(1-\pi_n)}{m + z_{1-\frac{\alpha}{2}}^2},$$

pelo que se usa como estimativa do erro padrão de $\widetilde{\pi}_n$ a expressão $\sqrt{\frac{\widetilde{\pi}_n(1-\widetilde{\pi}_n)}{m + z_{1-\frac{\alpha}{2}}^2}}$.

Assim, fazendo $\widetilde{\pi}_n = \frac{\widehat{\pi}_n + \frac{z_{1-\frac{\alpha}{2}}^2}{2m}}{1 + \frac{z_{1-\frac{\alpha}{2}}^2}{m}}$ e $\widetilde{m} = m + \frac{z_{1-\frac{\alpha}{2}}^2}{2}$ obtém-se finalmente

$$\widetilde{\pi}_n \pm z_{1-\frac{\alpha}{2}}^2 \sqrt{\frac{\widetilde{\pi}_n(1-\widetilde{\pi}_n)}{\widetilde{m}}}. \quad (4.12)$$

Como vimos anteriormente, o intervalo de score de Wilson e o intervalo de Agresti-Coull são centrados no mesmo ponto. Contudo, é fácil de constatar que o intervalo de Agresti-Coull tem uma amplitude superior. Dado que na pratica é usual usar-se intervalos com coeficiente de confiança 95% e como $z_{0,975}^2 = 1.96 \approx 2$; no caso de o valor observado de X_n^+ ser x_n^+ , o ponto médio do intervalo é aproximado por

$$\widetilde{\pi}_n = \frac{x_n^+}{m+4} + \frac{2}{m+4} = \frac{x_n^+ + 2}{m+4} \quad (4.13)$$

e por conseguinte o intervalo será da forma

$$\widetilde{\pi}_n \pm z_{1-\frac{\alpha}{2}}^2 \sqrt{\frac{\widetilde{\pi}_n(1-\widetilde{\pi}_n)}{\widetilde{m}}} \quad (4.14)$$

onde $\widetilde{m} = m + 4$ e $\widetilde{\pi}_n = \frac{x_n^+}{m+4}$.

4.3.4 Intervalo de confiança assintótico

Os intervalos de confiança baseados na normalidade assintótica são pouco precisos para prevalências reduzidas e amostras de dimensão moderada em virtude da distribuição de \widehat{p} ser muito assimétrica. Refira-se contudo que a variância do estimador de p é uma função decrescente com n (dimensão do grupo) e, por conseguinte, quanto maior for a dimensão do grupo mais preciso será o estimador de p .

Designando por \widehat{p} o estimador de máxima verosimilhança de p , já foi referido que $\widehat{p} = 1 - (1 - \widehat{\pi}_n)^{\frac{1}{n}}$ onde $\widehat{\pi}_n = \frac{X_n^+}{m}$ representa a proporção de testes positivos quando m grupos de dimensão n são analisados. Recorrendo às propriedades assintóticas dos estimadores de máxima verosimilhança, para amostras de dimensão elevada tem-se que

- $\mathbb{E}(\widehat{p}) \approx p$

- $\hat{p} \overset{o}{\sim} \text{Gaussiana} \left(p, \sqrt{\text{var}(\hat{p})} \right)$

e, por conseguinte,

$$z = \frac{\hat{p} - p}{\sqrt{\text{var}(\hat{p})}} \overset{o}{\sim} \text{Gaussiana}(0, 1).$$

Denotando por $z_{1-\frac{\alpha}{2}}$ o percentil $1 - \frac{\alpha}{2}$ da distribuição normal padrão, o intervalo de $(1 - \alpha) \times 100\%$ de confiança para a prevalência é dado por

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} [\text{var}(\hat{p})]^{\frac{1}{2}}$$

Dado que $\text{var}(\hat{\pi}_n) = \frac{p(1-p)}{m}$ e $\hat{p} = g(\hat{\pi}_n) = 1 - (1 - \hat{\pi}_n)^{1/n}$ vamos recorrer ao método Delta, que seguidamente se enuncia, para determinarmos a distribuição assintótica de \hat{p} .

Teorema 4.2. (Método delta)

Seja $\{Y_n\}_{n \geq 1}$ uma sucessão de variáveis aleatórias e suponha-se que

$$\sqrt{n}(Y_n - \theta) \xrightarrow[n \rightarrow \infty]{d} Z \sim \text{Gaussiana}(0, \sigma).$$

e, fixado θ , que para uma dada função g existe $g'(\theta) \neq 0$. Então também

$$\sqrt{n}[g(Y_n) - g(\theta)] \xrightarrow[n \rightarrow \infty]{d} Z \sim \text{Gaussiana}\left(0, \sigma \left|g'(\theta)\right|\right)$$

Demonstração. A expansão de $g(Y_n)$ em série de Taylor em torno de $Y_n = \theta$, de primeira ordem, é

$$g(Y_n) = g(\theta) + g'(\theta)(Y_n - \theta) + R_1,$$

onde $R_1 \rightarrow 0$ quando $Y_n \rightarrow \theta$.

Como $Y_n \xrightarrow[n \rightarrow \infty]{\mathbb{P}} \theta$, segue-se que também $R_1 \xrightarrow[n \rightarrow \infty]{\mathbb{P}} 0$ e, por conseguinte, recorrendo ao teorema de Slutsky tem-se

$$\sqrt{n}[g(Y_n) - g(\theta)] \approx g'(\theta) \sqrt{n}(Y_n - \theta).$$

□

Teorema 4.3. (Distribuição assintótica de \hat{p})

A distribuição assintótica de \hat{p} é dada por

$$\hat{p} \overset{o}{\sim} \text{Gaussiana}(p, \sqrt{\text{var}(\hat{p})}),$$

onde

$$\text{var}(\hat{p}) = \frac{1 - (1 - p)^n}{mn^2(1 - p)^{n-2}}.$$

Demonstração. :

Pelas propriedades assintóticas dos estimadores de máxima verosimilhança, sabemos que \hat{p} tem distribuição assintótica normal com valor médio p . Para determinar uma expressão para a sua variância considere-se:

$$\begin{aligned} g(\pi_n) &= 1 - (1 - \pi_n)^{\frac{1}{n}} \Leftrightarrow \\ \Leftrightarrow g'(\pi_n) &= \frac{1}{n}(1 - \pi_n)^{\frac{1}{n}-1} \Leftrightarrow \\ \Leftrightarrow [g'(\pi_n)]^2 &= \frac{1}{n^2}(1 - \pi_n)^{\frac{2}{n}-2} \Leftrightarrow \\ \Leftrightarrow \sigma^2 [g'(\pi_n)]^2 &= \frac{\pi_n(1 - \pi_n)}{m} \frac{1}{n^2} (1 - \pi_n)^{\frac{2}{n}-2} \\ &= \frac{\pi_n(1 - \pi_n)(1 - \pi_n)^{\frac{2}{n}-2}}{mn^2}. \end{aligned}$$

Dado que $\pi_n = 1 - (1 - p)^n$ obtém-se:

$$\sigma^2 [g'(\pi_n)]^2 = \frac{1 - (1 - p)^n}{mn^2 (1 - p)^{n-2}},$$

e, por conseguinte,

$$var(\hat{p}) = \frac{1 - (1 - p)^n}{mn^2 (1 - p)^{n-2}}.$$

Assim, a distribuição assintótica de \hat{p} é

$$\hat{p} \sim \text{Gaussiana} \left(p, \sqrt{\frac{1 - (1 - p)^n}{mn^2 (1 - p)^{n-2}}} \right). \quad (4.15)$$

□

4.3.5 Intervalo com correções de segunda ordem

Já foi referido anteriormente que as excelentes probabilidades de coberturas do intervalo de confiança de score de Wilson contribuem para que este seja recomendado como uma das alternativas ao intervalo de Wald. Brown *et al.* (2003) ampliaram os resultados obtidos no modelo binomial à família exponencial natural cuja variância é função quando muito quadrática do valor médio (FEN-FVQ) e concluíram que os problemas e soluções existentes no modelo binomial são comuns a todos os membros da referida família.

Mais concretamente, o mau desempenho ao nível da cobertura do intervalo de Wald e as excelentes probabilidades de cobertura do intervalo de score de Wilson são extensíveis à FEN-FVQ que inclui, para além da binomial, as secantes hiperbólicas, gaussianas, gamas, Poissons e binomiais negativas. Cai (2005) faz um tratamento unificado dos intervalos de confiança unilaterais para a família exponencial discreta com função de variância quadrática (como função do valor médio) que engloba as distribuições binomial, Poisson e binomial negativa. Apesar

da existência de características comuns, os problemas associados à estimação intervalar unilateral diferem significativamente dos problemas associados aos intervalos bilaterais. De facto, o bom desempenho do intervalo de score de Wilson na estimação bilateral não é transportado para o caso unilateral para qualquer dos três modelos discretos referidos, não apenas a nível de probabilidades de cobertura mas também no que respeita à amplitude média dos intervalos. Cai (2005) mostra que as probabilidades de cobertura dos intervalos unilaterais de Wald e de score de Wilson padecem de um pronunciado viés sistemático embora de gravidade e direção distintas.

Este mau desempenho dos intervalos unilaterais de Wald e de score de Wilson conduz à pesquisa de intervalos unilaterais alternativos que apresentem melhores probabilidades de cobertura.

Cai (2005) compara os termos não oscilantes das probabilidades de cobertura de diversos intervalos recorrendo às expansões de Edgeworth ⁽¹⁾ e deriva o chamado second-order corrected interval (SOC) que apresenta um melhor desempenho já que o viés sistemático das probabilidades de cobertura é quase inexistente para os três modelos discretos.

Deste modo, os intervalos de confiança unilaterais serão

$$\left(\tilde{p} - z_{1-\alpha} \frac{1}{\sqrt{m}} \sqrt{\hat{p}(1-\hat{p}) + \frac{\gamma_1 \hat{p}(1-\hat{p}) + \gamma_2}{m}}, 1 \right)$$

e

$$\left(0, \tilde{p} + z_{1-\alpha} \frac{1}{\sqrt{m}} \sqrt{\hat{p}(1-\hat{p}) + \frac{\gamma_1 \hat{p}(1-\hat{p}) + \gamma_2}{m}} \right)$$

onde: $\tilde{p} = \frac{x_n^+ - \tau}{m + 2\tau}$, $\tau = \frac{1}{3}z_{1-\alpha}^2 + \frac{1}{6}$, $\gamma_1 = \frac{-13}{18}z_{1-\alpha}^2 - \frac{17}{18}$, $\gamma_2 = \frac{1}{18}z_{1-\alpha}^2 + \frac{7}{36}$ e $z_{1-\alpha}$ é o quantil $1 - \alpha$ da gaussiana padrão. Estes limites podem ser aplicados a intervalos bilaterais para um número de grupos relativamente pequeno substituindo $z_{1-\alpha}^2$ por $z_{1-\frac{\alpha}{2}}^2$ obtendo-se

$$\left(\tilde{p} - z_{1-\frac{\alpha}{2}} \frac{1}{\sqrt{m}} \sqrt{\hat{p}(1-\hat{p}) + \frac{\gamma_1 \hat{p}(1-\hat{p}) + \gamma_2}{m}}, \tilde{p} + z_{1-\frac{\alpha}{2}} \frac{1}{\sqrt{m}} \sqrt{\hat{p}(1-\hat{p}) + \frac{\gamma_1 \hat{p}(1-\hat{p}) + \gamma_2}{m}} \right)$$

Caso o número testes seja reduzido, o limite superior do intervalo pode ser superior a 1 e a estimativa $\hat{p} = 1$ pode ser excluída no caso de $X_n^+ = m$ para m elevado. Por forma a que o limite superior fique restrito ao intervalo $[0, 1]$ o seu valor deverá ser igual a 1 quando $X_n^+ = m$.

⁽¹⁾As expansões de Edgeworth surgiram no final do século XIX, consistindo numa expansão em termos da função de distribuição gaussiana e suas derivadas. Consulte-se por exemplo Barndorff-Nielsen & Cox(1994) onde as mesmas são apresentadas numa perspetiva atual.

4.3.6 Intervalo de confiança de Clopper-Pearson

Para evitar recorrer a métodos aproximados alguma bibliografia recomenda o intervalo de confiança “exato” de Clopper-Pearson para π_n (Clopper & Pearson, 1934). Este intervalo é baseado na inversão do teste bilateral para o modelo binomial onde a hipótese nula é $H_0 : \pi_n = \pi_0$ versus a hipótese alternativa $H_1 : \pi_n \neq \pi_0$. Este intervalo é retratado como um intervalo *exato* apenas por recorrer à distribuição exata (binomial) pois pelo facto de se estar a proceder a contagens leva a que, de todos os números reais, apenas possamos observar os naturais. Se forem observados x sucessos, ou seja, se ocorrer o acontecimento $X_n^+ = x$ então para $x = 1, 2, \dots, m - 1$, o intervalo de Clopper-Pearson é definido por $(\pi_n(l), \pi_n(u))$, onde $\pi_n(l)$ e $\pi_n(u)$ são soluções do seguinte conjunto de equações

$$\begin{cases} \sum_{i=x}^m \binom{m}{i} [\pi_n(l)]^i [1 - \pi_n(l)]^{m-i} = \frac{\alpha}{2} \\ \sum_{i=0}^x \binom{m}{i} [\pi_n(u)]^i [1 - \pi_n(u)]^{m-i} = \frac{\alpha}{2} \end{cases} \quad (4.16)$$

sendo que quando $x = 0$ o limite inferior será 0 e quando $x = m$ o limite superior será igual a 1.

Recorrendo à conhecida relação existente entre a cauda direita de uma variável com distribuição binomial e a cauda esquerda de uma variável com distribuição beta, podemos obter os valores de $\pi_n(l)$ e de $\pi_n(u)$ que satisfazem as condições referidas através dos percentis $\frac{\alpha}{2}$ e $1 - \frac{\alpha}{2}$ dos modelos $Beta(x, m - x + 1)$ e $Beta(x + 1, m - x)$ respetivamente.

Assim, podemos escrever as igualdades (4.16) da seguinte forma

$$\begin{cases} \frac{1}{Be(x, m-x+1)} \int_0^{\pi_n(l)} t^{x-1} (1-t)^{m-x} dt = \frac{\alpha}{2} \\ \frac{1}{Be(x+1, m-x)} \int_0^{\pi_n(u)} t^x (1-t)^{m-x-1} dt = 1 - \frac{\alpha}{2} \end{cases} \quad (4.17)$$

onde

$$Be(a, b) = \int_0^1 t^{a-1} (1-t)^{b-1} dt$$

representa a função beta com parâmetros a e b .

A distribuição beta está também relacionada com a distribuição F de Fisher-Snedecor e, por conseguinte, os limites (inferior e superior) do intervalo de Clopper-Pearson podem ser escritos na forma

$$\pi_n(l) = \left(1 + \frac{m-x+1}{xF\left(1 - \frac{\alpha}{2}; 2x, 2(m-x+1)\right)} \right)^{-1}$$

e

$$\pi_n(u) = \left(1 + \frac{m-x}{(x+1)F\left(\frac{\alpha}{2}; 2(x+1), 2(m-x)\right)} \right)^{-1},$$

onde $F(\alpha, m, n)$ representa o quantil α da distribuição F de Snedecor com m e n graus de liberdade.

4.3.7 Algoritmo de Blaker

Dos muitos intervalos de confiança existentes para o parâmetro π_n da binomial, o clássico intervalo de confiança de Wald e o intervalo de Clopper-Pearson são os mais utilizados em aplicações. O primeiro, publicado há 200 anos atrás por Laplace (1812) baseia-se na aproximação da distribuição binomial à distribuição normal e tem uma probabilidade de cobertura que oscila em torno do nível de confiança nominal $1 - \alpha$. O segundo, publicado pela primeira vez em 1934 por Clopper-Pearson, é derivado diretamente da distribuição binomial e é um intervalo de confiança conservador pelo facto de ter probabilidade de cobertura sempre igual ou superior ao nível de confiança nominal. Os limites inferior e superior do intervalo de Clopper-Pearson para a obtenção de x sucessos em m provas são dados por

$$\pi_n(l) = \inf \{ \pi_n \in [0, 1] : \mathbb{P}_{\pi_n}(X_n^+ \geq x) \geq \alpha/2 \}$$

e

$$\pi_n(u) = \sup \{ \pi_n \in [0, 1] : \mathbb{P}_{\pi_n}(X_n^+ \leq x) \geq \alpha/2 \},$$

o que garante que $\forall \pi_n \in [0, 1]$

$$\mathbb{P}_{\pi_n}(\pi_n < \pi_n(l)) \leq \alpha/2 \text{ e } \mathbb{P}_{\pi_n}(\pi_n > \pi_n(u)) \leq \alpha/2 \quad (4.18)$$

e consequentemente

$$\mathbb{P}_{\pi_n}(\pi_n < \pi_n(l)) + \mathbb{P}_{\pi_n}(\pi_n > \pi_n(u)) \leq \alpha. \quad (4.19)$$

O intervalo de Clopper-Pearson tem sido frequentemente criticado por ser excessivamente conservador. Desde 1950 várias alternativas menos conservadoras têm sido sugeridas – Sterne (1954), Crow (1956), Blyth and Still (1983), Casela (1986) e Blaker (2000).

Todas elas mantêm válida a igualdade (4.19) no intervalo $[0, 1]$ mas violam (4.18) permitindo que $\mathbb{P}_{\pi_n}(\pi_n < \pi_n(l))$ ou $\mathbb{P}_{\pi_n}(\pi_n > \pi_n(u))$ exceda $\alpha/2$. Desta forma a probabilidade de cobertura aproxima-se do nível de confiança nominal sendo o seu valor mínimo igual a $1 - \alpha$.

A proposta de Blaker (2000) foi alvo de muita atenção devido às seguintes virtudes:

1. O intervalo de confiança de Blaker é sempre um subconjunto, não necessariamente próprio, do intervalo de Clopper-Pearson e, por conseguinte, permite uma precisão mais elevada (ao mesmo nível de confiança e baseado nos mesmos dados),
2. Quando $\alpha_1 > \alpha_2$, o intervalo com confiança $(1 - \alpha_1)$ é sempre um subconjunto (não necessariamente próprio) do intervalo com confiança $(1 - \alpha_2)$,

3. Fácil de calcular através da utilização de um pequeno programa em R incluído em Blaker (2000).

Para que seja possível efetuar uma analogia entre os intervalos de Blaker e Clopper-Pearson comecemos por notar que este pode ser definido alternativamente como $\{\pi_n : \beta_{m,x}^{CP}(\pi_n) > \alpha\}$ onde

$$\beta_{m,x}^{CP}(\pi_n) = \min \{2\mathbb{P}(X_n^+ \geq x); 2\mathbb{P}(X_n^+ \leq x); 1\}.$$

Os intervalos de confiança de Blaker são baseados na chamada função de aceitabilidade $\beta_{m,x}$ definida por

$$\min [\mathbb{P}_{\pi_n}(X_n^+ \geq x) + \mathbb{P}_{\pi_n}(X_n^+ \leq x^*), \mathbb{P}_{\pi_n}(X_n^+ \leq x) + \mathbb{P}_{\pi_n}(X_n^+ \geq x^{**}), 1]$$

onde

$$x^* = \max \{y : \mathbb{P}_{\pi_n}(X_n^+ \leq y) \leq \mathbb{P}_{\pi_n}(X_n^+ \geq x)\}$$

e

$$x^{**} = \min \{y : \mathbb{P}_{\pi_n}(X_n^+ \geq y) \leq \mathbb{P}_{\pi_n}(X_n^+ \leq x)\}.$$

Blaker (2000) mostrou que $C_\alpha(x) = \{\pi_n : \beta_{m,x}(\pi_n) > \alpha\}$ é um conjunto de confiança $(1 - \alpha)$ exato, isto é, $\mathbb{P}_{\pi_n}(\pi_n \in C_\alpha(X)) \geq \alpha, \forall \pi_n \in [0, 1]$. A situação é contudo mais complicada que no caso de Clopper-Pearson. Enquanto $\{\pi_n : \beta_{m,x}^{CP}(\pi_n) > \alpha\}$ é sempre um intervalo, $C_\alpha(x)$ pode não ser um intervalo mas apenas uma união de intervalos disjuntos. Isto deve-se ao facto de β , ao contrário de β^{CP} , não ser necessariamente uma função unimodal.

O algoritmo de Blaker é constituído por duas funções, nomeadamente *acceptbin* e *acceptinterval*. A primeira calcula a função de aceitabilidade num ponto π_n e a segunda efetua uma pesquisa numérica do ponto mais à esquerda e do ponto mais à direita onde a função de aceitabilidade é superior a α . A pesquisa numérica do limite inferior $\pi_n^{(l)}$ pode ser descrita do seguinte modo (a pesquisa do limite superior é análoga):

1. Entrada de dados: número de testes m , número de infetados x , nível de confiança $1 - \alpha$, tolerância Δ (real positivo que por defeito é igual a 10^{-10}). Escrever $\pi_n^{(l)}$.
2. Se $x = 0$, $\pi_n^{(l)} := 0$, escrever $\pi_n^{(l)}$ e terminar, caso contrário continuar.
3. $\pi_n := \pi_n^{(l)CP}$.
4. Enquanto $\beta(\pi_n + \Delta) < \alpha$, repetir $\pi_n = \pi_n + \Delta$.
5. $\pi_n^{(l)} := \pi_n$, escrever $\pi_n^{(l)}$ e terminar.

4.3.8 Comparação de resultados via simulação

Para avaliar as vantagens decorrentes da utilização de amostras conjuntas na estimação da taxa de prevalência, recorreu-se ao package *binGroup* para o R (Bilder *et al.* 2010) que engloba a generalidade dos estimadores intervalares anteriormente referidos. Foram efetuadas 100000 réplicas e o nível de confiança utilizado foi de 0.95. Para cada valor de p , foram registados, para grupos de dimensão ótima e para amostras individuais, os valores dos seguintes parâmetros:

$$\begin{aligned} ML &\rightarrow \text{Comprimento médio do intervalo} \\ \widehat{\alpha}_L &\rightarrow \text{taxa de erro inferior} \\ \widehat{\alpha}_U &\rightarrow \text{taxa de erro superior} \end{aligned}$$

Os estimadores intervalares utilizados foram:

$$\begin{aligned} CP &\rightarrow \text{Clopper-Pearson} \\ B &\rightarrow \text{Blaker} \\ AC &\rightarrow \text{Agresti-Coull} \\ S &\rightarrow \text{score de Wilson} \\ SOC &\rightarrow \text{Second order corrected - I.C. assintótico para problemas unilaterais} \\ W &\rightarrow \text{Wald} \end{aligned}$$

Ao analisar os resultados que figuram na Tabela 4.7 verificamos que, para todo e qualquer valor de p , o comprimento médio do intervalo de confiança (ML) para grupos de dimensão ótima é sempre inferior ao valor obtido para amostras individuais. Esta conclusão é mais do que espetável pois o número de observações utilizadas na construção do intervalo recorrendo a amostras conjuntas, $m \times n$, é sempre superior ao número de observações recorrendo amostras individuais, m , e como é sabido a amplitude dum intervalo de confiança varia inversamente com a dimensão da amostra (quanto mais informação existir menor será a amplitude do intervalo e conseqüentemente maior será a sua precisão). Por outro lado, a referida redução do comprimento médio do intervalo é mais acentuada para taxas de prevalência mais pequenas já que a dimensão ótima do grupo aumenta com a redução da taxa de prevalência.

No que respeita à taxa de erro inferior, $\widehat{\alpha}_L$, e à taxa de erro superior, $\widehat{\alpha}_U$, constatamos que na generalidade dos casos, recorrendo a amostras conjuntas, a taxa de erro, $\widehat{\alpha}_L + \widehat{\alpha}_U$, se aproxima do nível de significância utilizado ($\alpha = 0.05$) e, conseqüentemente, as probabilidades de cobertura dos intervalos $[1 - (\widehat{\alpha}_L + \widehat{\alpha}_U)]$ estão mais próximas do nível de confiança nominal, $1 - \alpha = 0.95$. Nas Tabelas 4.8 e 4.9 figuram os resultados obtidos para $m = 100$ e $m = 1000$ que veem reforçar tudo o que foi previamente referido.

Assim, para taxas de prevalências reduzidas, o recurso a testes conjuntos permite não apenas minimizar os custos inerentes à classificação dos elementos da população (assunto tratado no capítulo 1) mas também a redução do erro quadrático médio do estimador, obtendo-se assim estimativas mais precisas.

Relativamente à comparação dos resultados obtidos para os diversos intervalos podemos constatar que para $m = 30$ (ver Tabela 4.7) as probabilidades de cobertura do intervalo de Wald oscilam em torno do nível de confiança nominal. Para $p = 0.5$ observa-se uma taxa de erro superior a $0.05 \widehat{\alpha}_L + \widehat{\alpha}_U = 0.0615$ o que significa que a probabilidade de cobertura é inferior ao coeficiente de confiança nominal e, por conseguinte, o intervalo é liberal.

Ao invés, para $p = 0.1$ obtém-se ($\widehat{\alpha}_L + \widehat{\alpha}_U = 0.0371$) o que significa que a probabilidade de cobertura é superior ao nível de confiança nominal e conseqüentemente o intervalo de confiança é conservador. As frequentes oscilações de grande magnitude da taxa de cobertura deste intervalo juntamente com o facto de ser o intervalo aproximado com maior comprimento médio, fazem com que este intervalo tenha por sistema um mau desempenho.

Os intervalos de score de Wilson e de Agresti-Coull apresentam probabilidades de cobertura muito superiores ao intervalo de Wald. Como se pode constatar (ver Tabela 4.7), apenas para $p = 0.05$ e $p = 0.1$ se obtém uma taxa de cobertura inferior a 0.95. Refira-se que apesar de as probabilidades de cobertura serem iguais o intervalo de Agresti-Coull tem menor comprimento médio.

No que respeita aos intervalos exatos (Clopper-Pearson e Blaker), que são os únicos que apresentam probabilidades de cobertura iguais ou superiores ao nível de confiança nominal para todos os valores de p , constata-se que o intervalo de Blaker tem uma amplitude média menor. As diferenças ao nível das probabilidades de coberturas não são significativas

p	.50		.25		.10		.05		.01		.001		
	1	2	1	4	1	10	1	20	1	100	1	100	
<i>CP</i>	<i>ML</i>	36.85	33.14	32.34	21.74	23.58	9.89	18.56	5.14	13.18	1.06	11.74	0.27
	$\widehat{\alpha}_L$	2.10	1.09	2.17	1.97	0.76	2.41	1.52	1.87	0.36	1.49	0.05	2.00
	$\widehat{\alpha}_U$	2.13	2.26	1.10	1.12	0	1.16	0	1.62	0	2.07	0	0
<i>B</i>	<i>ML</i>	35.29	31.03	30.61	20.41	22.09	9.31	17.33	4.85	12.19	1.00	10.79	0.26
	$\widehat{\alpha}_L$	2.10	1.09	2.17	1.97	2.54	2.41	1.52	1.87	3.73	1.49	2.98	2.00
	$\widehat{\alpha}_U$	2.13	2.26	1.10	2.72	0	1.16	0	1.62	0	2.07	0	0
<i>AC</i>	<i>ML</i>	33.24	30.06	29.72	19.40	23.09	8.78	19.20	4.56	14.81	0.94	13.61	0.27
	$\widehat{\alpha}_L$	2.10	1.09	2.17	1.97	2.54	2.41	1.52	1.87	0.36	1.49	0.05	2.00
	$\widehat{\alpha}_U$	2.13	2.26	1.10	2.72	0	2.95	0	3.79	0	2.07	0	0
<i>S</i>	<i>ML</i>	33.18	29.29	29.15	19.10	21.50	8.68	17.25	4.51	12.72	0.93	11.49	0.25
	$\widehat{\alpha}_L$	2.10	1.09	2.17	1.97	2.54	2.41	6.00	1.87	3.73	1.49	2.98	2.00
	$\widehat{\alpha}_U$	2.13	2.26	1.10	2.72	0	2.95	0	3.79	0	2.07	0	0
<i>SOC</i>	<i>ML</i>	33.96	30.73	29.54	20.13	20.87	9.15	15.67	4.78	9.52	0.98	7.71	0.24
	$\widehat{\alpha}_L$	2.10	3.89	2.17	1.97	2.54	2.41	1.52	1.87	0.36	4.03	0.05	2.00
	$\widehat{\alpha}_U$	2.13	2.26	3.80	2.72	4.18	2.95	0	3.79	0	2.07	0	5.00
<i>W</i>	<i>ML</i>	35.17	30.94	30.25	20.00	20.15	9.02	13.21	4.67	3.54	0.96	0.39	0.22
	$\widehat{\alpha}_L$	2.10	3.89	2.17	1.97	0.77	0.76	0.33	0.60	0.02	0.44	0	0.13
	$\widehat{\alpha}_U$	2.13	2.26	3.80	2.72	18.30	2.95	21.49	3.79	74.01	4.63	97.02	5.00

Tabela 4.7: Resultados de simulação de I. C. na ausência de erros de classificação ($m = 30$)

Estimação da taxa de prevalência na ausência de erros de classificação

p	n	.50		.25		.10		.05		.01		10^{-3}	
		1	2	1	4	1	10	1	20	1	100	1	100
<i>CP</i>	<i>ML</i>	20.24	17.79	17.67	11.39	12.57	5.11	9.43	2.64	5.23	0.54	3.80	0.14
	$\widehat{\alpha}_L$	1.73	2.06	1.64	2.28	2.05	2.21	1.12	2.31	1.91	1.67	0.50	1.36
	$\widehat{\alpha}_U$	1.82	1.61	2.10	1.91	2.41	2.29	0.56	2.37	0	2.11	0	1.12
<i>B</i>	<i>ML</i>	19.82	17.30	17.20	11.10	12.05	4.98	8.89	2.58	4.95	0.53	3.71	0.13
	$\widehat{\alpha}_L$	1.73	2.06	2.78	2.28	2.05	2.21	2.76	2.31	1.91	2.85	0.50	2.78
	$\widehat{\alpha}_U$	1.82	2.70	2.10	1.91	2.41	2.29	0.56	2.37	0	2.12	0	1.12
<i>AC</i>	<i>ML</i>	19.14	16.79	16.76	10.73	12.14	4.81	9.45	2.49	5.86	0.51	4.60	0.13
	$\widehat{\alpha}_L$	2.82	2.06	2.78	2.28	2.05	2.21	2.76	2.31	1.91	2.85	0.50	2.78
	$\widehat{\alpha}_U$	2.94	2.70	2.10	3.09	0.78	2.29	0.56	2.37	0	2.12	0	1.12
<i>S</i>	<i>ML</i>	19.14	16.68	16.65	10.69	11.77	4.80	8.85	2.48	5.10	0.51	3.86	0.13
	$\widehat{\alpha}_L$	2.82	2.06	2.78	2.28	3.94	2.21	2.76	2.31	7.99	2.85	9.54	2.78
	$\widehat{\alpha}_U$	2.94	2.70	2.10	3.09	2.41	2.29	0.56	2.37	0	2.12	0	1.12
<i>SOC</i>	<i>ML</i>	19.30	16.91	16.73	10.85	11.65	4.87	8.54	2.52	4.20	0.52	2.52	0.13
	$\widehat{\alpha}_L$	2.82	2.06	2.78	2.28	2.05	2.21	1.12	2.31	0.33	2.85	0.02	2.78
	$\widehat{\alpha}_U$	2.94	2.70	2.10	3.09	2.41	2.29	3.66	2.37	0	2.11	0	3.33
<i>W</i>	<i>ML</i>	19.50	16.97	16.86	10.85	11.59	4.85	8.29	2.51	3.00	0.52	0.38	0.13
	$\widehat{\alpha}_L$	2.82	2.06	1.64	2.28	0.98	1.28	0.42	1.33	0.04	1.67	0.001	0.65
	$\widehat{\alpha}_U$	2.94	2.70	3.77	3.09	5.74	3.68	11.68	3.75	36.79	3.40	90.4	7.72

Tabela 4.8: Resultados de simulação de I.C. na ausência de erros de classificação ($m = 100$)

verificando-se em ambos uma taxa de cobertura muito elevada (0.98) para $p = 0.001$.

Nos intervalos aproximados, à medida que o número de testes aumenta as oscilações da taxa de cobertura em torno do nível de confiança vão diminuindo sendo, por conseguinte, menos visíveis as vantagens de um intervalo relativamente aos restantes. As amplitudes médias dos intervalos também se tornam quase constantes para um determinado valor de p (ver Tabelas 4.8 e 4.9).

Relativamente aos intervalos exatos, o intervalo de Blaker consegue ser bastante menos conservador que o intervalo de Clopper-Pearson e consequentemente apresentar um comprimento médio inferior.

No que concerne aos intervalos unilaterais as conclusões são significativamente diferentes. Assim, para $m = 30$ (ver Tabela 4.7) podemos constatar o seguinte:

- Relativamente ao intervalo de Wald, o limite superior é liberal para prevalências próximas de zero. As probabilidades de cobertura vão aumentando à medida que p se aproxima de 0.5. Ao invés observa-se taxas de cobertura muito altas para o limite inferior quando p está próximo de 0. Essas taxas de cobertura diminuem drasticamente à medida que p se aproxima de 0.5.
- Os resultados obtidos para os intervalos de Wilson e Agresti-Coull são praticamente coincidentes. Verifica-se que o limite superior é conservativo para valores de p próximos de zero. Para valores de p superiores, a taxa de cobertura tem pequenas oscilações em torno do nível de confiança nominal. Relativamente ao limite inferior podemos constatar que é conservador para todos os valores de p .

- Relativamente ao SOC, o limite superior é liberal para valores de p próximos de zero e conservativo para valores de p próximos de 0.5. Para valores intermédios existe uma oscilação em torno no nível de confiança nominal. Ao invés, o limite inferior é conservativo para valores de p próximos de zero e liberal para valores de p próximos de 0.5.
- Os testes exatos (Clopper-Pearson e Blaker) exibem as mesmas probabilidades de cobertura. Verifica-se que o limite superior é ultra-conservador para valores de p próximos de zero. Para valores de p superiores existe um oscilação da taxa de cobertura. Refira-se que neste intervalo a taxa de cobertura nunca é inferior ao nível de confiança nominal. O mesmo se passa para o limite inferior para todos os valores de p . A amplitude média do intervalo de Blaker é inferior à do intervalo de Clopper-Pearson.

p	n	.50		.25		.10		.05		.01		10^{-3}	
		1	2	1	4	1	10	1	20	1	100	1	100
CP	ML	6.29	5.46	5.46	3.48	3.81	1.55	2.80	0.80	1.34	0.16	0.53	0.04
	$\widehat{\alpha}_L$	2.33	2.10	2.20	2.32	2.17	2.14	2.01	2.34	1.37	2.24	1.91	2.11
	$\widehat{\alpha}_U$	2.40	2.35	2.08	2.43	2.24	2.39	2.17	2.32	1.02	2.30	0	1.93
B	ML	6.22	5.42	5.42	3.45	3.78	1.54	2.76	0.80	1.30	0.16	0.53	0.04
	$\widehat{\alpha}_L$	2.33	2.52	2.20	2.32	2.17	2.14	2.78	2.34	2.65	2.58	1.91	2.11
	$\widehat{\alpha}_U$	2.40	2.35	2.51	2.43	2.24	2.39	2.17	2.32	1.02	2.30	0	2.51
AC	ML	6.18	5.36	5.36	3.41	3.73	1.52	2.73	0.79	1.33	0.16	0.61	0.04
	$\widehat{\alpha}_L$	2.69	2.52	2.63	2.32	2.71	2.49	2.78	2.34	2.65	2.58	1.91	2.66
	$\widehat{\alpha}_U$	2.77	2.78	2.51	2.43	2.24	2.39	2.17	2.32	1.02	2.66	0	1.93
S	ML	6.18	5.36	5.36	3.41	3.72	1.52	2.71	0.79	1.27	0.16	0.53	0.04
	$\widehat{\alpha}_L$	2.69	2.52	2.63	2.32	2.71	2.49	2.78	2.34	2.65	2.58	7.98	2.66
	$\widehat{\alpha}_U$	2.77	2.78	2.51	2.43	2.24	2.39	2.17	2.32	1.02	2.66	0	2.51
SOC	ML	6.18	5.36	5.36	3.42	3.72	1.53	2.70	0.79	1.24	0.16	0.43	0.06
	$\widehat{\alpha}_L$	2.69	2.52	2.20	2.32	2.17	2.49	2.01	2.75	1.37	2.58	0.36	2.11
	$\widehat{\alpha}_U$	2.77	2.35	2.51	2.43	2.93	2.39	3.07	2.32	2.92	2.30	0	2.51
W	ML	6.19	5.37	5.36	3.42	3.71	1.53	2.69	0.79	1.22	0.16	0.30	0.04
	$\widehat{\alpha}_L$	2.69	2.52	2.20	2.32	1.73	2.14	1.46	2.34	0.69	2.24	0.06	1.65
	$\widehat{\alpha}_U$	2.77	2.35	2.99	2.87	2.93	2.78	4.34	2.70	6.66	2.66	37	3.24

Tabela 4.9: Resultados de simulação de I.C. na ausência de erros de classificação ($m = 1000$)

Para $m = 1000$ (ver Tabela 4.9) podemos concluir que:

1. Relativamente ao intervalo de Wald, o limite superior mantém-se liberal praticamente para todos os valores de p com exceção de $p = 0.5$. Ao invés o limite inferior é conservativo para todos os valores de p com exceção de $p = 0.5$.
2. Relativamente aos intervalos de Wilson e Agesti-Coull, que mais uma vez apresentam taxas de cobertura praticamente idênticas, observa-se que tanto o limite inferior como o limite superior têm pequenas oscilações em torno do coeficiente de confiança nominal. Os intervalos apresentam a mesma amplitude média para todos os valores de p .
3. O SOC tem um comportamento praticamente idêntico aos intervalos de Wilson e Agesti-Coull.

4. No que diz respeito aos intervalos exatos é notório o carácter menos conservador do limite superior do intervalo de Blaker para valores de p próximos de zero.

Capítulo 5

Estimação da taxa de prevalência na presença de erros de classificação

5.1 Estimação por verosimilhança máxima

Consideremos agora que os testes podem estar sujeitos a erro. Seja $\varphi_s \in [0, 1]$ a sensibilidade do teste (probabilidade de teste positivo numa amostra infectada) e $\varphi_e \in [0, 1]$ a especificidade do teste (probabilidade de teste negativo numa amostra limpa). Vamos assumir que a miscigenação de unidades não altera as características operacionais do teste, isto é, $\varphi_{s_n} = \varphi_s$ e $\varphi_{e_n} = \varphi_e$. Denotando por X_n^+ o número de testes positivos quando m grupos (com n indivíduos cada) são analisados, temos que $X_n^+ \sim \text{binomial}(m, \pi_n)$, onde

$$\begin{aligned}\pi_n &= [1 - (1 - p)^n] \varphi_s + (1 - p)^n (1 - \varphi_e) \\ &= \varphi_s + (1 - p)^n (1 - \varphi_e - \varphi_s)\end{aligned}$$

representa a probabilidade do teste conjunto dar positivo. Note-se que no caso de $\varphi_e = \varphi_s = 1$ se obtém $\pi_n = 1 - (1 - p)^n$, que representa a probabilidade da amostra combinada conter pelo menos um elemento contaminado na situação clássica, de ausência de erros de classificação, analisada no capítulo anterior. Refira-se ainda que no caso especial de $n = 1$ se obtém $\pi_1 = p$ que representa a probabilidade de um teste individual dar positivo na ausência de erros de classificação.

Facilmente se verifica que π_n é uma função monótona crescente com p caso $\varphi_e + \varphi_s > 1$. Note-se que tal só acontece quando $(1 - \varphi_e - \varphi_s) < 0$. Assim, quanto mais elevada for a taxa de prevalência maior será a probabilidade do teste à amostra combinada dar positivo. Atendendo ao facto de p, φ_s e φ_e tomarem valores no intervalo $[0, 1]$, tem-se que

$$1 - \varphi_e \leq \pi_n \leq \varphi_s,$$

em que a igualdade no limite inferior do intervalo ocorre quando $p = 0$ e no limite superior quando $p = 1$.

Para estimar a taxa de prevalência, vamos determinar o estimador de verosimilhança máxima de p , isto é encontrar o maximizante de

$$L(p) \propto x_n^+ \ln[\pi_n] + (m - x_n^+) \ln[1 - \pi_n].$$

Dado que π_n é uma função monótona crescente de p , o estimador de máxima verosimilhança \hat{p} , pode ser determinado calculando previamente o estimador de máxima verosimilhança de π_n . Assim, pela propriedade da invariância dos estimadores de máxima verosimilhança tem-se

$$\widehat{\pi}_n = \varphi_s [1 - (1 - \hat{p})^n] + (1 - \hat{p})^n (1 - \varphi_e)$$

donde

$$\begin{aligned} \varphi_s - \widehat{\pi}_n &= (1 - \hat{p})^n (\varphi_s + \varphi_e - 1) \Leftrightarrow \\ (1 - \hat{p})^n &= \left(\frac{\varphi_s - \widehat{\pi}_n}{\varphi_s + \varphi_e - 1} \right) \Leftrightarrow \\ \hat{p} &= 1 - \left(\frac{\varphi_s - \widehat{\pi}_n}{\varphi_s + \varphi_e - 1} \right)^{1/n}. \end{aligned}$$

Dado que $\widehat{\pi}_n = \frac{x_n^+}{m}$ tem-se finalmente que

$$\hat{p} = 1 - \left(\frac{\varphi_s - x_n^+/m}{\varphi_s + \varphi_e - 1} \right)^{1/n}.$$

Para se obter um valor pertencente ao intervalo $[0, 1]$ é contudo necessário que

$$1 - \varphi_e \leq \frac{x_n^+}{m} \leq \varphi_s.$$

Caso $m \rightarrow \infty$, $\widehat{\pi}_n = \frac{x_n^+}{m} \rightarrow \pi_n$ pela lei forte dos grandes números. Caso contrário, a condição anterior depende dos valores de $n, m, \varphi_s, \varphi_e$ e p . No caso de uma doença rara, p é pequeno e o valor da especificidade é crucial para que $\frac{x_n^+}{m} \geq 1 - \varphi_e$. Refira-se que $1 - \varphi_e$ representa a taxa de falsos positivos e por conseguinte caso $\frac{x_n^+}{m} < 1 - \varphi_e$ obtemos uma taxa de testes positivos inferior à probabilidade de um grupo saudável ser considerado infetado.

Já foi referido que $\mathbb{P}[\widehat{\pi}_n \geq 1 - \varphi_e]$ cresce com m (número de grupos) mas esta solução implica um elevado aumento de recursos. O nosso objetivo é mostrar que incrementando n (dimensão dos grupos) estamos também a aumentar a $\mathbb{P}[\widehat{\pi}_n \geq 1 - \varphi_e]$.

O argumento anterior mostra que os testes compostos aumentam a probabilidade de obter uma estimativa para a prevalência, dado que a condição $\widehat{\pi}_n \geq 1 - \varphi_e$ se verifica com maior probabilidade. Contudo, no caso de a prevalência ser elevada, os testes compostos já não são aconselháveis, pois a probabilidade de cada grupo ser positivo é maior e, por conseguinte, a condição $\widehat{\pi}_n \leq \varphi_s$ pode não ser satisfeita.

Assim, $\mathbb{P}[\widehat{\pi}_n \geq 1 - \varphi_e]$ é uma medida quantitativa útil para averiguar se a prevalência é ou não estimável – um valor próximo de um dá-nos a “esperança” de conseguirmos estimar a prevalência ao passo que um valor próximo de zero indica-nos um risco de desperdício de recursos.

Para mostrar que $\mathbb{P}[\widehat{\pi}_n \geq 1 - \varphi_e]$ é uma função crescente com a dimensão do grupo, n , comecemos por notar que

$$\begin{aligned} \mathbb{P}[\widehat{\pi}_n \geq 1 - \varphi_e] &= \mathbb{P}[X_n^+ \geq \lceil m(1 - \varphi_e) \rceil] \\ &= \sum_{i=\lceil m(1-\varphi_e) \rceil}^m \binom{m}{i} [\pi_n]^i [1 - \pi_n]^{m-i}, \end{aligned}$$

onde $\lceil x \rceil$ representa o menor valor inteiro igual ou superior a x .

Recorrendo à conhecida relação entre a função beta incompleta e a soma de binomiais tem-se

$$\sum_{i=\lceil m(1-\varphi_e) \rceil}^m \binom{m}{i} [\pi_n]^i [1 - \pi_n]^{m-i} = \frac{1}{Be(l, m - l + 1)} \int_0^{\pi_n} x^{l-1} (1 - x)^{m-l} dx$$

onde $l = \lceil m(1 - \varphi_e) \rceil$ e $Be(a, b)$ é a função beta definida por

$$Be(a, b) = \int_0^1 x^{a-1} (1 - x)^{b-1} dx. \quad (5.1)$$

Derivando o integral em ordem a n obtém-se

$$[\pi_n]^{l-1} [1 - \pi_n]^{m-l} \frac{d}{dn} \pi_n,$$

onde o termo $\frac{d}{dn} \pi_n$ é positivo se e só se π_n for uma função crescente com n .

Concluimos assim que $\mathbb{P}[\widehat{\pi}_n \geq 1 - \varphi_e]$ pode ser determinada através da função de distribuição do modelo $Be(\lceil m(1 - \varphi_e) \rceil, m - \lceil m(1 - \varphi_e) \rceil + 1)$ calculada no ponto π_n .

O valor médio, a variância e o erro quadrático médio de \widehat{p} podem ser determinados através das seguintes expressões

$$\mathbb{E}(\widehat{p}) = \sum_{x=0}^m \left[1 - \left(\frac{\varphi_s - \frac{x}{m}}{\varphi_s + \varphi_e - 1} \right)^{\frac{1}{n}} \right] \binom{m}{x} \pi_n^x (1 - \pi_n)^{m-x}; \quad (5.2)$$

$$var(\widehat{p}) = \sum_{x=0}^m \left[1 - \left(\frac{\varphi_s - \frac{x}{m}}{\varphi_s + \varphi_e - 1} \right)^{\frac{1}{n}} \right]^2 \binom{m}{x} \pi_n^x (1 - \pi_n)^{m-x} - \mathbb{E}^2(\widehat{p}) \quad (5.3)$$

e

$$EQM(\widehat{p}) = var(\widehat{p}) + (B(\widehat{p}))^2. \quad (5.4)$$

Estimação da taxa de prevalência na presença de erros de classificação

p	.50	.25	.15	.1	.05	.025	.01	.005	.001
$n = 1$	$-3e^{-8}$	$2e^{-5}$.0004	.0016	.0057	.0104	.0146	.0163	.0178
$n = 2$.0084	.0031	.0018	.0013	.0016	.0034	.0061	.0075	.0089
$n = 3$	-.0041	.0050	.0026	.0017	.0010	.0016	.0034	.0046	.0058
$n = 4$	-.0831	.0071	.0032	.0019	.0010	.0010	.0021	.0031	.0043
$n = 5$	-.1833	.0095	.0038	.0022	.0010	.0007	.0015	.0023	.0034
$n = 6$	-.2577	.0102	.0044	.0024	.0011	.0006	.0011	.0017	.0028
$n = 7$	-.3046	.0059	.0052	.0026	.0011	.0006	.0008	.0014	.0023
$n = 8$	-.3341	-.0061	.0061	.0029	.0012	.0006	.0006	.0011	.0020
$n = 9$	-.3540	-.0250	.0070	.0032	.0012	.0006	.0005	.0009	.0017
$n = 10$	-.3684	-.0477	.0076	.0035	.0012	.0006	.0004	.0008	.0015
$n = 15$	-.4087	-.1392	-.0076	.0054	.0015	.0006	.0003	.0004	.0009
$n = 20$	-.4296	-.1757	-.0466	.0025	.0019	.0007	.0003	.0002	.0006
$n = 25$	-.4427	-.1919	-.0776	-.0106	.0024	.0007	.0003	.0002	.0005
$n = 50$	-.4703	-.2203	-.1202	-.0677	-.0041	.0012	.0003	.0001	.0002
$n = 100$	-.4849	-.2349	-.1349	-.0849	-.0334	-.0017	.0004	.0001	$5e^{-5}$
$n = 1000$	-.4985	-.2485	-.1485	-.0985	-.0485	-.0235	-.0085	-.0033	$4e^{-5}$

Tabela 5.1: Valores do viés ($m = 30, \varphi_s = \varphi_e = 0.95$)

A presença de erros de classificação provoca alterações no viés e no erro quadrático médio de \hat{p} . Para amostras de dimensão um (ver Tabela 5.1), ao contrário do que sucede no modelo binomial, \hat{p} não é um estimador centrado de p . Por outro lado, para prevalências altas e amostras de dimensão elevada, observa-se a existência de elevados valores negativos do viés. Relativamente ao erro quadrático médio (ver Tabela 5.2), podemos constatar que mesmo na presença de erros de classificação, é possível obter uma redução significativa do erro quadrático médio recorrendo à utilização de testes conjuntos sendo, para tal, necessário construir grupos de dimensão inferior à dimensão ótima utilizada no modelo binomial. Os valores a negrito identificam a dimensão ótima para cada valor de p . Aumentando o número de testes (valor de m) observamos que tanto o viés como o erro quadrático médio sofrem uma redução significativa (ver Tabelas 5.3 a 5.6)

p	.50	.25	.15	.1	.05	.025	.01	.005	.001
$n = 1$.0103	.0082	.0061	.0046	.0028	.0018	.0014	.0013	.0012
$n = 2$.0093	.0047	.0031	.0022	.0013	.0007	.0004	.0004	.0003
$n = 3$.0186	.0038	.0022	.0015	.0008	.0005	.0002	.0002	.0001
$n = 4$.0469	.0038	.0018	.0012	.0006	.0003	.0002	.0001	$8e^{-5}$
$n = 5$.0831	.0042	.0017	.0010	.0004	.0003	.0001	$8e^{-5}$	$6e^{-5}$
$n = 6$.0112	.0051	.0016	.0009	.0004	.0002	.0001	$6e^{-5}$	$4e^{-5}$
$n = 7$.1306	.0064	.0016	.0008	.0004	.0002	$8e^{-5}$	$5e^{-5}$	$3e^{-5}$
$n = 8$.1428	.0082	.0017	.0008	.0003	.0002	$7e^{-5}$	$4e^{-5}$	$2e^{-5}$
$n = 9$.1512	.0108	.0019	.0008	.0003	.0001	$6e^{-5}$	$4e^{-5}$	$2e^{-5}$
$n = 10$.1575	.0141	.0021	.0008	.0003	.0001	$6e^{-5}$	$3e^{-5}$	$2e^{-5}$
$n = 15$.1780	.0297	.0036	.0010	.0002	$9e^{-5}$	$4e^{-5}$	$2e^{-5}$	$8e^{-6}$
$n = 20$.1911	.0374	.0068	.0013	.0002	$7e^{-5}$	$3e^{-5}$	e^{-5}	$5e^{-6}$
$n = 25$.2003	.0412	.0100	.0019	.0002	$6e^{-5}$	$2e^{-5}$	e^{-5}	$3e^{-6}$
$n = 50$.2224	.0497	.0156	.0057	.0005	$6e^{-5}$	e^{-5}	$5e^{-6}$	e^{-6}
$n = 100$.2354	.0555	.0185	.0075	.0014	.0001	$9e^{-6}$	$3e^{-6}$	$6e^{-7}$
$n = 1000$.2485	.0617	.0220	.0097	.0024	.0006	$7e^{-5}$	e^{-5}	$9e^{-8}$

Tabela 5.2: Valores do EQM ($m = 30, \varphi_s = \varphi_e = 0.95$)

p	.50	.25	.15	.1	.05	.025	.01	.005	.001
$n = 1$	$-e^{-15}$	e^{-11}	$2e^{-7}$	e^{-5}	.0005	.0026	.0058	.0075	.0091
$n = 2$.0025	.0009	.0005	.0004	.0002	.0004	.0018	.0030	.0044
$n = 3$.0066	.0014	.0007	.0005	.0003	.0002	.0007	.0016	.0028
$n = 4$.0117	.0020	.0009	.0006	.0003	.0002	.0004	.0009	.0020
$n = 5$	-.0128	.0027	.0011	.0006	.0003	.0002	.0002	.0006	.0015
$n = 6$	-.0677	.0036	.0012	.0007	.0003	.0002	.0001	.0004	.0012
$n = 7$	-.1182	.0051	.0014	.0008	.0003	.0002	.0001	.0003	.0010
$n = 8$	-.1534	.0082	.0017	.0008	.0003	.0002	$9e^{-5}$.0002	.0008
$n = 9$	-.1767	.0145	.0019	.0009	.0003	.0002	$9e^{-5}$.0001	.0007
$n = 10$	-.1926	.0242	.0023	.0010	.0004	.0002	$8e^{-5}$.0001	.0006
$n = 15$	-.2328	.0358	.0093	.0016	.0004	.0002	$8e^{-5}$	$8e^{-5}$.0003
$n = 20$	-.2525	.0041	.0516	.0032	.0005	.0002	$7e^{-5}$	$4e^{-5}$.0002
$n = 25$	-.2650	-.0135	.0822	.0155	.0006	.0002	$7e^{-5}$	$4e^{-5}$.0001
$n = 50$	-.2914	-.0414	.0589	.1103	.0116	.0003	$8e^{-5}$	$4e^{-5}$	$2e^{-5}$
$n = 100$	-.3054	-.0554	.0446	.0946	.1428	.0096	.0001	$4e^{-5}$	$9e^{-6}$
$n = 1000$	-.3185	-.0685	.0315	.0815	.1315	.1565	.1715	.1707	e^{-5}

Tabela 5.3: Valores do viés ($m = 100, \varphi_s = \varphi_e = 0.95$)

p	.50	.25	.15	.1	.05	.025	.01	.005	.001
$n = 1$.0031	.0025	.0019	.0015	.0010	.0006	.0004	.0004	.0003
$n = 2$.0025	.0014	.0009	.0007	.0004	.0003	.0001	.0001	$9e^{-5}$
$n = 3$.0035	.0011	.0006	.0004	.0002	.0002	$8e^{-5}$	$6e^{-5}$	$4e^{-5}$
$n = 4$.0142	.0010	.0005	.0003	.0002	.0001	$6e^{-5}$	$4e^{-5}$	$2e^{-5}$
$n = 5$.0530	.0010	.0005	.0003	.0001	$8e^{-5}$	$4e^{-5}$	$3e^{-5}$	$2e^{-5}$
$n = 6$.0975	.0011	.0004	.0003	.0001	$6e^{-5}$	$3e^{-5}$	$2e^{-5}$	e^{-5}
$n = 7$.1279	.0015	.0004	.0002	.0001	$5e^{-5}$	$3e^{-5}$	$2e^{-5}$	$8e^{-6}$
$n = 8$.1459	.0031	.0004	.0002	$9e^{-5}$	$5e^{-5}$	$2e^{-5}$	e^{-5}	$7e^{-6}$
$n = 9$.1567	.0089	.0004	.0002	$8e^{-5}$	$4e^{-5}$	$2e^{-5}$	e^{-5}	$5e^{-6}$
$n = 10$.1638	.0216	.0005	.0002	$8e^{-5}$	$4e^{-5}$	$2e^{-5}$	e^{-5}	$4e^{-6}$
$n = 15$.1829	.1056	.0048	.0002	$6e^{-5}$	$3e^{-5}$	e^{-5}	$6e^{-6}$	$2e^{-6}$
$n = 20$.1946	.1274	.0546	.0007	$5e^{-5}$	$2e^{-5}$	$8e^{-6}$	$4e^{-6}$	$2e^{-6}$
$n = 25$.2030	.1324	.1104	.0119	$5e^{-5}$	$2e^{-5}$	$6e^{-6}$	$3e^{-6}$	e^{-6}
$n = 50$.2236	.1404	.1420	.1467	.0096	e^{-5}	$3e^{-6}$	$2e^{-6}$	$4e^{-7}$
$n = 100$.2360	.1458	.1447	.1517	.1589	.0086	$2e^{-6}$	$8e^{-7}$	$2e^{-8}$
$n = 1000$.2485	.01518	.1481	.1537	.1644	.1716	0.1765	0.1721	$2e^{-8}$

Tabela 5.4: Valores do EQM ($m = 100, \varphi_s = \varphi_e = 0.95$)

p	.50	.25	.15	.1	.05	.025	.01	.005	.001
$n = 1$	$-2e^{-14}$	$-e^{-15}$	$-2e^{-15}$	$-6e^{-15}$	e^{-10}	$5e^{-6}$.0004	.0012	.0026
$n = 2$.0002	$9e^{-5}$	$5e^{-5}$	$4e^{-5}$	$2e^{-5}$	e^{-5}	$2e^{-5}$.0002	.0011
$n = 3$.0006	.0001	$7e^{-5}$	$5e^{-5}$	$3e^{-5}$	$2e^{-5}$	e^{-5}	$5e^{-5}$.0006
$n = 4$.0014	.0002	$9e^{-5}$	$6e^{-5}$	$3e^{-5}$	$2e^{-5}$	e^{-5}	e^{-5}	.0004
$n = 5$.0040	.0003	.0001	$6e^{-5}$	$3e^{-5}$	$2e^{-5}$	$9e^{-6}$	$8e^{-6}$.0003
$n = 6$	-.0016	.0003	.0001	$7e^{-5}$	$3e^{-5}$	$2e^{-5}$	$9e^{-6}$	$7e^{-6}$.0002
$n = 7$	-.0634	.0005	.0001	$7e^{-5}$	$3e^{-5}$	$2e^{-5}$	$9e^{-6}$	$6e^{-6}$.0001
$n = 8$	-.1423	.0006	.0002	$8e^{-5}$	$3e^{-5}$	$2e^{-5}$	$8e^{-6}$	$6e^{-6}$	$9e^{-5}$
$n = 9$	-.1986	.0009	.0002	9^{-5}	3^{-5}	2^{-5}	8^{-6}	5^{-6}	$6e^{-5}$
$n = 10$	-.2342	.0012	.0002	.0001	$4e^{-5}$	$2e^{-5}$	$8e^{-6}$	$5e^{-6}$	$5e^{-5}$
$n = 15$	-.3045	.0042	.0005	.0001	$4e^{-5}$	$2e^{-5}$	$7e^{-6}$	$5e^{-6}$	e^{-5}
$n = 20$	-.3347	-.0567	.0014	.0002	$5e^{-5}$	$2e^{-5}$	$7e^{-6}$	$4e^{-6}$	e^{-6}
$n = 25$	-.3540	-.0975	.0069	.0004	$6e^{-5}$	$2e^{-5}$	$7e^{-6}$	$4e^{-6}$	$2e^{-6}$
$n = 50$	-.3957	-.1457	-.0444	.0142	.0002	$3e^{-5}$	$8e^{-6}$	$4e^{-6}$	e^{-6}
$n = 100$	-.4183	-.1683	-.0683	.0183	.0294	.0001	e^{-5}	$4e^{-6}$	$9e^{-7}$
$n = 1000$	-.4398	-.1898	-.0898	-.0398	.0102	.0352	.0502	.0394	e^{-6}

Tabela 5.5: Valores do viés ($m = 1000, \varphi_s = \varphi_e = 0.95$)

p	.50	.25	.15	.1	.05	.025	.01	.005	.001
$n = 1$.0003	.0002	.0002	.0001	.0001	$8e^{-5}$	$6e^{-5}$	$4e^{-5}$	3^{-5}
$n = 2$.0002	.0001	$9e^{-5}$	$7e^{-5}$	$4e^{-5}$	$3e^{-5}$	$2e^{-5}$	e^{-5}	$8e^{-6}$
$n = 3$.0003	.0001	$6e^{-5}$	$4e^{-5}$	$2e^{-5}$	$2e^{-5}$	e^{-5}	$8e^{-6}$	$4e^{-6}$
$n = 4$.0005	.0001	$5e^{-5}$	$3e^{-5}$	$2e^{-5}$	e^{-5}	$6e^{-6}$	$5e^{-6}$	$2e^{-6}$
$n = 5$.0012	.0001	$4e^{-5}$	$3e^{-5}$	e^{-5}	$8e^{-6}$	$5e^{-6}$	$3e^{-6}$	$2e^{-6}$
$n = 6$.0115	.0001	$4e^{-5}$	$2e^{-5}$	e^{-5}	$6e^{-6}$	$3e^{-6}$	$3e^{-6}$	e^{-6}
$n = 7$.0493	.0001	$4e^{-5}$	$2e^{-5}$	e^{-5}	$5e^{-6}$	$3e^{-6}$	$2e^{-6}$	e^{-6}
$n = 8$.0881	.0001	$4e^{-5}$	$2e^{-5}$	$9e^{-6}$	$5e^{-6}$	$2e^{-6}$	$2e^{-6}$	$8e^{-7}$
$n = 9$.1129	.0002	$4e^{-5}$	$2e^{-5}$	$8e^{-6}$	$4e^{-6}$	$2e^{-6}$	e^{-6}	$7e^{-7}$
$n = 10$.1273	.0002	$4e^{-5}$	$2e^{-5}$	$8e^{-6}$	$4e^{-6}$	$2e^{-6}$	e^{-6}	$6e^{-7}$
$n = 15$.1538	.0126	$6e^{-5}$	$2e^{-5}$	$6e^{-6}$	$2e^{-6}$	e^{-6}	$6e^{-7}$	$3e^{-7}$
$n = 20$.1679	.0511	.0001	$2e^{-5}$	$5e^{-6}$	$2e^{-6}$	$8e^{-7}$	$4e^{-7}$	$2e^{-7}$
$n = 25$.1789	.0621	.0058	$3e^{-5}$	$5e^{-6}$	$2e^{-6}$	$6e^{-7}$	$3e^{-7}$	e^{-7}
$n = 50$.2083	.0729	.0535	.0404	$8e^{-6}$	e^{-6}	$3e^{-7}$	$2e^{-7}$	$5e^{-8}$
$n = 100$.2274	.0807	.0570	.0527	.0400	$2e^{-6}$	$2e^{-7}$	$8e^{-8}$	$2e^{-8}$
$n = 1000$.2476	.0902	.0622	.0558	.0543	.0554	.0567	.0403	$2e^{-9}$

Tabela 5.6: Valores do EQM ($m = 1000, \varphi_s = \varphi_e = 0.95$)

5.2 Estimação intervalar

Na presença de erros de classificação onde a sensibilidade e a especificidade são inferiores a 1, surge a necessidade de recalculer os limites de confiança apresentados no capítulo anterior. Tendo em conta que a variável aleatória $X_n^+ \sim \text{binomial}(m, \pi_n)$ com

$$\pi_n = \varphi_s + (1 - p)^n (1 - \varphi_s - \varphi_e)$$

e que o estimador de máxima verosimilhança de π_n é dado por $\frac{X_n^+}{m}$ obtém-se

$$\hat{p} = 1 - \left(\frac{\varphi_s - \frac{X_n^+}{m}}{\varphi_s + \varphi_e - 1} \right)^{\frac{1}{n}}. \quad (5.5)$$

Denotando os limites de confiança dum intervalo sem a presença de erros de classificação por $[L_1, U_1]$ vamos escrever o estimador de máxima verosimilhança de π_n à custa dos limites desse intervalo. Assim obtemos:

$$L_1 = 1 - \left(1 - \frac{X_n^+}{m} \right)^{\frac{1}{n}} \Leftrightarrow \frac{X_n^+}{m} = 1 - (1 - L_1)^n, \quad (5.6)$$

$$U_1 = 1 - \left(1 - \frac{X_n^+}{m} \right)^{\frac{1}{n}} \Leftrightarrow \frac{X_n^+}{m} = 1 - (1 - U_1)^n. \quad (5.7)$$

Substituindo em (5.5) $\frac{X_n^+}{m}$ por $1 - (1 - L_1)^n$ e por $1 - (1 - U_1)^n$ obtém-se respetivamente:

$$L_2 = 1 - \left(\frac{\varphi_s - 1 + (1 - L_1)^n}{\varphi_s + \varphi_e - 1} \right)^{\frac{1}{n}}, \quad (5.8)$$

$$U_2 = 1 - \left(\frac{\varphi_s - 1 + (1 - U_1)^n}{\varphi_s + \varphi_e - 1} \right)^{\frac{1}{n}}, \quad (5.9)$$

onde L_2 e U_2 são os novos limites do intervalo de confiança.

5.2.1 Intervalo de confiança assintótico

Designando por

$$\hat{p} = 1 - \left(\frac{\varphi_s - x_n^+/m}{\varphi_s + \varphi_e - 1} \right)^{1/n}$$

o estimador de máxima verosimilhança de p , tem-se o seguinte resultado.

Teorema 5.1. (Distribuição assintótica de \hat{p})

Para um número fixo de testes, a distribuição assintótica de \hat{p} é dada por

$$\hat{p} \overset{o}{\sim} \text{Gaussiana}(p, \sqrt{\text{var}(\hat{p})})$$

onde

$$\text{var}(\hat{p}) = \frac{1}{mn^2} [\varphi_s - \pi_n]^{\frac{2}{n}-2} \frac{\pi_n (1 - \pi_n)}{(\varphi_e + \varphi_s - 1)^{\frac{2}{n}}}.$$

Demonstração. .

Denotando por $\widehat{\pi}_n = \frac{X_n^+}{m}$ o estimador de verosimilhança máxima de π_n , onde X_n^+ representa o número de testes positivos quando m grupos de dimensão n são analisados, tem-se que $X_n^+ \sim$ binomial (m, π_n) . Sabemos que $\widehat{\pi}_n$ converge em probabilidade para π_n e, para valores elevados de m $\sqrt{I_m(\pi_n)}(\widehat{\pi}_n - \pi_n) \overset{o}{\sim} \text{Gaussiana}(0, 1)$ onde $I_m(\pi_n) = \frac{m}{\pi_n(1-\pi_n)}$ denota a quantidade de informação estatística numa amostra de dimensão m . Seja $\hat{p} = g[\widehat{\pi}_n]$ uma função de $\widehat{\pi}_n$. Atendendo a que, para valores elevados de m , $\widehat{\pi}_n$ está próximo de π_n , podemos aproximar $g[\widehat{\pi}_n]$ pelos dois primeiros termos da série de Taylor em torno de π_n obtendo

$$g[\widehat{\pi}_n] \approx v[\widehat{\pi}_n] = g[\pi_n] + (\widehat{\pi}_n - \pi_n) g'(\pi_n),$$

onde $g'(\pi_n)$ existe e é diferente de zero. Dado que $v[\widehat{\pi}_n]$ é uma função linear de $\widehat{\pi}_n$, tem-se que:

- $\mathbb{E}[v(\widehat{\pi}_n)] = g[\pi_n] + \mathbb{E}(\widehat{\pi}_n - \pi_n) g'(\pi_n) = g(\pi_n) = p$

$$\begin{aligned} \text{var} [v(\widehat{\pi}_n)] &= [g'(\pi_n)]^2 \text{var}(\widehat{\pi}_n - \pi_n) \\ &= [g'(\pi_n)]^2 \left(\frac{\pi_n(1 - \pi_n)}{m} \right) \end{aligned}$$

onde

$$\begin{aligned} [g'(\pi_n)]^2 &= \left\{ \frac{d}{d\pi_n} \left[1 - \left(\frac{\varphi_s - \pi_n}{\varphi_s + \varphi_e - 1} \right)^{1/n} \right] \right\}^2 \\ &= \left[\frac{-1}{n} \left(\frac{\varphi_s - \pi_n}{\varphi_s + \varphi_e - 1} \right)^{\frac{1}{n}-1} \left(\frac{-1}{\varphi_s + \varphi_e - 1} \right) \right]^2 \\ &= \frac{1}{n^2} \times \frac{[\varphi_s - \pi_n]^{\frac{2}{n}-2}}{(\varphi_s + \varphi_e - 1)^{\frac{2}{n}}} \end{aligned}$$

Para valores elevados de m , $g(\widehat{\pi}_n)$ é aproximadamente igual a $v[\widehat{\pi}_n]$ e por conseguinte têm a mesma distribuição limite. Conclui-se assim que

$$\widehat{p} \overset{o}{\sim} \text{Gaussiana} \left(p, \sqrt{\frac{1}{mn^2} (\varphi_s - \pi_n)^{\frac{2}{n}-2} \frac{\pi_n(1 - \pi_n)}{(\varphi_e + \varphi_s - 1)^{\frac{2}{n}}} \right). \quad (5.10)$$

□

Uma estimativa da variância do estimador de p obtém-se substituindo π_n por $\widehat{\pi}_n = \frac{X_n^+}{m}$.

Se em (5.10) fizermos $\varphi_s = \varphi_e = 1$ e substituirmos π_n por $1 - (1 - p)^n$ obtém-se (4.15), presente na página 57.

5.2.2 Comparação de intervalos de confiança via simulação

Na Tabela 5.7 podemos constatar que para um determinado número de teste ($m = 30$), a existência de erros de classificação provoca alterações significativas nos valores dos parâmetros analisados. É visível logo à partida um aumento significativo do comprimento médio dos intervalos se compararmos os resultados patentes na Tabela 5.7 que figura na página 77. O aumento do número de testes (ver Tabelas 5.8 e 5.9) tem como consequência a diminuição significativa das amplitudes dos intervalos bem como a estabilização dos valores dos parâmetros em torno do nível de confiança nominal.

No que respeita aos intervalos bilaterais refira-se que:

1. O intervalo de Wald é altamente conservador para prevalências próximas de zero e menos liberal para prevalências elevadas.

2. Relativamente aos intervalos de Wilson e Agresti-Coull salta à vista a diminuição significativa da taxa de cobertura para prevalências próximas de zero
3. No que diz respeito aos intervalos exatos constata-se um aumento das taxas de cobertura para prevalências próximas de zero e uma ligeira diminuição para prevalências mais altas.

Relativamente a intervalos unilaterais note-se que:

1. Na presença de erros de classificação o limite inferior do intervalo de Wald é liberal para todos os valores de p . O limite superior é ultra conservador para $p = 0.001$ e para prevalências superiores as probabilidades de cobertura apresentam grandes oscilações em torno do nível de confiança nominal.
2. Os limites superiores dos intervalos de Agresti-Coull e de score de Wilson não sofrem alterações significativas. O limite inferior é liberal para prevalências pequenas e conservador para prevalências elevadas.
3. A existência de erros de classificação faz com que o limite superior dos intervalos de Clopper-Pearson e Blaker sejam mais conservativos para prevalências baixas.

Em forma de conclusão podemos referir que:

- Para determinar a dimensão ótima do grupo é necessário dispor de uma estimativa prévia da taxa de prevalência
- Caso existam erros de classificação é necessário conhecer o valor da sensibilidade e da especificidade ou pelo menos boas estimativas dos mesmos.
- Nesta análise apenas foi considerado o custo de cada teste e, por conseguinte, o custo relativo à junção das amostras não foi considerado.
- Fixando o número de testes (m fixo), os estimadores baseados em testes compostos são mais precisos que os estimadores baseados em testes simples.
- É necessária a obtenção de estimadores quando $1 - \varphi_e \leq \frac{X_n^+}{m} \leq \varphi_s$, uma vez que os estimadores utilizados neste trabalho não permitem efetuar a estimação de p nesta situação. Deste modo, quando tal ocorreu nas simulações, esses valores foram retirados da análise.

Estimação da taxa de prevalência na presença de erros de classificação

p	n	.50		.25		.10		.05		.01		10^{-3}	
		1	2	1	4	1	10	1	20	1	100	1	100
<i>CP</i>	<i>ML</i>	40.93	39.67	36.96	26.10	30.38	12.05	27.87	6.44	25.99	1.55	25.61	0.36
	$\widehat{\alpha}_L$	2.13	1.84	1.92	1.17	1.96	1.59	2.62	1.32	1.35	1.07	3.64	1.66
	$\widehat{\alpha}_U$	2.17	1.92	1.88	2.02	0	1.98	0	2.37	0	1.20	0	0
<i>B</i>	<i>ML</i>	39.21	37.09	35.03	24.39	28.55	11.28	26.11	6.03	24.29	1.46	23.93	0.34
	$\widehat{\alpha}_L$	2.13	1.84	1.92	1.17	1.96	1.59	2.62	1.32	5.47	1.07	3.64	1.66
	$\widehat{\alpha}_U$	2.17	1.92	1.88	2.02	0	1.98	0	2.37	0	3.00	0	0
<i>AC</i>	<i>ML</i>	36.93	35.59	33.81	23.02	28.82	10.58	26.98	5.66	25.63	1.38	25.36	0.34
	$\widehat{\alpha}_L$	2.13	1.84	1.92	1.17	1.96	1.59	2.62	1.32	5.47	2.94	3.64	4.50
	$\widehat{\alpha}_U$	2.17	1.92	1.88	2.02	0	1.98	0	2.37	0	3.00	0	0
<i>S</i>	<i>ML</i>	36.87	34.95	33.30	22.71	27.48	10.46	25.30	5.60	23.67	1.37	23.34	0.33
	$\widehat{\alpha}_L$	2.13	1.84	1.92	1.17	5.33	1.59	7.68	1.32	5.47	2.94	13.96	4.50
	$\widehat{\alpha}_U$	2.17	1.92	1.88	2.02	0	1.98	0	2.37	0	3.00	0	0
<i>SOC</i>	<i>ML</i>	37.73	36.73	33.82	24.15	27.37	11.14	24.91	5.96	23.07	1.44	22.70	0.32
	$\widehat{\alpha}_L$	2.13	1.84	1.92	3.50	1.96	4.28	2.62	3.43	1.35	2.94	3.64	1.66
	$\widehat{\alpha}_U$	2.17	1.92	1.88	2.02	0	1.98	0	2.37	0	3.00	0	0
<i>W</i>	<i>ML</i>	39.07	36.99	34.74	23.72	27.46	10.77	24.65	5.74	22.53	1.39	22.11	0.31
	$\widehat{\alpha}_L$	2.13	1.84	1.92	1.17	0.62	1.59	0.72	1.32	0.26	1.07	0.82	0.54
	$\widehat{\alpha}_U$	2.17	1.92	1.88	4.59	0	1.98	0	5.29	0	6.35	0	0

Tabela 5.7: Resultados de simulação de I.C. utilizando $m = 30$ e $\varphi_s = \varphi_e = 0.95$

p	n	.50		.25		.10		.05		.01		10^{-3}	
		1	2	1	4	1	10	1	20	1	100	1	100
<i>CP</i>	<i>ML</i>	22.49	20.54	20.20	12.97	15.97	5.79	13.97	2.99	12.62	0.61	12.38	0.18
	$\widehat{\alpha}_L$	1.74	1.88	2.43	1.54	1.97	1.77	1.43	2.01	2.77	2.42	3.12	1.40
	$\widehat{\alpha}_U$	1.72	2.40	1.84	1.79	1.98	1.90	0	1.84	0	1.56	0	0.85
<i>B</i>	<i>ML</i>	22.02	19.99	19.69	12.64	15.43	5.64	13.39	2.92	11.97	0.60	11.70	0.17
	$\widehat{\alpha}_L$	1.74	1.88	2.43	2.62	1.97	1.77	2.90	2.01	2.77	2.42	7.71	2.65
	$\widehat{\alpha}_U$	1.72	2.40	1.84	1.79	1.98	3.10	0	2.94	0	2.58	0	0.85
<i>AC</i>	<i>ML</i>	21.27	19.35	19.15	12.20	15.29	5.44	13.49	2.81	12.30	0.58	12.09	0.17
	$\widehat{\alpha}_L$	2.82	1.88	2.43	2.62	3.54	1.77	2.90	2.01	6.25	2.42	7.71	2.65
	$\widehat{\alpha}_U$	2.79	2.40	1.84	2.89	1.98	3.10	0	2.94	0	2.58	0	0.85
<i>S</i>	<i>ML</i>	21.27	19.25	19.06	12.17	15.00	5.43	13.09	2.81	11.81	0.58	11.58	0.17
	$\widehat{\alpha}_L$	2.82	1.88	2.43	2.62	3.54	1.77	2.90	2.01	6.25	2.42	7.71	2.65
	$\widehat{\alpha}_U$	2.79	2.40	1.84	2.89	1.98	3.10	0	2.94	0	2.58	0	0.85
<i>SOC</i>	<i>ML</i>	21.44	19.54	19.16	12.35	14.95	5.52	12.96	2.85	11.61	0.59	11.38	0.17
	$\widehat{\alpha}_L$	2.82	3.39	2.43	2.62	1.97	3.03	2.90	3.31	2.77	2.42	3.121	2.65
	$\widehat{\alpha}_U$	2.79	2.40	3.31	2.89	1.98	3.10	0	2.94	0	2.58	0	2.55
<i>W</i>	<i>ML</i>	21.67	19.60	19.32	12.33	14.97	5.48	12.89	2.83	11.49	0.58	11.24	0.16
	$\widehat{\alpha}_L$	2.82	3.39	2.43	1.54	1.05	1.77	1.43	2.01	1.12	1.42	1.19	0.73
	$\widehat{\alpha}_U$	2.79	2.40	3.31	2.89	1.98	3.10	0	2.94	0	4.04	0	5.66

Tabela 5.8: Resultados de simulação de I.C. utilizando $m = 100$ e $\varphi_s = \varphi_e = 0.95$

p	n	.50		.25		.10		.05		.01		10^{-3}	
		1	2	1	4	1	10	1	20	1	100	1	100
<i>CP</i>	<i>ML</i>	6.99	6.26	6.25	3.92	4.88	1.74	4.14	0.90	3.40	0.18	3.29	0.053
	$\widehat{\alpha}_L$	2.42	2.45	2.22	2.45	2.25	2.19	1.97	2.40	2.48	2.14	4.02	2.29
	$\widehat{\alpha}_U$	2.32	2.31	2.47	2.43	2.25	2.16	2.02	2.23	0	2.31	0	2.40
<i>B</i>	<i>ML</i>	6.91	6.22	6.21	3.89	4.84	1.73	4.10	0.89	3.36	0.18	3.25	0.053
	$\widehat{\alpha}_L$	2.42	2.45	2.22	2.45	2.74	2.19	2.54	2.40	2.48	2.14	4.02	2.29
	$\widehat{\alpha}_U$	2.32	2.31	2.47	2.43	2.25	2.54	2.02	2.60	0	2.31	0	2.40
<i>AC</i>	<i>ML</i>	6.87	6.15	6.14	3.85	4.78	1.71	4.05	0.88	3.32	0.18	3.20	0.052
	$\widehat{\alpha}_L$	2.80	2.45	2.60	2.45	2.74	2.19	2.54	2.40	3.30	2.50	5.49	2.86
	$\widehat{\alpha}_U$	2.68	2.71	2.47	2.43	2.25	2.54	2.02	2.60	0	2.64	0	2.40
<i>S</i>	<i>ML</i>	6.87	6.14	6.14	3.85	4.77	1.71	4.04	0.88	3.30	0.18	3.19	0.052
	$\widehat{\alpha}_L$	2.80	2.45	2.60	2.45	2.74	2.19	2.54	2.40	3.30	2.50	5.49	2.86
	$\widehat{\alpha}_U$	2.68	2.71	2.47	2.43	2.25	2.54	2.02	2.60	0	2.64	0	2.40
<i>SOC</i>	<i>ML</i>	6.88	6.15	6.14	3.86	4.77	1.71	4.04	0.88	3.29	0.18	3.18	0.052
	$\widehat{\alpha}_L$	2.80	2.45	2.22	2.45	2.25	2.59	2.54	2.40	2.48	2.50	4.02	2.29
	$\widehat{\alpha}_U$	2.68	2.31	2.47	2.43	2.83	2.54	2.65	2.60	0	2.64	0	2.96
<i>W</i>	<i>ML</i>	6.88	6.15	6.15	3.86	4.77	1.71	4.03	0.88	3.29	0.18	3.17	0.052
	$\widehat{\alpha}_L$	2.80	2.45	2.22	2.45	2.25	2.19	1.97	2.40	1.82	2.14	2.87	1.86
	$\widehat{\alpha}_U$	2.68	2.71	2.93	2.87	2.83	2.54	3.41	3.00	0	3.04	0	3.68

Tabela 5.9: Resultados de simulação de I.C. utilizando $m = 1000$ e $\varphi_s = \varphi_e = 0.95$

5.3 Robustez do modelo binomial

Um dos pressupostos que é usual impor no modelo binomial é o teste de diagnóstico ter probabilidade nula de erro, isto é, a sensibilidade e a especificidade do teste serem iguais a 100%. Com o intuito de estudar a robustez do modelo binomial, Hung & Swallow (1999) abordaram as alterações provocadas no erro quadrático médio do estimador de p , quando o referido pressuposto é violado ($\varphi_s \neq 1$) e propuseram como alternativa dois modelos para o efeito de diluição (todavia continuam a considerar $\varphi_e = 1$).

O efeito de diluição existe quando a diminuição do número de contaminados no grupo tem como consequência um aumento da taxa de falsos negativos e, por conseguinte, uma diminuição da sensibilidade. Denotando verdadeiro número de contaminados no grupo j por $D_j \sim \text{binomial}(n, p)$ e por H_j uma variável indicatriz que assume o valor 1 caso o j -ésimo grupo seja classificado como limpo ou o valor 0 caso seja classificado como contaminado, tem-se:

$$\begin{aligned}
 \mathbb{P}(H_j = 0) &= \mathbb{P}(H_j = 0, D_j \geq 1) \\
 &= \sum_{\gamma=1}^n \mathbb{P}(H_j = 0 | D_j = \gamma) \mathbb{P}(D_j = \gamma) \\
 &= \sum_{\gamma=1}^n \varphi_s^{[\gamma, n]} \mathbb{P}(D_j = \gamma)
 \end{aligned}$$

	$f_2 = 0$	$f_2 = 0.1$	$f_2 = 1$	$f_2 = 10$	$f_2 = 100$
$n = 6$	1.00	$8.53e^{-1}$	$3.68e^{-1}$	$5.50e^{-2}$	$5.79e^{-3}$
$n = 7$	1.00	$6.1e^{-1}$	$1.35e^{-1}$	$1.54e^{-2}$	$1.56e^{-3}$
$n = 8$	1.00	$3.44e^{-1}$	$4.98e^{-2}$	$5.21e^{-3}$	$5.24e^{-4}$
$n = 9$	1.00	$1.57e^{-1}$	$1.83e^{-2}$	$1.86e^{-3}$	$1.87e^{-4}$
$n = 10$	1.00	$6.35e^{-2}$	$6.74e^{-3}$	$6.78e^{-4}$	$6.78e^{-5}$
$n = 15$	1.00	$4.54e^{-4}$	$4.54e^{-5}$	$4.54e^{-6}$	$4.54e^{-7}$
$n = 20$	1.00	$3.06e^{-6}$	$3.06e^{-7}$	$3.06e^{-8}$	$3.06e^{-9}$
$n = 30$	1.00	$1.39e^{-10}$	$1.39e^{-11}$	$1.39e^{-12}$	$1.39e^{-13}$

Tabela 5.10: Valores assumidos pelo primeiro modelo de diluição

e, por conseguinte, o número de grupos classificados como limpos, H , é modelado por

$$H = \sum_{j=1}^m H_j \sim \text{binomial} \left(m, 1 - \sum_{\gamma=1}^n \varphi_s^{[\gamma, n]} \mathbb{P}(D_j = \gamma) \right).$$

Assim, na presença do efeito de diluição, o parâmetro π_n que figura nas expressões (4.4) (4.5) e (4.6) da página 43 deverá ser substituído por $\sum_{\gamma=1}^n \varphi_s^{[\gamma, n]} \mathbb{P}(D_j = \gamma)$. Hung & Swallow (1999) propõem 2 modelos de diluição para $\varphi_s^{[\gamma, n]}$.

5.3.1 Primeiro modelo de diluição

Neste primeiro modelo assume-se que um *falso negativo* ocorre quando a proporção de indivíduos contaminados no grupo (γ/n) é inferior a uma determinada fasquia ($1/f_1$) e que a probabilidade de ocorrência de um *falso negativo* é inversamente proporcional à fração de contaminados no grupo. Assim, para $d = 1, 2, \dots, n$, tem-se

$$\varphi_s^{[\gamma, n]} = I\left(\frac{\gamma}{n} \geq \frac{1}{f_1}\right) + \frac{\gamma}{(e^{n-f_1\gamma} - 1) f_2 + \gamma} \times I\left(\frac{\gamma}{n} < \frac{1}{f_1}\right), \quad (5.11)$$

onde f_1 e f_2 são os parâmetros do modelo e I é uma função indicatriz. f_1 é um parâmetro conhecido cujo valor depende do teste utilizado para a classificação dos grupos. Note-se que quando $n < f_1$ não existe efeito de diluição dado que $\frac{1}{f_1} \leq \frac{1}{n} \leq \frac{\gamma}{n}$ e, por conseguinte, $\varphi_s^{[\gamma, n]} = 1$ para $\gamma \geq 1$. O parâmetro f_2 representa os diferentes níveis de diluição e, por conseguinte, quanto maior for o valor de f_2 maior será o efeito de diluição. No caso de $f_2 = 0$ tem-se $\varphi_s^{[\gamma, n]} = 1$ (ver Tabela 5.10).

Considerando $f_1 = 5$ e $f_2 = 100$, a probabilidade de um grupo com um contaminado ser declarado como positivo é aproximadamente igual a 0.0058 quando $n = 6$ e 4.54×10^{-7} quando $n = 15$ (ver Tabela 5.10).

5.3.2 Segundo modelo de diluição

Neste modelo não se estipula um limiar para a proporção de defeituosos no grupo abaixo do qual se considere a existência do efeito de diluição. Cada elemento não contaminado contribui com uma quantidade específica f (factor de diluição) tal que quando $f > 0$, a probabilidade de ocorrência de um falso negativo aumenta com o número $(n - \gamma)$ de não contaminados no grupo. Este modelo é definido por

$$\varphi_s^{[\gamma, n]} = \frac{\gamma}{(n - \gamma) \times f + \gamma}. \quad (5.12)$$

Cada um dos γ elementos contaminados tem peso 1 e cada um dos $n - \gamma$ elementos saudáveis tem peso f . Para garantir que a contribuição de um elemento saudável para a positividade do teste seja inferior à contribuição de um elemento contaminado considera-se que $f \leq 1$. Na presença do efeito de diluição é ainda possível manter os benefícios do modelo binomial essencialmente quando p é pequeno e se recorre a grupos de dimensão inferior à dimensão ótima para o modelo binomial.

Para concluir podemos dizer que o modelo binomial é robusto na presença do efeito de diluição dado que optando por grupos de dimensão inferior à dimensão ótima, consegue-se em muitas situações obter grande parte dos benefícios alcançados pelo modelo.

Capítulo 6

Testes quantitativos discretos

6.1 Introdução

Suponhamos que se pretende detetar a presença de uma bactéria para testar a contaminação de iogurtes numa unidade fabril. Para tal, assuma-se que a taxa de contaminação é constante e igual a p . O número de iogurtes contaminados em n iogurtes selecionados é uma variável aleatória $I^{[n]} \sim \text{binomial}(n, p)$. Assuma-se que o número de bactérias presentes num mililitro de iogurte é representado por Y_i^* . No caso de o iogurte não estar contaminado então $Y_i^* = 0$. Caso contrário, a quantidade de bactérias por ml de iogurte será modelada por uma variável aleatória discreta Y_i com distribuição \mathbf{D} , vetor de parâmetros θ e suporte \mathbb{N}_0 . Quando se misturam as n amostras individuais para posteriormente retirar um mililitro dessa mistura para teste, o número de bactérias numa amostra de n ml é uma variável aleatória $B_n = \sum_{i=1}^n Y_i^* = \sum_{i=1}^I Y_i$. Assim, assumindo que o processo de mistura é perfeito, o número de bactérias numa amostra de um mililitro de iogurte é uma variável aleatória $B_1 \sim \text{binomial}(B_n, \frac{1}{n})$ (Santos, Pestana e Martins, 2012). O mesmo resultado pode ser obtido considerando $B_1 = \sum_{i=1}^n W_i^* = \sum_{i=1}^I W_i$ onde as variáveis aleatórias W_i são descritas pelo modelo hierárquico $W_i \sim \text{binomial}(T, \frac{1}{n})$ com $T \sim \mathbf{D}(\theta)$.

Santos, Pestana e Martins (2012), deduziram expressões para o cálculo da sensibilidade e da especificidade no caso de o número de bactérias por mililitro de iogurte ser bem modelado, caso o iogurte esteja contaminado, por uma variável aleatória Y com distribuição de Poisson de parâmetro λ . As expressões referidas foram determinadas quer assumindo a ausência de erros de classificação associados ao teste (metodologia de Dorfman) quer considerando a existência de um erro associado ao próprio teste. Assim, neste capítulo, estendemos estes resultados a populações sobre-dispersas como a geométrica e mais geralmente qualquer binomial negativa. Condicionalmente ao número de iogurtes contaminados, é determinado o número de bactérias numa amostra de um mililitro de iogurte (essencial para o cálculo da sensibilidade e especificidade em testes compostos) para a família de distribuições de contagem não degeneradas que verificam a fórmula recursiva de Panjer (1981) (binomial, Poisson e binomial negativa), bem como para o modelo logarítmico (uma das suas extensões). Para in-

formação mais detalhada consulte-se Sundt & Jewell (1981), Willmot (87), Hess *et al.* (2002) e Pestana e Velosa (2004).

6.2 Classes de Panjer

As funções massa de probabilidade das variáveis aleatórias de contagem mais usuais – binomiais, binomiais negativas e de Poisson – verificam a equação

$$p_{n+1} = \left(a + \frac{b}{n+1} \right) p_n ; n = 0, 1, 2, \dots ; a, b \in \mathbb{R}. \quad (6.1)$$

Esta expressão recursiva foi originalmente usada por Katz (1965) para definir uma família ampla de distribuições, que Johnson, Kotz & Kemp (2005) consideram ter entre as variáveis aleatórias discretas um papel análogo aos que têm as curvas de Pearson no caso contínuo. Esta fórmula ganhou particular importância quando Panjer (1981) descobriu que facilitava o cálculo da função de distribuição de indenizações agregadas no processo de risco, em teoria de seguros.

Na sequência do trabalho seminal de Panjer, Sundt & Jewell (1981) provaram que binomial, Poisson e binomial negativa são os únicos modelos de contagem (não degenerados) que verificam aquela expressão recursiva.

A igualdade (6.1) pode ser escrita na forma

$$(n+1)p_{n+1} = anp_n + (a+b)p_n. \quad (6.2)$$

Seja $P(s) = \sum_{n=0}^{\infty} s^n p_n$ a função geradora de probabilidades da variável aleatória $X_{a,b}$ calculada no ponto s . Multiplicando ambos os membros de (6.2) por s^{n+1} e somando para $n = 0, 1, 2, \dots$, obtém-se:

$$\begin{aligned} \sum_{n=0}^{\infty} (n+1)p_{n+1}s^{n+1} &= a \sum_{n=0}^{\infty} np_n s^{n+1} + (a+b) \sum_{n=0}^{\infty} p_n s^{n+1} \\ \Leftrightarrow \sum_{n=0}^{\infty} (n+1)p_{n+1}s^{n+1} &= as^2 \sum_{n=0}^{\infty} np_n s^{n-1} p_n + (a+b)s \sum_{n=0}^{\infty} p_n s^n \\ \Leftrightarrow \sum_{n=0}^{\infty} (n+1)p_{n+1}s^{n+1} &= as^2 P'(s) + (a+b)sP(s) \\ \Leftrightarrow \sum_{n=1}^{\infty} np_n s^n &= as^2 P'(s) + (a+b)sP(s) \\ \Leftrightarrow sP'(s) &= as^2 P'(s) + (a+b)sP(s) \\ \Leftrightarrow P'(s) &= asP'(s) + (a+b)P(s) \\ \Leftrightarrow P'(s)(1-as) &= (a+b)P(s) \end{aligned}$$

donde se conclui

$$\frac{P'(s)}{P(s)} = \frac{a+b}{1-as}. \quad (6.3)$$

Teorema 6.1. :

A equação diferencial (6.3) admite as seguintes soluções absolutamente monótonas com $P(1)=1$:

1. $X_{0,0} = 0$ se $a = b = 0$;
2. $X_{0,b} \curvearrowright \text{Poisson}(b)$ se $a=0$ e $b>0$;
3. $X_{a,b} \curvearrowright \text{binomial negativa} \left(\frac{a+b}{a}, 1-a\right)$ se $a \in (0, 1)$ e $a+b > 0$;
4. $X_{a,b} \curvearrowright \text{binomial} \left(-1 - \frac{b}{a}, \frac{a}{a-1}\right)$ se $a < 0$ e $\frac{-b}{a} \in \mathbb{N}^+$.

Demonstração. :

1. se $a = b = 0$ tem-se

$$\frac{P'(s)}{P(s)} = 0 \Leftrightarrow \ln[P(s)] = k \Leftrightarrow P(s) = e^k.$$

Atendendo à condição $P(1) = 1$ vem que $e^k = 1 \Leftrightarrow k = 0$ e, por conseguinte, tem-se $P(s) = e^0 = 1$, que representa a função geradora de probabilidades de uma variável degenerada na origem. Pelo teorema da unicidade tem-se $X_{0,0} = 0$.

2. Se $a = 0$ e $b > 0$ tem-se:

$$\frac{P'(s)}{P(s)} = b \Leftrightarrow \ln[P(s)] = bs + k \Leftrightarrow P(s) = e^{bs+k}.$$

Atendendo à condição $P(1) = 1$ vem que $e^{b+k} = 1 \Leftrightarrow b+k=0 \Leftrightarrow k = -b$ e por conseguinte tem-se $P(s) = e^{b(s-1)}$, que representa a função geradora de probabilidades de uma variável aleatória com distribuição Poisson de parâmetro b . Pelo teorema da unicidade tem-se $X_{0,b} \curvearrowright \text{Poisson}(b)$.

3. Se $a \in (0, 1)$ e $a+b > 0$ tem-se:

$$\begin{aligned} \frac{P'(s)}{P(s)} = \frac{a+b}{1-as} &\Leftrightarrow \ln[P(s)] = -\ln(1-as) - \frac{b}{a} \ln(1-as) + c \\ &\Leftrightarrow P(s) = e^{\ln\left(\frac{1}{1-as}\right)} e^{\ln(1-as)\frac{-b}{a}} e^c \\ &\Leftrightarrow P(s) = \frac{e^c}{1-as} (1-as)^{\frac{-b}{a}}. \end{aligned}$$

A condição $P(1) = 1 \Rightarrow \frac{e^c}{1-a} (1-a)^{\frac{-b}{a}} = 1 \Rightarrow e^c = (1-a)^{1+\frac{b}{a}}$ e, consequentemente,

$$P(s) = \left(\frac{1-a}{1-as}\right)^{\frac{a+b}{a}}, \quad (6.4)$$

que representa a função geradora de probabilidades de uma variável aleatória com distribuição binomial negativa $(\frac{a+b}{a}, 1-a)$. Pelo teorema da unicidade tem-se $X_{a,b} \curvearrowright$ binomial negativa $(\frac{a+b}{a}, 1-a)$.

4. Se $a < 0$ e $\frac{-b}{a} \in \mathbb{N}^+$, e reescrevendo a função geradora de probabilidades obtida em (6.4), obtém-se

$$P(s) = \left(1 - \frac{a}{a-1} + \frac{a}{a-1}s\right)^{-(1+\frac{b}{a})},$$

que representa a função geradora de probabilidades de uma variável aleatória com distribuição binomial $(-1 - \frac{b}{a}, \frac{a}{a-1})$. Pelo teorema da unicidade tem-se $X_{a,b} \curvearrowright$ binomial $(-1 - \frac{b}{a}, \frac{a}{a-1})$.

□

Os resultados de Panjer que – decerto não por acaso, apanham os modelos de contagem mais tradicionais – foram imediatamente generalizados por Sundt & Jewell (1981) e mostraram que a variável logarítmica verificava a expressão recursiva

$$p_{n+1} = \left(a + \frac{b}{n+1}\right) p_n ; n = 1, 2, 3, \dots \quad (6.5)$$

onde se admite que $p_0 = 0$ e se exige que a recorrência é válida apenas para $n \geq 1$. Sundt & Jewell (1981) mostraram que a variável aleatória logarítmica gozava como subordinadora de somas aleatórias, de características tão boas como as clássicas distribuições de Panjer.

Para obter uma equação funcional para a função geradora de probabilidades há que multiplicar ambos os membros da igualdade (6.2) por s^{n+1} e somar em n obtendo-se

$$\begin{aligned} & \sum_{n=1}^{\infty} (n+1)p_{n+1}s^{n+1} = a \sum_{n=1}^{\infty} ns^{n+1}p_n + (a+b) \sum_{n=1}^{\infty} s^{n+1}p_n \\ \Leftrightarrow & \sum_{n=0}^{\infty} (n+1)p_{n+1}s^{n+1} - p_1s = a \left(\sum_{n=0}^{\infty} ns^{n+1}p_n - 0 \right) + \\ & + (a+b) \left(\sum_{n=0}^{\infty} s^{n+1}p_n - p_0s \right) \\ \Leftrightarrow & \underbrace{\sum_{n=1}^{\infty} ns^{n-1}p_n}_{=P'(s)} - p_1s = as^2 \underbrace{\sum_{n=1}^{\infty} np_n s^{n-1}}_{=P'(s)} + (a+b)s \sum_{n=0}^{\infty} s^n p_n - \underbrace{(a+b)p_0s}_{=0} \\ \Leftrightarrow & sP'(s) = as^2P'(s) + (a+b)sP(s) + p_1s \\ \Leftrightarrow & P'(s) = asP'(s) + (a+b)P(s) + p_1s \\ \Leftrightarrow & p_1 = P'(s)(1-as) - (a+b)P(s). \end{aligned}$$

Esta equação funcional tem um conjunto mais vasto de soluções, tais como a binomial negativa generalizada de Engen (1974) de que a $N \curvearrowright$ Logarítmica(α) é um caso limite, quando

$\alpha \in (0, 1)$ e $a + b = 0$. Mais precisamente, verificam a expressão recursiva (6.5) as variáveis aleatórias cuja função geradora de probabilidades $\tilde{P}(s)$ é da forma

$$\tilde{P}(s) = \tau + (1 - \tau)P(s)$$

onde $\tau \in [\frac{P(0)}{P(0)-1}, 1)$, sendo $P(s)$ uma das seguintes funções geradoras de probabilidades:

1. $P(s) = \frac{1-(1-as)^{-\left(1+\frac{b}{a}\right)}}{1-(1-a)^{-\left(1+\frac{b}{a}\right)}}$, função geradora de probabilidades de:
 - a) binomial negativa truncada em zero se $a \in (0, 1)$ e $a + b > 0$;
 - b) binomial negativa generalizada de Engen truncada em zero se $a \in (0, 1]$ e $b \in (-2a, -a)$;
 - c) logarítmica(a) se $-b = a \in (0, 1)$.
2. $P(s) = \frac{e^{-b}}{1-e^{-b}} (e^{bs} - 1)$, Poisson truncada em zero, se $a = 0$ e $a + b \rightarrow 0, b > 0$;
3. $P(s) = \frac{(1-as)^{-\left(1+\frac{b}{a}\right)} - 1}{(1-a)^{-\left(1+\frac{b}{a}\right)} - 1}$, binomial truncada em zero, se $a < 0$ e $\frac{-b}{a} \in \mathbb{N}$.

Assim, para além da Poisson, binomial negativa e binomial, verificam a expressão recursiva (6.5) os referidos modelos truncados na origem, a logarítmica e a binomial negativa generalizada de Engen truncada em zero.

6.3 O modelo de Poisson

Santos, Pestana e Martins (2012) modelaram a sensibilidade e a especificidade para testes quantitativos discretos no caso de $\mathbf{D} \sim \text{Poisson}(\lambda)$ quer considerando que os testes não estão sujeitos a erros de classificação (metodologia de Dorfman) quer considerando que a diluição e conseqüente rarefação alteram as características dos testes conjuntos.

Vamos supor que o número de bactérias numa amostra aleatória de um mililitro de iogurte é bem modelado, caso o iogurte esteja contaminado, por uma variável aleatória Y_i com distribuição de Poisson de parâmetro λ , isto é, $Y_i \sim \text{Poisson}(\lambda)$, cuja massa de probabilidade é dada por

$$Y = \begin{cases} y \\ p_y = e^{-\lambda} \frac{\lambda^y}{y!} \end{cases}, \quad y = 0, 1, \dots \quad (6.6)$$

Considerando que o teste não comete erros (identifica sempre a bactéria caso o mililitro analisado contenha alguma bactéria), a sensibilidade do teste será dada pela probabilidade de haver pelo menos uma bactéria num iogurte contaminado, isto é, por $\varphi_s = \mathbb{P}(Y_i > 0) = 1 - e^{-\lambda}$; sendo a especificidade igual a um (nos iogurtes não contaminados não existe a bactéria em estudo). Assim, atendendo a que a probabilidade de um iogurte estar contaminado é igual

a p , a quantidade de bactérias Y^* por mililitro de iogurte é modelada por uma distribuição de Poisson inflacionada em zero com massa de probabilidade dada por

$$f_{Y^*}(x) \begin{cases} (1-p) + pe^{-\lambda}; & x = 0 \\ p \frac{e^{-\lambda} \lambda^x}{x!} & ; x \in \mathbb{N} \end{cases} .$$

Caso se testem n iogurtes simultaneamente, começamos por misturar as n amostras para posteriormente retirar um mililitro da mistura para o teste. Nesta amostra conjunta, considerando que o processo de mistura é perfeito e que as variáveis são independentes, tem-se que o número de iogurtes infetados na amostra é descrito por uma binomial com parâmetros n e p . Por outro lado, caso existam $\gamma \leq n$ iogurtes contaminados nos n que compõem o grupo, o número de bactérias numa amostra de n mililitros de iogurte, é descrito por uma variável aleatória $B_n|_{I=\gamma}$ com distribuição Poisson $(\gamma\lambda)$ e, por conseguinte, o número de bactérias numa amostra de 1 mililitro de iogurte será descrito por um modelo de Poisson filtrado com massa de probabilidade dada por

$$\begin{aligned} \mathbb{P}(B_1 = k|I = \gamma) &= \sum_{j=0}^{\infty} \mathbb{P}(B_1 = k|B_n = k+j) \mathbb{P}(B_n|_{I=\gamma} = k+j) \\ &= \sum_{j=0}^{\infty} \binom{k+j}{k} \left(\frac{1}{n}\right)^k \left(1 - \frac{1}{n}\right)^j \frac{(\gamma\lambda)^{k+j}}{(k+j)!} e^{-\gamma\lambda} \\ &= e^{-\gamma\lambda} \frac{(\frac{\gamma\lambda}{n})^k}{k!} \underbrace{\sum_{j=0}^{\infty} \frac{1}{j!} \left[\gamma\lambda \left(1 - \frac{1}{n}\right)\right]^j}_{e^{\gamma\lambda(1-\frac{1}{n})}} \\ &= \frac{(\frac{\gamma\lambda}{n})^k}{k!} e^{-\frac{\gamma\lambda}{n}}, \end{aligned}$$

ou seja, o número de bactérias numa amostra de um mililitro de iogurte proveniente de uma mistura de n iogurtes dos quais γ estão contaminados é modelado por $B_1|_{I=\gamma} \sim \text{Poisson}(\frac{\gamma\lambda}{n})$. Consequentemente, a quantidade de bactérias Y_n^* por mililitro de iogurte numa amostra conjunta de n iogurtes será dada por

$$f_{Y_n^*}(x) = \begin{cases} \sum_{i=0}^n I^{[i,n]} e^{-\frac{i\lambda}{n}} & ; x = 0 \\ \sum_{i=1}^n I^{[i,n]} \frac{e^{-\frac{i\lambda}{n}} (\frac{i\lambda}{n})^x}{x!} & ; x \in \mathbb{N} \end{cases} ,$$

que no caso particular de $n = 1$ será igual à função massa de probabilidade Y^* . Denotando por C o conjunto dos iogurtes contaminados e continuando a considerar a ausência de erros no teste temos que $\varphi_{e_n} = 1$ (se os iogurtes não estiverem contaminados o teste nunca será

positivo) e

$$\begin{aligned}\varphi_{s_n} &= \mathbb{P}(Y_n^* \geq 1 | Y_1 \in C) \times \mathbb{P}(Y_1 \geq 1) \\ &= \sum_{i=0}^{n-1} I^{[i, n-1]} \left(1 - e^{-\frac{(i+1)\lambda}{n}}\right) (1 - e^{-\lambda}).\end{aligned}$$

A fórmula anterior é consequência do facto de a identificação correta de um elemento infetado na metodologia de Dorfman obrigar a que o grupo a que esse elemento pertence tenha um resultado positivo (o que ocorre com probabilidade $\varphi_s^{[n]}$) e o teste individual terá que ter também um resultado positivo (com probabilidade φ_s).

Considerando que existe um erro associado ao próprio teste, com probabilidade igual a $1 - \varphi_{e_T}$ de falso positivo e uma probabilidade igual a $1 - \varphi_{s_T}$ de falso negativo obtém-se

$$\begin{aligned}\varphi_{e_n} &= \varphi_{e_T} \mathbb{P}(Y_n^* = 0 | Y_1 \notin C) + (1 - \varphi_{s_T}) \mathbb{P}(Y_n^* > 0 | Y_1 \notin C) + \\ &+ \varphi_{e_T} [(1 - \varphi_{e_T}) \mathbb{P}(Y_n^* = 0 | Y_1 \notin C) + \varphi_{s_T} \mathbb{P}(Y_n^* > 0 | Y_1 \notin C)] \\ &= \varphi_{e_T} \varphi_0 + (1 - \varphi_{s_T}) (1 - \varphi_0) + \varphi_{e_T} [(1 - \varphi_{e_T}) \varphi_0 + \varphi_{s_T} (1 - \varphi_0)] \\ &= 1 - (1 - \varphi_{e_T}) \varphi_{s_T} - (1 - \varphi_{e_T}) (1 - \varphi_{e_T} - \varphi_{s_T}) \varphi_0,\end{aligned}$$

com $\varphi_0 = \sum_{i=0}^{n-1} I^{[i, n-1]} \left(e^{-i\frac{\lambda}{n}}\right)$, fórmula semelhante à previamente deduzida com q^{n-1} no lugar de φ_0 (ver fórmula (3.3) na página 34).

Em relação à sensibilidade, teremos que obter quer o teste conjunto quer o individual positivos e, por conseguinte,

$$\begin{aligned}\varphi_{s_n} &= [(1 - \varphi_{e_T}) \mathbb{P}(Y_n^* = 0 | Y_1 \in C) + \varphi_{s_T} \mathbb{P}(Y_n^* > 0 | Y_1 \in C)] \times \\ &\times [\varphi_{s_T} \mathbb{P}(Y_1 \geq 1) + (1 - \varphi_{e_T}) \mathbb{P}(Y_1 = 0)] \\ &= [(1 - \varphi_{e_T}) \varphi_1 + \varphi_{s_T} (1 - \varphi_1)] \times \\ &\times [\varphi_{s_T} (1 - e^{-\lambda}) + (1 - \varphi_{e_T}) (e^{-\lambda})] \\ &= \varphi_{s_T}^2 + \varphi_1 e^{-\lambda} (1 - \varphi_{e_T} - \varphi_{s_T})^2 + \\ &+ \varphi_{s_T} (\varphi_1 + e^{-\lambda}) (1 - \varphi_{e_T} - \varphi_{s_T}),\end{aligned}$$

onde $\varphi_1 = \sum_{i=0}^{n-1} I^{[i, n-1]} \left(e^{-(i+1)\frac{\lambda}{n}}\right)$.

Assim, além da contabilização do erro associado ao teste teremos igualmente a contabilização do erro de amostragem (dado pela probabilidade de numa amostra proveniente de iogurtes contaminados não haver qualquer bactéria). A utilização de testes compostos tem a vantagem de poupar no número de testes utilizados mas pode dar origem a grandes perdas de sensibilidade.

Vamos supor que existem em média 10 bactérias por mililitro de iogurte contaminado e, por conseguinte, $\lambda = 10$. Considerando uma taxa de prevalência de 0.01, a dimensão ótima

do teste composto é de $n^* = 11$. A sensibilidade do teste simples é dada por

$$\varphi_s = 1 - e^{-\lambda}.$$

Considerando que γ dos n elementos do grupo estão infetados a sensibilidade do teste composto é

$$\varphi_s^{[\gamma, n]} = 1 - e^{-\frac{\gamma}{n}\lambda}.$$

Recorrendo à fórmula (3.1) da página 32 tem-se que

$$k_\gamma^{[n]} = 1 - \frac{1 - e^{-\frac{\gamma}{n}\lambda}}{1 - e^{-\lambda}}, \quad (6.7)$$

podendo a diferença $\varphi_s - \varphi_s^{[n]}$ ser aproximada (ver igualdade (3.2) da página 32) por

$$\frac{11 \times (0.01) (0.99)^{10} \left(1 - \frac{1 - e^{-\frac{10}{11}\lambda}}{1 - e^{-\lambda}}\right) + \binom{11}{2} (0.01)^2 (0.99)^9 \left(1 - \frac{1 - e^{-\frac{20}{11}\lambda}}{1 - e^{-\lambda}}\right)}{1 - (0.99)^{11}} \varphi_s$$

donde se obtém

$$\varphi_s - \varphi_s^{[n]} \approx 0.39\varphi_s.$$

Assim, a redução do número de testes traduz-se numa perda de sensibilidade de aproximadamente 40% provocada pelo efeito de diluição.

Fazendo $\gamma = n$ na igualdade (6.7) obtém-se $k_n^{[n]} = 1 \Rightarrow \varphi_s^{[n, n]} = \varphi_s$. Assim, no modelo de Poisson a sensibilidade do teste composto quando todos os elementos estão infetados coincide com a sensibilidade do teste simples. No capítulo 7 veremos que esta característica não é extensível ao modelo Gaussiano nem ao modelo exponencial.

6.4 Mistura Poisson-exponencial – o modelo geométrico

Quando se pretende modelar o número de ocorrências num processo cujo comportamento médio é estável a Poisson é um modelo adequado. Contudo caso se pretenda flexibilizar o modelo de Poisson, admitindo a existência de variabilidade individual, o modelo geométrico e mais geralmente o binomial negativo são duas alternativas interessantes a ter em conta. Assumindo que o número de bactérias numa amostra aleatória de um mililitro de iogurte é bem modelado, caso o iogurte esteja contaminado, por uma variável aleatória Y_i com distribuição de Poisson de parâmetro λ , $Y \sim \text{Poisson}(\lambda)$, com massa de probabilidade dada em (6.6) e para o qual se tem $\mathbb{E}(Y) = \text{var}(Y) = \lambda$. No entanto, por estarmos a modelar uma população que consideramos mais variável que a Poisson, vamos admitir que λ é um valor observado de uma variável aleatória Λ com distribuição exponencial de valor médio β cuja densidade é dada por

$$f_\Lambda(\beta) = \begin{cases} \frac{1}{\beta} e^{-\frac{\lambda}{\beta}} & \lambda > 0 \\ 0 & \lambda \leq 0 \end{cases} ; \quad \beta > 0.$$

A massa de probabilidade do modelo mistura é dada por

$$\begin{aligned}
 \mathbb{P}(Y = y) &= \int_0^{+\infty} e^{-\lambda} \frac{\lambda^y}{y!} \frac{1}{\beta} e^{-\frac{\lambda}{\beta}} d\lambda \\
 &= \frac{1}{\beta y!} \int_0^{+\infty} \lambda^y e^{-\lambda(1+\frac{1}{\beta})} d\lambda \\
 &= \frac{1}{\beta y!} \left(\frac{\beta}{\beta+1}\right)^{y+1} \underbrace{\int_0^{+\infty} \left(\frac{\beta+1}{\beta}\right)^{y+1} \lambda^y e^{-\lambda(1+\frac{1}{\beta})} d\lambda}_{=\Gamma(y+1)} \\
 &= \frac{1}{\beta+1} \left(\frac{\beta}{\beta+1}\right)^y I_{\{0,1,\dots\}}(y); \quad \beta > 0,
 \end{aligned}$$

ou seja $Y \sim \text{geométrica}\left(\frac{1}{\beta+1}\right)$. Assim, a quantidade de bactérias Y^* por mililitro de iogurte é modelada por uma geométrica inflacionada em zero com massa de probabilidade

$$f_{Y^*}(x) \begin{cases} (1-p) + p\left(\frac{1}{\beta+1}\right) & ; \quad x = 0 \\ p\left(\frac{1}{\beta+1}\right) \left(\frac{\beta}{\beta+1}\right)^x & ; \quad x \in \mathbb{N} \end{cases} .$$

Caso existam $\gamma \leq n$ iogurtes contaminados nos n que compõe o grupo, o número de bactérias numa amostra de n mililitros de iogurte é descrito por uma variável aleatória $B_n|_{I=\gamma}$ com distribuição binomial negativa $\left(\gamma, \frac{1}{\beta+1}\right)$ e, por conseguinte, o número de bactérias numa amostra de um mililitro de iogurte será descrito por uma binomial negativa filtrada. Assim, aplicando um filtro binomial, tem-se

$$\begin{aligned}
 \mathbb{P}(B_1 = k|I = \gamma) &= \sum_{j=0}^{+\infty} \mathbb{P}(B_1 = k|B_n = k+j) \times \mathbb{P}(B_n|_{I=\gamma} = k+j) \\
 &= \sum_{j=0}^{+\infty} \frac{(k+j)!}{k!j!} \left(\frac{1}{n}\right)^k \left(1-\frac{1}{n}\right)^j \binom{\gamma+k+j-1}{k+j} \left(\frac{1}{\beta+1}\right)^\gamma \left(\frac{\beta}{\beta+1}\right)^{k+j} \\
 &= \frac{\beta^k}{k!n^k (\beta+1)^{k+\gamma}} \sum_{j=0}^{+\infty} \frac{(\gamma+k+j-1)!}{j! (\gamma-1)!} \left(\frac{\beta}{\beta+1}\right)^j \left(\frac{n-1}{n}\right)^j \\
 &= \frac{(\gamma+k-1)! \beta^k}{k!n^k (\gamma-1)! (\beta+1)^{\gamma+k}} \sum_{j=0}^{+\infty} \frac{(\gamma+k+j-1)!}{j! (\gamma+k-1)!} \left[\frac{(n-1)\beta}{n(\beta+1)}\right]^j \\
 &= \frac{(\gamma+k-1)! \beta^k}{k!n^k (\gamma-1)! (\beta+1)^{\gamma+k}} \underbrace{\sum_{j=0}^{+\infty} \binom{\gamma+k+j-1}{j} \left[\frac{(n-1)\beta}{n(\beta+1)}\right]^j}_{= \left(\frac{n+\beta}{n(\beta+1)}\right)^{-\gamma-k}} \\
 &= \binom{\gamma+k-1}{k} \left(\frac{\beta}{n(\beta+1)}\right)^k \left(\frac{1}{\beta+1}\right)^\gamma \left(\frac{n+\beta}{n(\beta+1)}\right)^{-\gamma-k} \\
 &= \binom{\gamma+k-1}{k} \left(\frac{\beta}{n+\beta}\right)^k \left(\frac{n}{n+\beta}\right)^\gamma,
 \end{aligned}$$

ou seja, o número de bactérias numa amostra de um mililitro de iogurte proveniente de uma mistura de n iogurtes dos quais γ estão contaminados é modelado por $B_1|_{I=\gamma} \curvearrowright$ binomial negativa $\left(\gamma, \frac{n}{n+\beta}\right)$. Consequentemente, a quantidade de bactérias Y_n^* por mililitro de iogurte numa amostra conjunta de n iogurtes será dada por

$$f_{Y_n^*}(x) = \begin{cases} \sum_{i=0}^n I^{[i,n]} \left(\frac{n}{n+\beta}\right)^i & ; x = 0 \\ \sum_{i=1}^n I^{[i,n]} \binom{i+x-1}{x} \left(\frac{n}{n+\beta}\right)^i \left(\frac{\beta}{n+\beta}\right)^x & ; x \in \mathbb{N} \end{cases},$$

que coincidirá com Y^* no caso $n = 1$. Considerando a ausência de erros no teste temos que $\varphi_{e_n} = 1$ (se os iogurtes não estiverem contaminados o teste nunca será positivo) e

$$\begin{aligned} \varphi_{s_n} &= \mathbb{P}(Y_n^* \geq 1 | Y_1 \in \mathcal{C}) \times \mathbb{P}(Y_1 \geq 1) \\ &= \sum_{i=0}^{n-1} I^{[i,n-1]} \left[1 - \left(\frac{n}{n+\beta}\right)^{i+1}\right] \left(\frac{\beta}{\beta+1}\right). \end{aligned}$$

que corresponde à fórmula (3.3) com $\varphi_s^{[i+1,n]} = \left[1 - \left(\frac{n}{n+\beta}\right)^{i+1}\right] \mathbf{e}$ $\varphi_s = \frac{\beta}{\beta+1}$.

Considerando que existe um erro associado ao próprio teste obtém-se

$$\varphi_{e_n} = 1 - (1 - \varphi_{e_T}) \varphi_{s_T} - (1 - \varphi_{e_T}) (1 - \varphi_{e_T} - \varphi_{s_T}) \varphi_0,$$

com

$$\varphi_0 = \sum_{i=0}^{n-1} I^{[i,n-1]} \left(\frac{n}{n+\beta}\right)^i.$$

Em relação à sensibilidade, teremos que obter quer o teste conjunto quer o individual positivos e por conseguinte

$$\begin{aligned} \varphi_{S_n} &= [(1 - \varphi_{e_T}) \mathbb{P}(Y_n^* = 0 | Y_1 \in \mathcal{C}) + \varphi_{s_T} \mathbb{P}(Y_n^* > 0 | Y_1 \in \mathcal{C})] \times \\ &\times [\varphi_{s_T} \mathbb{P}(Y_1 \geq 1) + (1 - \varphi_{e_T}) \mathbb{P}(Y_1 = 0)] \\ &= [(1 - \varphi_{e_T}) \varphi_1 + \varphi_{s_T} (1 - \varphi_1)] \times \\ &\times \left[\varphi_{s_T} \left[1 - \left(\frac{1}{\beta+1}\right)\right] + (1 - \varphi_{e_T}) \left(\frac{1}{\beta+1}\right) \right] \\ &= \varphi_{s_T}^2 + \varphi_1 \left(\frac{1}{\beta+1}\right) (1 - \varphi_{e_T} - \varphi_{s_T})^2 + \\ &+ \varphi_{s_T} \left[\varphi_1 + \left(\frac{1}{\beta+1}\right) \right] (1 - \varphi_{e_T} - \varphi_{s_T}), \end{aligned}$$

onde $\varphi_1 = \sum_{i=0}^{n-1} I^{[i,n-1]} \left(\frac{n}{n+\beta}\right)^{i+1}$.

Vamos admitir que a quantidade de bactérias por mililitro de iogurte varia com a temperatura e que esta não é homogénea dentro da unidade fabril. Admitindo a existência de

variabilidade individual, vamos recorrer ao modelo geométrico como alternativa ao modelo de Poisson. Assumindo uma taxa de prevalência $p = 0.2$ e conseqüentemente grupos de dimensão três ($n^* = 3$), o número médio de bactérias por mililitro de iogurte utilizando o modelo de Poisson é $\lambda = np = 0.6$. Para determinar o valor de β (valor médio do modelo geométrico) e tendo em conta que o modelo geométrico é uma variável aleatória Poisson composta $S_N = \sum_{k=1}^N X_k$ com parcelas i.i.d. $X_k \sim \text{exponencial}(\lambda)$, considere-se o seguinte resultado (informação mais detalhada pode ser consultada em Allan Gut (1995)).

Teorema 6.2. (Valor médio de uma soma aleatória)

Considere-se $S_N = X_1 + X_2 + \dots + X_N$ uma soma aleatória de variáveis aleatórias com parcelas $\{X_k\}_{k \geq 1}$ variáveis i.i.d., sendo o seu número, N , independente de cada uma delas. O valor médio de S_N é dado por

$$\mathbb{E}[S_N] = \mathbb{E}\left[\sum_{k=1}^N X_k\right] = \mathbb{E}\left[\mathbb{E}\left(\sum_{k=1}^N X_k \mid N\right)\right] = \mathbb{E}[N\mathbb{E}(X)] = \mathbb{E}(N)\mathbb{E}(X).$$

Tendo em conta que $\mathbb{E}(S_N) = \beta + 1$, $\mathbb{E}(N) = 0.6$ e que $\mathbb{E}(X) = \beta$, obtém-se $\beta = \frac{1}{1-\lambda} = 2.5$. Assumindo, por exemplo, $\varphi_s = 0.99$ e $\varphi_e = 0.95$ obtém-se um custo relativo 0.8213 e especificidade $\varphi_{e_n} \approx 0.9893$. Refira-se que utilizando o modelo de Poisson obteríamos $\varphi_{e_n} = 0.9942$ e, por conseguinte, a variabilidade individual provoca inevitavelmente uma redução na especificidade do teste composto (φ_{e_n}).

Pressupondo a existência de 3 zonas dentro da unidade fabril onde a temperatura é significativamente diferente, e considerando a existência de 3 subpopulações com pesos iguais a 0.5, 0.3 e 0.2 e taxas de prevalência iguais a 0.25, 0.2 e 0.075 respetivamente obtém-se $CR = 0.8057$ e $\varphi_{e_n}^{[S]} = 0.9883$. Constata-se assim mais uma vez que a inclusão das subpopulações conduziu a um aumento da eficiência do teste originada pela diminuição do custo relativo.

6.5 Mistura Poisson-gama – o modelo binomial negativo

Na situação em que no modelo hierárquico se adota uma gama para modelar a variabilidade do parâmetro, obtém-se, como atrás se viu, um modelo binomial negativo para as contagens. Com o intuito de generalizar os resultados obtidos vamos considerar que $Y \sim \text{binomial negativa}(\alpha, p_1)$ com massa de probabilidade dada por

$$f_Y(x) = \binom{\alpha + x - 1}{x} (p_1)^\alpha (1 - p_1)^x; \quad x = 0, 1, \dots$$

e, por conseguinte, o número de bactérias Y^* por mililitro de iogurte é modelado por uma binomial negativa inflacionada em zero com massa de probabilidade dada por

$$f_{Y^*}(x) = \begin{cases} (1 - p) + p(p_1)^\alpha & ; \quad x = 0 \\ p \binom{\alpha + x - 1}{x} (p_1)^\alpha (1 - p_1)^x & ; \quad x \in \mathbb{N} \end{cases}.$$

Seguindo o mesmo raciocínio que foi aplicado anteriormente é fácil concluir que caso existam $\gamma \leq n$ iogurtes contaminados nos n que compõe o grupo, o número de bactérias numa amostra de n mililitros de iogurte é modelado por uma variável aleatória $B_n|_{I=\gamma}$ com distribuição binomial negativa $(\gamma\alpha, p_1)$ e, por conseguinte, o número de bactérias num mililitro de iogurte condicional a $I = \gamma$ tem massa de probabilidade dada por

$$\begin{aligned}
 \mathbb{P}(B_1 = k|I = \gamma) &= \sum_{j=0}^{+\infty} \mathbb{P}(B_1 = k|B_n = k+j) \times \mathbb{P}(B_n|_{I=\gamma} = k+j) \\
 &= \sum_{j=0}^{+\infty} \binom{k+j}{k} \left(\frac{1}{n}\right)^k \left(1 - \frac{1}{n}\right)^j \binom{\gamma\alpha + k + j - 1}{k+j} (p_1)^{\gamma\alpha} (1-p_1)^{k+j} \\
 &= \sum_{j=0}^{+\infty} \frac{(k+j)!}{k!j!} \left(\frac{1}{n}\right)^k \left(1 - \frac{1}{n}\right)^j \frac{(\gamma\alpha + k + j - 1)!}{(k+j)!(\gamma\alpha - 1)!} (p_1)^{\gamma\alpha} (1-p_1)^{k+j} \\
 &= \frac{\left[\frac{1}{n}(1-p_1)\right]^k (\gamma\alpha + k - 1)!}{k! (\gamma\alpha - 1)!} (p_1)^{\gamma\alpha} \sum_{j=0}^{+\infty} \frac{(\gamma\alpha + k + j - 1)!}{(\gamma\alpha + k - 1)!j!} \times \\
 &\times \left[(1-p_1) \left(1 - \frac{1}{n}\right) \right]^j \\
 &= \binom{\gamma\alpha + k - 1}{k} (p_1)^{\gamma\alpha} \left[(1-p_1) \frac{1}{n} \right]^k \times \\
 &\times \underbrace{\sum_{j=0}^{+\infty} \binom{\gamma\alpha + k + j - 1}{j} \left[(1-p_1) \left(1 - \frac{1}{n}\right) \right]^j}_{\left[p_1 + \frac{1}{n}(1-p_1)\right]^{-\gamma\alpha - k}} \\
 &= \binom{\gamma\alpha + k - 1}{k} \left(\frac{np_1}{np_1 + 1 - p_1}\right)^{\alpha\gamma} \left(\frac{1-p_1}{np_1 + 1 - p_1}\right)^x.
 \end{aligned}$$

Assim, o número de bactérias numa amostra de um mililitro de iogurte proveniente de uma mistura de n iogurtes dos quais γ estão contaminados é modelado por uma variável aleatória $B_1|_{I=\gamma} \sim$ binomial negativa $\left(\gamma\alpha, \frac{np_1}{np_1+1-p_1}\right)$ e, conseqüentemente, a quantidade de bactérias Y_n^* por mililitro numa amostra conjunta de n iogurtes será dada por

$$f_{Y_n^*}(x) = \begin{cases} \sum_{i=0}^n I^{[i,n]} \left(\frac{np_1}{np_1+1-p_1}\right)^{\alpha i} & ; x = 0 \\ \sum_{i=1}^n I^{[i,n]} \binom{\alpha i + x - 1}{x} \left(\frac{np_1}{np_1+1-p_1}\right)^{\alpha i} \left(\frac{1-p_1}{np_1+1-p_1}\right)^x & ; x \in \mathbb{N} \end{cases}$$

que coincidirá com Y^* no caso de $n = 1$. No que respeita à sensibilidade tem-se

$$\begin{aligned}
 \varphi_{s_n} &= \mathbb{P}(Y_n^* \geq 1|Y_1 \in C) \times \mathbb{P}(Y_1 \geq 1) \\
 &= \sum_{i=0}^{n-1} I^{[i,n-1]} \left[1 - \left(\frac{np_1}{np_1 + 1 - p_1}\right)^{(i+1)\alpha} \right] [1 - (p_1)^\alpha].
 \end{aligned}$$

Considerando que um falso positivo ocorre com probabilidade igual a $1 - \varphi_{e_T}$ e que a probabilidade de ocorrência de um falso negativo é de $1 - \varphi_{S_T}$ tem-se

$$\varphi_{e_n} = 1 - (1 - \varphi_{e_T}) \varphi_{S_T} - (1 - \varphi_{e_T}) (1 - \varphi_{e_T} - \varphi_{S_T}) \varphi_0$$

onde

$$\varphi_0 = \sum_{i=0}^{n-1} I^{[i, n-1]} \left(\frac{np_1}{np_1 + 1 - p_1} \right)^{\alpha i},$$

fórmula semelhante à equação (3.3) da página 34 com q^{n-1} no lugar de φ_0 . Em relação à sensibilidade teremos

$$\varphi_{S_n} = \varphi_{S_T}^2 + \varphi_1 (p_1)^\alpha (1 - \varphi_{e_T} - \varphi_{S_T})^2 + \varphi_{S_T} [\varphi_1 + (p_1)^\alpha] (1 - \varphi_{e_T} - \varphi_{S_T}) \quad (6.8)$$

onde $\varphi_1 = \sum_{i=0}^{n-1} I^{[i, n-1]} \left(\frac{np_1}{np_1 + 1 - p_1} \right)^{(i+1)\alpha}$.

Para os modelos logarítmico e binomial será apresentado apenas a massa de probabilidade de $B_1|_{I=\gamma}$. As expressões para a sensibilidade e para a especificidade são determinadas de forma análoga.

6.6 O modelo logarítmico

Considere-se que o número de bactérias numa amostra aleatória de um mililitro de iogurte será, no caso de o iogurte estar contaminado, descrito por uma variável aleatória com distribuição logarítmica de parâmetro θ com massa de probabilidade dada por

$$f_Y(x) = -\frac{1}{\ln(1 - \theta)} \frac{\theta^x}{x}; \quad x \in \mathbb{N}.$$

Dado que a distribuição da soma de variáveis independentes com distribuição logarítmica não é conhecida, não é possível determinar a distribuição de $B_n = \sum_{i=1}^n Y_i^* = \sum_{i=1}^I Y_i$. Contudo, a determinação de B_1 pode ser obtida considerando que $B_1 = \sum_{i=1}^n W_i^* = \sum_{i=1}^I W_i$ onde W_i é um modelo hierárquico, ou seja, $W_i \sim \text{binomial}(T, \frac{1}{n})$ com $T \sim \text{logarítmica}(\theta)$.

Escolhendo agora um mililitro deste liquido vamos obter

$$\begin{aligned}
 \mathbb{P}(W_i = x) &= \sum_{j=x}^{\infty} \binom{j}{x} \left(\frac{1}{n}\right)^x \left(1 - \frac{1}{n}\right)^{j-x} \left(\frac{-1}{\ln(1-\theta)}\right) \frac{\theta^j}{j} \\
 &= \frac{-1}{x \ln(1-\theta)} \sum_{j=x}^{\infty} \frac{(j-1)!}{(x-1)!(j-x)!} \left(\frac{\theta}{n}\right)^x \left[\left(1 - \frac{1}{n}\right)\theta\right]^{j-x} \\
 &= \frac{-1}{x \ln(1-\theta)} \sum_{j=x}^{\infty} \binom{j-1}{x-1} \left(\frac{\theta}{n}\right)^x \left[\left(1 - \frac{1}{n}\right)\theta\right]^{j-x} \\
 &= \frac{-1}{x \ln(1-\theta)} \left(\frac{\theta}{n}\right)^x \underbrace{\sum_{j=x}^{\infty} \binom{j-1}{x-1} \left[\left(1 - \frac{1}{n}\right)\theta\right]^{j-x}}_{\left[1 - \left(1 - \frac{1}{n}\right)\theta\right]^{-x}} \\
 &= \frac{-1}{x \ln(1-\theta)} \left[\frac{\frac{\theta}{n}}{\left[1 - \left(1 - \frac{1}{n}\right)\theta\right]} \right]^x ; \quad x \in \mathbb{N}.
 \end{aligned}$$

No caso $x = 0$ obtém-se

$$\begin{aligned}
 \mathbb{P}(W_i = 0) &= \sum_{j=1}^{\infty} \binom{j}{0} \left(\frac{1}{n}\right)^0 \left(1 - \frac{1}{n}\right)^j \left(\frac{-1}{\ln(1-\theta)}\right) \frac{\theta^j}{j} \\
 &= \frac{-1}{\ln(1-\theta)} \underbrace{\sum_{j=1}^{\infty} \frac{\left(\theta - \frac{\theta}{n}\right)^j}{j}}_{-\ln\left(1 - \theta + \frac{\theta}{n}\right)} \\
 &= \frac{\ln\left(1 - \theta + \frac{\theta}{n}\right)}{\ln(1-\theta)},
 \end{aligned}$$

e, por conseguinte, conclui-se

$$f_{W_i}(x) \begin{cases} \frac{\ln\left(1 - \theta + \frac{\theta}{n}\right)}{\ln(1-\theta)} & ; \quad x = 0 \\ \frac{-1}{x \ln(1-\theta)} \left(\frac{\frac{\theta}{n}}{1 - \theta \left(1 - \frac{1}{n}\right)}\right)^x & ; \quad x \in \mathbb{N} \end{cases} ;$$

sendo que $B_1 = \sum_{i=1}^I W_i$.

6.7 O modelo binomial

Considere-se que, caso o iogurte esteja contaminado, o número de bactérias numa amostra aleatória de um mililitro de iogurte é bem modelado por $Y_i \sim \text{binomial}(\alpha, p_1)$. Caso existam

γ iogurtes contaminados, o número de bactérias numa amostra aleatória de n mililitros (um mililitro de cada iogurte) é bem modelada por $B_n|_{I=\gamma} \sim \text{binomial}(\gamma\alpha, p_1)$. Escolhendo agora um mililitro deste liquido vamos obter

$$\begin{aligned}
 \mathbb{P}(B_1 = k|I = \gamma) &= \sum_{j=k}^{\gamma\alpha} \mathbb{P}(B_n = j|I=\gamma) \mathbb{P}(B_1 = k|B_n = j) \\
 &= \sum_{j=k}^{\gamma\alpha} \binom{\gamma\alpha}{j} (p_1)^j (1-p_1)^{\gamma\alpha-j} \binom{j}{k} \left(\frac{1}{n}\right)^k \left(1-\frac{1}{n}\right)^{j-k} \\
 &= \frac{(\gamma\alpha)!}{k!(\gamma\alpha-k)!} \sum_{j=k}^{\gamma\alpha} \frac{(\gamma\alpha-k)!}{(j-k)!(\gamma\alpha-j)!} (p_1)^j (1-p_1)^{\gamma\alpha-j} \left(\frac{1}{n}\right)^k \left(1-\frac{1}{n}\right)^{j-k} \\
 &= \binom{\gamma\alpha}{k} \left(\frac{1}{n^k}\right) \sum_{x=0}^{\gamma\alpha-k} \binom{\gamma\alpha-k}{x} (p_1)^{x+k} (1-p_1)^{\gamma\alpha-k-x} \left(1-\frac{1}{n}\right)^x \\
 &= \binom{\gamma\alpha}{k} \left(\frac{p_1}{n}\right)^k \left(1-\frac{p_1}{n}\right)^{\gamma\alpha-k} \sum_{x=0}^{\gamma\alpha-k} \binom{\gamma\alpha-k}{x} (p_1)^x (1-p_1)^{\gamma\alpha-k-x} \times \\
 &\times \left(\frac{n-1}{n}\right)^x \left(1-\frac{p_1}{n}\right)^{k-\gamma\alpha} \\
 &= \binom{\gamma\alpha}{k} \left(\frac{p_1}{n}\right)^k \left(1-\frac{p_1}{n}\right)^{\gamma\alpha-k} \sum_{x=0}^{\gamma\alpha-k} \binom{\gamma\alpha-k}{x} \left(p_1 \frac{n-1}{n}\right)^x \times \\
 &\times (1-p_1)^{\gamma\alpha-k-x} \left(\frac{n}{n-p_1}\right)^{\gamma\alpha-k} \\
 &= \binom{\gamma\alpha}{k} \left(\frac{p_1}{n}\right)^k \left(1-\frac{p_1}{n}\right)^{\gamma\alpha-k} \sum_{x=0}^{\gamma\alpha-k} \binom{\gamma\alpha-k}{x} \left(p_1 \frac{n-1}{n} \frac{n}{n-p_1}\right)^x \times \\
 &\times \left(\frac{n(1-p_1)}{n-p_1}\right)^{\gamma\alpha-k-x} \\
 &= \binom{\gamma\alpha}{k} \left(\frac{p_1}{n}\right)^k \left(1-\frac{p_1}{n}\right)^{\gamma\alpha-k}.
 \end{aligned}$$

Logo $B_1|_{I=\gamma} \sim \text{binomial}(\gamma\alpha, \frac{p_1}{n})$

A escolha inicial para modelar a variável aleatória $B_1|_{I=\gamma}$ recaiu na distribuição de Poisson, modelo apropriado para proceder a contagens nos casos em que a aleatoriedade é moderada por uma grande regularidade (estacionaridade e independência dos incrementos) e que se traduz na estabilidade dos valores esperados em cada unidade de tempo. A necessidade de flexibilizar o modelo de Poisson, admitindo variabilidade individual conduziu-nos ao modelo binomial negativo, modelo sobre-disperso, isto é, com variância superior à média. A binomial surge como o modelo onde a variância inferior à média. Refira-se ainda que a binomial negativa pode ser obtida como uma soma aleatória de logarítmicas com subordinadora com distribuição de Poisson. Assim se justifica o recurso à família de distribuições de contagem que verificam a fórmula recursiva de Panjer e suas extensões.

Capítulo 7

Testes quantitativos contínuos

7.1 Introdução

A ideia original de Dorfman (1943) aplicada inicialmente a análises qualitativas conjuntas de sangue, é uma estratégia adequada caso se pretenda determinar a presença (positividade) ou ausência (negatividade) de determinada característica nas unidades amostrais. A extensão da metodologia de Dorfman não é contudo imediata quando a positividade é determinada por uma certa quantidade na análise exceder um determinado patamar, um ponto de corte previamente determinado (ver Sousa (2008)). De facto, ao efetuar um teste conjunto quantitativo, pretendemos determinar se existe algum elemento infetado no grupo e, por conseguinte, estamos interessados em averiguar se o máximo (ou o mínimo) do grupo é superior (ou inferior) a determinado limiar. Uma vez que temos acesso apenas ao valor da média do grupo, a correlação existente entre o máximo e a média terá um papel preponderante na avaliação da qualidade do teste.

Deste modo, neste capítulo começamos por determinar analiticamente a potência do teste simples para o modelo Gaussiano (para populações homocedásticas e heterocedásticas) e para o modelo exponencial. O cálculo da potência do teste conjunto é também determinado para os dois modelos. Para quantificar a qualidade dos testes conjuntos foram efetuadas simulações para os modelos gaussiano, exponencial e Pareto, com o intuito de, para cada um deles, quantificar a quantidade de informação que a média do grupo contém sobre o máximo. A comparação dos resultados é feita com base na comparação dos valores da sensibilidade e da especificidade do teste conjunto.

7.2 Testes de hipóteses para amostras individuais

Num teste individual pretende-se testar se a quantidade de substância de um determinado indivíduo ultrapassa um limiar crítico l . Uma regra para optar pela positividade ou negatividade da amostra é determinada sob H_0 . Suponhamos que a quantidade de substância nos indiví-

duos saudáveis é modelada por uma determinada distribuição contínua $Y \sim D_{\theta_0}$ e que a quantidade de substância nos indivíduos doentes é modelada por $Y \sim D_{\theta_1}$ onde θ_0 e θ_1 são vectores de parâmetros distintos. Assim, para um nível de significância α a hipótese nula é rejeitada (ver Martins, Santos e Sousa 2012) caso a quantidade de substância Y exceda o limiar $l = F_{D_{\theta_0}}^{-1}(1 - \alpha)$ onde $F_{D_{\theta_0}}^{-1}(1 - \alpha)$ representa o quantil de probabilidade $(1 - \alpha)$ do modelo estipulado na hipótese nula. Assim, para testar as hipóteses:

$$H_0 : X_i = 0 \text{ versus } H_1 : X_i = 1 \quad (7.1)$$

a hipótese nula será rejeitada caso $y > F_{D_{\theta_0}}^{-1}(1 - \alpha)$.

7.2.1 Sensibilidade do teste simples para o modelo Gaussiano

Comecemos por considerar que existe homocedasticidade, ou seja, $D_{\theta_0} \sim$ Gaussiana (μ_0, σ) e $D_{\theta_1} \sim$ Gaussiana (μ_1, σ) e assumamos, sem perda de generalidade, que $\mu_1 > \mu_0$. Assim, pretende-se determinar o teste mais potente de nível α para testar

$$H_0 : \mu = \mu_0 \text{ versus } H_1 : \mu = \mu_1, \text{ com } \mu_1 > \mu_0.$$

Se $Y \sim$ Gaussiana (μ, σ) a sua função densidade de probabilidade é

$$f_Y(y|\mu) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}; \quad y \in \mathbb{R}$$

e por conseguinte

$$\begin{aligned} \frac{f_Y(y|\mu_0)}{f_Y(y|\mu_1)} &= \frac{\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y-\mu_0)^2}}{\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y-\mu_1)^2}} \\ &= e^{-\frac{1}{2\sigma^2}[(y-\mu_0)^2 - (y-\mu_1)^2]}. \end{aligned}$$

Assim, pelo lema de Neyman-Pearson, a região de rejeição é dada por

$$\begin{aligned} R &= \left\{ y \in \mathbb{R} : \frac{f_Y(y|\mu_0)}{f_Y(y|\mu_1)} < c \right\} \\ &= \left\{ y \in \mathbb{R} : e^{-\frac{1}{2\sigma^2}[(y-\mu_0)^2 - (y-\mu_1)^2]} < c \right\} \end{aligned}$$

e conseqüentemente

$$-2y(\mu_1 - \mu_0) - (\mu_0^2 - \mu_1^2) < 2\sigma^2 \ln(c),$$

donde se conclui que

$$y > \underbrace{\frac{-2\sigma^2 \ln(c) - (\mu_0^2 - \mu_1^2)}{2(\mu_1 - \mu_0)}}_{=l}.$$

A região de rejeição é então

$$R = \{y \in \mathbb{R} : y > l\};$$

onde

$$l = \mathbb{P}(\text{Rejeitar } H_0 | H_0 \text{ Verdadeira}) = \alpha.$$

Assim tem-se

$$\mathbb{P}(\text{Rejeitar } H_0 | H_0 \text{ Verdadeiro}) = \alpha$$

$$\Leftrightarrow \mathbb{P}(Y > l | \mu = \mu_0) = \alpha$$

$$\Leftrightarrow \mathbb{P}\left(\frac{Y - \mu_0}{\sigma} > \frac{l - \mu_0}{\sigma}\right) = \alpha$$

$$\Leftrightarrow 1 - \Phi\left(\frac{l - \mu_0}{\sigma}\right) = \alpha$$

$$\Leftrightarrow \Phi\left(\frac{l - \mu_0}{\sigma}\right) = 1 - \alpha$$

$$\Leftrightarrow \frac{l - \mu_0}{\sigma} = \Phi^{-1}(1 - \alpha)$$

$$\Leftrightarrow l = \mu_0 + \sigma \Phi^{-1}(1 - \alpha),$$

onde $\Phi^{-1}(1 - \alpha)$ representa o quantil $1 - \alpha$ da Gaussiana padrão.

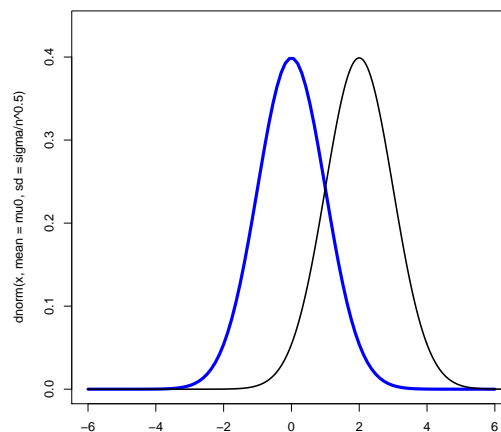


Figura 7.1: Densidade de \bar{Y} condicional a H_0 e H_1

Uma vez determinado o limiar crítico $l = \mu_0 + \sigma_0 \Phi^{-1}(1 - \alpha)$, a hipótese nula é rejeitada caso a quantidade de substância y seja estritamente superior a esse valor. Denotando por β a probabilidade de cometer um erro de segunda espécie, a potência do teste é dada por

$$\begin{aligned}
 1 - \beta &= \mathbb{P}(\text{Rejeitar } H_0 | H_0 \text{ Falso}) \\
 &= \mathbb{P}(Y > \mu_0 + \sigma \Phi^{-1}(1 - \alpha) | \mu = \mu_1) \\
 &= \mathbb{P}\left(Z > \frac{\mu_0 + \sigma \Phi^{-1}(1 - \alpha) - \mu_1}{\sigma}\right)
 \end{aligned}$$

e, por conseguinte,

$$1 - \beta = 1 - \Phi\left[\frac{\mu_0 - \mu_1}{\sigma} + \Phi^{-1}(1 - \alpha)\right]. \quad (7.2)$$

Denotando por $\Delta = \mu_1 - \mu_0$ tem-se

$$1 - \beta = 1 - \Phi\left[\Phi^{-1}(1 - \alpha) - \frac{\Delta}{\sigma}\right]. \quad (7.3)$$

Note-se que a potência do teste é uma função crescente com $\Delta = \mu_1 - \mu_0$, ou seja, quanto maior for a discrepância entre μ_1 e μ_0 maior será a potência do teste.

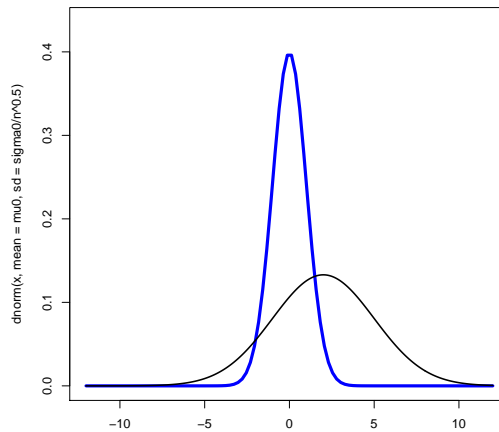


Figura 7.2: Densidade de \bar{Y} condicional a H_0 e H_1 com $\mu_1 > \mu_0$ e $\sigma_1 > \sigma_0$

Considere-se que $Y \sim \text{Gaussiana}(\mu, 1)$ e com base numa observação se pretende testar $H_0 : \mu = 0$ versus $H_1 : \mu = 3$. Facilmente se constata que quanto mais afastado μ_1 estiver de μ_0 menor será a sobreposição das duas curvas e conseqüentemente, maior será a potência do teste (ver Figura 7.1).

Suponhamos agora que $D_{\theta_0} \sim \text{Gaussiana}(\mu_0, \sigma_0)$ e $D_{\theta_1} \sim \text{Gaussiana}(\mu_1, \sigma_1)$. Considere-se que $K = \frac{\sigma_1}{\sigma_0}$ e $\Delta = \mu_1 - \mu_0 > 0$. Para determinar o teste mais potente de nível α para testar (7.1) comecemos por observar a Figura 7.2 onde é visível que a razão de verosimilhanças não é monótona. Contudo, intuitivamente sabemos que a região de rejeição será do tipo

$$R = \{y \in \mathbb{R} : y > l\},$$

uma vez que consideramos $\mu_1 > \mu_0$, e onde

$$l = \mathbb{P}(\text{Rejeitar } H_0 | H_0 \text{ Verdadeira}) = \alpha$$

Assim tem-se

$$\begin{aligned} \mathbb{P}(\text{Rejeitar } H_0 | H_0 \text{ Verdadeira}) &= \alpha \\ \Leftrightarrow \mathbb{P}(Y > l | \mu = \mu_0, \sigma = \sigma_0) &= \alpha \\ \Leftrightarrow \mathbb{P}\left(\frac{Y - \mu_0}{\sigma_0} > \frac{l - \mu_0}{\sigma_0}\right) &= \alpha \\ \Leftrightarrow 1 - \Phi\left(\frac{l - \mu_0}{\sigma_0}\right) &= \alpha \\ \Leftrightarrow \Phi\left(\frac{l - \mu_0}{\sigma_0}\right) &= 1 - \alpha \\ \Leftrightarrow \frac{l - \mu_0}{\sigma_0} &= \Phi^{-1}(1 - \alpha) \\ \Leftrightarrow l &= \mu_0 + \sigma_0 \Phi^{-1}(1 - \alpha) \end{aligned}$$

A potência do teste é dada por

$$\begin{aligned} 1 - \beta &= \mathbb{P}(\text{Rejeitar } H_0 | H_0 \text{ Falso}) \\ &= \mathbb{P}(Y > \mu_0 + \sigma_0 \Phi^{-1}(1 - \alpha) | \mu = \mu_1, \sigma = \sigma_1) \\ &= \mathbb{P}\left(Z > \frac{\mu_0 + \sigma_0 \Phi^{-1}(1 - \alpha) - \mu_1}{\sigma_1}\right), \end{aligned}$$

e, por conseguinte,

$$1 - \beta = 1 - \Phi\left[\frac{\mu_0 - \mu_1}{\sigma_1} + \frac{\Phi^{-1}(1 - \alpha)}{K}\right],$$

ou ainda

$$1 - \beta = 1 - \Phi\left[\frac{\Phi^{-1}(1 - \alpha)}{K} - \frac{\Delta}{\sigma_1}\right]. \quad (7.4)$$

Quanto maiores forem os valores de σ_0 e de σ_1 em relação a $\Delta = \mu_1 - \mu_0$ maior será a redução da potência do teste. Se considerarmos uma situação de homocedasticidade onde, por exemplo, $D_{\theta_0} \sim \text{Gaussiana}(\mu_0 = 2, \sigma_0 = 5)$ e $D_{\theta_1} \sim \text{Gaussiana}(\mu_1 = 5, \sigma_1 = 5)$ temos que $\sigma = \frac{\sigma_1}{\sigma_0} = 1$ e $\Delta = \mu_1 - \mu_0 = 3$. Recorrendo à igualdade (7.2) obtém-se um valor de 0.1469 para a sensibilidade do teste

Considerando agora uma situação em que as variâncias são diferentes, considere-se que $D_{\theta_0} \sim$ Gaussiana ($\mu_0 = 2, \sigma_0 = 1$) e $D_{\theta_1} \sim$ Gaussiana ($\mu_1 = 5, \sigma_1 = 3$). Substituindo na igualdade (7.4) $K = 3, \Delta = 3$ e $\alpha = 0.05$ obtém-se uma potência de aproximadamente 0.3228. Esta redução da potência do teste pode explicar a não utilização do teste conjunto, uma vez que estes só devem ser aplicados caso a potência do teste simples seja elevada (uma vez que a sensibilidade do teste conjunto é inferior à do teste individual).

7.2.2 Sensibilidade do teste simples para o modelo exponencial

Consideremos agora o caso em que $D_{\theta_0} \sim$ exponencial (δ_0) e $D_{\theta_1} \sim$ exponencial (δ_1) e assumamos, sem perda de generalidade, que $\delta_1 > \delta_0$. Assim, pretende-se determinar o teste mais potente de nível α para testar

$$H_0 : \mu = \mu_0 \text{ versus } H_1 : \mu = \mu_1, \text{ com } \mu_1 > \mu_0, \quad (7.5)$$

onde $Y \sim$ exponencial (δ) e portanto com função densidade dada por

$$f_Y(y|\delta) = \frac{1}{\delta} e^{-\frac{y}{\delta}}, \quad y > 0.$$

O valor médio deste modelo é dado por $\mathbb{E}(Y) = \delta$ e por conseguinte testar (7.5) é equivalente a testar

$$H_0 : \delta = \delta_0 \text{ versus } H_1 : \delta = \delta_1, \text{ com } \delta_1 > \delta_0. \quad (7.6)$$

Considere-se o quociente

$$\frac{f_Y(y|\delta_0)}{f_Y(y|\delta_1)} = \frac{\frac{1}{\delta_0} e^{-\frac{y}{\delta_0}}}{\frac{1}{\delta_1} e^{-\frac{y}{\delta_1}}} = \frac{\delta_1}{\delta_0} e^{-y\left(\frac{1}{\delta_0} - \frac{1}{\delta_1}\right)}.$$

O lema de Neyman-Pearson conduz-nos à seguinte região de rejeição

$$\begin{aligned} R &= \left\{ y \in \mathbb{R} : \frac{f_Y(y|\delta_0)}{f_Y(y|\delta_1)} < c \right\} \\ &= \left\{ y \in \mathbb{R} : \frac{\delta_1}{\delta_0} e^{-y\left(\frac{1}{\delta_0} - \frac{1}{\delta_1}\right)} < c \right\}, \end{aligned}$$

e portanto

$$e^{-y\left(\frac{\delta_1 - \delta_0}{\delta_1 \delta_0}\right)} < \frac{c \delta_0}{\delta_1},$$

donde se conclui que

$$y > \underbrace{-\left(\frac{\delta_0 \delta_1}{\delta_1 - \delta_0}\right) \ln\left(\frac{c \delta_0}{\delta_1}\right)}_l.$$

A região de rejeição é então

$$R = \{y \in \mathbb{R} : y > l\},$$

onde

$$l = \mathbb{P}(\text{Rejeitar } H_0 | H_0 \text{ Verdadeiro}) = \alpha.$$

Atendendo a que $Y \sim \text{exponencial}(\delta) \Rightarrow W = \frac{2Y}{\delta_0} \sim \chi_{(2)}^2$ obtém-se

$$\mathbb{P}(\text{Rejeitar } H_0 | H_0 \text{ Verdadeiro}) = \alpha$$

$$\Leftrightarrow \mathbb{P}(Y > l | \delta = \delta_0) = \alpha$$

$$\Leftrightarrow \mathbb{P}\left(\frac{2Y}{\delta_0} > \frac{2l}{\delta_0}\right) = \alpha$$

$$\Leftrightarrow \mathbb{P}\left(W > \frac{2l}{\delta_0}\right) = \alpha$$

$$\Leftrightarrow F_W\left(\frac{2l}{\delta_0}\right) = 1 - \alpha$$

$$\Leftrightarrow l = \frac{\delta_0}{2} \chi_{(2)}^2(1 - \alpha)$$

onde $\chi_{(2)}^2(1 - \alpha)$ representa o quantil $1 - \alpha$ do modelo qui-quadrado com dois graus de liberdade. Uma vez determinado o limiar crítico l que figura em (7.1), rejeitamos a hipótese nula caso a quantidade de substância y seja estritamente superior a esse limiar. A potência do teste é dada por

$$\begin{aligned} 1 - \beta &= \mathbb{P}(\text{Rejeitar } H_0 | H_0 \text{ Falsa}) \\ &= \mathbb{P}\left(Y > \frac{\delta_0}{2} \chi_{(2)}^2(1 - \alpha) | \delta = \delta_1\right) \\ &= \mathbb{P}\left(\frac{2Y}{\delta_1} > \frac{\delta_0}{\delta_1} \chi_{(2)}^2(1 - \alpha)\right) \\ &= \mathbb{P}\left(W > \frac{\delta_0}{\delta_1} \chi_{(2)}^2(1 - \alpha)\right), \end{aligned}$$

obtendo-se finalmente

$$1 - \beta = 1 - F_{\chi_{(2)}^2}\left(\frac{\delta_0}{\delta_1} \chi_{(2)}^2(1 - \alpha)\right). \quad (7.7)$$

Refira-se mais uma vez que quanto maior for a distância entre δ_1 e δ_0 (em termos de quociente $\frac{\delta_1}{\delta_0}$) maior será o valor da potência do teste.

7.3 Testes de hipóteses para amostras conjuntas

7.3.1 Sensibilidade do teste composto para o modelo gaussiano considerando grupos de dimensão n dos quais γ estão infetados

Suponhamos que com base numa amostra aleatória de dimensão n extraída de uma população $Y \sim \text{Gaussiana}(\mu, \sigma)$ se pretende testar $H_0 : \mu = \mu_0$ versus $H_1 : \mu = \mu_1$ com $\mu_1 > \mu_0$.

A função de verosimilhança da amostra é dada por

$$\begin{aligned} L_\mu(\mathbf{y}) = L(\mu|\mathbf{y}) &= \prod_{k=1}^n \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2\sigma^2}(y_k - \mu)^2} \\ &= \frac{1}{(2\pi)^{\frac{n}{2}} \sigma^n} e^{\frac{-1}{2\sigma^2} \sum_{k=1}^n (y_k - \mu)^2}, \end{aligned}$$

onde

$$\begin{aligned} \sum_{k=1}^n (y_k - \mu)^2 &= \sum_{k=1}^n [(y_k - \bar{y}) + (\bar{y} - \mu)]^2 \\ &= \sum_{k=1}^n (y_k - \bar{x})^2 + 2 \underbrace{\sum_{k=1}^n (y_k - \bar{x})(\bar{x} - \mu)}_{=0} + \sum_{k=1}^n (\bar{y} - \mu)^2 \\ &= (n-1)s^2 + n(\bar{y} - \mu)^2. \end{aligned}$$

Assim ter-se-á

$$L(\mu_0|\mathbf{y}) = \frac{1}{\sigma^n (2\pi)^{\frac{n}{2}}} e^{\frac{-1}{2} \frac{n(\bar{y} - \mu_0)^2 + (n-1)s^2}{\sigma^2}}$$

e

$$L(\mu_1|\mathbf{y}) = \frac{1}{\sigma^n (2\pi)^{\frac{n}{2}}} e^{\frac{-1}{2} \frac{n(\bar{y} - \mu_1)^2 + (n-1)s^2}{\sigma^2}}$$

e, consequentemente,

$$\begin{aligned} \frac{L(\mu_0|\mathbf{y})}{L(\mu_1|\mathbf{y})} &= e^{\frac{-n}{2} \frac{(\bar{y} - \mu_0)^2 - (\bar{y} - \mu_1)^2}{\sigma^2}} \\ &= e^{\frac{n}{\sigma^2} (\mu_0 - \mu_1) (\bar{y} - \frac{\mu_0 + \mu_1}{2})}. \end{aligned}$$

Assim, pelo lema de Neyman-Pearson, a região de rejeição é dada por

$$\begin{aligned} R &= \left\{ \mathbf{y} \in \mathbb{R}^n : \frac{L(\mu_0|\mathbf{y})}{L(\mu_1|\mathbf{y})} < c \right\} \\ &= \left\{ \mathbf{y} \in \mathbb{R}^n : e^{\frac{n}{\sigma^2} (\mu_0 - \mu_1) (\bar{y} - \frac{\mu_0 + \mu_1}{2})} < c \right\}, \end{aligned}$$

e, por conseguinte,

$$(\mu_0 - \mu_1) \left(\bar{y} - \frac{\mu_0 + \mu_1}{2} \right) < \ln(c) \frac{\sigma^2}{n}.$$

Atendendo a que $\mu_1 > \mu_0$ e, conseqüentemente, $\mu_0 - \mu_1 < 0$ obtém-se

$$\bar{y} > \underbrace{\frac{\mu_0 + \mu_1}{2} + \frac{1}{\mu_0 - \mu_1} \frac{\sigma^2}{n} \ln(c)}_{=c'}.$$

A região de rejeição é então

$$R = \left\{ \mathbf{y} \in \mathbb{R}^n : \bar{y} > c' \right\},$$

onde

$$c' = \mathbb{P}(\text{Rejeitar } H_0 | H_0 \text{ Verdadeira}) = \alpha.$$

Assim tem-se

$$\mathbb{P}(\text{Rejeitar } H_0 | H_0 \text{ Verdadeira}) = \alpha$$

$$\Leftrightarrow \mathbb{P}(\bar{Y} > c' | \mu = \mu_0) = \alpha$$

$$\Leftrightarrow \mathbb{P}\left(\frac{\bar{Y} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > \frac{c' - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right) = \alpha$$

$$\Leftrightarrow 1 - \Phi\left(\frac{c' - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right) = \alpha$$

$$\Leftrightarrow \Phi\left(\frac{c' - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right) = 1 - \alpha$$

$$\Leftrightarrow \frac{c' - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \Phi^{-1}(1 - \alpha)$$

$$\Leftrightarrow c' = \mu_0 + \frac{\sigma_0}{\sqrt{n}} \Phi^{-1}(1 - \alpha),$$

onde $\Phi^{-1}(1 - \alpha)$ representa o quantil $1 - \alpha$ da Gaussiana padrão. Considerando que há γ infetados tem-se $\bar{Y} \sim \text{Gaussiana}\left(\mu', \frac{\sigma}{\sqrt{n}}\right)$ onde $\mu' = \mu_0 + \frac{\gamma}{n}(\mu_1 - \mu_0)$ e condicionalmente à existência de γ infetados ter-se-á

$$\begin{aligned} 1 - \beta &= \mathbb{P}(\text{Rejeitar } H_0 | H_0 \text{ Falsa}) \\ &= \mathbb{P}\left[\bar{Y} > \mu_0 + \frac{\sigma}{\sqrt{n}} \Phi^{-1}(1 - \alpha) | \mu = \mu'\right] \\ &= \mathbb{P}\left[Z > \left(\mu_0 + \frac{\sigma}{\sqrt{n}} \Phi^{-1}(1 - \alpha) - \mu'\right) \frac{\sqrt{n}}{\sigma}\right] \\ &= 1 - \Phi\left[\frac{(\mu_0 - \mu')}{\sigma} \sqrt{n} + \Phi^{-1}(1 - \alpha)\right]. \end{aligned}$$

Atendendo a que $\mu' = \mu_0 + \frac{\gamma}{n}(\mu_1 - \mu_0)$ obtém-se

$$1 - \beta = 1 - \Phi \left[\Phi^{-1}(1 - \alpha) - \frac{\gamma}{\sqrt{n}} \left(\frac{\Delta}{\sigma} \right) \right]. \quad (7.8)$$

Facilmente se constata que a potência do teste aumenta com a diferença entre as médias $\Delta = (\mu_1 - \mu_0)$. Por outro lado quanto maior for o valor de γ (número de contaminados no grupo), maior será o valor da potência do teste e por conseguinte tem-se

$$0 \leq \varphi_s^{[1,n]} \leq \varphi_s^{[2,n]} \leq \dots \leq \varphi_s^{[n,n]}. \quad (7.9)$$

Se em (7.8) fizermos $\gamma = n$ obtém-se

$$1 - \beta = 1 - \Phi \left[\Phi^{-1}(1 - \alpha) - \sqrt{n} \left(\frac{\Delta}{\sigma} \right) \right]. \quad (7.10)$$

Da equação 3.1 presente na página 32 obtém-se

$$k_\gamma^{[n]} = 1 - \frac{\varphi_s^{[\gamma,n]}}{\varphi_s}$$

e, por conseguinte, quando $\gamma = n$ tem-se

$$k_n^{[n]} = 1 - \frac{\varphi_s^{[n,n]}}{\varphi_s}.$$

Atendendo às expressões (7.2) e (7.10) obtém-se finalmente

$$k_n^{[n]} = 1 - \frac{1 - \Phi \left[\Phi^{-1}(1 - \alpha) - \sqrt{n} \left(\frac{\Delta}{\sigma} \right) \right]}{1 - \Phi \left[\Phi^{-1}(1 - \alpha) - \left(\frac{\Delta}{\sigma} \right) \right]} < 0$$

Podemos assim concluir que no modelo Gaussiano a sensibilidade do teste conjunto quando todos os indivíduos estão contaminados é superior à sensibilidade do teste simples. Mais especificamente, (ver equação 7.8), podemos ainda concluir que

$$\begin{cases} \varphi_s^{[\gamma,n]} < \varphi_s & ; \quad \gamma < \sqrt{n} \\ \varphi_s^{[\gamma,n]} = \varphi_s & ; \quad \gamma = \sqrt{n} \quad (\text{se } \sqrt{n} \in \mathbb{N}) \\ \varphi_s^{[\gamma,n]} > \varphi_s & ; \quad \gamma > \sqrt{n} \end{cases}$$

Na sequência do exemplo dado na página 102 suponhamos agora que estamos perante um grupo de 3 indivíduos. Observando a Figura 7.3 observamos que à medida que aumenta o número de infetados no grupo maior será a diferença $\mu_1 - \mu_0$ e por conseguinte maior será a sensibilidade do teste.

7.3.2 Sensibilidade do teste composto para o modelo exponencial considerando grupos de dimensão n dos quais γ estão infetados

Considerando agora que o teste é efetuado a grupos de dimensão n com $D_{\theta_0} \sim$ exponencial (δ_0) e $D_{\theta_1} \sim$ exponencial (δ_1), vamos continuar a assumir que $\delta_1 > \delta_0$. Pretende-se determinar o teste mais potente de nível α para testar

$$H_0 : \mu = \mu_0 \text{ versus } H_1 : \mu = \mu_1, \text{ com } \mu_1 > \mu_0, \quad (7.11)$$

que, como já foi referido, equivale a testar

$$H_0 : \delta = \delta_0 \text{ versus } H_1 : \delta = \delta_1, \text{ com } \delta_1 > \delta_0. \quad (7.12)$$

O quociente das verosimilhanças é dado por

$$\frac{L(\delta_0|\mathbf{y})}{L(\delta_1|\mathbf{y})} = \frac{\left(\frac{1}{\delta_0}\right)^n e^{-\frac{1}{\delta_0} \sum_{i=1}^n y_i}}{\left(\frac{1}{\delta_1}\right)^n e^{-\frac{1}{\delta_1} \sum_{i=1}^n y_i}} = \left(\frac{\delta_1}{\delta_0}\right)^n e^{-\left(\frac{1}{\delta_0} - \frac{1}{\delta_1}\right) \sum_{i=1}^n y_i},$$

e, por conseguinte, a região de rejeição será da forma

$$\begin{aligned} R &= \left\{ \mathbf{y} \in \mathbb{R}^n : \left(\frac{\delta_1}{\delta_0}\right)^n e^{-\left(\frac{1}{\delta_0} - \frac{1}{\delta_1}\right) \sum_{i=1}^n y_i} < c \right\} \\ &= \left\{ \mathbf{y} \in \mathbb{R}^n : -\left(\frac{1}{\delta_0} - \frac{1}{\delta_1}\right) \sum_{i=1}^n y_i < \ln \left[c \left(\frac{\delta_0}{\delta_1}\right)^n \right] \right\}, \end{aligned}$$

donde se obtém

$$\sum_{i=1}^n y_i > \underbrace{-\left(\frac{1}{\delta_0} - \frac{1}{\delta_1}\right) \ln \left[c \left(\frac{\delta_0}{\delta_1}\right)^n \right]}_{=l}$$

Assim, rejeita-se a hipótese nula quando a soma das observações for superior a l sendo l determinado por forma a que

$$\mathbb{P}(\text{Rejeitar } H_0 | H_0 \text{ Verdadeiro}) = \alpha.$$

Comecemos por recordar que se

$$Y \sim \text{exponencial}(\delta) \Rightarrow \sum_{i=1}^n Y_i \sim \text{gama}(n, \delta) \Rightarrow W = \frac{2}{\delta} \sum_{i=1}^n Y_i \sim \chi^2_{(2n)},$$

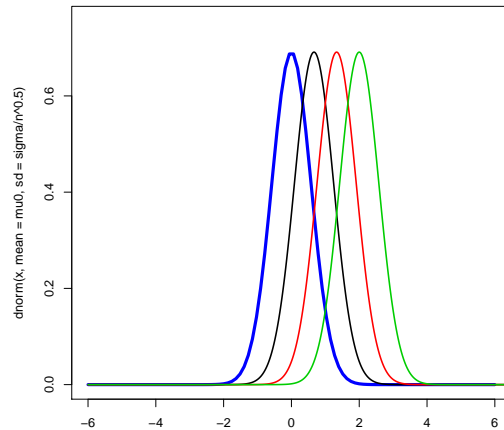


Figura 7.3: Densidade de \bar{Y} condicional a $\gamma = m$ com $m = \{0, 1, 2, 3\}$

e portanto tem-se

$$\begin{aligned}
 \mathbb{P}(\text{Rejeitar } H_0 | H_0 \text{ Verdadeira}) &= \alpha \\
 \Leftrightarrow \mathbb{P}\left(\sum_{i=1}^n Y_i > l | \delta = \delta_0\right) &= \alpha \\
 \Leftrightarrow \mathbb{P}\left(\frac{2}{\delta_0} \sum_{i=1}^n Y_i > \frac{2}{\delta_0} l\right) &= \alpha \\
 \Leftrightarrow \mathbb{P}\left(W > \frac{2}{\delta_0} l\right) &= \alpha \\
 \Leftrightarrow F_W\left(\frac{2}{\delta_0} l\right) &= 1 - \alpha \\
 \Leftrightarrow \frac{2}{\delta_0} l &= F_W^{-1}(1 - \alpha) \\
 \Leftrightarrow l &= \frac{\delta_0}{2} \chi_{(2n)}^2 (1 - \alpha)
 \end{aligned}$$

onde $\chi_{(2n)}^2 (1 - \alpha)$ representa o quantil de probabilidade $1 - \alpha$ do modelo qui-quadrado com $2n$ graus de liberdade.

Neste caso, o cálculo da potência do teste é mais complexo. A potência do teste é a probabilidade de rejeitar a hipótese nula quando esta é falsa, e, conseqüentemente, a hipótese alternativa é verdadeira. Se levarmos em linha de conta que na hipótese alternativa o número de infetados, γ , pode variar entre 1 e a dimensão do grupo, n , ou seja, se atendermos a que $\delta' = \delta_0 + \frac{\gamma}{n} (\delta_1 - \delta_0)$, concluímos que

$$\begin{aligned}
 1 - \beta &= \mathbb{P}(\text{Rejeitar } H_0 | H_0 \text{ Falsa}) \\
 &= \mathbb{P}\left(\sum_{i=1}^n Y_i > \frac{\delta_0}{2} \chi_{(2n)}^2 (1 - \alpha) \mid \delta = \delta'\right) \\
 &= \mathbb{P}\left[\left(\sum_{i=1}^{n-\gamma} Y_i + \sum_{i=1}^{\gamma} Y_i^*\right) > \frac{\delta_0}{2} \chi_{(2n)}^2 (1 - \alpha)\right]
 \end{aligned}$$

onde $Y_i \sim \text{exponencial}(\delta_0)$ e $Y_i^* \sim \text{exponencial}(\delta_1)$ e por conseguinte há que determinar a distribuição duma soma de exponenciais com parâmetros de escala distintos (Informação detalhada sobre este resultado pode ser encontrada em Brilhante (1999)). Recorrendo ao teorema 7.1 podemos determinar a potência do teste para qualquer valor de γ

Teorema 7.1. (*Convolução generalizada de exponenciais*)

Sejam X_1, X_2, \dots, X_n , n variáveis aleatórias independentes com função densidade de probabilidade $f_{X_i}(x) = \frac{1}{\delta_i} e^{-\frac{x}{\delta_i}}$, $x > 0$, $i = 1, 2, \dots, n$, onde $\delta_i > 0$ e $\delta_i \neq \delta_j$. A variável aleatória $S_n = \sum_{i=1}^n X_i$ tem função densidade de probabilidade $f_{S_n}(s) = \sum_{i=1}^n C_{i,n} \frac{1}{\delta_i} e^{-\frac{s}{\delta_i}}$, $s > 0$ em que $C_{i,n} = \prod_{\substack{j=1 \\ j \neq i}}^n \frac{\delta_i}{\delta_i - \delta_j}$.

No entanto, quando $\gamma = n$ tem-se que $\delta' = \delta_1$ e por conseguinte a potência do teste quando todos os elementos do grupo estão infetados é dada por

$$\begin{aligned}
 1 - \beta &= \mathbb{P}(\text{Rejeitar } H_0 | H_0 \text{ Falsa}) \\
 &= \mathbb{P}\left(\sum_{i=1}^n Y_i > \frac{\delta_0}{2} \chi_{(2n)}^2 (1 - \alpha) \mid \delta = \delta_1\right) \\
 &= \mathbb{P}\left(\frac{2}{\delta_1} \sum_{i=1}^n Y_i > \frac{\delta_0}{\delta_1} \chi_{(2n)}^2 (1 - \alpha)\right) \\
 &= 1 - F_W\left(\frac{\delta_0}{\delta_1} \chi_{(2n)}^2 (1 - \alpha)\right) \\
 &= 1 - F_{\chi_{(2n)}^2}\left(\frac{\delta_0}{\delta_1} \chi_{(2n)}^2 (1 - \alpha)\right).
 \end{aligned}$$

Atendendo à expressão anterior e à igualdade (7.7) da página 104, podemos determinar o valor da contante

$$k_n^{[n]} = 1 - \frac{1 - F_{\chi^2_{(2n)}}\left(\frac{\delta_0}{\delta_1} \chi^2_{(2n)}(1 - \alpha)\right)}{1 - F_{\chi^2_{(2)}}\left(\frac{\delta_0}{\delta_1} \chi^2_{(2)}(1 - \alpha)\right)} < 0.$$

Assim, no modelo exponencial, a sensibilidade do teste composto quando todos os seus elementos estão infetados é superior à sensibilidade do teste simples.

7.3.3 Proposta de duas metodologias

O nosso objetivo consiste em estender a metodologia de Dorfman a análises quantitativas. Para tal, proceda-se à análise de uma amostra composta (Y_1, Y_2, \dots, Y_n) de n unidades amostrais com média \bar{Y}_n e máximo $Y_{n:n}$. Para concluir se uma unidade amostral está ou não infetada, recorrendo à metodologia de Dorfman, há que efetuar previamente um teste ao grupo.

Para a obtenção de uma amostra composta retira-se uma determinada quantidade de substância de cada um dos elementos do grupo que posteriormente é misturada de forma homogénea. Se o teste ao grupo der negativo conclui-se que todos os elementos que o compõem estão limpos, caso contrário, é sinal de que existe pelo menos um elemento contaminado no grupo. O principal problema que se coloca nesta fase, reside em detetar a presença de pelo menos um contaminado no grupo, ou seja, determinar uma forma de quantificar a probabilidade de existência no grupo de pelo menos um elemento com quantidade de substância superior a l . As hipóteses a testar são:

$$H_0 : \sum_{i=1}^n X_i = 0 \text{ versus } H_1 : \sum_{i=1}^n X_i \geq 1 \quad (7.13)$$

Já foi referido que num teste composto pretende-se averiguar a existência de pelo menos um indivíduo infetado e, por conseguinte, utilizando a mesma metodologia que foi aplicada nos testes individuais (na qual um indivíduo é classificado como infetado se $y_i > l$), a análise conjunta vai identificar se o máximo do grupo é superior ao limiar l . Dado que só dispomos de informação da média do grupo, os testes efetuados baseiam-se na quantificação da informação sobre o máximo dada pela média do grupo. Note-se que a eventual existência de pelo menos um infetado no grupo implica que o máximo tome um valor elevado.

Na situação referida tanto as amostras referentes a indivíduos saudáveis como as referentes a indivíduos doentes possuem uma determinada quantidade de substância de interesse. Suponhamos que a quantidade de substância nos indivíduos saudáveis é modelada por uma determinada distribuição contínua $Y \sim D_{\theta_0}$ e que a quantidade de substância nos indivíduos doentes é modelada por $Y^* = \beta_0 + \beta_1 Y$ (ou $Y \sim D_{\theta_1}$) onde θ_0 e θ_1 são vectores de parâmetros distintos. Vamos considerar os casos em que D engloba os modelos gaussiano, exponencial e Pareto. Naturalmente que outras formas para D_{θ_1} poderiam ser aplicadas, con-

tudo iremos restringir às situações em que D_{θ_1} é do mesmo tipo que D_{θ_0} com apenas alteração de localização e escala.

Para concluir se a hipótese nula deve ser rejeitada ou mantida são propostas duas metodologias:

Primeira metodologia

Nesta metodologia inicial – designada por T_1 – determina-se a distribuição de \bar{Y}_n e utiliza-se o seu percentil $1 - \alpha$ como limiar crítico.

Quando se misturam n amostras saudáveis Y_1, Y_2, \dots, Y_n e se extrai uma porção $1/n$ da quantidade total, a quantidade de substância de interesse é dada pela variável aleatória $C_{0,n}$ onde $C_{\gamma,n}$ é dado por

$$C_{\gamma,n} = \frac{\sum_{i=1}^{n-\gamma} Y_i + \sum_{i=1}^{\gamma} Y_i^*}{n}. \tag{7.14}$$

Deste modo, a variável aleatória $C_{\gamma,n}$ representa a quantidade de substância de interesse numa amostra de dimensão n com γ amostras individuais infectadas. A hipótese nula do teste de hipóteses (7.13) é rejeitada se $C_{0,n} > q_{1-\alpha}$ onde $F_{C_{0,n}}(q_{1-\alpha}) = 1 - \alpha$ e $F_{C_{0,n}}$ representa a função de distribuição da variável aleatória $C_{0,n}$.

Naturalmente que ao efetuar a análise conjunta observamos apenas a média do grupo. Contudo pretende-se analisar o máximo do grupo e não a média (apesar da correlação existente entre elas).

n	Gaussiana	Exponencial	Pareto		
			$\theta = 5$	$\theta = 3$	$\theta = 1$
2	0.8546	0.9441	0.9680	0.9815	0.999944
3	0.7653	0.9018	0.9417	0.9652	0.999982
4	0.7128	0.8749	0.9185	0.9518	0.9999782
5	0.6618	0.8395	0.8947	0.9382	0.9999734
10	0.5296	0.7390	0.8260	0.8892	0.9997459
20	0.4260	0.6224	0.7462	0.8377	0.9999075
50	0.3227	0.4931	0.6347	0.7612	0.999982
100	0.2295	0.4012	0.5468	0.6892	0.9999732

Tabela 7.1: Correlação entre a soma e o máximo

Caso exista uma amostra infectada no grupo, o problema principal consiste em saber de que forma é que o valor observado da variável aleatória $C_{0,n}$ é influenciado pela presença da amostra infectada. A quantidade de informação sobre o máximo da amostra

$M_n = \max(Y_1, Y_2, \dots, Y_n)$ fornecido pela média amostral $C_{k,n}$ será provavelmente elevada quando a correlação entre o máximo da amostra e a média amostral for alta. Para as três distribuições mencionadas anteriormente foi calculada a correlação entre o máximo da amostra e a média amostral para varias dimensões amostrais.

Todos os valores apresentados na Tabela 7.1 foram obtidos por simulação apesar de se conseguirem obter os mesmos resultados analiticamente no caso da distribuição exponencial e por aproximação numérica no caso gaussiano. Refira-se que no caso da Pareto(1) não existe média, no entanto a sequência de correlações converge para 1 quando $\theta \rightarrow 1$. Verificamos que à medida que a dimensão da amostra aumenta a correlação diminui. No caso da Pareto, talvez por ser um modelo de caudas pesadas, a correlação mantém-se elevada mesmo para valores de n elevados, tais como 20 ou 50. Para o modelo exponencial e para o modelo gaussiano os valores obtidos para a correlação são muito reduzidos o que nos leva a crer que a potência dos testes utilizando esses modelos deverá ser bastante reduzida.

Segunda metodologia

Uma outra metodologia – designada por T_2 – consiste em determinar, sob H_0 , a distribuição de \overline{Y}_n condicionada ao máximo.

Seja α o nível de significância do teste (7.13). A hipótese nula será rejeitada se pelo menos um indivíduo ultrapassar o limiar crítico l . Desta forma, sob H_0

$$\begin{aligned} \mathbb{P}(M_n \leq l) &= \mathbb{P}(Y_1 \leq l, \dots, Y_n \leq l) = F_{D_{\theta_0}}^n(l) = 1 - \alpha \Leftrightarrow \\ &\Leftrightarrow l = F_{D_{\theta_0}}^{-1}\left((1 - \alpha)^{\frac{1}{n}}\right). \end{aligned}$$

Assim, l representa o limiar crítico para amostras conjuntas de dimensão n . O cálculo do quantil $F_{D_{\theta_0}}^{-1}$ não é usualmente linear contudo a utilização de valores simulados fornece boas aproximações.

Para a obtenção do ponto crítico para a média amostral, procedeu-se da seguinte forma:

1. Simulou-se, sob H_0 , $N = 100000$ amostras de dimensão n que posteriormente foram ordenados pelo seu máximo;
2. Determinou-se a probabilidade de nenhuma observação ser superior ao percentil $(1 - \alpha)$ da distribuição \mathbf{D} , ou seja, determinou-se o percentil $(1 - \alpha)^{\frac{1}{n}}$ da distribuição \mathbf{D} que caracteriza um indivíduo sob H_0 .
3. Selecionaram-se as $m = 1000$ amostras cujo máximo está mais próximo de l . A média amostral dessas amostras é calculada e utilizada como o limiar l^* para o teste composto. Assim, se a média exceder l^* a hipótese nula é rejeitada e a amostra composta é declarada infetada.

Comparação das metodologias via simulação

Ao juntar n porções de uma determinada substância num grupo, a sensibilidade do teste conjunto sofre uma redução significativa. Com o intuito de comparar as duas metodologias apresentadas relativamente à magnitude dessas perdas, recorreu-se a métodos de Monte Carlo para obter valores para a sensibilidade $\varphi_s^{[n]}$ e para a especificidade $\varphi_e^{[n]}$ do teste composto (ver Tabelas 7.2 a 7.6). Todos os valores simulados foram obtidos considerando-se grupos de dimensão ótima e assumindo-se que cada modelo D_{θ_1} é apenas uma translação de D_{θ_0} . A translação foi determinada por forma a manter a sensibilidade e a especificidade do teste individual iguais a 0.95 (caso 1) e 0.995 (caso 2). Da análise dos valores simulados verificamos facilmente que a metodologia T_2 é mais eficiente já que a redução operada na sensibilidade do teste composto é significativamente menor comparativamente à metodologia T_1 . No que respeita à especificidade, as perdas são em ambas as metodologias muito pouco significativas.

		caso 1: $\varphi_s^{[1]} = \varphi_e^{[1]} = 0.95$				caso 2: $\varphi_s^{[1]} = \varphi_e^{[1]} = 0.995$			
		T_1		T_2		T_1		T_2	
p	$1 - \alpha$	$\varphi_s^{[n]}$	$\varphi_e^{[n]}$	$\varphi_s^{[n]}$	$\varphi_e^{[n]}$	$\varphi_s^{[n]}$	$\varphi_e^{[n]}$	$\varphi_s^{[n]}$	$\varphi_e^{[n]}$
0.15 ($n^* = 3$)	0.9	.7746	.8994	.8573	.8283	.9623	.9002	.9830	.8231
	0.95	.6627	.9504	.8012	.8818	.9229	.9501	.9728	.8707
	0.99	.4310	.9900	.6902	.9408	.7835	.9901	.9319	.9427
0.10 ($n^* = 4$)	0.90	.6946	.9009	.8141	.8085	.9182	.9004	.9598	.8183
	0.95	.5707	.9499	.7617	.8557	.8518	.9501	.9450	.8561
	0.99	.3393	.9900	.6507	.9209	.6603	.9899	.8931	.9235
0.05 ($n^* = 5$)	0.90	.6127	.9003	.7636	.7971	.8628	.8992	.9371	.7960
	0.95	.4804	.9501	.7109	.8417	.7716	.9500	.9202	.8320
	0.99	.2534	.9901	.9490	.9082	.5431	.9900	.8445	.9137
0.01 ($n^* = 11$)	0.90	.4066	.8997	.6600	.7343	.6261	.9001	.8277	.7404
	0.95	.2759	.9497	.5917	.7874	.4880	.9502	.8182	.7606
	0.99	.1061	.9899	.5266	.8338	.2491	.9900	.7337	.8347

Tabela 7.2: Modelo Gaussiano (100000 replicas)

Testes quantitativos contínuos

		caso 1: $\varphi_s^{[1]} = \varphi_e^{[1]} = 0.95$				caso 2: $\varphi_s^{[1]} = \varphi_e^{[1]} = 0.995$			
		T_1		T_2		T_1		T_2	
p	$1 - \alpha$	$\varphi_s^{[n]}$	$\varphi_e^{[n]}$	$\varphi_s^{[n]}$	$\varphi_e^{[n]}$	$\varphi_s^{[n]}$	$\varphi_e^{[n]}$	$\varphi_s^{[n]}$	$\varphi_e^{[n]}$
0.15 ($n^* = 3$)	0.9	.9423	.8984	.9923	.8825	1.00	.9019	1.00	.8856
	0.95	.5496	.9492	.6673	.9339	1.00	.9554	1.00	.9429
	0.99	.1626	.9894	.2464	.9792	.6413	.9895	.9596	.9820
0.10 ($n^* = 4$)	0.90	.8641	.9032	.9743	.8785	1.00	.9049	1.00	.8860
	0.95	.4788	.9458	.6585	.9256	1.00	.9517	1.00	.9324
	0.99	.1174	.9908	.2473	.9781	.4929	0.9903	.8980	0.9792
0.05 ($n^* = 5$)	0.90	.8206	.9005	.9644	.8698	1.00	.8975	1.00	.8667
	0.95	.4240	.9486	.6660	.9198	1.00	.9494	1.00	.9229
	0.99	.0792	.9900	.2195	.9746	.4019	.9903	.8979	.9737
0.01 ($n^* = 11$)	0.90	.7668	.9015	.9907	.8371	1.00	.8957	1.00	.8220
	0.95	.3514	.9487	.8710	.8842	.9521	.9472	1.00	.8896
	0.99	.0256	.9910	.3292	.9525	.1896	.9902	.9367	.9512

Tabela 7.3: Modelo exponencial (100000 replicas)

		caso 1: $\varphi_s^{[1]} = \varphi_e^{[1]} = 0.95$				caso 2: $\varphi_s^{[1]} = \varphi_e^{[1]} = 0.995$			
		T_1		T_2		T_1		T_2	
p	$1 - \alpha$	$\varphi_s^{[n]}$	$\varphi_e^{[n]}$	$\varphi_s^{[n]}$	$\varphi_e^{[n]}$	$\varphi_s^{[n]}$	$\varphi_e^{[n]}$	$\varphi_s^{[n]}$	$\varphi_e^{[n]}$
0.15 ($n^* = 3$)	0.9	.5813	.9010	.63	.8883	1.0	.8995	1.0	.8928
	0.95	.3616	.9473	.3994	.9408	1.0	.9495	1.0	.9422
	0.99	.0646	.9896	.0802	.9874	.4166	.9898	.5075	.9873
0.10 ($n^* = 4$)	0.90	.5014	.8996	.5359	.8886	1.0	.8997	1.0	.8866
	0.95	.2908	.9497	.3314	.9414	.9503	.9503	.9410	.9381
	0.99	.0556	.9910	.0752	.9874	.3345	0.9902	.4070	.9875
0.05 ($n^* = 5$)	0.90	.4189	.9001	.4725	.8827	.9924	.8861	.9987	.8861
	0.95	.2248	.9507	.2875	.9353	.8087	.9514	.9084	.9362
	0.99	.0406	.9901	.0583	.9862	.2526	.9892	.3350	.9842
0.01 ($n^* = 11$)	0.90	.2585	.90	.3555	.8551	.6771	.9004	.7738	.8687
	0.95	.1365	.9504	.2113	.9207	.4294	.9497	.5940	.9168
	0.99	.0263	.99	.0554	.98	.1118	.9902	.2054	.9881

Tabela 7.4: Modelo Pareto(5) (100000 replicas)

A redução da sensibilidade do teste composto provocada pelo efeito de diluição pode ser minimizada caso se opte por grupos de dimensão inferior à dimensão ótima. Nas tabelas 7.7, 7.8 e 7.9 é visível o ganho significativo na sensibilidade do teste composto.

		caso 1: $\varphi_s^{[1]} = \varphi_e^{[1]} = 0.95$				caso 2: $\varphi_s^{[1]} = \varphi_e^{[1]} = 0.995$			
		T_1		T_2		T_1		T_2	
p	$1 - \alpha$	$\varphi_s^{[n]}$	$\varphi_e^{[n]}$	$\varphi_s^{[n]}$	$\varphi_e^{[n]}$	$\varphi_s^{[n]}$	$\varphi_e^{[n]}$	$\varphi_s^{[n]}$	$\varphi_e^{[n]}$
0.15 ($n^* = 3$)	0.9	.5691	.9019	.5883	.8978	1.00	.8994	1.00	.8905
	0.95	.2900	.9518	.3216	.9479	1.00	.9534	1.00	.9494
	0.99	.0287	.9891	.0325	.9886	.3409	.9904	.3840	.9893
0.10 ($n^* = 4$)	0.90	.3667	.8985	.4994	.8882	1.00	.8983	1.00	.8917
	0.95	.2319	.9561	.2595	.9512	1.00	.9479	1.00	.9420
	0.99	.0254	.9903	.0294	.9893	.2622	0.9897	.2851	.9887
0.05 ($n^* = 5$)	0.90	.3669	.9047	.3996	.8966	1.00	.8984	1.00	.8873
	0.95	.1836	.9529	.2189	.9457	.9359	.9529	.9793	.9472
	0.99	.0286	.9901	.0339	.9887	.01924	.9903	.2304	.9886
0.01 ($n^* = 11$)	0.90	.2265	.9032	.2714	.8803	.7715	.9002	.08551	.8768
	0.95	.1030	.9490	.1337	.9382	.4418	.9492	.5234	.9355
	0.99	.0183	.9888	.0220	.9863	.0772	.9904	.1089	.9875

Tabela 7.5: Modelo Pareto(3) (100000 replicas)

		caso 1: $\varphi_s^{[1]} = \varphi_e^{[1]} = 0.95$				caso 2: $\varphi_s^{[1]} = \varphi_e^{[1]} = 0.995$			
		T_1		T_2		T_1		T_2	
p	$1 - \alpha$	$\varphi_s^{[n]}$	$\varphi_e^{[n]}$	$\varphi_s^{[n]}$	$\varphi_e^{[n]}$	$\varphi_s^{[n]}$	$\varphi_e^{[n]}$	$\varphi_s^{[n]}$	$\varphi_e^{[n]}$
0.15 ($n^* = 3$)	0.9	.3581	.8962	.3581	.8962	1.00	.8946	1.00	.8943
	0.95	.0794	.9513	.0822	.9508	1.00	.9490	1.00	.9483
	0.99	.0100	.9895	.0101	.9895	.1835	.9903	.1827	.9906
0.10 ($n^* = 4$)	0.90	.2398	.8969	.2437	.8959	1.00	.8950	1.00	.8926
	0.95	.0740	.9465	.0751	.9460	1.00	.9503	1.00	.9498
	0.99	.0100	.9901	.0100	.9901	.0764	0.9883	.0703	.9883
0.05 ($n^* = 5$)	0.90	.1612	.9013	.1674	.8982	1.00	.9042	1.00	.9021
	0.95	.0551	.9482	.0559	.9480	1.00	.9522	1.00	.9522
	0.99	.0099	.9899	.0103	.9899	.0240	.9908	.0227	.9909
0.01 ($n^* = 11$)	0.90	.1110	.8973	.1110	.8960	1.00	.9016	1.00	.8986
	0.95	.0592	.9488	.0592	.9478	.3154	.9522	.3281	.9517
	0.99	.0107	.9909	.0107	.9909	.0123	.9808	.0123	.9893

Tabela 7.6: Modelo Pareto(1) (100000 replicas)

		caso 1: $\varphi_s^{[1]} = \varphi_e^{[1]} = 0.95$				caso 2: $\varphi_s^{[1]} = \varphi_e^{[1]} = 0.995$			
		T_1		T_2		T_1		T_2	
p	$1 - \alpha$	$\varphi_s^{[n]}$	$\varphi_e^{[n]}$	$\varphi_s^{[n]}$	$\varphi_e^{[n]}$	$\varphi_s^{[n]}$	$\varphi_e^{[n]}$	$\varphi_s^{[n]}$	$\varphi_e^{[n]}$
0.15 ($n = 2$) $n^* = 3$	0.9	.7555	.9006	.7785	.8970	1	.9009	1	.8968
	0.95	.3994	.9495	.4114	.9478	1	.95	1	.9465
	0.99	.0698	.99	.0799	.9889	.6155	.9901	.7203	.9888
0.10 ($n = 2$) $n^* = 4$	0.90	.7492	.8998	.7704	.8967	1	.9018	1	.9
	0.95	.3785	.9499	.4061	.9465	1	.9503	1	.9472
	0.99	.0637	.9901	.0707	.9892	.6296	.9898	.72	.9888
0.05 ($n = 3$) $n^* = 5$	0.90	.5321	.9013	.5650	.8936	1	.8992	1	.8909
	0.95	.2859	.9499	.3195	.9438	1	.9507	1	.9440
	0.99	.0454	.9899	.0564	.9877	.3410	.9893	.4145	.9872
0.01 ($n = 9$) $n^* = 11$	0.90	.2842	.8991	.3372	.8759	.7489	.8995	.8530	.8635
	0.95	.1495	.9497	.2153	.9255	.4911	.9508	.6603	.9219
	0.99	.0284	.99	.0544	.9813	.1276	.9904	.2198	.9820

Tabela 7.7: Modelo Pareto(5) com $1 < n < n^*$ (100000 replicas)

		caso 1: $\varphi_s^{[1]} = \varphi_e^{[1]} = 0.95$				caso 2: $\varphi_s^{[1]} = \varphi_e^{[1]} = 0.995$			
		T_1		T_2		T_1		T_2	
p	$1 - \alpha$	$\varphi_s^{[n]}$	$\varphi_{s_1}^{[n]}$	$\varphi_s^{[n]}$	$\varphi_{s_1}^{[n]}$	$\varphi_s^{[n]}$	$\varphi_{s_1}^{[n]}$	$\varphi_{s_1}^{[n]}$	$\varphi_e^{[n]}$
0.15 ($n = 2$) $n^* = 3$	0.9	.7442	.9019	.7672	.8989	1	.9006	1	.8982
	0.95	.3275	.9502	.3341	.9492	1	.9884	1	.9475
	0.99	.0328	.9897	.0356	.9891	.6258	.9898	.7068	.9894
0.10 ($n = 2$) $n^* = 4$	0.90	.7371	.9012	.7607	.8982	1	.8993	1	.8974
	0.95	.3109	.9492	.3248	.9473	1	.9502	1	.9466
	0.99	.0318	.9902	.034	.9898	.6375	.9899	.6936	.9896
0.05 ($n = 3$) $n^* = 5$	0.90	.4936	.8995	.5343	.8911	1	.9004	1	.8943
	0.95	.2216	.9498	.2437	.9455	1	.9504	1	.9458
	0.99	.025	.9904	.0287	.9891	.2573	.9905	.3010	.9895
0.01 ($n = 9$) $n^* = 11$	0.90	.2378	.8985	.2730	.8832	.8575	.8989	.9202	.8799
	0.95	.1158	.9496	.1438	.9373	.5192	.9504	.6398	.9371
	0.99	.0154	.9896	.0205	.9869	.0784	.99	.1087	.9871

Tabela 7.8: Modelo Pareto(3) com $1 < n < n^*$ (100000 replicas)

		caso 1: $\varphi_s^{[1]} = \varphi_e^{[1]} = 0.95$				caso 2: $\varphi_s^{[1]} = \varphi_e^{[1]} = 0.995$			
		T_1		T_2		T_1		T_2	
p	$1 - \alpha$	$\varphi_s^{[n]}$	$\varphi_{s_1}^{[n]}$	$\varphi_s^{[n]}$	$\varphi_{s_1}^{[n]}$	$\varphi_s^{[n]}$	$\varphi_{s_1}^{[n]}$	$\varphi_{s_1}^{[n]}$	$\varphi_e^{[n]}$
0.15 ($n = 2$) $n^* = 3$	0.9	.7053	.8993	.7084	.8992	1	.9012	1	.9005
	0.95	.1211	.9491	.1247	.9488	1	.9505	1	.9507
	0.99	.0109	.9905	.0109	.9905	.4357	.9903	.4228	.9903
0.10 ($n = 2$) $n^* = 4$	0.90	.6879	.9013	.7150	.9005	1	.9005	1	.9005
	0.95	.1151	.9502	.1159	.9500	1	.9504	1	.9506
	0.99	.0108	.9897	.0110	.9896	.3361	.9901	.4818	.99
0.05 ($n = 3$) $n^* = 5$	0.90	.2730	.8988	.2740	.8986	1	.90	1	.8990
	0.95	.0723	.9508	.0741	.9501	1	.9495	1	.9486
	0.99	.0115	.9902	.0117	.99	.0776	.9898	.0782	.9906
0.01 ($n = 9$) $n^* = 11$	0.90	.1225	.8988	.1249	.8974	1	.8987	1	.8965
	0.95	.0577	.9501	.0589	.9494	.8738	.9501	.9351	.9493
	0.99	.0095	.9903	.0096	.9902	.0109	.9899	.0110	.9898

Tabela 7.9: Modelo Pareto(I) com $1 < n < n^*$ (100000 replicas)

7.4 Conclusões

A rarefação tem um efeito preponderante na qualidade do teste composto. Quando a média e o máximo da amostra são altamente correlacionados este efeito é minimizado e a utilização de amostras compostas pode ser recomendada. Ao invés, isto é, quando a correlação é fraca, a presença de uma amostra individual infetada no grupo tem um reduzido efeito na quantidade de substância da amostra conjunta e conseqüentemente a detecção da amostra infetada é extremamente difícil.

Capítulo 8

Comentários finais

A escolha da metodologia de Dorfman (que foi, em si mesma, um salto qualitativo importante na gestão da saúde pública) para tema de investigação esteve relacionada com o facto de lecionar numa escola de saúde – Escola Superior de Tecnologias da Saúde de Lisboa, onde a temática da saúde pública é obviamente prioritária.

Neste trabalho pretendeu-se expor apenas resultados coerentes e de formulação clara, tendo sido omitido uma coleção de resultados dispersos e de comentários a simulações que se considera não terem atingido ainda o estatuto de publicáveis.

Uma das questões fundamentais abordadas neste trabalho, onde se conseguiu um avanço significativo, tem a ver com os erros nas análises combinadas, isto é, sensibilidade e especificidade de análises a uma amostra compósita usando parcelas de unidades amostrais independentes. No que toca ao caso contínuo constatámos ser um problema de difícil resolução que necessita de um horizonte temporal apreciável para se obter resultados mais consistentes.

Cada vez mais, os resultados simples do tipo positivo/negativo ou um determinado número de cruces para quantificar o estado de degradação de um órgão tendem a ser substituídos por resultados fornecidos por um número e um intervalo de referência correspondente a “dentro dos padrões normais”. Isto leva naturalmente a pensar que um “varrimento” de modelos com diferentes pesos de caudas, proporcionado por Paretos generalizadas (em que índice negativo corresponde a uma subfamília beta, com caudas leves, e o limite quando o índice tende para zero ao modelo exponencial, de cauda direita “neutra”) é um problema árduo mas que merece esforço, nomeadamente quando cada vez mais é necessário poupar recursos, usando testes com amostras combinadas.

BIBLIOGRAFIA

- AGRESTI, A. , COULL, B.A. (1998). Approximate is better than “exact” for interval estimation of binomial proportions, *The American Statistician*, **52** (2), 119-126.
- BILDER, C.R., ZHANG, B., SCHAARSEHMIDT, F., TEBBS, J. (2010). Bingroup: a package for group testing, *The R Journal*, **2**, 56-60.
- BLAKER, H. (2000). Confidence curves and improved exact confidence intervals for discrete distributions, *The Canadian Journal of Statistics*, **28** (4), 783-798.
- BLYTH, C.R., STILL, H.A. (1983). Binomial confidence intervals. *Journal of the American Statistical Association*, **78**, 108-116.
- BOSWELL, M.T., GORE, S.D., LOVISON, G. E PATIL, G.P. (1996). Annotated bibliography of composite sampling, Part A: 1936–92, *Environmental and Ecological Statistics*, **3**, 1-50.
- BRILHANTE, M.F. (1999). Inferência Estatística em Modelos não Gaussianos com Recurso a Spacings e Outras Funções de Estatísticas Ordinais, Universidade dos Açores, Ponta Delgada.
- BROWN, L.D., CAI, T., DASGUPTA, A. (2003). Interval estimation in exponential families, *Statist. Sinica*, **13**, 19-49.
- BROWN, L.D., CAI, T., DASGUPTA, A. (2002). Confidence Intervals for a Binomial Proportion and Edgeworth expansions, *The Annals of Statistics*, **30** (1), 160-201.
- BROWN, L.D., CAI, T., DASGUPTA, A. (2001). Interval Estimation for a binomial proportion, *Statistical Science*, **16** (2), 101-133.
- CAI, T.T. (2005). One-sided confidence intervals in discrete distributions, *Journal of Statistical Planning and Inference*, **131**, 63-88.
- CASELLA, G. (1986) Refining binomial confidence intervals. *Canadian Journal of statistics*, **14**, 113-129.
- CHEN, C.L. , SWALLOW, W.H. (1990). Using group testing to estimate a proportion, and to test the binomial model, *Biometrics*, 1035-1046.
- CLOPPER, C.J. , PEARSON, E.S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial, *Biometrika*, **26**, 404-413.
- CROW, E.L. (1956). Confidence Intervals for a proportion, *Biometrika*, **43**, 423-435
- DORFMAN, R. (1943). The detection of defective members in large populations, *Ann. Math. Statistics*, **14**, 436–440.
- ENGEN, S. (1974). On species frequency models, *Biometrika*, **61**, 263–270.

- FELLER, W. (1968). An Introduction to Probability Theory and Some of its Applications, vol. II, Willey, New York.
- FINUCAN, H.M. (1964). The blood testing problem, *Applied Statistics*, **13**, 43-50.
- GARNER, F.C., STAPANIAN, M.A., YFANTIS, E.A., WILLIAMS, L.R. (1989). Probability estimation with sample compositing techniques, *Journal of Official Statistics* 5, 365-374.
- GILL, A., GOTTLIEB, D. (1974). The identification of a set by successive intersections, *In Information and control*. Ellis Horwood, Chichester, 20-35.
- GUT, ALLAN (1995) An Intermediate Course in Probability. Springer-Verlag.
- HESS, K.T.; LIEWALD, A. E SCHMIDT, K.D. (2002). An extension of Panjer's recursion, *Astin Bulletin* **32**, 283-297.
- HUNG, M., SWALLOW, W. H. (1999). Robustness of group testing in the estimation of proportions, *Biometrics*, **55**, 231-237.
- HWANG, F.K. (1976) Group testing with a dilution effect, *Biometrics*, **63**, 671-673.
- JOHNSON, N.L.; KEMP, A.W. AND KOTZ, S. (2005), *Univariate Discrete Distributions*, 3rd Edition, John Wiley & Sons.
- KATZ, L. (1965). Unified treatment of a broad class of discrete probability distribution, *Classical and Contagious Discrete Distribution*, G.P. Patil (ed.), Pergamon Press, Oxford, 175-182.
- KIM, H., HUDGENS, M., DREYFUSS, J., WESTREICH, D., PILCHER, C. (2007). Comparison of group testing algorithms for case identification in the presence of testing errors, *Biometrics*, **63**, 1152-1163.
- LAPLACE, P.S. (1812). *Théorie analytique des probabilités*, Paris: Courcier.
- LIU, S.C., CHIANG, K.S., LIN, C.H., CHUNG, W.C., LIN, S.H., YANG, T.C. (2011). Cost analysis in choosing group size when group testing for potato virus Y in the presence of classification errors, *Ann. Appl. Biology*, **159**, 491-502
- LOYER, M.W. (1983). Bad probability, good statistics, and group testing for binomial estimation, *The American Statistician* **37**, 57-59.
- MARTINS, J.P, SANTOS, R. E SOUSA, R.(2012) Testing the maximum by the mean in quantitative group tests. Selected Papers of SPE 2011 (aceite).
- PANJER, H.H.(1981). Recursive evaluation of a family of compound distributions, *Astin Bulletin* **12**, 22-26.

- PESTANA, D. E VELOSA, S. (2006). Introdução à Probabilidade e à Estatística, vol I, 2.^a edição, Fundação Calouste Gulbenkian.
- PESTANA, D., VELOSA, S. (2004). Extensions of Katz-Panjer families of discrete distributions, *REVSTAT* **2**, 145-162.
- PIRES, A.M., AMADO, C.. (2008). Interval for a Binomial proportion: Comparison of twenty methods. *REVSTAT- Statistical Journal*, **6**(2), 165-197.
- SAMUELS, S.M. (1978). The exact solution of the two-stage group-testing problem, *Technometrics*, **20**, 497-500
- SANTOS, R., PESTANA, D., MARTINS, J.P. (2012). Extensions of Dorfman's theory, *Selected Papers of SPE 2010* (aceite).
- SOBEL, K.M., ELASHOFF, R.M. (1975) Group testing with a new goal, estimation *Biometrika*, **62**, 181-193
- SOBEL, M., GROLL, P.A. (1959) Group testing to eliminate efficiently all defectives in a binomial sample, *Bell system Technical Journal*, **38**, 1179-1252
- SOUSA, R. (2005). Análises ao sangue conjuntas. Uma estratégia para redução de custos, *Actas da IV Conferência Estatística e Qualidade na Saúde*, 190–196.
- SOUSA, R. (2006). Análises ao sangue conjuntas. Uma consideração mais realista, *Actas da V Conferência Estatística e Qualidade na Saúde*, 129–133.
- SOUSA, R. (2008). O Problema de Dorfman Revisitado - Análises Quantitativas, *Actas da VI Conferência Estatística e Qualidade na Saúde*, 140–144.
- STERNE, T.E. (1954). Some remarks on confidence or fiducial limits, *Biometrika*, **41**, 117-129
- STERRET, A. (1957). On the detection of defective members of large populations, *Ann. Math. Statistics*, **28**, 1033–1036.
- SUNDT, B., JEWELL, W.S. (1981). Further results on recursive evaluation of compound distributions, *Astin Bulletin* **12**, 27-39.
- TU, X.M., LITVAK, E., PAGANO, M. (1995). On the informativeness and accuracy of pooled testing in the estimating prevalence of a rare disease: application to HIV screening, *Biometrika*, **82**, 287-297.
- TU, X.M., LITVAK, E., PAGANO, M. (1994). Studies of AIDS and HIV surveillance, screening tests: can you get more by doing less?, *Statistics in Medicine*, **13**, 1905-1919.
- Willmot, G.E. (1987). Sundt and Jewell's family of discrete distributions, *Astin Bulletin* **18**, 17-29.

WILSON, E.B. (1927). Probable Inference, the Law of Succession, and Statistical Inference, *Journal of the American Statistical Association*, **22**(158), 209-212.