UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

CONTRIBUTOS PARA O ESTUDO DE DADOS EM FALTA

Maria Fernanda Nunes Diamantino

DOUTORAMENTO EM ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL
(Especialidade de Probabilidades e Estatística)

2008

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

CONTRIBUTOS PARA O ESTUDO DE DADOS EM FALTA

Maria Fernanda Nunes Diamantino

Tese orientada pelo Professor Doutor Dinis Duarte Pestana

DOUTORAMENTO EM ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

(Especialidade de Probabilidades e Estatística)

2008

A ti... MÃE

Agradecimentos

Uma lista de agradecimentos engloba os que nela estão presentes e todos aqueles que, embora não presentes ("um dado em falta"), por alguma razão, lá deveriam estar... Não estão aqui no papel, mas estão no meu coração!

- ◇ Em primeiro lugar quero expressar os meus profundos agradecimentos ao meu orientador, Professor Doutor Dinis Pestana — de quem recebi os primeiros ensinamentos em Probabilidades e Estatística — que, para além da excelente orientação científica, me ofereceu, incondicionalmente, todo o apoio, incentivo e compreensão. Jamais esquecerei o seu empenho pessoal e as suas palavras amigas.

- ◇ Aos meus saudosos amigos Professor Doutor Orlando Oliveira e Professor Doutor José Rocha. A qualquer um deles devo a gentileza de aceitarem ser co-orientadores da minha investigação, facto que a vida breve de um e de outro veio a frustrar.

- ◇ Às Professoras Doutoras Fátima Brilhante e Sandra Mendonça e à Dra. Sandra Aleixo que participaram no desenvolvimento de resultados sobre rarefação e a sua pertinência em amostragem apresentados nesta tese.

- ◇ À Dra. Sandra Aleixo tenho ainda que agradecer a participação e incentivo no uso da família da gaussiana generalizada para estudos de simulação e as muitas palavras amigas que nunca esquecerei.

- ◇ Ao Dr. João Paulo Martins a partilha de resultados usando desenvolvimentos assintóticos, cujo estudo iniciei há alguns anos na perspectiva de trabalhar com o Professor Doutor José Rocha, mas em que o Dr. João Paulo Martins se adiantou.

- ◇ À Professora Doutora Carla Kulberg pela disponibilidade de dados administrativos.

- ◇ À Dra. Aldina Vieira pela simpatia e disponibilidade com que sempre me recebeu.

- ◇ Ao aluno de Cabo Verde que foi contratado e que prefere não ser identificado.

◇ Um agradecimento especial a todos os que foram meus alunos e que, sem o saberem, me deram um incentivo muito particular para continuar o meu trabalho de investigação.

◇ A todos os meus colegas e amigos do Departamento de Estatística e Investigação Operacional que me apoiaram, das mais diversas formas.

◇ Ao Centro de Estatística e Aplicações da Universidade de Lisboa pelos meios computacionais e documentação disponibilizados.

◇ Às Professoras Doutoras Lisete Sousa e Marília Antunes pelas longas conversas que me ajudaram, não só na realização desta tese, mas também nos tempos menos bons que atravessei.

◇ A todos os meus Amigos e Amigas, uns sempre por perto, outros mais afastados, pelas suas palavras de incentivo e de apoio.

◇ Um agradecimento muito especial ao meu pai, à Olívia e à minha irmã pelo incentivo, apoio e carinho.

◇ Ao meu filho Jorge e ao meu filho José, um agradecimento do fundo do meu coração — foram eles a minha fonte de força e coragem — pela partilha diária, pelo apoio e, principalmente, por todo o seu Amor.

A todos o meu,
Muito Obrigada!

Resumo

O Capítulo 1 (Preliminares I — O método delta; amostragem) apresenta o método delta e algumas questões de amostragem, essenciais para abordar as implicações de não-resposta.

No Capítulo 2 (Rarefação e não-resposta) o método delta é usado para analisar a eficiência de diversos estimadores que procuram contornar a não-resposta global a inquéritos, nomeadamente recorrendo à teoria da rarefação de Rényi.

No capítulo 3 (A família GLE) é estudada a família GLE (Gaussiana — Laplace Estendida). São apresentadas propriedades estruturais desta família tendo em conta o respectivo suporte e descrevem-se os vários métodos usados para obter números pseudo-aleatórios de populações GLE.

O Capítulo 4 (Preliminares II — Cumulantes, localização, escala, (as)simetria e curtose) é um capítulo de preliminares, com uma resenha necessária de conhecimentos sobre parâmetros importantes no estudo da robustez da localização e escala.

No Capítulo 5 (A inferência sobre localização revisitada) apresenta-se um estudo visando investigar a robustez na studentização em populações não gaussianas, bem como a análise de escala.

No Capítulo 6 (Bem estar e progresso académico dos alunos da FCUL provenientes de PALOP), um capítulo de índole mais prática, estudamos alguns aspectos que envolvem os estudantes africanos da FCUL. Abordam-se algumas directrizes que devemos ter em conta ao elaborar um questionário e comenta-se o inquérito piloto efectuado aos estudantes de PALOP. Finalmente, apresenta-se uma análise do percurso académico destes alunos feita com base em dados administrativos, e um breve estudo de imputação.

Palavras-chave: Amostragem, não-respostas, rarefação, localização e escala, imputação.

Abstract

Chapter 1 (Preliminaries I — The delta method; sampling) contains known results on the delta methods and on sampling, that are useful in the main body of the thesis. The presentation is lean, namely in the choice of sampling results, where the main purpose has been to enhance the theoretical importance of non-response.

In Chapter 2 (Rarefaction and nonresponse) the delta method is used to analyse the efficiency (in the sense of variance evaluation) of various estimators aimed at dealing with global survey nonresponse, namely when the fraction of nonresponse is high and Rényi's rarefaction theory seems to provide an useful alternative to the usual methods.

In Chapter 3 (The EGL family) the Extended Gaussian–Laplace family is studied. Structural properties are presented and exploited to generate pseudo-random numbers of EGL populations.

Chapter 4 (Preliminaries II — Cumulants, location, scale, (as)simetry and kurtosis) is a chapter of presenting useful information on moments and cumulants.

Chapter 5 (Location inference revisited) presents a study that investigate the robustness of studentization techniques in non gaussian populations; scale analysis is also briefly analyzed in the light of robustness.

Chapter 6 (Welfare and academic progress of PALOP students at FCUL) is a chapter where some aspects that involve our PALOP students are analysed. The elaboration of a questionnaire are presented and a pilot questionnaire applied to PALOP students are commented. Finally, an analysis of their academic progress and a preliminary evaluation of imputation techniques is performed.

Keywords: Sampling, nonresponse, rarefaction, localization and scale, imputation.

Conteúdo

Lista de Figuras

Lista de Tabelas

Antes de começar — um guia de leitura

Há uma forte componente de Matemática em todo o trabalho estatístico, mas a Estatística não é apenas um ramo da Matemática. Em particular, a obtenção de dados por amostragem, nomeadamente quando se pretende obter informação mais reservada sobre populações humanas é uma área em que a integração de saberes de muitos outros campos — nomeadamente Ciências Cognitivas, Psicologia, Sociologia, para citar apenas alguns — devidamente cimentados pelo bom senso, é indispensável.

Depois de muitos acidentes de percurso, cuja descrição de detalhe parece nesta fase extemporânea, o trabalho derivou para o estudo das dificuldades — sociais e estudantis — dos estudantes de PALOP.

No entanto, parte substancial da investigação feita tem em mente a avaliação da robustez de estimadores de médias — eventualmente de 0's e 1's, e portanto, proporções — ou de diferenças de efeitos medidas através de diferenças de médias. Portanto, inevitavelmente, a escala é também objecto de reflexão.

Acresce ainda que a construção desta tese teve que aproveitar trabalhos em que a autora colaborou (e daí uma dose, que se espera ser de dimensão aceitável, de repetições, em que tentámos de, qualquer modo, alguma variedade de apresentação).

Assim, este trabalho tem como objectivo estudar questões que têm que ver com o papel central que as médias ocupam em inferência estatística, uma questão vasta abordada com maior generalidade em Brilhante *et al.* (2001), e em que a nossa perspectiva se limita a questões associadas a studentização (com uma incursão na questão de análise de escala em populações não gaussianas), e às questões de não-resposta que se colocam em todas as sondagens.

No Capítulo 1 (Preliminares I) apresenta-se o essencial sobre o

"método delta", e uma breve resenha sobre estratégias de amostragem; estes preliminares apresentam questões relevantes para o exposto no Capítulo seguinte.

No Capítulo 2 (Rarefação e Não-Resposta) aborda-se de forma nova a questão da determinação da dimensão da amostra de forma a torná-la adequada para obtenção da precisão desejada, quando se sabe que muitos dos inquiridos não fornecem a informação solicitada, a ponto de se atingir situações que se adequam à rarefação de Rényi (1956). Estes resultados foram obtidos com S. Aleixo, F. Brilhante, S. Mendonça e D. Pestana.

Tendo em conta resultados de Rényi sobre rarefação e de Kozubowski sobre emagrecimento (*thinning*, filtragem), investigámos a estimação de parâmetros populacionais usando uma alternativa diferente da solução *ad-hoc* habitualmente explorada: em vez de aumentar a dimensão da amostra deterministicamente para contrariar o efeito de filtragem (não-respostas), estudámos a situação em que, em vez do T.L.C. clássico e aproximação por gaussiana, se pode usar $\mathcal{N}$-divisibilidade infinita no esquema de somas aleatórias subordinadas pela geométrica (geo-somas), e o correspondente limite exponencial ou, mais geralmente, Laplace (no caso de parente com variância finita).

O estudo da qualidade dos estimadores obtidos usa uma família "dobrada" (*folded*) que generaliza a exponencial e a gaussiana dobrada, associada a uma família "gaussiana generalizada" anteriormente usada por Diamantino e Pestana (1997) para investigar questões de robustez na estatística studentizada $T_{(n-1)} = \sqrt{n(n-1)}\frac{\overline{X}_n}{\sqrt{SS_n}}$.

Essas duas famílias "GLE" (de Gaussiana — Laplace Estendida) são o objecto do Capítulo 3, que discute formas de gerar números pseudo-aleatórios com aquelas distribuições.

No Capítulo 4 (Preliminares II) apresentamos questões sobre cumulantes, relevantes para o desenvolvimento do Capítulo 5, em que o método delta é de novo usado intensivamente para concluir que a studentização em populações não gaussianas funciona melhor na inferência sobre diferenças de médias do que na inferência sobre uma média. A explicação está decerto associada ao observado por Efron (1968): a simetrização tem um papel mais relevante do que lhe é, em geral, creditado. Este trabalho deve muito a J. Rocha, e tem por isso intersecção com resultados obtidos por J. P. Martins, por ele co-orientado nas provas de Mestrado.

No Capítulo 6 ocupamo-nos dos nossos alunos africanos, discutindo algumas vertentes não matemáticas da Estatística, o trabalho piloto feito com a colaboração de um recém licenciado do DEIO na fase em que

preparava o seu regresso a Cabo Verde, e questões que se prendem com o uso de fontes administrativas na recolha de informação; fazemos ainda uma abordagem preliminar ao importante tema de imputação, um dos tratamentos possíveis a dados omissos.

Capítulo 1

Preliminares I — O método delta; amostragem

No presente capítulo fazemos uma brevíssima apresentação do método delta, que será usado já no próximo capítulo, para obtenção de estimadores de médias no caso de anteciparmos não-resposta e quermos limitar os seus efeitos, por vezes devastadores, e em questões de amostragem que põem a tónica no facto de as probabilidades de inclusão de elementos da população na amostra ser fulcral em todo o tipo de amostragem. Havendo excelentes obras sobre estes temas, não houve necessidade de ir longe, apenas se referindo algumas disciplinas fundamentais; assim, optou-se por detalhar o estimador clássico de Horwitz-Thompson, que consideramos a matriz de todos os outros que surgem quando se escolhe uma estratégia de amostragem.

1.1 O método delta

O método delta consiste, em traços gerais, em truncar a expansão em série de Taylor de uma função com vista à obtenção de aproximações, nomeadamente, para os momentos de uma estatística de interesse. A prática habitual consiste em truncar a expansão após o primeira derivada. No caso de esta se anular no ponto em que é efectuada a expansão recorre-se ao termo correspondente à segunda derivada. Para demonstrar o principal teorema que suporta este método é necessário apresentar parte do enunciado do teorema de Slutsky.

Teorema 1.1. *(Slutsky)*

Sejam $\{X_n\}_{n\in\mathbb{N}}$ *e* $\{Y_n\}_{n\in\mathbb{N}}$ *duas sucessões de variáveis aleatórias e* $c \in \mathbb{R}$ *tais que* $X_n \underset{n\to\infty}{\overset{d}{\to}} X$ *e* $Y_n \underset{n\to\infty}{\overset{p}{\to}} c$ *então:*

1. $X_n + Y_n \underset{n\to\infty}{\overset{d}{\to}} X_n + c$
2. $X_n Y_n \underset{n\to\infty}{\overset{d}{\to}} cX$

A demonstração deste teorema pode ser consultada em Pestana e Velosa (2008, p. 997–998).

Passamos agora ao teorema principal.

Teorema 1.2. *(Método Delta)*

Dada $\{X_n\}_{n\in\mathbb{N}}$ *uma sucessão de variáveis aleatórias tal que*

$$\sqrt{n}\,(X_n - \theta) \underset{n\to\infty}{\overset{d}{\to}} Z \frown \text{Gau}\,(0, \sigma)\,,$$

uma função f *e* θ *fixo tal que* $f'(\theta)$ *existe e é diferente de zero, então*

$$\sqrt{n}\,[f(X_n) - f(\theta)] \underset{n\to\infty}{\overset{d}{\to}} f'(\theta)\, Z \frown \text{Gau}\,(0, \sigma\,|f'(\theta)|)\,.$$

Demonstração. A expansão de $f(X_n)$ em série de Taylor de ordem 1 em torno de θ é

$$f(X_n) = f(\theta) + f'(\theta)\,(X_n - \theta) + R_1(X_n)$$

onde $R_1(X_n) \underset{n\to\infty}{\overset{p}{\to}} 0$ quando $X_n \underset{n\to\infty}{\overset{p}{\to}} \theta$.
Logo, as seguintes expressões são assimptoticamente iguais:

$$\sqrt{n}\,[f(X_n) - f(\theta)] \approx \sqrt{n} f'(\theta)\,(X_n - \theta)\,. \tag{1.1}$$

Recorrendo ao Teorema de Slutsky é possível afirmar que

$$\sqrt{n} f'(\theta)\,(X_n - \theta) \underset{n\to\infty}{\overset{d}{\to}} f'(\theta)\, Z \frown \text{Gau}\,(0, \sigma\,|f'(\theta)|)$$

donde, utilizando a aproximação (1.1) obtém-se o resultado pretendido. □

As hipóteses do teorema podem ser relaxadas, nomeadamente, a constante normalizadora não tem de ser necessariamente $\sqrt{n}$ (Bishop *et al* (1975)).

Há duas extensões do método delta bastante importantes. Uma já foi mencionada no início e é relativa à suposição de $f'(\theta) = 0$. A outra relaciona-se com o caso em que f é uma função multivariada. Uma panorâmica sobre o tema pode ser encontrada em Chandra (1999).

O primeiro caso é suportado pelo teorema que se segue.

Teorema 1.3. *(Método Delta de Segunda Ordem)*
Dadas $\{X_n\}_{n\in\mathbb{N}}$ *uma sucessão de variáveis aleatórias tal que*

$$\sqrt{n}\,(X_n - \theta) \xrightarrow[n\to\infty]{d} Z \frown \text{Gau}\,(0, \sigma)\,,$$

uma função f *e* θ *fixo tal que* $f'(\theta) = 0$ *e* $f''(\theta)$ *existe e é diferente de zero, então*

$$n\,[f\,(X_n) - f\,(\theta)] \xrightarrow[n\to\infty]{d} \sigma^2 \frac{f''\,(\theta)}{2} K$$

onde $K \frown \chi^2_1$.

Demonstração. A expansão de $f\,(X_n)$ em série de Taylor de ordem 2 em torno de θ é

$$f\,(X_n) = f\,(\theta) + f'\,(\theta)\,(X_n - \theta) + \frac{f''\,(\theta)}{2}\,(X_n - \theta)^2 + R_2\,(X_n)$$

onde $R_2\,(X_n) \xrightarrow[n\to\infty]{p} 0$ quando $X_n \xrightarrow[n\to\infty]{p} \theta$.
Como $f'\,(\theta) = 0$,

$$f\,(X_n) - f\,(\theta) \approx \frac{f''\,(\theta)}{2}\,(X_n - \theta)^2 \tag{1.2}$$

Ora, o quadrado de uma variável aleatória com distribuição Gau$(0, 1)$ é uma variável aleatória com distribuição χ^2_1, donde

$$n\frac{(X_n - \theta)^2}{\sigma^2} \frown \chi^2_1.$$

Portanto, tal como na demonstração anterior, o Teorema de Slutsky garante que

$$\frac{n f''\,(\theta)}{2}\,(X_n - \theta)^2 \xrightarrow[n\to\infty]{d} \sigma^2 \frac{f''\,(\theta)}{2} K$$

onde $K \frown \chi^2_1$.
Logo, pela aproximação (1.2), obtém-se o resultado pretendido. □

Em algumas situações poderá ocorrer que $f'(\theta)$ bem como $f''(\theta)$ sejam simultaneamente nulas pelo que a aproximação $f(X_n) - f(\theta)$ deverá ser feita pela parcela da série de Taylor correspondente à primeira derivada não nula em θ. Obtém-se uma convergência em distribuição para uma potência de uma variável aleatória Normal padrão de expoente igual à ordem dessa derivada.

O tratamento do caso de uma função dependente de duas ou mais variáveis é estabelecido pelo próximo teorema. Utilizaremos a notação $\boldsymbol{X}_n = (X_{n1}, ..., X_{np})$ para representar um vector aleatório de dimensão p e a representação $\mu = (\mu_1, ..., \mu_p)$ para um vector de parâmetros da mesma dimensão.

Teorema 1.4. *Seja $\boldsymbol{X}_n$ um vector aleatório, μ vector de parâmetros com valores em D^p e $f : D^p \to \mathbb{R}^p$ uma função diferenciável. Se*

$$\sqrt{n}\left(\boldsymbol{X}_n^T - \mu^T\right) \xrightarrow[n\to\infty]{d} Z \frown \text{Gau}(\mathbf{0}, W)$$

em que $\mathbf{0}$ é o vector nulo e W é a matriz das covariâncias, então

$$\sqrt{n}\left[[f(\boldsymbol{X}_n)]^T - [f(\mu)]^T\right] \xrightarrow[n\to\infty]{d} \text{Gau}\left(\mathbf{0}, \left(\frac{\partial f}{\partial \mu}\right) W \left(\frac{\partial f}{\partial \mu}\right)^T\right)$$

Demonstração. Como

$$f_i(\boldsymbol{X}_n) = f_i(X_{n1}, ..., X_{np}) = f_i(\mu_1, ..., \mu_p) + \sum_{j=1}^{p} \frac{\partial f_i}{\partial \mu_j}(X_{nj} - \mu_j) + R_1(\boldsymbol{X}_n)$$

onde $R_1(\boldsymbol{X}_n) \to 0$ em probabilidade quando $(X_{n1}, ..., X_{np}) \to (\mu_1, ..., \mu_p)$ em probabilidade.
Em notação matricial pode-se representar

$$[f(\boldsymbol{X}_n)]^T = [f(\mu)]^T + (\boldsymbol{X}_n - \mu)\left(\frac{\partial f}{\partial \mu}\right)^T + R_1(\boldsymbol{X}_n).$$

Donde,

$$\sqrt{n}\left[[f(\boldsymbol{X}_n)]^T - [f(\mu)]^T\right] \approx \sqrt{n}(\boldsymbol{X}_n - \mu)\left(\frac{\partial f}{\partial \mu}\right)^T.$$

Portanto, por uma generalização imediata do teorema de Slutsky,

$$\sqrt{n}\left[[f(\boldsymbol{X}_n)]^T - [f(\mu)]^T\right] \xrightarrow[n\to\infty]{d} Gau\left(\mathbf{0}, \left(\frac{\partial f}{\partial \mu}\right) W \left(\frac{\partial f}{\partial \mu}\right)^T\right)$$

como se queria demonstrar. □

Mais detalhes sobre o Método Delta aplicado a funções multivariadas podem ser encontrados em Lehman e Casella (2003).

1.2 Uma Pequena Amostra (de Conveniência, Claro!) de Amostragem

1.2.1 Generalidades

Denotamos $\mathcal{P}$ uma população de dimensão N^*, e $\mathcal{S}$ um plano amostral, isto é, uma função que a cada subconjunto $\boldsymbol{s} \subset \mathcal{P}$ atribui uma probabilidade $\mathbb{P}(\boldsymbol{s})$ de ser seleccionado como amostra.

Consideramos apenas planos amostrais *próprios*, aqueles em que a probabilidade $\pi_{i^\star}$ de selecção do $i^\star$-ésimo elemento da população como elemento da amostra, $\forall i^\star \in \{1, \ldots, N^*\}$, é positiva (por outras palavras, num plano amostral *impróprio* há elementos da população que ficam definitivamente excluídos do estudo, por terem probabilidadde nula de serem incluídos na amostra, e tal não nos interessa). Os recíprocos das probabilidades de inclusão, $w_{i^\star} = \frac{1}{\pi_{i^\star}}$, são os *pesos de amostragem*.

Para simplicidade de exposição, admita-se que nos interessa estimar a média populacional μ da variável Y. Denotamos os seus valores na população $Y_{1^\star}, Y_{2^\star}, \ldots, Y_{N^\star}$, e os seus valores na amostra $y_1, y_2, \ldots, y_n$. Note-se que a dimensão n da amostra pode ser um valor fixo ou variável e, neste caso, inclusivamente aleatório. Por outro lado colocamo-nos numa perspectiva *design based* da amostragem (por contraposição a uma perspectiva *model based*), isto é não consideramos Y uma variável aleatória, a aleatoriedade da amostra deriva apenas das probabilidades de inclusão $\pi_{i^\star} = \mathbb{P}\left[Y_{i^\star} = y_k \in \boldsymbol{s}\right]$.

Denotando $\mathrm{I}(i^\star)$ o indicador de que o $i^\star$-ésimo membro da população é seleccionado para a amostra, a probabilidade de inclusão é

$$\pi_{i^\star} = \mathbb{E}\left[\, \mathrm{I}(i^\star)\right].$$

Por outro lado, a probabilidade de inclusão $\pi_{i^\star}$ é a probabilidade dos subconjuntos de $\mathcal{P}$ de que esse elemento é membro:

$$\pi_{_{i^\star}} = \sum_{\left\{\boldsymbol{s}\subset\mathcal{P}:\, Y_{_{i^\star}}\in\boldsymbol{s}\right\}} \mathbb{P}\left(\boldsymbol{s}\right).$$

Esta expressão geral tem pouca utilidade prática, excepto nas situações em que o plano amostral implica simetrias que tornam aquele cálculo fácil (por exemplo, em amostragem aleatória simples sem reposição, a probabilidade de inclusão de qualquer elemento de uma população de dimensão N numa amostra de dimensão n [(1)] é $\pi_{_{i^\star}} = \pi = \frac{n}{N}$).

Retomando o problema de estimar a média populacional $\mu = \frac{1}{N}\sum_{i^\star=1}^{N} Y_{_{i^\star}}$ com base na amostra $(y_1, \ldots, y_n)$. O estimador [(2)] intuitivo de μ é a média amostral $\tilde{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i$ que pode facilmente ser expresso em termos dos valores populacionais:

$$\tilde{\mu} = \frac{\sum_{i^\star=1}^{N} \mathrm{I}(i^\star)\, Y_{_{i^\star}}}{\sum_{i^\star=1}^{N} \mathrm{I}(i^\star)}$$

Sublinhamos desde já que o cálculo do valor médio e da variância deste estimador pode ser extremamente complicado quando o denominador é aleatório — e no nosso trabalho frequentemente teremos que recorrer ao método delta [(3)] para obter aproximações, quando tal é o caso. É frequente ver esta dificuldade "resolvida" condicionando na dimensão da amostra efectivamente recolhida ou arredondando às unidades o valor esperado do denominador.

Admita-se, de momento, que a dimensão da amostra é um valor fixo n (isto é, que apenas subconjuntos de dimensão n têm probabilidade positiva de ser seleccionados). Neste pressuposto, e denotando $\pi_{_{i^\star j^\star}}$ a

(1) Que se caracteriza por igual probabilidade $\frac{1}{\binom{N}{n}}$ de extracção de qualquer possível subconjunto de n elementos como amostra de dimensão n.

(2) Usamos o termo estimador e não estimativa porque consideramos que se trata de uma amostra aleatória, no sentido de haver probabilidades de inclusão.

(3) O método delta, que em termos simples se pode descrever como a aplicação do teorema de Slutsky à truncatura de uma expansão em série de Taylor, é abordado de forma simples em por exemplo Casella e Berger (2002); para exposições mais completas, veja-se por exemplo Bishop *et al.* (1975) ou Chandra (1999).

probabilidade de inclusão conjunta dos $i^\star$-ésimo e $j^\star$-ésimo elementos da população na amostra, com a notação simplificada $\pi_{_{i^\star}} = \pi_{_{i^\star i^\star}}$,

$$\mathbb{E}\,(\tilde{\mu}) = \frac{1}{n}\sum_{i^\star=1}^{N} \pi_{_{i^\star}}\, Y_{_{i^\star}}$$

e

$$n^2\,\mathrm{var}(\tilde{\mu}) = \sum_{i^\star=1}^{N} \pi_{_{i^\star}}\,(1-\pi_{_{i^\star}})\,Y_{_{i_\star}}^2 + \sum_{i^\star=1}^{N}\sum_{j^\star=1}^{N}(\pi_{_{i^\star j^\star}} - \pi_{_{i^\star}}\pi_{_{j^\star}})\,Y_{_{i^\star}}\,Y_{_{j^\star}} =$$

$$= \sum_{i^\star=1}^{N}\sum_{j^\star=1}^{N} \pi_{_{i^\star j^\star}}\,Y_{_{i^\star}}\,Y_{_{j^\star}} - \left(\sum_{\ell^\star=1}^{N} \pi_{_{\ell^\star}}\,Y_{_{\ell^\star}}\right)^2 .$$

O viés B do estimador $\tilde{\mu}$ é

$$B(\tilde{\mu}) = \frac{1}{n}\sum_{i^\star=1}^{N}\left(\pi_{_{i^\star}} - \frac{\mathbb{E}(n)}{N}\right) Y_{_{i^\star}}$$

quando a variável indicatriz I da inclusão dos elementos da população na amostra e Y são correlacionadas. Note-se, consequentemente, que $\tilde{\mu}$ é centrado quando no plano amostral há equiprobabilidade de inclusão de qualquer elemento da população na amostra de dimensão fixa n. Observe-se, contudo, que o preço da amostragem é um argumento que, em geral, tem precedência nas decisões sobre o plano amostral, e que há planos amostrais que proporcionam uma boa eficiência (no sentido de erro quadrático médio moderado) do estimador $\tilde{\mu}$.

1.2.2 O estimador de Horwitz-Thompson

Como $\mathbb{E}\left[\,\mathrm{I}(i^\star)\,w_{_{i^\star}}\right] = 1$, da expressão $\mathbb{E}\,(\tilde{\mu}) = \frac{1}{n}\sum_{i^\star=1}^{N} \pi_{_{i^\star}}\,Y_{_{i^\star}}$ conclui-se que se I e Y forem não correlacionadas o estimador de Horwitz-Thompson

$$\hat{\mu} = \frac{1}{N}\sum_{i^\star=1}^{N}\,\mathrm{I}(i^\star)\,w_{_{i^\star}}\,Y_{_{i^\star}} \tag{1.3}$$

é centrado.

Supondo que $Y_{i^\star}$ é o k-ésimo elemento da população $\mathcal{P}$ seleccionado para a amostra $\boldsymbol{s}$, isto é que $Y_{i^\star} = y_k$, e denotando $\pi_{i^\star} = \mathbb{P}\left[\ \mathrm{I}(i^\star)\right] = p_k$ e $w_{i^\star} = w_k$, o estimador de Horwitz-Thompson pode ser reformulado como uma soma envolvendo apenas os n elementos da amostra:

$$\hat{\mu} = \frac{1}{N}\sum_{k=1}^{n}\frac{y_k}{p_k} = \frac{1}{N}\sum_{k=1}^{n} w_k y_k. \tag{1.4}$$

A variância amostral do estimador de Horwitz-Thompson é

$$\begin{aligned}\operatorname{var}(\hat{\mu}) &= \frac{1}{N^2}\sum_{i^\star=1}^{N}\sum_{j^\star=1}^{N}\left(\frac{\pi_{i^\star j^\star}}{\pi_{i^\star}\pi_{j^\star}} - 1\right) Y_{i^\star} Y_{j^\star} \\ &= \frac{1}{N^2}\sum_{i^\star=1}^{N}\sum_{j^\star=1}^{N}\frac{\pi_{i^\star j^\star}}{\pi_{i^\star}\pi_{j^\star}} Y_{i^\star} Y_{j^\star} - \mu^2.\end{aligned}$$

No caso de planos amostrais em que a amostra tem dimensão fixa,

$$\operatorname{var}(\hat{\mu}) = \frac{1}{2\,N^2}\sum_{i^\star=1}^{N}\sum_{j^\star=1}^{N}\left(\pi_{i^\star j^\star} - \pi_{i^\star}\pi_{j^\star}\right)\left(\frac{Y_{i^\star}}{\pi_{i^\star}} - \frac{Y_{j^\star}}{\pi_{j^\star}}\right)^2, \tag{1.5}$$

porque em amostras em que a dimensão é fixa, $\displaystyle\sum_{j^\star=1}^{N}\left(\pi_{i^\star j^\star} - \pi_{i^\star}\pi_{j^\star}\right) = 0.$

A expressão (1.5) mostra que em planos amostrais em que as probabilidades de inclusão $\pi_{i^\star}$ são proporcionais à magnitude de $Y_{i^\star}$ a variância do estimador de Horwitz-Thompson é baixa.

No caso de $Y_{i^\star} = y_k$ e $Y_{j^\star} = y_\ell$ e no caso de planos amostrais em que as probabilidades de inclusão conjunta de $(Y_{i^\star}, Y_{j^\star})$ são positivas

$$\pi_{i^\star j^\star} = \mathbb{P}\left[\ \mathrm{I}(i^\star)\ \mathrm{I}(j^\star)\right] = p_{k\ell} > 0$$

(por outras palavras: o plano amostral em $\mathcal{P}\times\mathcal{P}$ é próprio), o estimador de $\operatorname{var}(\hat{\mu})$ é

$$\widehat{\text{var}\,(\hat{\mu})} = \frac{1}{N^2} \sum_{i^\star=1}^{N} \sum_{j^\star=1}^{N} \frac{1}{\pi_{i^\star j^\star}} \left(\frac{\pi_{i^\star j^\star}}{\pi_{i^\star} \pi_{j^\star}} - 1 \right) \mathrm{I}(i^\star) Y_{i^\star} \, \mathrm{I}(j^\star) Y_{j^\star}$$

ou seja

$$\widehat{\text{var}\,(\hat{\mu})} = \frac{1}{N^2} \sum_{k=1}^{n} \sum_{\ell=1}^{n} \left(\frac{1}{p_k p_\ell} - \frac{1}{p_{k\ell}} \right) y_k y_\ell. \tag{1.6}$$

Indubitavelmente o estimador de Horwitz-Thompson tem uma universalidade que o torna atraente, no sentido em que é facilmente usável com qualquer plano amostral próprio (em $\mathcal{P} \times \mathcal{P}$). Mas, por outro lado, é fácil exibir algumas patologias pouco agradáveis deste estimador. Por exemplo, se $Y_{i^\star} = \mu$ e considerarmos um plano amostral com probabilidades de inclusão desiguais e dimensão amostral $n = 1$,

$$\hat{\mu} = \frac{\mu}{N p_1}$$

e

$$\text{var}\,(\hat{\mu}) = \frac{\mu^2}{N^2} \left(\sum_{k=1}^{N} 1_{p_k} - 1 \right) > 0$$

(porque com probabilidades de inclusão desiguais a média aritmética e a média harmónica são necessariamente diferentes).

Evidentemente, se na amostra figurarem elementos da população com probabilidades de inclusão muito baixas, a estimativa da média pode ser desmedidamente inflacionada, e o efeito de probabilidades de inclusão conjunta ainda mais baixas leva a que o peso que esses pares têm na estimação da variância possa ter um papel semelhante ao que nos faz temer a presença de *outliers* em muitas áreas de análise de dados. É este, provavelmente, o maior óbice à utilização geral do estimador de Horwitz-Thompson e a consequente limitação do seu uso a planos amostrais em que os pesos de amostragem sejam equilibradamente controlados.

1.2.3 Comentários breves sobre algumas estratégias amostrais e soluções engenhosas para problemas amostrais complexos

A *amostragem aleatória simples sem reposição* tem sobre a *amostragem aleatória simples com reposição* a vantagem de menor variância obtendo-se, portanto, a precisão pretendida com menor custo; é, por outro lado, a

única forma de dar igual probabilidade de selecção a todos os subconjuntos com a mesma dimensão n da população. Claro que ocasionalmente importa dar maior probabilidade de selecção a subconjuntos especiais, por exemplo, quando se deseja usar *amostragem com probabilidades proporcionais* à importância, por exemplo, quando a população de cidades ou as áreas de terrenos não devem ser escamoteados.

Por vezes, a dimensão da população não é conhecida e a decisão importante de base deixa de ser a determinação da dimensão da amostra, como número, e o que ganha relevo é assegurarmo-nos de que uma determinada proporção de unidades amostrais é obtida; nesse caso, a *amostragem sistemática* assegura igual probabilidade de selecção de qualquer elemento da população (mas não de qualquer amostra!); o estudo das propriedades dos estimadores decorre facilmente do estudo dos *estimadores de rácio*, o caso porventura mais importante da classe dos estimadores de regressão, uma forma engenhosa de usar a sombra para avaliar a presa, se nos é permitida a metáfora que traduz o âmago da Estatística.

Amostragem e recenseamento podem ser usadas em simbiose aproveitando o facto de a variabilidade inter-grupos e entre grupos poder contribuir para baixar o custo da recolha da informação necessária para atingir a precisão que se pretende. Na *amostragem estratificada* começa-se pelo recenseamento de uma particção do universo num número, em geral, diminuto de estratos coesos, dentro dos quais a variância do estimador que se vai usar é diminuta, usando-se depois o teorema da probabilidade total para caracterizar o estimador global; na *amostragem por grupos*, pelo contrário, usa-se em primeiro lugar amostragem, por vezes multietápica, para se chegar a grupos com um número diminuto de unidades amostrais, que são recenseadas. É uma excelente estratégia para baixar o custo por unidade amostral.

Em problemas complexos, como a avaliação de *stocks* em pescas ou estudo de populações móveis e, eventualmente, fugidias (por exemplo, sem abrigo) há técnicas de *amostragem adaptativa* que permitem abandonar estratégias que não estão a dar frutos, reforçando a recolha de unidades amostrais em circunstâncias propícias. Mesmo assim, em questões particularmente complexas, como avaliação de sociedades secretas ou de imigrantes ilegais, torna-se necessário o recurso a estratégias sofisticadas, como *amostragem usando covariáveis* ou *amostragem indirecta*. Em Biologia a *amostragem por distâncias* tem conhecido desenvolvimentos com aplicações espectaculares.

A amostragem clássica tem como eixo de força a admissão de que a probabilidade a ter em conta é a probabilidade de selecção de unidades da população para a amostra (ainda que se possa argumentar que os estimadores de regressão vão mais longe do que isso no recurso a ideias probabilistas). Modernamente, em contraposição a esta amostragem *design based* tem vindo a ter sucesso, entre os investigadores de índole mais matematizante, uma abordagem *model based*, representada, por exemplo, em Särndal *et al* (1992).

1.3 Dados em Falta e Problema das Não-Respostas

A noção de não-respostas refere-se a várias situações em que não foi obtida informação (respostas, medições) dos elementos seleccionados para uma amostra. É importante aceder à taxa de não-respostas para se poder entender a sua origem, para se controlarem e se reduzirem as não-respostas em sondagens futuras, e para se poder estimar os seus efeitos nas sondagens.

Em qualquer procedimento de amostragem existem erros que podem ter diversas proveniências, estando uma delas relacionada com as não-respostas. O problema das não-respostas existe em quase todos os inquéritos, embora a sua extensão e os seus efeitos sejam diferentes de inquérito para inquérito, o que leva à necessidade de usar técnicas especiais de estimação. A definição seguinte caracteriza o conceito de não-resposta.

Definição 1.1. Seleccionada uma amostra, dizemos que estamos perante um problema de não-respostas desde que exista, pelo menos, um dos indivíduos da amostra que não forneceu toda a informação que dele se pretendia, ou seja, o indivíduo não originou um valor para as variáveis de interesse do inquérito.

Segundo Barnett (2002) as não-respostas podem surgir por diversas razões relacionadas com o tipo de informação pedida (por exemplo, factos ou opiniões), com as características dos indivíduos inquiridos (por exemplo, pessoas ou unidades administrativas) e com o método usado para obter informação (por exemplo, entrevista, questionário, telefone

ou e-mail). Tais factores estão muitas vezes inter-relacionados.

As principais causas que levam a baixas taxas de resposta são as recusas e os "fora de casa". Kish (1965) refere outras causas como a incapacidade em responder e o facto de "não encontrar"os respondentes. As recusas resultam da falta de abilidade ou falta de vontade de participar das pessoas incluídas na amostra. Várias medidas podem ser tomadas para reduzir as recusas, tais como:

* Notificação *a priori*
* Motivação dos respondentes
* Incentivos
* Desenho e implementação do questionário

Dado que as não-respostas são um problema em qualquer inquérito por amostragem é importante ter em conta o aumento das taxas de resposta e o tratamento (ajustamento) das não-respostas. Assim, uma boa estratégia de amostragem para um dado inquérito deve passar por algumas acções a serem tomadas de acordo com a fase em que se intervém.

Na fase de planeamento devem ser definidas as acções a tomar de forma a minimizar as não-respostas, ou seja, tentar reduzi-las a níveis insignificantes e/ou minimizar o enviesamento. Deverão ser tomadas em conta algumas recomendações, tais como, a insistência junto do inquirido e o cuidado com a extensão do questionário, assim como, algumas atitudes que possam minimizar os factores responsáveis pelas não-respostas. Um deles é a selecção, formação e supervisionamento dos entrevistadores que devem ser alvo do maior cuidado, uma vez que, em geral, a colaboração do inquirido depende muito da atitude do entrevistador. Outro factor é a forma como são colocadas as questões e, como já referimos, a própria extensão do questionário, bem como a escolha do método de recolha de informação, isto é, por via postal, por telefone ou por entrevista directa. Outro factor muito importante é a relação de confiança entre a instituição que é responsável pela sondagem e os inquiridos.

Na fase de recolha, ao constatar a existência de não-respostas, poderá ser efectuada uma subamostragem dos não respondentes, dado que continua desconhecida uma parcela da estimativa desejada. Mais

adiante veremos que se trata de um procedimento para controlar o problema das não-respostas possibilitando a estimação não enviesada. Apesar de ser uma boa técnica,é preciso ter em atenção que a sua implementação tem elevados custos, sobretudo quando se trata de inquéritos com alguma dimensão, dado que se exige o conhecimento da matriz de resposta completa da subamostra.

Na fase de estimação, o conhecimento do mecanismo (aleatório) das não-respostas permite-nos contornar o enviesamento devido às não-respostas, mediante a consideração das probabilidades com que cada inquirido responde ao inquérito. A modelização do mecanismo desconhecido de resposta permite, em função de informação auxiliar, proceder à estimação das referidas probabilidades e reduzir o enviesamento do estimador. Em certo sentido, a imputação é um processo semelhante na medida em que, através do uso de informação auxiliar relevante, isto é, relacionada com a variável desconhecida e com base nas semelhanças entre indivíduos , poder-se-á obter "bons" valores a imputar (estimativas para indivíduos).

Várias técnicas têm sido utilizadas para contornar os efeitos das não-respostas:

* Subamostragem dos não respondentes
* Reposição de não respondentes de um inquérito anterior
* Substituição dos não respondentes por outros elementos que se espera que venham a responder
* Estimativas subjectivas
* Análise de tendência
* Modelação do mecanismo de respostas
* Imputação de respostas

A modelação do mecanismo de respostas é o procedimento técnico que permite, aquando da estimação, incorporar e minimizar as contrariedades do processo de resposta mediante o uso das potencialidades da amostragem em duas fases.

1.4 Enviesamento Devido às Não-Respostas

Para examinar o enviesamento devido às não-respostas, consideremos a população em estudo dividida em respondentes e não respondentes, Rao (2000).

	Dimensão	Média	Variância
Respondentes	N_1	$\overline{Y}_1$	S_1^2
Não respondentes	$N_2 = N - N_1$	$\overline{Y}_2$	S_2^2

Desta forma, a média é $\overline{Y} = W_1\overline{Y}_1 + W_2\overline{Y}_2$, sendo $W_1 = \frac{N_1}{N}$ e $W_2 = \frac{N_2}{N} = 1 - W_1$ as proporções dos respondentes e dos não respondentes, respectivamente. O total populacional é $Y = Y_1 + Y_2$.

Consideremos uma amostra aleatória simples de n elementos retirada desta população. A amostra também reflecte a divisão entre respondentes e não respondentes,

	Dimensão	Média	Variância
Respondentes	n_1	$\overline{y}_1$	s_1^2
Não respondentes	$n_2 = n - n_1$	$\overline{y}_2$	s_2^2

Note-se que para os n_2 não respondentes não temos $\overline{y}_2$ e s_2^2. A média amostral $\overline{y}_1$ é um estimador centrado de $\overline{Y}_1$, mas usar a média amostral dos respondentes, $\overline{y}_1$, para estimar $\overline{Y}$ leva a um enviesamento, $B(\overline{y}_1) = \overline{Y}_1 - \overline{Y} = W_2(\overline{Y}_1 - \overline{Y}_2)$.

O valor absoluto deste enviesamento depende da diferença entre $\overline{Y}_1$ e $\overline{Y}_2$ e da dimensão N_2 dos não respondentes, mas não depende da dimensão da amostra.

A variância de $\overline{y}_1$ é $V(\overline{y}_1) = (1 - f_1)\frac{S_1^2}{n_1}$ com $f_1 = \frac{n_1}{n}$ e o seu erro médio quadrático é $MSE(\overline{y}_1) = V(\overline{y}_1) + B^2(\overline{y}_1)$.

O número n_1 de respondentes pode crescer com o aumento da dimensão n da amostra e, como consequência, $V(\overline{y}_1)$ e $MSE(\overline{y}_1)$ diminuiriam.

Capítulo 2

Rarefação e Não-Resposta

2.1 Introdução

A determinação da dimensão da amostra necessária para obter a precisão que necessitamos quando estimamos parâmetros populacionais é um tema chave na teoria da amostragem e suas aplicações como ferramenta metodológica nas ciências experimentais. Por exemplo, se o nosso objectivo é estimar a média populacional a partir da média amostral, de modo a que a amplitude do intervalo de confiança $(1-\alpha)\times 100\%$ seja limitado por B, a dimensão da amostra $n = n_{_G}$ é o menor inteiro maior do que

$$\frac{4z^{^2}_{1-\frac{\alpha}{2}}\sigma^{^2}}{B^2}$$

no caso da amostragem com independência, e

$$\frac{\nu}{1+\frac{(\nu-1)B^2}{4z^2_{1-\frac{\alpha}{2}}\sigma^{^2}}}$$

na amostragem aleatória simples sem reposição (isto é, permutabilidade em vez de independência) a partir de uma população finita de dimensão ν; na prática, o $\sigma^{^2}$ desconhecido é substituído por uma estimativa $s^{^2}$, e o uso de quantis gaussianos $z_{_{1-\frac{\alpha}{2}}}$ é justificado pelo teorema limite central clássico no caso de independência, e sua extensão para parcelas permutáveis (Erdös e Rényi, 1959), quando a amostragem é feita sem reposição a partir de populações finitas. No entanto, em muitas

situações de amostragem, algumas das unidades seleccionadas para a amostra aleatória acabam por sair dela. A fracção de saídas pode ser bastante elevada, e em estudos por correio, por exemplo, o questionário é enviado a um vasto número de indivíduos — uma regra prática *ad hoc* é $\frac{n_G}{\tilde{p}}$, onde $\tilde{p}$ é a percentagem prevista de formulários devolvidos, usualmente uma estimativa grosseira baseada em estudos similares e populações alvo —, uma vez que a experiência acumulada mostra que apenas uma pequena percentagem p deles devolverá os formulários.

Vamos considerar o caso da rarefação (filtragem aleatória) em que, cada unidade incluída na amostra na etapa de planeamento permanece nela com probabilidade p, ou sai dela com probabilidade $1-p$, independentemente de qualquer outra.

Na Secção 2.2 investigamos os resultados obtidos ao usar uma amostra com dimensão aleatória $N \frown BinomialNegativa(n_G, p)$, em vez da regra *ad hoc* de dimensão $\frac{n_G}{\tilde{p}}$.

Para valores muito pequenos de p, o limite desse processo de filtragem é um processo de rarefação de Rényi (1956) da amostra inicial, e no ponto de vista dos resultados de Kovalenko (1965) e de Kozubovsky (1994), o processo de rarefação de Rényi é equivalente a parar aleatoriamente a soma de variáveis aleatórias i.i.d., com subordinador independente $V \frown$ Geométrica(p); a distribuição assintótica de $T = \sum_{k=1}^{V} X_k$, assumindo a existência da variância da distribuição "mãe", é Laplace generalizada, como se verá na Secção 2.3 e, em particular, Exponencial quando as parcelas são positivas. Este pode ser um resultado muito útil quando se faz amostragem de acontecimentos raros, um assunto que será discutido na Secção 2.5 usando uma família de variáveis aleatórias simétricas, contendo a Laplace e a Gausssiana, cuja curtose tem um vasto intervalo de variação com o parâmetro de forma, e uma família de variáveis aleatórias positivas, contendo a Exponencial, cuja assimetria e curtose — e, portanto, o comportamento da cauda — muda com o parâmetro de forma.

2.2 Estimando a média

Considerem-se as duas situações seguintes:

1. $(X_1, \ldots, X_n)$ é uma amostra aleatória de dimensão n, onde os X_i

independentes são tais que $X_i \stackrel{\mathrm{d}}{=} X$, com $\mathbb{E}(X) = \mu$ e $\mathrm{var}(X) = \sigma^2$. Nesta situação, podemos usar $\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$ para estimar μ com um limite de erro (erro padrão) B, com confiança $(1-\alpha)\times 100\%$, usando uma amostra de dimensão n_G, o menor inteiro maior do que $\frac{4z^2_{1-\frac{\alpha}{2}}\sigma^2}{B^2}$.

2. $(X_1, \ldots, X_\nu)$ é uma população finita com média $\mu = \frac{1}{\nu}\sum_{k=1}^{\nu} X_k$ e variância $\sigma^2 = \frac{1}{\nu-1}\sum_{k=1}^{\nu}(X_k - \mu)^2$. A amostragem aleatória simples sem reposição garante que todas as $\binom{\nu}{n}$ amostras de dimensão n são equiprováveis. Note que, nesta situação, os X_k já não são independentes, mas a sua dependência mútua é fraca, e o teorema limite central para variáveis aleatórias permutáveis pode ser usado para determinar a dimensão da amostra necessária para obter a precisão que desejamos: $\overline{X} = \dfrac{1}{n}\sum_{k=1}^{n} X_k$ estima μ com um limite de erro padrão B, com confiança $(1-\alpha)\times 100\%$, usando uma amostra de dimensão n_G, o menor inteiro maior do que $\dfrac{\nu}{1 + \frac{(\nu-1)B^2}{4z^2_{1-\frac{\alpha}{2}}\sigma^2}}$. (Como não pode surgir confusão, usamos o mesmo símbolo n_G quer no caso independente, quer no caso permutável).

   Assuma-se, no entanto, que sabemos que a amostra será sujeita a uma filtragem-p, isto é, cada X_k será efectivamente observado com probabilidade p, independentemente de cada um dos outros. Precisamos, portanto, de uma amostra maior de dimensão N, de modo a que a amostra filtrada tenha aproximadamente dimensão n_G. No que se segue iremos comparar os resultados usando uma amostra aleatória de dimensão $N \frown \mathit{BinomialNegativa}(n_G, p)$ com os resultados obtidos usando dimensão determinística $\frac{n_G}{p}$.

   Observe que se $(Y_1, \ldots, Y_N)$ é uma amostra de Y_k independentes tais que $Y_k \stackrel{\mathrm{d}}{=} Y \frown \mathit{Bernoulli}(p)$, independente de $(X_1, \ldots, X_N)$, a amostra $(Z_1, \ldots, Z_{n^*})$ onde os Z_k são os $X_k Y_k$ não nulos, é uma amostra filtrada-p, e $T = \sum_{i=1}^{n^*} Z_i = \sum_{k=1}^{N} X_k Y_k$. Observe que $\mathbb{E}(X_k Y_k) = p\mu$ e que $\mathrm{var}(X_k Y_k) = p(\sigma^2 + (1-p)\mu^2)$.

2.2.1 Dimensão da amostra determinística usando a regra *ad hoc* — estimador $\tilde{\mu}_1 = \dfrac{T}{n_G}$

Se usarmos $N = \frac{n_G}{p}$, o valor esperado e a variância do estimador $\tilde{\mu}_1 = \frac{T}{n_G}$ são:

1. No caso da amostragem independente,

$$\mathbb{E}(\tilde{\mu}_1) = \mu \qquad \text{e} \qquad \text{var}(\tilde{\mu}_1) = \frac{\sigma^2 + (1-p)\mu^2}{n_G}.$$

Portanto, a precisão da estimativa será muito pior do que a esperada sempre que $\mu \gg 0$.

2. Na amostragem aleatória simples sem repetição a partir de uma população com dimensão ν, com a correcção da dimensão da amostra finita para a variância, obtemos resultados semelhantes:

$$\mathbb{E}(\tilde{\mu}_1) = \mu \qquad \text{e} \qquad \text{var}(\tilde{\mu}_1) = \frac{\nu - N}{\nu - 1}\,\frac{\sigma^2 + (1-p)\mu^2}{n_G},$$

e, portanto, há um valor $\dfrac{\nu - N}{\nu - 1}\dfrac{(1-p)\mu^2}{n_G}$ a mais na variância do estimador, quando comparado com a situação de não filtragem.

2.2.2 Dimensão da amostra aleatória Binomial Negativa — estimador $\tilde{\mu}_2 = \dfrac{T}{pN}$

Seja $N \frown \mathit{BinomialNegativa}(n_G, p)$, $\mathbb{E}(N) = \dfrac{n_G}{p}$, $\text{var}(N) = \frac{n_G(1-p)}{p^2}$, e $\mathbb{E}(N^2) = \dfrac{n_G(n_G + 1 - p)}{p^2}$, e consideremos o estimador $\tilde{\mu}_2 = \dfrac{T}{pN}$.

Usando a expansão de Taylor linearmente truncada,

$$\tilde{\mu}_2 = \frac{T}{pN} \approx \frac{\mathbb{E}(T)}{p\mathbb{E}(N)} + \frac{1}{p\mathbb{E}(N)}\left[T - \mathbb{E}(T)\right] - \frac{\mathbb{E}(T)}{p[\mathbb{E}(N)]^2}\left[N - \mathbb{E}(N)\right].$$

$\mathbb{E}(T) = \mathbb{E}\left[\mathbb{E}(T|N)\right] = \mathbb{E}(Np\mu) = \frac{n_G}{p}\, p\mu = n_G\mu$, e assim $\mathbb{E}(\tilde{\mu}_2) \approx \dfrac{n_G\mu}{p\frac{n_G}{p}} = \mu$. No que respeita à variância do estimador $\tilde{\mu}_2$:

1. No caso da amostragem independente, $\mathrm{var}(T|N) = Np(\sigma^2 + (1-p)\mu^2)$, e assim

$$\mathrm{var}(T) = \mathbb{E}\left[\mathrm{var}(T|N)\right] + \mathrm{var}\left[\mathbb{E}(T|N)\right] = n_G(\sigma^2 + 2(1-p)\mu^2).$$

Portanto,

$$\mathrm{var}(\tilde{\mu}_2) \approx \frac{1}{n_G^2} n_G(\sigma^2 + 2(1-p)\mu^2) + \left(\frac{n_G\mu}{p\left(\frac{n_G}{p}\right)^2}\right)^2 \frac{n_G(1-p)}{p^2} = \frac{\sigma^2 + 3(1-p)\mu^2}{n_G}.$$

2. Na população finita, com dimensão ν, temos o factor de correcção para populações finitas, $\mathrm{var}(T|N) = \frac{\nu - N}{\nu - 1} Np(\sigma^2 + (1-p)\mu^2)$, e assim

$$\mathrm{var}(T) = n_G\left[\frac{(\nu+1)p - (n_G+1)}{p(\nu-1)}\right]\sigma^2 + n_G\left[1 + \frac{(\nu+1)p - (n_G+1)}{p(\nu-1)}\right](1-p)\mu^2.$$

Portanto

$$\mathrm{var}(\tilde{\mu}_2) \approx \frac{\left[\frac{(\nu+1)p-(n_G+1)}{p(\nu-1)}\right]\sigma^2 + \left[2 + \frac{(\nu+1)p-(n_G+1)}{p(\nu-1)}\right](1-p)\mu^2}{n_G}.$$

Em consequência, $\tilde{\mu}_2$ é menos eficiente do que $\tilde{\mu}_1$. Observe que o valor esperado do denominador de $\tilde{\mu}_2$ é n_G e, portanto, estamos a dividir uma soma mais variável T (com um número aleatório de parcelas) por um valor próximo do mesmo n_G que usámos no denominador de $\tilde{\mu}_1$, e assim este resultado faz sentido.

2.2.3 Dimensão da amostra aleatória Binomial Negativa — estimador $\tilde{\mu}_3 = \dfrac{T}{W}$

Uma abordagem mais sofisticada seria contar o número de parcelas não nulas, $W = \sum_{k=1}^{N} Y_k$. Sendo $W|N \frown Binomial(N, p)$, para dividir a soma $T = \sum_{k=1}^{N} X_k Y_k$, isto é, para usar o e stimador $\tilde{\mu}_3 = \frac{T}{W}$.

No entanto, neste cenário $\mathbb{E}(W) = n_{_G}$, $\mathrm{var}(W) = 2n_{_G}(1-p)$ e a variância de $\frac{T}{W} \approx \mu + \frac{1}{n_{_G}}(T - n_{_G}\mu) - \frac{\mu}{n_{_G}}(W - n_{_G})$ é

1. No esquema de amostragem independente

$$\mathrm{var}(\tilde{\mu}_{_3}) \approx \frac{1}{n_{_G}^{^2}} n_{_G}(\sigma^{^2} + 2(1-p)\mu^{^2}) + \left(\frac{\mu}{n_{_G}}\right)^{^2} 2n_{_G}(1-p) = \frac{\sigma^{^2} + 4(1-p)\mu^{^2}}{n_{_G}}.$$

2. Na população finita de dimensão ν, a variância de $\tilde{\mu}_{_3}$ quando a amostragem é sem reposição é

$$\mathrm{var}(\tilde{\mu}_{_3}) \approx \frac{\left[\frac{(\nu+1)p-(n_{_G}+1)}{p(\nu-1)}\right]\sigma^{^2} + \left[3 + \frac{(\nu+1)p-(n_{_G}+1)}{p(\nu-1)}\right](1-p)\mu^{^2}}{n_{_G}}.$$

Portanto, quanto mais variabilidade introduzimos, menor eficiência obtemos do estimador. Parece que a única forma de alcançar os nossos objectivos seria usar um esquema de amostragem inversa desajeitado, continuando a amostragem até a dimensão da amostra atingir $n_{_G}$.

Nas secções seguintes tentaremos uma abordagem alternativa, quando o parâmetro de filtragem p está próximo de zero. Esta filtragem radical foi designada "rarefação" por Rényi (1956).

2.3 Filtragem Geométrica

Seja $Y = \sum_{k=1}^{V} X_{_k}$, onde os $X_{_k}$ variáveis aleatórias independentes tais que $X_{_k} \stackrel{\mathrm{d}}{=} X \geq 0$, $k = 1, 2, \ldots$, são independentes de $V \frown$ Geométrica$(p_{_n})$, e $\mathbb{E}(X) = \mu$, e portanto a média da soma parada geométrica Y é $\delta = \frac{\mu}{p_{_n}}$. A função característica de Y é $\varphi_{_Y}(t) = \mathcal{G}_{_V}(\varphi_{_X}(t))$, onde $\mathcal{G}_{_V}$ é a função geradora de probabilidade de V. Então

$$\varphi_{_Y}(p_{_n}t) = \frac{1}{1 + \frac{1-\varphi_{_X}(p_{_n}t)}{p_{_n}\varphi_{_X}(p_{_n}t)}} = \frac{1}{1 + \frac{1-\varphi_{_X}(p_{_n}t)}{p_{_n}t}\,\frac{t}{\varphi_{_X}(p_{_n}t)}}.$$

Como $\frac{1-\varphi_{_X}(p_{_n}t)}{p_{_n}t} \xrightarrow[p_{_n}\to 0]{} -\varphi'_{_X}(0) = -\mathrm{i}\mu$ e $\varphi_{_X}(p_{_n}t) \xrightarrow[p_{_n}\to 0]{} \varphi_{_X}(0) = 1$, o limite do lado direito em cima é $\frac{1}{1-\mathrm{i}\mu t}$ e consequentemente $\varphi_{_Y}(t) = \frac{1}{1-\mathrm{i}\delta t}$,

que é a função característica de uma variável aleatória exponencial com média δ.

Assim, o limite de um processo rarefeito geometricamente com parcelas positivas com valor esperado finito é exponencial. Este resultado assintótico para o processo rareficado foi inicialmente descoberto por Rényi (1956); Kovalenko (1965) estabeleceu que as transformadas de Laplace de variáveis aleatórias positivas que são estáveis com respeito à rarefação elementar são da forma $L(s) = \dfrac{1}{1+cs^{\delta}}$, $c > 0,\ \delta \in (0,1]$, o caso $\delta = 1$ — isto é, limite exponencial — correspondendo a variância finita. Os resultados de Kovalenko mostram que isto coincide com a classe de somas paradas aleatórias $\sum_{k=0}^{V} X_{k}$ com parcelas positivas i.i.d., independentes da variável subordinadora $V \frown$ Geométrica(p). A teoria geral das somas paradas geométricas (Kozubowski, 1994) conduz a resultados semelhantes para a função característica de parcelas cujo suporte não é necessariamente positivo, e em particular a lei limite simétrica para somas geométricas de parcelas independentes de variância finita é a distribuição Laplace.

Não estamos munidos de qualquer investigação sistemática sobre até onde a rarefação deve ir de modo a que o resultado assintótico possa ser tomado como uma boa aproximação. A exponencial é estável relativamente à filtragem-p, ou seja, a exponencial filtrada-p é ainda exponencial, qualquer que seja o valor de $p \in (0,1]$, no entanto, isto é uma situação excepcional. Na Secção 2.4 descrevemos duas famílias dependendo de um parâmetro de forma, que iremos usar para fazer uma abordagem preliminar desta questão.

Assuma agora que $Y = \sum_{k=1}^{V} X_{k} \overset{\circ}{\frown} Exponencial(\delta)$; de

$$W = \frac{Y}{V} \approx \mu + p(Y-\delta) - p^2\delta\left(V - \frac{1}{p}\right)$$

obtemos que $\mathbb{E}(W) \approx \mu$ e $\mathrm{var}(W) \approx (2-p)\mu^2$.

Sejam $(W_{1}, \ldots, W_{n})$ réplicas independentes de W, $\tilde{\mu}_{4} = \overline{W} = \frac{1}{n}\sum_{k=1}^{n} W_{k}$, para as quais o teorema do limite central se verifica. Como a variância de $\tilde{\mu}_{4}$ é $\dfrac{(2-p)\mu^2}{n}$, se pretendermos que o intervalo de confiança a $(1-\alpha)\times 100\%$ seja limitado por B, devemos tomar uma amostra

de dimensão $n_{_E}$, o menor inteiro maior do que $\dfrac{4z^{^2}_{1-\frac{\alpha}{2}}(2-p)\mu^{^2}}{B^2}$.

Na Secção 2.5 usaremos as famílias de funções introduzidas na Secção 2.4 para comparar $\tilde{\mu}_{_4}$ com $\tilde{\mu}_{_1}$.

2.4 Densidades Laplace, Gaussiana e GLE

Como para $\beta > -1$

$$\int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left|x\right|^{\frac{2}{1+\beta}}\right\} \mathrm{d}x = 2^{\frac{\beta+1}{2}}(\beta+1)\Gamma\left(\frac{\beta+1}{2}\right) = 2^{\frac{\beta+3}{2}}\Gamma\left(\frac{\beta+3}{2}\right),$$

a função

$$f(x \mid \beta,\, \lambda,\, \delta) = \frac{1}{2^{\frac{\beta+3}{2}}\Gamma\left(\frac{\beta+3}{2}\right)\delta} \exp\left\{-\frac{1}{2}\left|\frac{x-\lambda}{\delta}\right|^{\frac{2}{1+\beta}}\right\} \mathrm{I}_{\mathbb{R}}(x)$$

é a função densidade de probabilidade de uma variável aleatória $X_{\beta,\,\lambda,\,\delta}$ para qualquer $\beta > -1$, $\lambda \in \mathbb{R}$ e $\delta > 0$. Temos assim uma família parametrizada a qual contém em particular as variáveis aleatórias Gaussiana($\beta = 0$) e Laplace ($\beta = 1$). Denotamos $X_{\beta,\,0,\,1} = X_{\beta}$.

Como

$$\mathbb{E}\left(X_{\beta}^{2k}\right) = 2^{k(\beta+1)}\,\frac{\Gamma\left(k(\beta+1)+\frac{\beta+1}{2}\right)}{\Gamma\left(\frac{\beta+1}{2}\right)},$$

o valor médio μ de $X_{\beta,\,\lambda,\,\delta}$ é $\mu = \lambda$, a variância $\sigma^{^2}$ é

$$\sigma^{^2} = \frac{2^{\beta+1}\Gamma\left(3\,\frac{\beta+1}{2}\right)}{\Gamma\left(\frac{\beta+1}{2}\right)}\,\delta^{^2},$$

e a curtose $\gamma_{_2}$ é

$$\gamma_{_2} = \frac{\Gamma\left(\frac{\beta+1}{2}\right)\,\Gamma\left(5\,\frac{\beta+1}{2}\right)}{\left[\Gamma\left(3\,\frac{\beta+1}{2}\right)\right]^{^2}} - 3\,,$$

a qual aumenta para ∞ com β, cf. Figura 2.1.

Para $\beta \approx -1$ variável aleatória $X_{\beta,\,\lambda,\,\delta}$ tem uma curtose muito baixa (por exemplo, para $\beta = -0.999$ a curtose é -1.1999, não muito distante do limite inferior -2 que a curtose pode atingir).

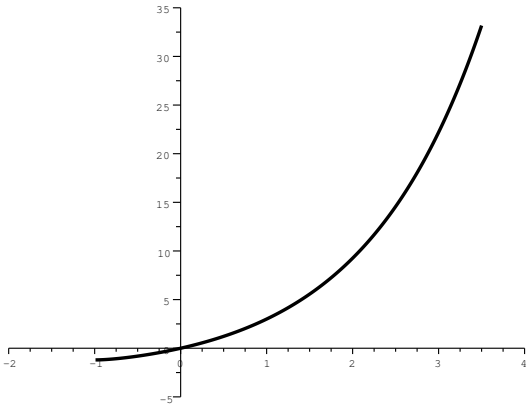

Figura 2.1: γ_2 em função de β

Assim esta família parametrizada, da qual as variáveis aleatórias Gaussiana($\beta = 0$) e Laplace ($\beta = 1$) são casos especiais, abarca uma vasta variedade de caudas leves e pesadas; a curtose desempenha um papel importante em estudos sobre a velocidade de convergência através do limite central, ver Barndorff-Nielsen e Cox (1989). Esta é a principal razão que nos conduz ao uso desta família numa avaliação preliminar dos benefícios de determinar a dimensão da amostra considerando que, sob filtragem, a lei limite é Laplace.

Para o caso de populações com suporte positivo, usamos a família de funções densidade de probabilidade de $W_{\beta,\lambda,\delta} = \lambda + \delta W_\beta$, $\beta > 0$, com

$$f_{W_\beta}(x) = \frac{\exp\left(-x^\beta\right)}{\Gamma\left(1+\frac{1}{\beta}\right)}\ \mathrm{I}_{(0,\infty)}(x).$$

Observe que $W_1 \frown \mathit{Exponencial}(1)$ e $\sqrt{2}W_2$ é a variável aleatória gaussiana dobrada ("*folded Gaussian*"), que o valor médio de $W_{\beta,\lambda,\delta}$ é

$$\mu = \lambda + \delta\,\frac{\Gamma\left(\frac{2}{\beta}\right)}{\Gamma\left(\frac{1}{\beta}\right)} = \lambda + \frac{2^{\frac{2}{\beta}-1}\delta}{\sqrt{\pi}}\,\Gamma\left(\frac{2+\beta}{2\beta}\right),$$

e a variância é $\sigma^2 = \left[\frac{\Gamma\left(\frac{3}{\beta}\right)}{\Gamma\left(\frac{1}{\beta}\right)} - \left(\frac{\Gamma\left(\frac{2}{\beta}\right)}{\Gamma\left(\frac{1}{\beta}\right)}\right)^2\right]\delta^2$.

Seja $Y_{\alpha} \frown Gama(\alpha, 1)$; então a função densidade de probabilidade de $V_{\alpha} = Y_{\alpha}^{\alpha}$ é $f_{V_{\alpha}}(x) = \frac{\mathrm{e}^{-x^{\frac{1}{\alpha}}}}{\Gamma(\alpha+1)}\, \mathrm{I}_{(0,\infty)}(x)$, e portanto a variável aleatória W_{β} é

$$W_{\beta} = V_{\frac{1}{\beta}} = Y_{\frac{1}{\beta}}^{\frac{1}{\beta}}.$$

Por outro lado, com B uma variável aleatória Bernoulli com suporte $\{-1, 1\}$, independente de V_{α}, a função densidade de probabilidade de $T_{\alpha} = 2^{\alpha}\, B V_{\alpha}$ é $f_{T_{\alpha}}(x) = \dfrac{\mathrm{e}^{-\frac{1}{2}|x|^{\frac{1}{\alpha}}}}{2^{\alpha+1}\Gamma(\alpha+1)}\, \mathrm{I}_{\mathbb{R}}(x)$, e consequentemente a variável aleatória $X_{\beta,\lambda,\delta}$ é

$$X_{\beta,\lambda,\delta} = \lambda + 2^{\frac{1+\beta}{2}}\, \delta\, B\, Y_{\frac{1+\beta}{2}}^{\frac{1+\beta}{2}}.$$

Assim a geração de números aleatórios das populações que escolhemos investigar — e que será mais detalhada no Capítulo 3 — pode usar os métodos eficientes de geração de números aleatórios gama. A qualidade das populações geradas foi verificada comparando os momentos empíricos e populacionais de menor ordem, uma estimativa de máxima verosimilhança $\hat{\beta}$ com o verdadeiro β, e o teste de ajustamento de Kolmogorov-Smirnov, com resultados altamente satisfatórios em todos os casos.

Para esta fase da investigação gerámos populações finitas de dimensão $\nu = 5000(5000)20000$ elementos x_k de $X_{\beta,}$, para $\beta = -0.75(0.25)1.5(0.5)3$, e de W_{β}, para $\beta = 0.25(0.25)1.5(0.5)6$; em cada caso calculámos a verdadeira média populacional gerada $\mu^* = \frac{1}{\nu}\sum_{k=1}^{\nu} x_k$ e a variância $\sigma^{*2} = \frac{1}{\nu-1}\sum_{k=1}^{\nu}(x_k - \mu^*)^2$.

2.5 Métodos, critérios e conclusões

Com o objectivo de investigar a proximidade das somas filtradas$-p$ definidas na Secção 3 com o limite Exponencial , comparámos a média e o desvio padrão empíricos de $\sum_{k=1}^{V} W_{\beta,k}$, onde os independentes $W_{\beta,k} \stackrel{\mathrm{d}}{=} W_{\beta}$, tal como foram definidos na Secção 3, são independentes de

$V \frown$Geométrica(p). Esta é uma avaliação muito preliminar, apenas uma indicação grosseira de quanto longe estamos da situação ideal.

Como esperado, a aproximação deteriora-se com o afastamento de β a partir de 1, no intervalo $\beta \in (0.25, 0.95) \cup (1.1, 4)$ precisamos de $p \approx 0.01$ para obter resultados razoáveis.

Como a rarefação de Rényi e a geo-estabilidade de Kovalenko são dois modos de obter as mesmas leis assintóticas, os resultados foram verificados com investigação semelhante de amostras aleatórias provenientes de $\sum W_{\beta,k} Y_k$, onde $Y_k \frown Bernoulli(p)$.

Resultados de simulação para comparação da eficiência de $\tilde{\mu}_4$ com $\tilde{\mu}_1$ estão ainda muito incompletos. Em termos teóricos, pensamos que $\tilde{\mu}_4$ é mais fidedigno, tendo em conta que estamos a lidar com acontecimentos raros, uma vez que a variância de $\tilde{\mu}_1$ é maior do que a que tem sido usada no cálculo da dimensão da amostra necessária para ter um limite de erro B. Mas até agora, não temos evidência de simulação sistemática para apoiar isto.

Capítulo 3

A família GLE

3.1 Introdução

Na investigação dos estimadores de parâmetros populacionais no caso de não-resposta por parte da amostra inquirida e, nomeadamente, para comparar os resultados recorrendo às técnicas de rarefação que desenvolvemos, e conveniente usar dados provenientes de uma família de populações coerente, mas exibindo curtose diversificada. Gerámos, para o efeito, números pseudo-aleatórios de uma família de variáveis aleatórias que contém a exponencial e a gaussiana dobrada (*folded*).

Considere-se a família de variáveis aleatórias W_β com funções densidade de probabilidade

$$f_{W_\beta}(x) = \frac{\exp\left(-x^\beta\right)}{\Gamma\left(1+\frac{1}{\beta}\right)}\ \mathrm{I}_{(0,\infty)}(x),\ \beta > 0.$$

Adiante estabelecemos relações simples de W_β com produtos de potências de gamas independentes que usamos para a geração de números pseudo-aleatórios daquelas populações.

Começamos por descrever as propriedades estruturais importantes, e os métodos usados para geração dos números aleatórios.

Descrevemos ainda os métodos usados para avaliar se os números pseudo-aleatórios assim gerados se ajustam convenientemente às leis que pretendemos usar.

O mesmo fazemos com uma família de variáveis aleatórias simétricas X_{β^*}, que inclui a gaussiana e a Laplace, e que se obtém de forma simples

multiplicando uma variável positiva da classe anteriormente descrita por uma Bernoulli simétrica independente. Por uma questão de conveniência, optamos por uma parametrização em que o parâmetro de forma é

$$\beta^* = \frac{2-\beta}{\beta} > -1$$

e

$$f_{X_{\beta^*}}(x \mid \beta^*,\, \lambda,\, \delta) = \frac{1}{2^{\frac{\beta^*+3}{2}}\,\Gamma\left(\frac{\beta^*+3}{2}\right)\delta} \exp\left\{ -\frac{1}{2}\left|\frac{x-\lambda}{\delta}\right|^{\frac{2}{1+\beta^*}} \right\} \mathrm{I}_{\mathbb{R}}(x),$$

com $\beta^* > -1,\ \lambda \in \mathbb{R},\ \delta > 0$.

A família $\left\{W_{\beta}\right\}_{\beta>0}$ é adiante referida como família GLE com suporte positivo e a famíla $\left\{X_{\beta^*}\right\}_{\beta^*>-1}$ como família GLE com suporte real.

Sempre que, não estando em causa a relação $X_{\beta^*} = BW_{\frac{2-\beta}{\beta}}$ (onde $W_{\frac{2-\beta}{\beta}}$ e $B \frown \mathit{Bernoulli}\left(\frac{1}{2},\, \{-1,1\}\right)$ são variáveis independentes) entre variáveis daquelas famílias usaremos a notação X_{β} em lugar de X_{β^*} — e a concomitante alteração em todas as expressões em que surja o parâmetro de forma — a fim de tornar as fórmulas mais simples.

A título de exemplo, incluímos em apêndice uma escolha de números pseudo-aleatórios destas famílias. Admitindo que podem ser úteis em outros trabalhos de investigação da robustez de estimadores, fizemos a publicação electrónica nas páginas de Sandra Aleixo — com quem desenvolvemos este trabalho — e de Fernanda Diamantino em www.ceaul.fc.ul.pt.

3.2 Propriedades estruturais da família GLE com suporte positivo

Para o caso de populações com suporte positivo, usamos a família de funções densidade de probabilidade de

$$W_{\beta,\lambda,\delta} = \lambda + \delta W_{\beta}, \qquad \beta > 0,$$

com

$$f_{W_{\beta}}(x) = \frac{\exp\left(-x^{\beta}\right)}{\Gamma\left(1+\frac{1}{\beta}\right)}\ \mathrm{I}_{(0,\infty)}(x).$$

Observe que W_1 é a variável aleatória Exponencial(1) (ou Laplace dobrada) e $\sqrt{2}W_2$ é a variável aleatória gaussiana dobrada.

O valor médio de $W_{\beta,\lambda,\delta}$ é

$$\mu = \lambda + \delta \frac{\Gamma\left(\frac{2}{\beta}\right)}{\Gamma\left(\frac{1}{\beta}\right)} = \lambda + \frac{2^{\frac{2}{\beta}-1}\delta}{\sqrt{\pi}}\Gamma\left(\frac{2+\beta}{2\beta}\right),$$

e a variância de $W_{\beta,\lambda,\delta}$ é

$$\sigma^2 = \left[\frac{\Gamma\left(\frac{3}{\beta}\right)}{\Gamma\left(\frac{1}{\beta}\right)} - \left(\frac{\Gamma\left(\frac{2}{\beta}\right)}{\Gamma\left(\frac{1}{\beta}\right)}\right)^2\right]\delta^2.$$

Seja $Y_\alpha \frown Gama(\alpha, 1)$; então a função densidade de probabilidade de $V_\alpha = Y_\alpha^\alpha$ é $f_{V_\alpha}(x) = \frac{e^{-x^{\frac{1}{\alpha}}}}{\Gamma(\alpha+1)} I_{(0,\infty)}(x)$, e portanto a variável aleatória W_β é

$$W_\beta = V_{\frac{1}{\beta}} = Y_{\frac{1}{\beta}}^{\frac{1}{\beta}}.$$

3.3 Propriedades estruturais da família GLE simétrica com suporte real

Como, para $\beta > -1$, se tem

$$\int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}|x|^{\frac{2}{1+\beta}}\right\} dx = 2^{\frac{\beta+1}{2}}(\beta+1)\Gamma\left(\frac{\beta+1}{2}\right) = 2^{\frac{\beta+3}{2}}\Gamma\left(\frac{\beta+3}{2}\right)$$

então a função

$$f(x \mid \beta, \lambda, \delta) = \frac{1}{2^{\frac{\beta+3}{2}}\Gamma\left(\frac{\beta+3}{2}\right)\delta} \exp\left\{-\frac{1}{2}\left|\frac{x-\lambda}{\delta}\right|^{\frac{2}{\beta+1}}\right\} I_{\mathbb{R}}(x)$$

é a função densidade de probabilidade de uma variável aleatória $X_{\beta,\lambda,\delta}$ para qualquer $\beta > -1$, $\lambda \in \mathbb{R}$ e $\delta > 0$. Temos assim uma família parametrizada a qual contém em particular as variáveis aleatórias Gaussiana($\beta = 0$) e Laplace ($\beta = 1$). Denotamos $X_{\beta,0,1} = X_\beta$.

Como

$$\mathbb{E}\left(X_{\beta}^{2k}\right) = 2^{k(\beta+1)}\,\frac{\Gamma\left(k(\beta+1)+\frac{\beta+1}{2}\right)}{\Gamma\left(\frac{\beta+1}{2}\right)},$$

o valor médio μ de $X_{\beta,\,\lambda,\,\delta}$ é

$$\mu = \lambda,$$

a variância σ^2 de $X_{\beta,\,\lambda,\,\delta}$ é

$$\sigma^2 = \frac{2^{\beta+1}\Gamma\left(3\,\frac{\beta+1}{2}\right)}{\Gamma\left(\frac{\beta+1}{2}\right)}\,\delta^2,$$

e a curtose γ_2 de $X_{\beta,\,\lambda,\,\delta}$ é

$$\gamma_2 = \frac{\Gamma\left(\frac{\beta+1}{2}\right)\,\Gamma\left(5\,\frac{\beta+1}{2}\right)}{\left[\Gamma\left(3\,\frac{\beta+1}{2}\right)\right]^2} - 3,$$

a qual aumenta para ∞ com β.

Para $\beta \approx -1$ variável aleatória $X_{\beta,\,\lambda,\,\delta}$ tem uma curtose muito baixa (por exemplo, para $\beta = -0.999$ a curtose é -1.1999, não muito distante do limite inferior -2 que a curtose pode atingir).

Assim, esta família parametrizada da qual as variáveis aleatórias Gaussiana($\beta = 0$) e Laplace ($\beta = 1$) são casos especiais, abarca uma vasta variedade de caudas leves e pesadas; a curtose desempenha um papel importante em estudos sobre a velocidade de convergência da soma para o seu limite (central), ver Barndorff-Nielsen e Cox (1989). Assim, pode usar-se esta família numa avaliação preliminar dos benefícios de determinar a dimensão da amostra considerando que, sob filtragem, a lei limite é Laplace.

3.4 Geração de números pseudo-aleatórios de populações GLE

Ao pretendermos gerar directamente os números pseudo-aleatórios das populações com distribuição pertencente à família GLE que pretendemos obter, experimentamos vários métodos sugeridos em Devroye (1986), Kundu e Gupta (2007), Law e Kelton (1982) e Ross (1997). Concluimos que o método mais adequado em cada caso dependia do valor do parâmetro β considerado. Obtivemos resultados bastante satisfatórios (veremos como se avalia o grau de satisfação mais à frente) usando os seguintes métodos de geração:

* Para $0 < \beta < 1$, método da rejeição utilizando a distribuição $Pareto(\alpha)$, com α óptimo.

* Para $\beta = 1$, método da inversão para gerar exponenciais unitárias.

* Para $1 < \beta < 6$, método da rejeição utilizando a distribuição $Exponencial(1)$.

* Para $\beta \geq 6$, método da rejeição utilizando a distribuição $Weibull(0, \alpha, \gamma)$ (a exponencial unitária deixa de funcionar).

Verificámos por exemplo que, usando $\alpha = 0.7$ e $\gamma = 1.3$, se obtêm bons resultados para os números pseudo-aleatórios gerados usando o método da rejeição com a Weibull, para $6 \leq \beta \leq 15$. Já para $16 \leq \beta \leq 20$, pode usar-se o método da rejeição com a $Weibull(0, 0.75, 1.7)$. No entanto, não é possivel obter uma expressão geral para os parâmetros de escala e forma da Weibull para cada valor de β. Os valores adequados destes parâmetros dependem do parâmetro β considerado.

Mas, tendo em conta as relações existentes entre as variáveis aleatórias com distribuição pertencente à família GLE e variáveis aleatórias com distribuição Gama, a geração de números aleatórios das populações que pretendemos obter, com $\beta \neq 1$, pode ser feita indirectamente, usando os métodos eficientes de geração de números aleatórios gama já bem conhecidos e testados. Assim vamos optar por este processo de geração, para $\beta \neq 1$.

Uma vez que as relações com produtos de potências de gamas independentes são usadas para a geração no caso de suporte positivo, as correspondentes populações simétricas são obtidas multiplicando por uma Bernoulli com $p = 0.5$ e com suporte $\{-1, 1\}$, independente da anteriormente gerada para suporte positivo. No entanto, para $-1 < \beta \leq 0$, estas populações simétricas têm que ser geradas com base na relação existente entre as variáveis aleatórias com distribuição pertencente à família GLE de suporte $\mathbb{R}$ e variáveis aleatórias com distribuição Gama.

Vamos portanto começar por expor os métodos utilizados para a geração no caso de suporte positivo. Qualquer um destes métodos tem como base o seguinte algoritmo genérico, para gerar uma população com a distribuição GLE de dimensão N, ou seja, N números pseudo-aleatórios da família GLE, para um dado valor do parâmetro $\beta \neq 1$:

Para $i = 1$ até N fazer:

* Gerar $Y_{i\ \beta} \frown Gama(\frac{1}{\beta}, 1)$.

* Determinar $W_{i\ \beta} = Y_{i\ \beta}^{\frac{1}{\beta}} \frown GLE(\beta)$.

Uma questão interessante é a escolha do algoritmo mais adequado para gerar $Y_{\beta_i} \frown Gama(\frac{1}{\beta}, 1)$ em cada uma das i iterações do algoritmo anterior, de acordo com o valor do parâmetro β. Após investigar este assunto, de entre os vários algoritmos distintos existentes (geralmente, os algoritmos para gerar gamas funcionam só para $\beta < 1$ ou só para $\beta > 1$), escolheram-se os mais rápidos e eficientes consoante o valor do parâmetro β considerado em cada caso, veja-se Devroye (1986), Kundu e Gupta (2006), Law e Kelton (1982) e Ross (1997).

* Para $0 < \beta < 1$ optámos pelo algoritmo GB de Cheng.

* Para $\beta > 1$ e $\beta \neq 2$ escolhemos o algoritmo GS de Ahrens e Dieter.

* Para $\beta = 2$ usámos um algoritmo que utiliza o método da rejeição com a $Exponencial(1)$ para gerar uma variável $Normal(0, 1)$. A esta variável aplica-se uma tranformação adequada para obter a variável GLE com $\beta = 2$.

Algoritmo para gerar $\mathbf{Y}_\beta \frown \mathbf{Gama}(\frac{\mathbf{1}}{\beta}, \mathbf{1})$, com $\mathbf{0 < \beta < 1}$

* Calcular o valor das seguintes constantes:

$$\begin{aligned} a &= \left(\frac{2}{\beta} - 1\right)^{-\frac{1}{2}}; \\ c &= \frac{1}{\beta} - \ln 4; \\ q &= \frac{1}{\beta} + \frac{1}{a}; \\ \theta &= 4.5; \\ d &= 1 + \ln \theta. \end{aligned}$$

1. Gerar dois números pseudo-aleatórios $U_1, U_2 \frown Uniforme(0, 1)$ independentes.

2. Fazer:

$$\begin{aligned} V &= a\ \ln\left(\frac{U_1}{1-U_1}\right); \\ Y &= \frac{e^V}{\beta}; \\ Z &= U_1^2 U_2; \\ T &= c + qV - Y. \end{aligned}$$

3. Se $T + d - \theta Z \geq 0$, fazer $Y_\beta = Y$ e sair.
   Caso contrário, ir para o passo 4.

4. Se $T \geq lnZ$, fazer $Y_\beta = Y$ e sair.
   Caso contrário, voltar ao passo 1.

Este é um algoritmo de rejeição modificado com teste prévio para aceitação. No entanto, o método da rejeição com exponencial da mesma média que a gama que se pretende gerar, $\frac{1}{\beta}$, também é eficaz e pode ser uma alternativa interessante.

Algoritmo para gerar $\mathbf{Y_\beta \frown Gama(\frac{1}{\beta}, 1)}$, com $\beta > 1$ e $\beta \neq 2$

$*$ Calcular o valor da constante $b = \frac{e+\frac{1}{\beta}}{e}$.

1. Gerar um número pseudo-aleatório $U_1 \frown Uniforme(0,1)$ e fazer $P = b\,U_1$.
   Se $P > 1$, ir para o passo 3.
   Caso contrário, prosseguir para o passo 2.

2. Fazer $Y = P^\beta$ e gerar um número pseudo-aleatório $U_2 \frown Uniforme(0,1)$.
   Se $U_2 \leq e^{-Y}$, fazer $Y_\beta = Y$ e sair.
   Caso contrário, voltar ao passo 1.

3. Fazer $Y = \ln((b-P)\beta)$ e gerar um número pseudo-aleatório $U_2 \frown Uniforme(0,1)$.
   Se $U_2 \leq Y^{\frac{1}{\beta}-1}$, fazer $Y_\beta = Y$ e sair.
   Caso contrário, voltar ao passo 1.

Este é um algoritmo de rejeição com uma função majorante especial. Alternativamente, outros algoritmos tais como o de rejeição de Vaduva (1977) (veja-se Devroye (1986)), ou o algoritmo apresentado em Kundu e Gupta (2007), podem ser usados.

Algoritmo para gerar $\mathbf{Y}_\beta \frown \mathbf{Gama}(\frac{\mathbf{1}}{\beta}, \mathbf{1})$, com $\beta = \mathbf{2}$

Tendo em conta que

$$X \frown N(0,1) \Rightarrow X^2 \frown \chi^2_{(1)} \equiv Gama(\frac{1}{2}, 2) \Rightarrow \frac{X^2}{2} \frown Gama(\frac{1}{2}, 1),$$

temos

1. Gerar um número pseudo-aleatório $X \frown Normal(0,1)$ do seguinte modo:
   (a) Gerar um número pseudo-aleatório $Y \frown Exponencial(1)$:
      i. Gerar um número pseudo-aleatório $U_1 \frown Uniforme(0,1)$.
      ii. Fazer $Y = -\ln(U_1)$.
   (b) Gerar um número pseudo-aleatório $U_2 \frown Uniforme(0,1)$.
   (c) Se $U_2 \leq e^{-\frac{(Y-1)^2}{2}}$, fazer $X = Y$ e sair.
   Caso contrário, voltar ao passo 1.
2. Fazer $Y_2 = \frac{X^2}{2}$ e sair.

Optámos pelo algoritmo que usa o método da rejeição usando a $Exponencial(1)$ para gerar a variável aleatória $Normal(0,1)$. Para tal efeito poder-se-iam usar alternativamente, por exemplo, o método de Box-Muller ou o método polar.

Para gerar uma população com distribuição GLE de dimensão N e parâmetro $\beta = 1$ temos que gerar N exponenciais unitárias. Recorde-se que o algoritmo de inversão para gerar uma observação de uma variável aleatória $\mathbf{W_1} \frown Exponencial(1)$ é o seguinte:

1. Gerar um número pseudo-aleatório $U \frown Uniforme(0,1)$.
2. Fazer $W_1 = -ln(U)$ e sair.

Geração de populações da família GLE com suporte $\mathbb{R}$

Para gerar uma população com a distribuição GLE, com suporte $\mathbb{R}$, de dimensão N, para um dado valor do parâmetro $\beta > 0$ pode utilizar-se o seguinte algoritmo:

Para $i = 1$ até N fazer:

1. Gerar um número pseudo-aleatório $W_{i\ \beta}$ da família GLE com suporte positivo, usando o método adequado consoante o valor de β.
2. Gerar um número pseudo-aleatório $B_i \frown Bernoulli(0.5)$ com suporte $\{-1, 1\}$:
   (a) Gerar um um número pseudo-aleatório $U \frown Uniforme(0, 1)$.
   (b) Se $U \leq 0.5$, fazer $B_i = -1$.
   Caso contrário, fazer $B_i = 1$.
3. Fazer $X_{i\ \beta} = B_i W_{i\ \beta}$.

Mas, no caso de pretendermos gerar uma população com a distribuição GLE, com suporte $\mathbf{R}$, de dimensão N, para um dado valor do parâmetro $-1 < \beta < 0$ pode utilizar-se o seguinte algoritmo:

Para $i = 1$ até N fazer:

1. Gerar $Y_{i\ \beta} \frown Gama(\frac{\beta+1}{2}, 1)$ usando o algoritmo GS de Ahrens e Dieter.
2. Gerar um número pseudo-aleatório $B_i \frown Bernoulli(0.5)$ com suporte $\{-1, 1\}$:
   (a) Gerar um número pseudo-aleatório $U \frown Uniforme(0, 1)$.
   (b) Se $U \leq 0.5$, fazer $B_i = -1$.
   Caso contrário, fazer $B_i = 1$.
3. Determinar $X_{i\ \beta} = 2^{\frac{\beta+1}{2}} B_i\ Y_i^{\frac{\beta+1}{2}}{}_{\beta} \frown GLE(\beta)$.

Para $\beta = 0$ no passo 1, o algoritmo GS de Ahrens e Dieter é substituido pelo algoritmo que utiliza o método da rejeição com a $Exponencial(1)$ para gerar uma variável aleatória $Normal(0, 1)$. A esta variável aplica-se uma tranformação adequada para obter a variável GLE com $\beta = 0$.

3.5 Tabelas de números pseudo-aleatórios da família GLE

Gerámos populações finitas de dimensão $\nu = 5000$ elementos x_k de $X_{\beta,}$, para $\beta = -0.75(0.25)1.5(0.5)3$, e de W_β, para $\beta = 0.25(0.25)1.5(0.5)6$; em cada caso calculámos a verdadeira média populacional gerada

$$\mu^* = \frac{1}{\nu}\sum_{k=1}^{\nu} x_k$$

e variância

$$\sigma^{*2} = \frac{1}{\nu - 1}\sum_{k=1}^{\nu}(x_k - \mu^*)^2.$$

Programámos os algoritmos anteriores no programa Mathematica, tendo gerado os ficheiros que se encontram publicados electronicamente em www.ceaul.fc.ul.pt. Como exemplo, temos em anexo um deles para $\beta = 1.5$.

3.6 Validade das populações geradas

A qualidade das populações geradas foi verificada:

- comparando os momentos empíricos e populacionais de menor ordem, μ_{W_β} com $\mu^*_{W_\beta}$ e $\sigma^2_{W_\beta}$ com $\sigma^{*2}_{W_\beta}$;
- comparando uma estimativa de máxima verosimilhança $\hat{\beta}$ com o verdadeiro β;
- efectuando várias réplicas do teste de ajustamento de Kolmogorov-Smirnov.

Em todos os casos, os resultados encontrados foram satisfatórios, como podemos constatar observando a Tabela 3.1. Não há grandes diferenças entre os valores empíricos e os populacionais, quer em termos de valores médios, quer em termos de desvios padrões.

Nesta tabela apenas colocámos o pior valor observado (o mais elevado), obtido para a estatística de teste, nas várias réplicas realizadas do teste de Kolmogorov-Smirnov, para cada população gerada correspondente a um dado v alor do parâmetro β. Como o teste foi realizado

a partir de uma amostra de cem observações, ao nível de significância $\alpha = 5\%$, então o ponto crítico de rejeição é $\frac{1.36}{\sqrt{100}} = 0.136$. Como todos os valores observados da estatística do teste de Kolmogorov-Smirnov são inferiores a 0.136, não se rejeita a hipótese de que a distribuição da população subjacente à amostra recolhida pertença à família GLE com o parâmetro β c onsiderado. Caso se pretendam mais pormenores sobre o teste de Kolmogorov-Smirnov pode consultar-se Pestana e Velosa (2008).

Não se consegue explicitar a expressão do estimador de máxima verosimilhança para o parâmetro β da distribuição GLE com suporte positivo. No entanto, é possível implementar computacionalmente no Mathematica, a equação a partir da qual $\hat{\beta}$ é calculado.

A função de verosimilhança é dada por

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{e^{-x_i^{\beta}}}{\Gamma\left(1+\frac{1}{\beta}\right)} \right] = \frac{e^{-\sum_{i=1}^{n} x_i^{\beta}}}{\Gamma^n\left(1+\frac{1}{\beta}\right)}.$$

Logaritmizando vem

$$\ln L(\beta) = -\sum_{i=1}^{n} x_i^{\beta} - nln\left(\Gamma\left(1+\frac{1}{\beta}\right)\right).$$

O estimador de máxima verosimilhança de β, $\hat{\beta}$, é a solução da equação 3.1

$$\frac{d\ln L(\beta)}{d\beta} = 0. \tag{3.1}$$

Os cálculos dos momentos empíricos e populacionais, bem como o teste de ajustamento de Kolmogorov-Smirnov também foram implementados computacionalmente no programa Mathematica.

Tabela 3.1: Validação das populações geradas com distribuição GLE com parâmetro de forma β

β	$\hat{\beta}$	μ	σ	μ^*	σ^*	$K-S$
-0.75	-0.7319	0.0000	0.6117	-0.0019	0.6057	0.1278
-0.50	-0.4919	0.0000	0.6914	0.0017	0.6882	0.1255
-0.25	-0.2550	0.0000	0.8174	0.0159	0.8187	0.1045
0.00	0.0242	0.0000	1.0000	0.0002	1.0200	0.1116
0.25	0.2506	840.0000	2438.6900	811.7070	2375.5700	0.1144
0.50	0.4959	6.0000	9.1652	6.2544	9.8940	0.1034
0.75	0.7505	1.6849	1.9698	1.6840	1.9532	0.1111
1.00	1.0111	1.0000	1.0000	0.9924	0.9806	0.1084
1.25	1.2717	0.7675	0.6913	0.7572	0.6725	0.1161
1.50	1.5039	0.6595	.5510	0.6599	0.5491	0.1136
2.00	2.0089	0.5642	0.4263	0.5517	0.4249	0.1148
2.50	2.5095	0.5249	0.3721	0.5242	0.3721	0.1236
3.00	2.9590	0.5055	0.3432	0.5104	0.3485	0.1236
3.50	3.5080	0.4949	0.3259	0.4897	0.3241	0.0952
4.00	3.9659	0.4889	0.3146	0.4862	0.3143	0.1104
4.50	4.1715	0.4853	0.3070	0.4766	0.3133	0.1294
5.00	4.9462	0.4832	0.3015	0.4887	0.3031	0.1097
5.50	5.3100	0.4819	0.2976	0.4857	0.2927	0.1002
6.00	6.0160	0.4813	0.2946	0.4728	0.2927	0.1276

Capítulo 4

Preliminares II — cumulantes, localização, escala, (as)simetria e curtose

4.1 Introdução

Retomam-se os resultados de Diamantino e Pestana (1997) em que se investigam quantis nominais e quantis exactos da estatística de Student no caso da população parente pertencer a uma família de que a gaussiana é o caso $\gamma_2 = 0$, mostrando que estes resultados confirmam, no caso de simetria, o que a literatura, em geral, advoga sobre a "liberdade ou conservadorismo", da t de Student. Recorde-se que Montgomery (1997), por exemplo, invoca o facto de o teste t estar muito próximo do teste de permutações, para defender o seu uso mesmo que a população de base não seja gaussiana.

Desde o trabalho pioneiro de Student (1908), o problema da localização/escala tem estado no cerne do desenvolvimento da teoria estatística. Desde cedo começou a investigação sobre a teoria distribucional da estatística studentizada

$$T_{(n-1)} = \sqrt{n(n-1)}\frac{\overline{X}_n}{\sqrt{SS_n}} \tag{4.1}$$

em populações não gaussianas, onde $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ e $SS_n = \sum_{i=1}^{n}(X_i - \overline{X}_n)^2$ dão conteúdo aos conceitos vagos de localização e escala, respectiva-

mente, veja-se Iglewicz (1981). Sem perda de generalidade, consideramos em (4.1) a translação centrada da população parente; no caso de população parente gaussiana, $T_{(n-1)} \equiv t_{(n-1)}$, a t de Student com n-1 graus de liberdade.

No caso geral, as dificuldades analíticas — que advêm em grande parte da estrutura de dependência entre $\overline{X}_n$ e SS_n (a independência entre $\overline{X}_n$ e SS_n é uma característica da gaussiana) — impediram a obtenção de resultados exactos de interesse para amostras de dimensão $n \geq 3$, com excepção do caso "simples", estudado por Perlo (1933), de amostras de dimensão $n = 3$ em populações uniformes. No caso $n = 2$, $T_{(n-1)} \stackrel{d}{=} \frac{X_1+X_2}{|X_1-X_2|}$ e tem-se

$$f_{T_{(1)}} = 2 \int_0^\infty u f\left(\frac{t+1}{\sqrt{2}}u\right) f\left(\frac{t-1}{\sqrt{2}}u\right) du \tag{4.2}$$

Rocha (1995) apresenta grande variedade de exemplos do cálculo exacto de (4.2) em diversas populações, comentando que populações parentes bem diversas — por exemplo, uniformes e Pareto de índice 1 — podem originar estatísticas studentizadas com a mesma distribuição. Também Efron (1969) comentara, na sequência da sua investigação sobre a geometria da t de Student, que simetria radial de $(X_1, ..., X_n)$ levava a uma estatística studentizada idêntica à obtida por Student em populações gaussianas.

As dificuldades analíticas levaram muitos estatísticos a aceitar que $T_{(n-1)}$ seria razoavelmente aproximada por $t_{(n-1)}$, dado a estatística $\overline{X}_n$, que aparece no numerador, ter uma distribuição assintoticamente gaussiana, sob condições muito gerais.

Esta atitude, em certa medida uma seguidora das ideias de winsorização (que postulavam que na região central qualquer distribuição pode ser aproximada pela gaussiana) veio a ser criticada por Hotelling (1961), que chamou a atenção para alguns "pormaiores" convenientemente ignorados:

- o teorema limite central é um resultado geral e, por isso mesmo, nada estabelece sobre a velocidade de convergência, que pode ser bastante rápida (a soma de três uniformes é, praticamente, gaussiana) ou desesperadamente lenta (a soma de de um milhão de $P(10^{-27})$ está tão longe da gaussiana!);

- a aproximação que o teorema limite central propicia é "suspeita" para quantis extremos — ora são estes que, de um modo geral, são usados em inferência;

- mesmo ignorando as críticas anteriores, o interesse da studentização reside em viabilizar inferência sobre o parâmetro de localização por eliminação de um parâmetro perturbador em *pequenas amostras*, sendo dificilmente justificável, nessa situação, o recurso a argumentos assintóticos.

Como T. Huxley observou, as grandes ideias nascem como heresia, morrem como dogma.

Depois do trabalho de Hotteling (1961) e das investigações de Efron (1968) e de Logan, Mallows, Rice e Shepp (1973) aproximar de qualquer forma $T_{(n-1)}$ por $t_{(n-1)}$ parece não ser uma atitude correcta. Infelizmente esta constatação em nada simplificou a obtenção da distribuição exacta da estatística studentizada (4.1). Desenvolvimentos recentes viraram-se para o estudo de estatísticas studentizadas mais gerais, veja-se David (1981), em que o objectivo é eliminar um parâmetro perturbador de escala — não necessariamente SS_n — a fim de viabilizar inferência sobre um parâmetro de localização que, eventualmente, não é $\overline{X}_n$.

Pestana e Rocha (1995) retomaram o problema aceitando a filosofia de Tukey de que, mais vale a solução aproximada de um problema bem posto do que a solução exacta de um problema aproximado.

Com base em expressões recursivas para $\overline{X}_n$ e SS_n em amostras crescentes, e considerando a distribuição parente simétrica — hipótese que já Efron (1968) sublinhara — mostraram que, no caso de ser válido um teorema do valor médio generalizado,

$$f_{T_{(n-1)}}(t) = \frac{2^{n-1}}{\sqrt{n-1}} \prod_{i=3}^{n} (1-\xi_i^2)^{\frac{i-4}{2}} \times \\ \times \int_0^\infty u^{n-1} \prod_{i=1}^{n} f\left[u\left(\frac{t}{\sqrt{n(n-1)}} + \alpha_{in}\right)\right] du \quad (4.3)$$

onde $\xi_i = \sqrt{\frac{3}{i+1}}$, $i = 3, ..., n$ e $\alpha_{in} = \sqrt{3}\frac{n+1-2i}{\sqrt{n(n^2+1)}}$, $i = 1, 2, ..., n$.

Mostraram ainda que (4.3) leva à densidade

$$f_{t_{(n-1)}}(t) = \frac{1}{\sqrt{n-1}\beta\left(\frac{1}{2}, \frac{n-1}{2}\right)} \left(\frac{n-1}{n-1+t^2}\right)^{\frac{n}{2}} \quad \mathrm{I}_{\mathbb{R}}(t)$$

no caso de parente gaussiana e comentaram que, no caso do teorema do valor médio generalizado não ser exactamente válido, (4.3) fornece uma aproximação que merece ser investigada como alternativa à prespectiva tradicional de aceitar a aproximação por t_{n-1}.

Recorrendo aos conceitos de "conservadorismo" e "liberalidade" de testes, vamos investigar o comportamento da expressão (4.3) para a família gaussiana generalizada.

4.2 Estudo de robustez usando a Família Gaussiana Generalizada

A investigação da qualidade da aproximação apresentada em (4.3) no caso de a densidade parente não observar estritamente as hipóteses de Pestana e Rocha (1995) — nomeadamente a hipótese de validade do teorema do valor médio generalizado — mantendo-se, no entanto, a hipótese de simetria, foi abordada através das densidades parentes da família gaussiana generalizada. Estas têm a vantagem de ser uma perturbação uni-parametrizada da gaussiana, estando o parâmetro perturbador directamente relacionado com o achatamento ou curtose, e consequentemente avaliando o peso das caudas.

Considere-se a família de funções densidade de probabilidade

$$f(x|\mu,\sigma,\beta) = \frac{\exp\left[-\frac{1}{2}\left|\frac{x-\mu}{\sigma}\right|^{\frac{2}{1+\beta}}\right]}{2^{\frac{3+\beta}{2}}\sigma\Gamma(\frac{3+\beta}{2})} \quad \mathrm{I}_{\mathbb{R}}(x)$$

com $\mu \in \mathbb{R}, \sigma > 0, \beta \in (-1, 1]$. Sem perda de generalidade, considerámos a versão centrada ($\mu = 0$) e reduzida ($\sigma = 1$), uma vez que T_{n-1} é invariante relativamente à escala. Note-se que o parâmetro β cresce com a curtose, uma vez que

$$k = \frac{\Gamma\left(\frac{1+\beta}{2}\right)\Gamma\left(5\frac{1+\beta}{2}\right)}{\left[\Gamma\left(3\frac{1+\beta}{2}\right)\right]^2};$$

com $\beta = 0$ — isto é, parente gaussiana —, equacionando $\prod_{i=3}^{n} \left(1-\xi^2\right)^{\frac{i-4}{2}}$ com $\frac{\pi^{\frac{n-1}{2}}}{2^{n-2}\Gamma\left(\frac{n-1}{2}\right)}$ obtém-se a expressão

$$\xi_n = \sqrt{1 - \left(\frac{\sqrt{\pi}\Gamma\left(\frac{n-2}{2}\right)}{2\Gamma\left(\frac{n-1}{2}\right)} \right)^{\frac{2}{n-4}}}$$

que leva a que (4.3) coincida exactamente com

$$= \frac{1}{\sqrt{(n-1)}\beta\left(\frac{1}{2}, \frac{n-1}{2}\right)} \left(\frac{n-1}{n-1+t^2} \right)^{\frac{n}{2}} \quad \mathrm{I}_{\mathbb{R}}(t); \tag{4.4}$$

no caso $\beta \neq 0$, obtém-se

$$f_{T_{(n-1)}}(t|\beta) = \frac{(1+\beta)\Gamma\left(n\frac{1+\beta}{2}\right) \prod_{i=3}^{n} \left(1-\xi_i^2\right)^{\frac{i-4}{2}}}{4\sqrt{n-1}\left[\Gamma\left(\frac{3+\beta}{2}\right)\right]^n} \times \\ \times \frac{1}{\left\{ \sum_{i=1}^{n} \left| \frac{t}{\sqrt{n(n-1)}} + \alpha_{in} \right|^{\frac{2}{1+\beta}} \right\}^{n\left(\frac{1+\beta}{2}\right)}} \tag{4.5}$$

e, devido à simetria da densidade parente, a simetria de $f_{t_{(n-1)}}$ leva a que a expressão acima coincida com a que se obtém substituindo α_{in} por $-\alpha_{in}$.

A expressão (4.5) é tratável no caso $\frac{2}{1+\beta}$ par: expandindo o binómio e notando que os coeficientes dos termos ímpares são necessariamente nulos e $\sum_{i=1}^{n} \alpha_{in}^3 = 0$, o que leva a que reencontremos

$$\xi_n = 0 \qquad \text{ou} \qquad \xi_n^2 = \frac{3}{i+1} \qquad \text{e} \qquad \alpha_{in} = (n+1-2i)\sqrt{\frac{3}{n(n^2-1)}}$$

extensivamente discutidos em Rocha (1995) — mas agora como apro_ximação e não como expressão exacta, como é patente do facto de $f_{T_{(n-1)}}(t|\beta)$ com os referidos valores de ξ_i e α_{in} não ser uma função densidade de probabilidade.

Normalizando

$$f_{T_{(n-1)}}(t|\beta) = \frac{f^*_{T_{(n-1)}}(t|\beta)}{\int_{\mathbb{R}} f^*_{T_{(n-1)}}(t|\beta)}$$

com

$$f^*_{T_{(n-1)}}(t|\beta) =$$

$$= \frac{(1+\beta)\Gamma(\omega)\prod_{i=3}^{n}\left(\frac{i-2}{i+1}\right)^{\frac{i-4}{2}}}{4\sqrt{n-1}\left[\Gamma\left(\frac{3+\beta}{2}\right)\right]^n\left\{\sum_{i=1}^{n}\left|\frac{t}{\sqrt{n(n-1)}}+(n+1-2i)\sqrt{3}n(n^2-1)\right|^{\frac{2}{1+\beta}}\right\}^{\omega}}$$

sendo $\omega = n\frac{1+\beta}{2}$.

Tabela 4.1: $\varepsilon(n,\beta); \beta = -\frac{1}{2}, k = 2.2$

n	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$
2	0.1135	0.0590	0.0298	0.0120
3	0.1107	0.0621	0.0338	0.0145
4	0.1017	0.0572	0.0322	0.0147
5	0.0938	0.0517	0.0291	0.0137
6	0.0876	0.0470	0.0261	0.0123
7	0.0828	0.0432	0.0236	0.0109
8	0.0779	0.0402	0.0214	0.0098
9	0.0758	0.0377	0.0197	0.0088
10	0.0732	0.0356	0.0183	0.0080
11	0.0710	0.0339	0.0171	0.0073
16	0.0642	0.0285	0.0133	0.0052
21	0.0606	0.0256	0.0113	0.0041
26	0.0583	0.0239	0.0102	0.0035
31	0.0568	0.0227	0.0094	0.0031

A fim de avaliar o carácter liberal ou conservador da estatística, vamos observar as tabelas de

$$\varepsilon(n,\beta) = \mathbb{P}[T_{(n-1,\beta)} > t_{(n-1;1-\alpha)}] = \int_{t_{(n-1;1-\alpha)}}^{\infty} f_{T_{(n-1)}}(t|\beta)dt$$

onde $t_{(n-1;1-\alpha)}$ denota o quantil de probabilidade $(1-\alpha)$ da clássica t de Student. A probabilidade $\varepsilon(n,\beta)$ pode então ser interpretada como a verdadeira probabilidade de rejeitar H_0 dado que H_0 é verdadeira,

Tabela 4.2: $\varepsilon(n, \beta); \beta = -\frac{2}{3}, k = 2$

n	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$
2	0.1180	0.0628	0.0321	0.0129
3	0.1140	0.0664	0.0376	0.0168
4	0.1030	0.0603	0.0354	0.0170
5	0.0927	0.0532	0.0312	0.0154
6	0.0842	0.0470	0.0272	0.0135
7	0.0773	0.0418	0.0237	0.0116
8	0.0716	0.0376	0.0209	0.0100
9	0.0670	0.0341	0.0185	0.0087
10	0.0631	0.0313	0.0166	0.0076
11	0.0598	0.0289	0.0150	0.0067
16	0.0491	0.0212	0.0099	0.0040
21	0.0434	0.0172	0.0074	0.0027
26	0.0398	0.0148	0.0060	0.0020
31	0.0374	0.0132	0.0051	0.0016

Tabela 4.3: $\varepsilon(n, \beta); \beta = \frac{1}{2}, k = 4.2$

n	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$
2	0.0894	0.0439	0.0219	0.0087
3	0.0932	0.0422	0.0201	0.0078
4	0.1012	0.0464	0.0207	0.0076
5	0.1065	0.0507	0.0231	0.0080
6	0.1108	0.0542	0.0254	0.0089
7	0.1130	0.0570	0.0275	0.0099
8	0.1158	0.0589	0.0291	0.0109
9	0.1168	0.0607	0.0304	0.0117
10	0.1186	0.0618	0.0316	0.0124
11	0.1192	0.0631	0.0324	0.0130
16	0.1225	0.0663	0.0354	0.0150
21	0.1243	0.0680	0.0369	0.0162
26	0.1252	0.0691	0.0379	0.0169
31	0.1258	0.0698	0.0385	0.0174

ao nível α, usando abusivamente a hipótese de parente gaussiana (e, consequentemente, o ponto crítico $t_{(n-1;1-\alpha)}$ ao trabalhar com uma amostra de dimensão n.

Tabela 4.4: $\varepsilon(n, \beta); \beta = \frac{1}{4}, k = 3.6$

n	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$
2	0.0943	0.0467	0.0233	0.0093
3	0.0962	0.0457	0.0222	0.0087
4	0.1005	0.0479	0.0226	0.0086
5	0.1034	0.0502	0.0239	0.0089
6	0.1057	0.0521	0.0251	0.0094
7	0.1071	0.0536	0.0262	0.0099
8	0.1085	0.0547	0.0271	0.0104
9	0.1093	0.0597	0.0278	0.0108
10	0.1102	0.0363	0.0284	0.0112
11	0.1106	0.0570	0.0289	0.0115
16	0.1125	0.0588	0.0305	0.0126
21	0.1135	0.0598	0.0313	0.0132
26	0.1141	0.0604	0.0318	0.0135
31	0.1144	0.0608	0.0322	0.0138

Observando as Tabelas 4.1 e 4.2 concluímos que para valores de β tal que a curtose, k, é inferior a 3, o teste t revela-se conservador para valores de n superiores a 5 (note-se, no entanto, que para valores imediatamente a seguir a $n = 5$ o teste é conservador em níveis elevados de α, sendo liberal para níveis mais baixos). Para amostras de pequena dimensão ($n = 2$, 3 e 4) o teste é liberal, o valor de α é maior do que o declarado. Por outro lado, a análise das Tabelas 4.3 e 4.4 leva-nos a dizer, tal como Benjamini (1983), que para valores de β de forma que $k > 3$, ou seja, com uma distribuição com caudas mais longas do que a gaussiana standard, o teste t é conservador para amostras de pequena dimensão ($n < 7$). Repare-se que para $n = 4$, 5, 6 e 7 o teste é conservador para níveis de α razoavelmente pequenos. Para amostras com dimensão maior do que 7 o teste é liberal. O padrão detectado parece-nos geral, e em outra oportunidade apresentaremos evidência mais detalhada de que o peso das caudas, em situação de simetria, é o factor influente no comportamento do teste. As comparações parciais com o uso de $t_{(n-1)}$ levam-nos a crer que a aproximação $T_{(n-1)}$ é favorável.

4.3 Momentos e Cumulantes

Seja $\mathbf{X} = (X_1, ..., X_n)$ uma amostra de dimensão n, $X_j \stackrel{d}{=} X \frown F_X(x)$ e admita-se que $\mu_j{}' = E(X^j) = \int_{-\infty}^{\infty} x^j dF_X(x)$ (momento em relação à origem) existe (no sentido em que $\int_{-\infty}^{\infty} |x|^j dF_X(x) < \infty$). Como é usual, denota-se o valor médio $\mu_1{}' = E(X)$ por μ. Definem-se também os momentos centrais (ou momentos em relação à média), $\mu_j = E[(X-\mu)^j]$. Como é habitual, denotamos a variância $\mu_2 = E[(X-\mu)^2]$ por σ^2.

Da linearidade do operador de valor médio e do uso do binómio de Newton, facilmente se exprime μ_j em função de $\mu_k{}'$, $k \leq j$, sendo

$$\mu_j = \sum_{k=0}^{j} \binom{j}{k} (-1)^k \mu_1{}'^k \mu'_{j-k}$$

e $\mu_k{}'$ em função de μ_j, $j \leq k$, sendo

$$\mu_k{}' = \sum_{j=0}^{k} \binom{k}{j} \mu_1{}'^j \mu_{k-j}.$$

O valor médio e a variância, como parâmetros de localização e escala, não só têm uma interpretação imediata, como uma relevância especial conferida pela desigualdade de Chebycheff, tão apreciável pela sua generalidade, cujo preço é, naturalmente, fornecer limites muito grosseiros.

A localização e a escala determinam completamente a mais importante família de variáveis aleatórias, $X \frown Gaussiana(\mu, \sigma)$, $\mu \in \mathbb{R}$, $\sigma > 0$. As razões que nos levam a considerar esta família a mais importante são:

* $\{ X \frown Gaussiana(\mu, \sigma),\ \mu \in \mathbb{R},\ \sigma > 0\}$ é uma família de localização — escala e $\frac{X-\mu}{\sigma} \frown Gaussiana(0, 1)$, a gaussiana padrão.

* O Teorema Limite Central, T.L.C. (em condições muito genéricas, a soma de variáveis aleatórias convenientemente normalizadas, isto é, relocalizadas em 0 e com a variância estabilizada, habitualmente em 1, tem distribuição aproximadamente gaussiana). Muitas variáveis

aleatórias podem ser decompostas em somas e, consequentemente, aproximadas por uma gaussiana com igual valor médio e igual variância.
No T.L.C. clássico as parcelas são i.i.d. com variância finita. Ao longo do século XX qualquer daquelas hipóteses foi consideravelmente relaxada. O T.L.C. com as condições de Lindeberg-Feller relaxou a hipótese de identidade distribucional. Há formas do T.L.C. com diversas condições de dependência fraca (markoviana, de martingalas, estacionaridade, permutabilidade) — veja-se, apenas a título de exemplo, Fraser (1957, p. 219) — e a teoria dos domínios de atracção de Lévy-Doeblin-Gnedenko mostrou, num contexto muito mais geral, que basta que a sucessão de momentos truncados de segunda ordem seja uniformemente limitada (uma condição que pode ser reformulada em termos de variação regular da soma das caudas $(F(-x)+1-F(x))$ para se obter convergência da sucessão de somas parciais, convenientemente normalizadas, para uma gaussiana.

* Os momentos empíricos $\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$ e $S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X})^2$ são independentes, uma propriedade que caracteriza a família gaussiana. Esta propriedade tem consequências notáveis no cálculo da distribuição amostral de uma variável fulcral de grande importância em inferência estatística, a t de Student, $t = \sqrt{n}\,\frac{\overline{X}-\mu}{S}$.

* Se $\mathbf{X}$ for um vector aleatório multi-gaussiano — numa simplificação grosseira, se qualquer combinação linear de componentes de $\mathbf{X}$ for gaus siana — a correlação nula entre componentes implica independência, e a regressão de qualquer margem univariada noutra margem (e estamos apenas a referir uma situação particularmente simples) é linear.

* A família gaussiana pertence à intersecção de duas famílias mais vastas, a família exponencial de Fisher-Darmois-Koopman e a família das estáveis de Lévy.

Se a localização e a escala determinam completamente a família de variáveis aleatórias $X \frown \mathit{Gaussiana}(\mu, \sigma)$, $\mu \in \mathbb{R}$, $\sigma > 0$, o mesmo não acontece com muitas outras variáveis aleatórias importantes em que $\mu_3 = E[(X-\mu)^3]$ é um bom indicador do peso relativo das caudas direita e esquerda, e $\mu_4 = E[(X-\mu)^4]$ é um indicador suplementar sobre

o peso das caudas da distribuição.

É, em geral, mais interessante usar versões estandardizadas $\gamma_1 = \frac{\mu_3}{\sigma^3}$, como coeficiente de assimetria e $\gamma_2 = \frac{\mu_4}{\sigma^4} - 3$, como coeficiente de achatamento. Se $Z \frown Gaussiana(0,1)$, $E(Z^{2j-1}) = 0$, $j \geq 1$ e $E(Z^{2j}) = \frac{(2j)!}{2^j j!}$, de onde é fácil calcular, quer os momentos ordinários, quer os momentos centrais de $X \frown Gaussiana(\mu, \sigma)$. Assim,

$\mu_1{}' = \mu = E(X)$,
$\mu_2 = \sigma^2 = E((X-\mu)^2))$,
$\gamma_1 = 0$, (pois $\mu_{2j-1} = 0$, devido à simetria de X em relação ao eixo $x = \mu$) e
$\gamma_2 = 0$.

Tem também de grande interesse o valor médio de e^{tX}, $M_X(t) = E(e^{tX})$ que, sendo $e^{tX} = \sum_{j=0}^{\infty} \frac{(tX)^j}{j!}$, admite a expansão em série de MacLaurin,

$$M_X(t) = \sum_{j=0}^{\infty} E(X^j) \frac{t^j}{j!}$$

pelo que se chama função geradora dos momentos. Note-se que $M_X(t)$ está definida numa banda vertical do plano complexo que contém o eixo imaginário — e recorde-se que $E(e^{itX}) = \varphi_X(t)$, a função característica, existe sempre, pois $|E(e^{itX})| \leq \int_{\Re} |e^{itx}| dF_X(x) = 1$ —, podendo degenerar nesse eixo. A sucessão dos momentos não determina uma distribuição, existindo, no entanto, condições suficientes para que tal aconteça (veja-se o critério de Carleman em Feller (1971, p. 224), ou o tratamento exaustivo em Shohat e Tamarkin (1943). Por outro lado, se $M_X(t)$ estiver definida para $|t| \leq \tau > 0$, existem momentos de todas as ordens e $M_X(t)$ determina univocamente $F_X(x)$.

Em finais do século XIX, Thiele [1] definiu os cumulantes de uma distribuição. A ideia foi retomada por Fisher nos anos 30: os cumulantes são, simplesmente, os coeficientes κ_j da expansão em série de MacLaurin

[1] A tradução de *Theory of Observations*, originalmente publicada em 1897, foi reimpressa nos *Ann. Math. Statist.*, 1931, **2**, 165-308.

do logaritmo da função geradora de momentos

$$C_X(t) = \ln M_X(t) = \sum_{j=0}^{\infty} \kappa_j \frac{t^j}{j!}$$

com a convenção $\kappa_0 = 1$.

A partir da igualdade

$$\sum_{j=0}^{\infty} \kappa_j \frac{t^j}{j!} = ln(1 + \sum_{j=1}^{\infty} {\mu_j}' \frac{t^j}{j!})$$

obtemos as relações

i) Relação entre cumulantes e momentos em relação à origem
$\kappa_1 = {\mu_1}'$
$\kappa_2 = {\mu_2}' - ({\mu_1}')^2$
$\kappa_3 = {\mu_3}' - 3{\mu_2}'{\mu_1}' + 2({\mu_1}')^3$
$\kappa_4 = {\mu_4}' - 4{\mu_3}'{\mu_1}' - 3({\mu_2}')^2 + 12{\mu_2}'({\mu_1}')^2 - 6({\mu_1}')^4$
$\kappa_5 = {\mu_5}' - 5{\mu_4}'{\mu_1}' - 10{\mu_3}'{\mu_2}' + 20{\mu_3}'({\mu_1}')^2 + 30({\mu_2}')^2{\mu_1}' - 60{\mu_2}'({\mu_1}')^3 + 24({\mu_1}')^5$
...

ii) Relação entre cumulantes e momentos em relação à média
$\kappa_1 = \mu$
$\kappa_2 = \mu_2 = \sigma^2$
$\kappa_3 = \mu_3$
$\kappa_4 = \mu_4 - 3{\mu_2}^2$
$\kappa_5 = \mu_5 - 10\mu_3\mu_2$
$\kappa_6 = \mu_6 - 15\mu_4\mu_2 - 10{\mu_3}^2 + 30{\mu_2}^3$
...

iii) Relação entre momentos em relação à média e cumulantes
$\mu_1 = \kappa_1$
$\mu_2 = \kappa_2$
$\mu_3 = \kappa_3$
$\mu_4 = \kappa_4 + 3\kappa_2^2$
$\mu_5 = \kappa_5 + 10\kappa_3\kappa_2$
$\mu_6 = \kappa_6 + 15\kappa_4\kappa_2 + 10\kappa_3^2 + 15\kappa_2^3$
...

Assim, mesmo que não exista $M_X(t)$, no caso de existir ${\mu_n}'$ também existem os cumulantes $\kappa_j, j = 1, ..., n$.

Também se definem os cumulantes estandardizados

$$\rho_j(X) \;=\; \frac{\kappa_j(X)}{\sigma^j} \tag{4.6}$$

Note-se que

$$\rho_1(X) = \frac{\mu}{\sigma},\ \rho_2(X) = 1,\ \rho_3(X) = \gamma_1,\ \rho_4(X) = \gamma_2.$$

Os cumulantes não têm uma interpretação intuitiva imediata, como os momentos. No entanto, do ponto de vista operativo oferecem grande facilidade de cálculo. Em particular, definindo a função geradora de momentos do vector $\mathbf{X} = (X_1, ..., X_n)$ como

$$M_{\mathbf{X}}(\mathbf{t}) = E[e^{\mathbf{tX}}] = E[e^{t_1X_1+t_2X_2+...+t_nX_n}]$$

e denotando $S_n = \sum_{j=1}^{n} X_j$, tem-se que

$$M_{S_n}(t) = M_{\mathbf{X}}(t, t, ..., t)$$

e, caso as parcelas X_j sejam independentes

$$M_{S_n}(t) = M_{X_1}(t)M_{X_2}(t)...M_{X_n}(t).$$

Se, além de independentes, forem identicamente distribuídas, $X_j \stackrel{d}{=} X$, então,

$$M_{S_n}(t) = [M_X(t)]^n$$

e da definição da função geradora de cumulantes,

$$C_{S_n}(t) = ln[M_X(t)]^n = nC_X(t) \Rightarrow \kappa_j(S_n) = n\kappa_j(X),$$

uma das razões pelas quais consideramos o uso de cumulantes cómodo. Por outro lado, o efeito de centrar apenas altera o termo linear na expansão em série de MacLaurin

$$M_{X-\mu}(t) = e^{-\mu t}M_X(t) \Rightarrow C_{X-\mu}(t) = C_X(t) - \mu t$$

o que mostra, que para $n \geq 2$, os cumulantes são invariantes no que respeita a adição de constantes à variável parente X; por outras palavras, são funções dos momentos centrados. É também imediato que

$$\kappa_j(\alpha X) = \alpha^j \kappa_j(X).$$

Da expressão $C_{S_n}(t) = nC_X(t)$ decorre imediatamente que

$$\rho_j(S_n) = \frac{\rho_j(X)}{n^{\frac{j}{2}-1}},$$

sendo $\rho_j(X)$ dados por (4.6) e, em particular,

$$\rho_1(S_n) = \sqrt{n}\,\rho_1(X),\ \rho_2(S_n) = 1,\ \rho_3(S_n) = \frac{\rho_3(X)}{\sqrt{n}},\ \rho_4(S_n) = \frac{\rho_4(X)}{n}.$$

Então, em relação à média empírica $\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$, fica demonstrado que

$$E(\overline{X}) = \mu,\ var(\overline{X}) = \frac{\sigma^2}{n},\ \gamma_1(\overline{X}) = \frac{\gamma_1(X)}{\sqrt{n}},\ \gamma_2(\overline{X}) = \frac{\gamma_2(X)}{n}$$

o que evidencia os altos e baixos do poder regularizador das médias: o valor médio da média é o valor médio populacional, mas a variabilidade $\frac{\sigma^2}{n}$ de $\overline{X}$ em torno de μ é, consideravelmente, inferior à de X.
Também, no que respeita o achatamento, $\gamma_2(\overline{X})$ vai decrescendo para 0 (o achatamento da gaussiana) à velocidade de $\frac{1}{n}$. Mas, no que respeita o coeficiente de assimetria, $\gamma_1(\overline{X})$, o seu decrescimento para 0 é lento, acompanha $\frac{1}{\sqrt{n}}$.

Apresentamos alguns exemplos:

Exemplo 4.3.1. Sendo $X \frown Gaussiana(\mu, \sigma)$, $f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\ exp\ (-\frac{1}{2}\ (\frac{x-\mu}{\sigma})^2), I_{\Re}$. Facilmente se calcula a função geradora de cumulantes:

$$M_X(t) = E[e^{tX}] = e^{\mu t + \frac{1}{2}\sigma^2 t^2} \Rightarrow C_X(t) = \mu t + \frac{1}{2}\,\sigma^2 t^2.$$

Então, $\kappa_1 = \mu,\ \kappa_2 = \sigma^2$ e $\kappa_j = 0$, para $j \geq 3$ (característica usada em testes de gaussianidade). O facto de $\kappa_j = 0$, para $j \geq 3$ evidencia que $X \frown Gaussiana(\mu, \sigma)$ é o modelo perfeito de uma família de localização — escala.

Exemplo 4.3.2. Se $X \frown Poisson(\lambda)$ a sua f.m.p. é $P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!},\ k = 0, 1, 2, ...$
Tem-se

$$M_X(t) = E[e^{tX}] = e^{\lambda(e^t-1)} \Rightarrow C_X(t) = \lambda(e^t - 1) = \lambda \sum_{j=1}^{\infty} \frac{t^j}{j!}$$

concluindo-se que $\kappa_j = \lambda,\ j \geq 1$ (uma caracterização da Poisson).

Exemplo 4.3.3. Sendo $X \frown Binomial(n, p)$ com f.m.p. dada por $P(X = k) =$
$\binom{n}{k}\ p^k\ (1-p)^{n-k}\ k = 0, 1, ..., n$, a função geradora de momentos é

$$M_X(t) = E[e^{tX}] = [1 - p\ (1 - e^t)]^n \Rightarrow C_X(t) = nln[1 - p\ (1 - e^t)].$$

Depois de alguns cálculos obtém-se

$$C_X(t) = n \sum_{j=1}^{\infty} \frac{(-1)^{2j+1}}{j} \sum_{r=0}^{j} \binom{j}{r} (-1)^r p^j \sum_{k=0}^{\infty} \frac{(rt)^k}{k!}$$

sendo os três primeiros cumulantes dados por

$$\kappa_1 = np,\ \kappa_2 = np\ (1-p),\ \kappa_3 = np\ (1 - 3p + 2p^2).$$

Exemplo 4.3.4. Seja $X \frown Gama(\alpha, \delta)$. Sendo X uma soma de v.a. i.i.d. com $Y \frown Exponencial(\delta)$, cuja função geradora de momentos é $M_Y(t) = \frac{1}{1-\delta t}, t < \frac{1}{\delta}$, então

$$M_X(t) = (1 - \delta t)^{-\alpha}, t < \frac{1}{\delta} \Rightarrow C_X(t) = -\alpha \ln(1 - \delta t) = \alpha \sum_{j=1}^{\infty} \frac{(\delta t)^j}{j}.$$

Os cumulantes são então

$$\kappa_1 = \alpha\delta,\ \kappa_2 = \alpha\delta^2,\ \kappa_3 = 2\alpha\delta^3,\ \kappa_4 = 3!\alpha\delta^4,$$

ou seja, $\kappa_r = (r-1)!\alpha\delta^r$.

Note-se que $\rho_1\rho_3 = 2$ para qualquer distribuição gama, por outras palavras, o coeficiente de assimetria é sempre o dobro do coeficiente de variação.

Complementos:

1.

$$\begin{aligned} var[(X-\mu)^2] &= E[(X-\mu)^4] - \{E[(X-\mu)^2]\}^2 = \mu_4 - \sigma^4 \\ &= \sigma^4(\frac{\mu_4}{\sigma^4} - 1 - 2 + 2) = \sigma^4(\gamma_2 + 2) \\ &\Rightarrow \gamma_2 \geq -2 \end{aligned}$$

Por outro lado, considerando a matriz de covariâncias (necessariamente semi-definida positiva) do vector $(X-\mu, (X-\mu)^2)$

$$\begin{bmatrix} \sigma^2 & \sigma^3\gamma_1 \\ \sigma^3\gamma_1 & \sigma^4(\gamma_2+2) \end{bmatrix}$$

conclui-se que $\gamma_2 + 2 \geq {\gamma_1}^2$. A igualdade dá-se se $(X-\mu)^2)$ for q.c. uma função linear de $X-\mu$, isto é, se $X \frown Bernoulli(p)$.
As desigualdades envolvendo momentos centrais que se obtêm trabalhando os menores da matriz de covariâncias do vector $\big(X-\mu$, $(X-\mu)^2), ..., (X-\mu)^r\big)$ parecem excessivamente complicadas para merecerem exploração cuidada.

2. A função geradora de momentos $E(e^{tX})$ é usada em alguns textos elementares para estabelecer resultados importantes sobre somas de variáveis aleatórias independentes e seus limites, por exemplo, o T.L.C. Tem a desvantagem de poder não existir para $\mathrm{Re}(t) \neq 0$. Como atrás comentámos, no eixo imaginário é $E(e^{itX})$, a função característica, para a qual não se põe qualquer problema de existência para t real e que, por isso, é, em geral, preferida em textos mais avançados de Teoria da Probabilidade.
Mas, a importância central que a família exponencial de Fisher-Darmois-Koopman veio a ganhar na moderna inferência estatística recolocou a função geradora de momentos (que é, afinal, uma transformada de Laplace) num plano de importância teórica que convém sublinhar.

A família exponencial inclui todas as distribuições discretas, contínuas ou mistas, cuja função massa de probabilidade ou função densidade de probabilidade pode ser expressa na forma

$$f_X(x;\theta) = \exp\{a(\theta)b(x) + c(x) + d(\theta)\},$$

em que a(.), b(.), c(.), d(.) são funções reais conhecidas. Se $b(x) = x$, tem-se a representação natural ou canónica da família exponencial univariada, sendo $a(\theta)$ o parâmetro natural ou canónico. Por exemplo, se X é uma v.a. gaussiana de valor médio θ e variância unitária,

$$\begin{aligned} f_X(x;\theta) &= \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}(x-\theta)^2\} \\ &= \exp\{\theta x - \frac{1}{2}\,\theta^2 - \frac{1}{2}(x^2 + \log(2\pi))\}, \end{aligned}$$

com

$$a(\theta) = \theta; \quad d(\theta) = -\frac{1}{2}\,\theta^2; \quad c(x) = -\frac{1}{2}(x^2 + \log(2\pi)).$$

A função geradora de momentos é

$$\begin{aligned} M_X(t) &= \int_{-\infty}^{\infty} e^{tx} f_X(x;\theta)dx \\ &= \exp\{d(\theta)\} \int_{-\infty}^{\infty} \exp\{(\theta+t)x + c(x)\}dx \\ &= \exp\{d(\theta) - d(\theta+t)\}. \end{aligned}$$

Assim, quando X tem a densidade expressa na forma canónica, a sua função geradora de cumulantes é

$$C_X(t) = d(\theta) - d(\theta + t).$$

4.4 Análise de Escala em Pequenas Amostras

4.4.1 Introdução

A grande variedade de modelos usados em análise de variância não nos deve fazer esquecer as ideias simples que lhe estão subjacentes: a avaliação da variância comum de várias populações deve ser influenciada, de forma detectável, por diferentes localizações e, assim, o problema residirá na comparação de estimativas da variância que revelem diferentes fontes de variação.

Uma vez que o teste de forma gaussiana de Iglewicz (1981) mostra que a situação gaussiana é mais a excepção do que a regra, será importante analisar a escala em situações mais gerais. E, neste sentido mais geral, a escala deve ser tomada como uma ideia vaga, a substanciar em cada população particular. Claro que a extensão da análise da variância para uma análise de escala mais geral exige, de qualquer modo, que a concretização de ideia de escala tenha como referência uma localização e que heterogeneidade de localizações seja influente e, eventualmente detectável, nas diferenças observadas nas estimativas da escala.

A relevância dos resultados obtidos com as extensões da análise de variância a situações não gaussianas que conhecemos, perde-se, em nosso entender, pelo facto de não apresentar uma teoria distribucional exacta para pequenas amostras. Estas extensões recorrem ao quociente de diferentes estimadores da variabilidade obtendo-se apenas, em geral, a distribuição assintótica.

O presente trabalho insere-se no desenvolvimento de estatísticas pseudo-F's como abordagem geral ao problema de comparação de localizações de populações com base em $k \geq 2$ amostras, no caso de haver um parâmetro de escala cuja estimativa possa ser particionada de forma relevante para a análise. Este trabalho é, consequentemente, uma generalização dos resultados de Fisher ao lidar-se com pequenas amostras em situações não gaussianas e está no seguimento da abordagem à análise de escala proposta por Pestana e Rocha (1993). Procurámos estudar o quociente de dois estimadores independentes da escala em situações de pequenas amostras provenientes de populações uniforme e exponencial. Numa primeira abordagem usaram-se estimadores de máxima verosimilhança, quer da localização, quer da escala.

4.4.2 População Uniforme

Sendo a distribuição uniforme um caso especial das distribuições beta, obtida quando se considera p e q iguais a 1, é natural termos iniciado o estudo de estatísticas pseudo-F's supondo como distribuição parente a uniforme.

Para uma variável aleatória X com distribuição uniforme no intervalo

$(a-b, a+b)$, a função densidade de probabilidade (f.d.p.) é da forma

$$f_X(x) = \frac{1}{2b} \qquad a-b \leq x \leq a+b, b>0, \tag{4.7}$$

sendo a função de distribuição (f.d.)

$$F_X(x) = \begin{cases} 0 & , x < a-b \\ \frac{x-(a-b)}{2b} & , a-b \leq x \leq a+b \\ 1 & , x > a+b \end{cases}$$

Uma forma alternativa a (4.8) é

$$f_X(x) = \frac{1}{\beta - \alpha} \qquad \alpha \leq x \leq \beta, \beta > \alpha.$$

A forma standard da distribuição uniforme é obtida quando $\alpha = 0$ e $\beta = 1$.

Este caso particular é de grande interesse devido ao conhecido teorema: se X é uma variável aleatória contínua, com f.d. $F_X(x)$, então a variável aleatória $Y = F_X(x)$ tem distribuição uniforme no intervalo (0,1). A mudança de variável $Y = F_X(x)$ é denominada transformação uniformizante.

Seja $(X_1, X_2, ..., X_n)$ uma amostra aleatória extraída de uma população modelada por $X \frown U_{(a-b,a+b)}$. Representemos por $(X_{1:n}, X_{2:n}, ..., X_{n:n})$ as e.o. ascendentes associadas a $(X_1, X_2, ..., X_n)$.

O estimador de máxima verosimilhança do parâmetro b é

$$\hat{b} = \frac{1}{2}(X_{n:n} - X_{1:n})$$

e o estimador de máxima verosimilhança do desvio padrão populacional é

$$\hat{\sigma} = \frac{1}{\sqrt{12}}(X_{n:n} - X_{1:n})$$

o qual é um estimador robusto ajustado (Johnson and Kotz, 1995).

Estes estimadores têm a particularidade de serem função da amplitude e, segundo Pearson: "[...] When dealing with samples containing only a small number of observations the range may often be usefully employed as a measure of dispersion [...]" (Pearson, 1941–42).

Assim, a primeira estatística pseudo-F a estudar tem a forma de um quociente entre dois estimadores de escala obtidos a partir de duas amostras aleatórias $(X_1, ..., X_n)$ e $(Y_1, ..., Y_m)$ supostamente provenientes de população uniforme

$$F^*_{n-1,m-1} = \frac{X_{n:n} - X_{1:n}}{Y_{m:m} - Y_{1:m}} = \frac{W_n}{W_m} \tag{4.8}$$

No que se segue usamos, sem perda de generalidade, a forma standard da distribuição uniforme.

Para podermos calcular a f.d.p. de (4.9) necessitamos da f.d.p. de $W_n = X_{n:n} - X_{1:n}$. No caso de distribuição uniforme é da forma

$$f_{W_n}(w) = n(n-1)w^{n-2}(1-w), \qquad 0 \leq w \leq 1.$$

A densidade de $F^*_{n-1,m-1}$ é, então calculada a partir da expressão

$$f_{F^*}(t) = \int_{-\infty}^{\infty} |y| f_{W_n}(ty) f_{W_m}(y) dy. \tag{4.9}$$

Depois de alguns cálculos obtemos

$$f_{F^*_{n-1,m-1}}(t) = \begin{cases} nm(n-1)(m-1)\left(\dfrac{t^{n-2}}{(n+m-2)(n+m-1)} - \dfrac{t^{n-1}}{(n+m)(n+m-1)}\right), \\ \quad 0 \leq t \leq 1 \\ \\ nm(n-1)(m-1)\left(\dfrac{1}{(n+m-2)(n+m-1)t^m} - \dfrac{1}{(n+m)(n+m-1)t^{m+1}}\right), \\ \quad t \geq 1. \end{cases}$$

4.4.3 População Exponencial

Prosseguindo o estudo da estatística (4.9) foi usada a distribuição exponencial como distribuição parente. A opção por esta distribuição tem a ver com o facto de, além de ser um caso especial da distribuição gama(n, α), com $n = 1$, os cálculos matemáticos que lhe estão associados são muitas vezes de natureza simples.

Para uma variável aleatória X com distribuição exponencial a f.d.p. é

$$f_X(x) = \frac{1}{\sigma}\ exp\left(-\frac{x-\theta}{\sigma}\right), \quad x > \theta, \sigma > 0$$

sendo a f.d.

$$F_X(x) = 1 - exp\left(-\frac{x-\theta}{\sigma}\right), \quad x > \theta, \sigma > 0.$$

Obtém-se a forma standard quando $\theta = 0$ e $\sigma = 1$. Foi esta forma que usámos para calcular a distribuição de (4.9) uma vez que ao fazer as diferenças $X_{n:n} - X_{1:n}$ descartamos a localização e ao fazer o quociente é irrelevante a escala da população parente.

Neste caso a f.d.p. de $W_n = X_{n:n} - X_{1:n}$ é

$$f_{W_n}(w) = (n-1)exp(-w)[1 - exp(-w)]^{n-2}, \quad w > 0.$$

De facto, as propriedades markovianas das e.o. vêm, neste caso especial, reforçadas pela "falta de memória"da exponencial. Consequentemente, $X_{n:n} - X_{1:n} \stackrel{d}{=} X_{n-1:n-1}$.

A partir da expressão (4.10) obtemos

$$f_{F^*}(t) = (n-1)(m-1) \cdot \int_0^\infty y\ \ exp[-y(t+1)][1 - exp(-ty)]^{n-2}[1 - exp(-y)]^{m-2}dy.$$

Fazendo a transformação u=exp(-y) e atendendo a propriedades da função Beta,

$$f_{F^*}(t) = (n-1)(m-1)$$
$$\cdot \sum_{k=0}^{n-2} \binom{n-2}{k} (-1)^{k+1} B(m-1, (k+1)t+1)$$
$$\cdot [\psi((k+1)t+1) - \psi((k+1)t+m)].$$

Tendo em conta que

$$\psi(z+1) = \frac{1}{z} + \psi(z)$$

$$\psi(z+n) = \sum_{k=1}^{n} \frac{1}{z+(n-k)} + \psi(z)$$

vem, finalmente,

$$f_{F^*_{n-1,m-1}}(t) = (n-1)(m-1)$$
$$\cdot \sum_{k=0}^{n-2} \sum_{s=1}^{m-1} \binom{n-2}{k} (-1)^{k+2} \ \frac{B(m-1, (k+1)t+1)}{(k+1)t+(m-s)}, \ \ t > 0.$$

Como o grande biólogo inglês Huxley afirmou, também em Ciência as grandes ideias nascem como blasfémias e morrem como dogmas.
As ideias de Winsor, de que na parte central todas as distribuições são "normais", teve um impacto poderoso na evolução da Estatística, mas foi pouco a pouco ficando absoleta, por haver inúmeras situações em que o modelo gaussiano não é, de todo, adequado. E o advento de meios computacionais permitiu o desenvolvimento da estatística resistente e/ou robusta, em que se "deixa os dados falar por si mesmo", em vez de se lhes impor um modelo gaussiano. Já em 1961 Hotelling criticava acutilantemente as ideias de Winsor com uma argumentação irrebatível. A "ditadura" da gaussiana provocou uma revolta extrema: na Estatística não paramétrica apenas se pressupõe que os dados provêm de uma $F_X(x)$ contínua (apesar de na prática se observarem empates), e desprezarem-se as magnitudes dos dados, pesando x_i por $w_i = \frac{r_i}{x_i}$ — isto é, passamos aos *ranks*, uma forma extremista, quase terrorista, de

moderar a dispersão dos dados.

Felizmente, a abordagem posterior, procurando resultados resistentes e/ou robustos, e procurando novos desenvolvimentos teóricos que permitam resultados, quer exactos, quer assintóticos numa grande variedade de situações não gaussianas, tem trazido novas ideias, como a studentização interna e a análise de escala, com estudo de distribuições "pseudo-F´s".

Muitos temas estimulantes se põem nesta nova perspectiva. Nomeadamente, imergindo a rectangular na família das betas, tem interesse investigar a robustez destes resultados quando a parente é B(μ, ν), com $\mu \approx 1$ e $\nu = 1$, em vez de B(1,1).

Capítulo 5

A inferência sobre localização revisitada

5.1 Introdução

A Estatística Matemática teve um desenvolvimento espectacular quando Student, desenvolveu a teoria de como testar uma hipótese sobre o valor médio de uma população $X \frown Gaussiana(\mu, \sigma)$ — ou, alternativamente, construir um intervalo de confiança para μ — no caso da escala σ ser desconhecida. Por analogia ao que se faz no caso de σ ser conhecido, i. é, a normalização $\dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} \frown Gaussiana(0, 1)$, Student fez o que é justo designar por "studentização" da gaussiana, $\dfrac{\overline{X} - \mu}{S/\sqrt{n}}$, onde $\overline{X} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$ e $S^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(X_i - \overline{X})^2$ são estimadores centrados de μ e de σ , respectivamente.

Mais precisamente, Student (1908) deduziu a distribuição amostral daquela função dos momentos empíricos da gaussiana. Este trabalho pioneiro de Student foi, além de um golpe de génio, um golpe de sorte. $\overline{X}$ e S^2 são estatísticas independentes apenas se a população parente for gaussiana. Em todos os outros casos, os cálculos, mesmo em populações com uma caracterização tão simples como a uniforme, são uma tarefa hercúlea, e as expressões apresentadas por Perlo (1933) para a densidade de $\frac{\overline{X}-\mu}{S/\sqrt{n}}$ para o caso $n = 3$, em população uniforme, desencorajaram,

decerto, potenciais interessados, optando-se por alternativas à dedução da distribuição exacta daquela variável:

1. A solução mais "preguiçosa" foi considerar que

$$\frac{\overline{X}-\mu}{S/\sqrt{n}} = \sum_{i=1}^{n} \frac{X_i - \mu}{\sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(X_i - \overline{X})^2}}$$

   é uma soma de n v.a. i.i.d. e que, portanto, se deve invocar o T.L.C. (e recorrer à tabela da $Z \frown Gaussiana(0,1)$) e tomar decisões face aos valores observados e sua posição relativamente aos quantis extremos de referência. Quem defendia esta situação raramente se preocupava com questões de velocidade de convergência do T.L.C. e a crítica demolidora de Hotelling (1961) não deixou de sublinhar que, ainda que a convergência na zona central seja, em geral, razoável, as decisões são tomadas face a probabilidades de caudas.

2. Uma situação, um pouco menos "preguiçosa", defendia que se usasse parte da amostra para estimar $\overline{X}$ e a outra parte para estimar S "externamente", com o intuito de se ter um quociente de estatísticas independentes cuja densidade é mais fácil de deduzir. A objecção natural a esta solução é o desperdício de informação inerente à sua filosofia. Teve, no entanto, a vantagem de colocar a studentização numa perspectiva mais vasta: quando se pretende fazer inferência sobre um parâmetro de localização λ usando uma variável $g_1(\mathbf{X}) - \lambda$ cuja densidade depende de um parâmetro perturbador de escala δ, procura-se um estimador $g_2(\mathbf{X})$ de δ tal que

$$\frac{g_1(\mathbf{X}) - \lambda}{g_2(\mathbf{X})}$$

   tenha uma distribuição independente do parâmetro perturbador. Pelas razões, acima apontadas, é habitual falar de "studentização externa" quando $g_1(\mathbf{X})$ e $g_2(\mathbf{X})$ são independentes e de "studentização interna" quando $g_1(\mathbf{X})$ e $g_2(\mathbf{X})$ são dependentes. Veja-se David (1981) ou Brilhante *et al.* (2001) para mais detalhes.

3. Uma solução mais radical consistiu na procura de um paradigma totalmente diverso: se as contas são irremediavelmente horríveis mesmo para as populações com caracterizações mais simples,

abandona-se o paradigma paramétrico e procede-se a inferência sobre a localização central usando apenas considerações sobre aspectos combinatórios decorrentes da permutabilidade das componentes da amostra aleatória $\mathbf{X} = (X_1, ..., X_n)$ ou à estrutura de ordem do vector das estatísticas ordinais.

4. Sob a influência marcante de Tukey (cujos resultados estão sintetizados em Tukey (1977) e Mosteller e Tukey (1977)) e como resposta às críticas de Hotelling (1961), capitalizando nas novas capacidades computacionais disponíveis, a studentização enveredou por novos rumos usando estimadores robustos e/ou resistentes "aparando", por exemplo, os elementos extremos da amostra.

Uma vez que a questão parece indissociável da simetria da população de base, a secção seguinte é dedicada ao estudo de funções dos momentos do numerador e do denominador de $\dfrac{\overline{X} - \mu}{S/\sqrt{n}}$, e momentos desta variável com ênfase em γ_1 e γ_2. O resultado mais notável que encontrámos foi que a assimetria de t contraria a assimetria da população parente, o que arrasta erros na avaliação das probabilidades das regiões críticas (p-values), nomeadamente no caso de testes unilaterais.

5.2 Influência da assimetria e da curtose na velocidade de convergência de $t_{(n)}$ para $\mathbf{Z} \frown$ Gaussiana(0,1)

Seja $\mathbf{X} = (X_1, ..., X_n)$ uma amostra de dimensão n, e admita-se que $X_j \stackrel{d}{=} X \frown F_X(x)$ com $\mu_4{}' = \displaystyle\int_{-\infty}^{\infty} x^4 dF_X(x) < \infty$. Fica assim garantida a existência dos momentos centrais μ_j e dos cumulantes $\kappa_j,\ j = 0, 1, ..., 4$, com a convenção $\mu_0 = \kappa_0 = 1$.

Nestas condições não se põe qualquer dúvida na convergência de

$$t_{(n-1)} = \sqrt{n}\,\frac{\overline{X}_n - \mu}{S_n}$$

onde

$$\mu = \mu_1^{'} = \int_{\mathbb{R}} x dF_X(x), \quad \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad e \quad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

para $Z \frown Gaussiana(0,1)$ — bastaria, aliás, a existência de $\mu_2{}^{'}$ para assegurar o resultado.

Com respeito à estatística $\overline{X}_n$ que aparece no numerador, vimos já que

$$\mathbb{E}(\overline{X}_n) = \mu, \;\; var(\overline{X}_n) = \frac{\sigma^2}{n}, \;\; \gamma_1(\overline{X}_n) = \frac{\gamma_1(X)}{\sqrt{n}}, \;\; \gamma_2(\overline{X}_n) = \frac{\gamma_2(X)}{n},$$

enquanto que $\gamma_1(Z) = \gamma_2(Z) = 0$. Assim, a curtose $\gamma_2(\overline{X}_n) \underset{n\to\infty}{\longrightarrow} 0$ mais rapidamente do que a assimetria $\gamma_1(\overline{X}_n)$, que também tende para zero, mas, com a velocidade de $\frac{1}{\sqrt{n}}$. De qualquer modo, em muitas distribuições importantes a assimetria é moderada. As afirmações de van Belle (2008) — ainda que noutro contexto — levam-nos a admitir que o poder regularizador das somas funciona, no numerador, com bastante rapidez, e que para $n \geq 12$, em muitas situações, aproximar $\overline{X}_n$ por $X \frown Gaussiana(\mu, \sigma)$, não é descabido.

Já o comportamento do denominador é menos favorável, pois, apesar de S_n^2 ser um estimador centrado de σ^2, tem-se

1. $var(S_n^2) = \sigma^4 \, (\frac{2}{n-1} + \frac{\gamma_2}{n})$
2. $\mathbb{E}(S_n) \simeq \sigma - \frac{\sigma}{8} \, (\frac{2}{n-1} + \frac{\gamma_2}{n})$
3. $var(S_n) \simeq \frac{\sigma^2}{4} \, (\frac{2}{n-1} + \frac{\gamma_2}{n})$

de que se conclui que se $\gamma_2 > 0$ a convergência de S_n para σ é mais lenta do que no caso gaussiano, enquanto para $\gamma_2 < 0$ acontece o reverso.

A partir daqui passaremos a escrever simplesmente $\overline{X}$ e S^2 em vez de $\overline{X}_n$ e de S_n^2, sempre que desta simplificação de notações não advenha

ambiguidade.

Provemos os resultados acima enunciados:

Comprovemos, primeiro, que $\mathbb{E}(S^2) = \sigma^2$:
De $(n-1)S^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2$, resulta

$(n-1)\mathbb{E}(S^2) = \sum_{i=1}^{n}\mathbb{E}(X_i^2) - n\mathbb{E}(\overline{X}^2) = n(\sigma^2+\mu^2) - n(\frac{\sigma^2}{n}+\mu^2) =$
$(n-1)\sigma^2$,

donde $\mathbb{E}(S^2) = \sigma^2$.

Quanto ao resultado 1., $var(S_n^2) = \sigma^4\ (\frac{2}{n-1} + \frac{\gamma_2}{n})$, como

$$n^2\overline{X}^2 = \sum_{i=1}^{n} X_i^2 + 2\sum_{i<j} X_iX_j,$$

podemos reescrever

$$(n-1)S^2 = \frac{n-1}{n}\sum_{i=1}^{n} X_i^2 - \frac{2}{n}\sum_{i<j} X_iX_j,$$

donde

$$(n-1)^2S^4 = \left(\frac{n-1}{n}\right)^2 \sum_{i=1}^{n} X_i^4 + 2\left[\left(\frac{n-1}{n}\right)^2 + \frac{2}{n^2}\right]\sum_{i<j} X_i^2X_j^2 + \\ + \alpha\sum_{i<j} X_i^3X_j + \beta\sum_{i<j<l} X_i^2X_jX_l.$$

Como os X_j são independentes — e tendo em conta a propriedade da invariância da variância para translações, podemos, sem perda de generalidade, considerar os X_j centrados — segue-se que

$$\mathbb{E}(X_i^3X_j) = \mathbb{E}(X_i^2X_jX_l) = 0,$$

e assim

$$(n-1)^2\mathbb{E}(S^4) = \left(\frac{n-1}{n}\right)^2 n\mu_4 + 2\left[\left(\frac{n-1}{n}\right)^2 + \frac{2}{n^2}\right]\ \frac{n(n-1)}{2}\ (\mu^2+\sigma^2)^2,$$

ou seja,

$$\mathbb{E}(S^4) = \frac{\mu_4}{n} + \frac{n^2 - 2n + 3}{n(n-1)}\,\sigma^4$$

e, consequentemente,

$$\begin{aligned} var(S^2) &= \frac{\mu_4}{n} + \frac{n^2 - 2n + 3}{n(n-1)}\,\sigma^4 - (\sigma^2)^2 = \\ &= \frac{\mu_4}{n} + \frac{n-3}{n(n-1)}\,\sigma^4 = \frac{(n-1)(\gamma_2+3)\,\sigma^4 - (n-3)\sigma^4}{n(n-1)} = \\ &= \sigma^4\left(\frac{\gamma_2}{n} + \frac{3n-3-n+3}{n(n-1)}\right) = \\ &= \sigma^4\left(\frac{\gamma_2}{n} + \frac{2}{n-1}\right), \end{aligned}$$

como enunciado.

Convém-nos agora estabelecer o seguinte

Lema 5.1. $cov(\overline{X}, S^2) = \frac{\mu^3}{n} = \frac{\sigma^3\gamma_1}{n}$.

Demonstração.

$$\begin{aligned} cov(\overline{X}, S^2) &= \mathbb{E}[(\overline{X}-\mu)(S^2-\sigma^2)] = \mathbb{E}[(\overline{X}-\mu)S^2] = \\ &= \frac{1}{n(n-1)}\,\mathbb{E}[\sum_{i=1}^{n}(X_i-\mu)\sum_{j=1}^{n}(X_j-\mu-(\overline{X}-\mu))^2] = \\ &= \frac{1}{n(n-1)}\,\mathbb{E}[\sum_{i=1}^{n}(X_i-\mu)(\sum_{j=1}^{n}(X_j-\mu)^2 - n(\overline{X}-\mu)^2)] = \\ &= \frac{1}{n(n-1)}[\sum_{i=1}^{n}\mathbb{E}(X_i-\mu)^3 - \frac{1}{n}\sum_{j=1}^{n}(X_j-\mu)(\sum_{l=1}^{n}(X_l-\mu))^2] = \\ &= \frac{1}{n(n-1)}(n\mu_3 - \frac{1}{n}\,n\mu_3) = \frac{(n-1)\mu_3}{n(n-1)} = \frac{\mu_3}{n}. \end{aligned}$$

□

Continuemos agora a demonstração dos resultados anteriormente enunciados. A aproximação indicada no ponto 2. obtém-se fazendo a expansão de $S = \sqrt{W}$, com $W = S^2$, em torno de σ^2

$$f(W) = \sqrt{W}, \quad f^{'}(W) = \frac{1}{2\sqrt{W}}, \quad f^{''}(W) = \frac{1}{4W\sqrt{W}}$$
$$f(\sigma^2) = \sigma, \quad f^{'}(\sigma^2) = \frac{1}{2\sigma}, \quad f^{''}(\sigma^2) = -\frac{1}{4\sigma^{\frac{3}{2}}}$$

pelo que

$$S = f(W) = \sigma + \frac{1}{2\sigma}(S^2 - \sigma^2) - \frac{1}{8\sigma^3}(S^2 - \sigma^2)^2 + ...$$

e, tomando valores médios, como $\mathbb{E}(S^2 - \sigma^2) = 0$ e

$$\mathbb{E}(S^2 - \sigma^2)^2 = var(S^2) = \sigma^4(\frac{2}{n-1} + \frac{\gamma_2}{n}),$$

obtém-se

$$\mathbb{E}(S) \simeq \sigma - \frac{1}{8\sigma^3}\ \sigma^4(\frac{2}{n-1} + \frac{\gamma_2}{n}) = \sigma - \frac{\sigma}{8}(\frac{2}{n-1} + \frac{\gamma_2}{n}).$$

Quanto a 3., $var(S_n) \simeq \frac{\sigma^2}{4}\ (\frac{2}{n-1} + \frac{\gamma_2}{n})$, basta usar o teorema de König, desprezando os termos $0(\frac{1}{n})$

$$\begin{aligned} var(S) &= \mathbb{E}(S^2) - \mathbb{E}(S)^2 = \\ &= \sigma^2 - \sigma^2 + 2\ \frac{\sigma^2}{8}(\frac{2}{n-1} + \frac{\gamma_2}{n}) + 0(\frac{1}{n}) \simeq \\ &\simeq \frac{\sigma^2}{4}\ (\frac{2}{n-1} + \frac{\gamma_2}{n}) \end{aligned}$$

Vamos agora investigar $cov(\overline{X}, S)$.

Lema 5.2. $cov(\overline{X}, S) \simeq \frac{\sigma^2\gamma_1}{2n}$.

Demonstração.

$$\begin{aligned} cov(\overline{X}, S) &= \mathbb{E}(\overline{X}S) - \mathbb{E}(\overline{X})\mathbb{E}(S) = \\ &= \mathbb{E}(\overline{X}S) - \mu\ (\sigma - \frac{\sigma}{8}\ (\frac{2}{n-1} + \frac{\gamma_2}{n})) \end{aligned}$$

e podemos usar a técnica já atrás utilizada de expandir $\overline{X}\sqrt{T}$ com $T = S^2$ em série de Taylor para obter uma aproximação conveniente de $\mathbb{E}(\overline{X}S)$

$$f_{(\overline{X},W)}(\overline{X},T) = \overline{X}\sqrt{T} \Rightarrow f_{(\overline{X},W)}(\mu,\sigma^2) = \mu\sigma$$

$$f'_{\overline{X}}(\overline{X},W) = \sqrt{W}, \quad f'_{W}(\overline{X},W) = \frac{\overline{X}}{2\sqrt{W}}, \quad f''_{\overline{X}^2}(\overline{X},W) = 0$$

$$f''_{(\overline{X},W)}(\overline{X},W) = \frac{1}{2\sqrt{W}}, \quad f''_{W^2}(\overline{X},W) = -\frac{\overline{X}}{4\sqrt{W^3}}$$

$$f'_{\overline{X}}(\mu,\sigma^2) = \sigma, \quad f'_{T}(\mu,\sigma^2) = \frac{\mu}{2\sigma}, \quad f''_{\overline{X}^2}(\mu,\sigma^2) = 0$$

$$f''_{(\overline{X},T)}(\mu,\sigma^2) = \frac{1}{2\sigma}, \quad f''_{T^2}(\mu,\sigma^2) = -\frac{\mu}{4\sigma^3}$$

Então,

$$\overline{X}S = \overline{X}\sqrt{W} \sim \mu\sigma + \sigma(\overline{X}-\mu) + \frac{\mu}{2\sigma}(S^2-\sigma^2) + \\ + 2\frac{1}{2\sigma}\frac{(\overline{X}-\mu)(S^2-\sigma^2)}{2} - \frac{\mu}{4\sigma^3}\frac{(S^2-\sigma^2)}{2}$$

e tomando valores médios

$$\mathbb{E}(\overline{X}S) \sim \mu\sigma + \frac{1}{\sigma}cov(\overline{X},S^2) - \frac{\mu}{8\sigma^3}var(S^2).$$

Recordando os anteriores resultados

$$\mathbb{E}(\overline{X}S) \sim \mu\sigma + \frac{1}{2\sigma}\frac{\mu_3}{n} - \frac{\mu}{8\sigma^3}\sigma^4(\frac{\gamma_2}{n} + \frac{2}{n-1}).$$

Assim,

$$cov(\overline{X},S) \simeq \mu\sigma + \frac{\sigma^2\gamma_1}{2n} - \frac{\mu}{8}\sigma(\frac{\gamma_2}{n} + \frac{2}{n-1}) - \mu\sigma + \frac{\mu\sigma}{8}(\frac{2}{n-1} + \frac{\gamma_2}{n}) = \\ = \frac{\sigma^2\gamma_1}{2n}.$$

□

Lema 5.3. *A correlação assintótica entre* $\overline{X}$ *e* S *é* $\dfrac{\gamma_1}{\sqrt{\gamma_2+2}}$.

Impõe-se o comentário: sabemos já que $\overline{X}$ e S são independentes apenas no caso $X \frown Gaussiana(\mu, \sigma)$. No entanto, sempre que a população parente seja simétrica ($\gamma_1 = 0$), $\overline{X}$ e S são assintoticamente independentes. Assim, os bons resultados de Efron (1968) sobre studentização em populações simétricas (com caudas não excessivamente pesadas) não são contra-intuitivos.

Abordemos agora o cálculo dos momentos de

$$T = \sqrt{n}\frac{\overline{X} - \mu}{S} = \sqrt{n}\frac{\overline{X} - \mu}{\sqrt{W}}$$

com $W = S^2$, expandindo aquela expressão em série de Taylor em torno de (μ, σ^2),

$$\begin{array}{ll}
f_{(\overline{X},W)}(\overline{X}, W) = \sqrt{n}\frac{\overline{X}-\mu}{\sqrt{W}} & f_{(\overline{X},W)}(\mu, \sigma^2) = 0 \\
f'_{\overline{X}}(\overline{X}, W) = \sqrt{\frac{n}{W}} & f'_{\overline{X}}(\mu, \sigma^2) = \frac{\sqrt{n}}{\sigma} \\
f'_{W}(\overline{X}, W) = -\frac{1}{2}(\overline{X} - \mu)W^{-\frac{3}{2}} & f'_{W}(\mu, \sigma^2) = 0 \\
f''_{\overline{X}^2}(\overline{X}, W) = 0 & f''_{\overline{X}^2}(\mu, \sigma^2) = 0 \\
f''_{(\overline{X}W)}(\overline{X}, W) = -\frac{\sqrt{n}}{2\sqrt{W^3}} & f''_{(\overline{X}W)}(\mu, \sigma^2) = -\frac{\sqrt{n}}{2\sqrt{\sigma^3}} \\
f''_{W^2}(\overline{X}, W) = \frac{3}{4}(\overline{X} - \mu)W^{-\frac{5}{2}} & f''_{W^2}(\mu, \sigma^2) = 0
\end{array}$$

donde

$$T = \sqrt{n}\,\frac{\overline{X} - \mu}{\sqrt{W}} \sim \sqrt{n}\,\frac{\overline{X} - \mu}{\sigma} - 2\frac{\sqrt{n}}{2\sigma^3}\,\frac{(\overline{X} - \mu)(S^2 - \sigma^2)}{2!}$$

e tomando valores médios

$$\mathbb{E}(T) \sim -\frac{\sqrt{n}}{2\sigma^3}\,cov(\overline{X}, S^2) = -\frac{\sqrt{n}}{2\sigma^3}\,\frac{\sigma^3\gamma_1}{n} = -\frac{\gamma_1}{2\sqrt{n}},$$

com termo dominante sempre de sinal contrário ao do coeficiente de assimetria da população parente!

Assim,

$$T - \mathbb{E}(T) \simeq T + \frac{\gamma_1}{2\sqrt{n}} \sim \frac{\gamma_1}{2\sqrt{n}} + \frac{\sqrt{n}}{\sigma}\,(\overline{X} - \mu) - \frac{\sqrt{n}}{2\sigma^3}\,(\overline{X} - \mu)(S^2 - \sigma^2)$$

e

$$[T - \mathbb{E}(T)]^2 \simeq \frac{\gamma_1^2}{4n} + \frac{n}{\sigma^2}\,(\overline{X} - \mu)^2 + \frac{n}{4\sigma^6}\,(\overline{X} - \mu)^2(S^2 - \sigma^2)^2 +$$

$$+ \frac{\gamma_1}{\sigma}\,(\overline{X} - \mu) - \frac{\gamma_1}{2\sigma^3}\,(\overline{X} - \mu)(S^2 - \sigma^2) -$$

$$- \frac{n}{2\sigma^4}\,(\overline{X} - \mu)^2(S^2 - \sigma^2).$$

Tomando valores médios

$$var(T) \sim \frac{\gamma_1^2}{4n} + 1 + \frac{n}{4\sigma^6}\,\mathbb{E}[(\overline{X} - \mu)^2(S^2 - \sigma^2)^2] + 0 - \frac{\gamma_1}{2\sigma^3}\,\frac{\sigma^3\gamma_1}{n} -$$

$$- \frac{n}{2\sigma^4}\,\mathbb{E}[(\overline{X} - \mu)^2(S^2 - \sigma^2)^2]$$

ou seja

$$var(T) \sim 1 - \frac{\gamma_1^2}{4n} - \frac{n}{2\sigma^4}\,\mathbb{E}[(\overline{X} - \mu)^2(S^2 - \sigma^2)] + \frac{n}{4\sigma^6}\,\mathbb{E}[(\overline{X} - \mu)^2(S^2 - \sigma^2)^2]$$

onde

$$\mathbb{E}[(\overline{X} - \mu)^2(S^2 - \sigma^2)] = \mathbb{E}[(\overline{X} - \mu)^2 S^2] - \frac{\sigma^4}{n}$$

em que podemos reescrever

$$\mathbb{E}[(\overline{X} - \mu)^2 S^2] = \frac{1}{n^2(n-1)}\,\mathbb{E}[(n\sum_{i=1}^{n}(\overline{X} - \mu)^2(n-1)S^2]$$

Como

$$(n-1)S^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n}[(X_i - \mu) - (\overline{X} - \mu)]^2 =$$

$$= \sum_{i=1}^{n}(X_i - \mu)^2 - \frac{1}{n}[\sum_{i=1}^{n}(\overline{X} - \mu)]^2,$$

escrevendo

$$Y_i = X_i - \mu \Rightarrow \mathbb{E}(Y_i) = 0,\ \ var(Y_i) = \sigma^2,\ \ \mathbb{E}(Y_i^3) = \mu_3,\ \ \mathbb{E}(Y_i^4) = \mu_4,$$

vem

$$\mathbb{E}[(\overline{X} - \mu)^2 S^2] = \frac{1}{n^2(n-1)}\ \mathbb{E}[(\sum_{i=1}^{n} Y_i)^2 (\sum_{j=1}^{n} Y_j^2 - \frac{1}{n}(\sum_{k=1}^{n} Y_k)^2)] =$$

$$= \frac{1}{n^2(n-1)}\ \mathbb{E}[(\sum_{i=1}^{n} Y_i)^2 \sum_{j=1}^{n} Y_j^2 - \frac{1}{n}(\sum_{k=1}^{n} Y_k)^4]$$

em que todas as parcelas em que apareça algum factor Y_i têm valor esperado zero.

Assim,

$$\mathbb{E}[(\overline{X} - \mu)^2 S^2] = \frac{1}{n^2(n-1)}[n\mu_4 + n(n-1)\sigma^4 - \frac{1}{n}(n\mu_4 + 3n(n-1)\sigma^4)] =$$

$$= \frac{\mu_4}{n^2} + \frac{n-3}{n^2}\ \sigma^4$$

e na expressão de $var(T)$ a terceira parcela é então

$$-\frac{n}{2\sigma^4}\ (\frac{\mu_4}{n^2} + \frac{n-3}{n^2}\sigma^4 - \frac{\sigma^4}{n}) = -\frac{\gamma_2}{n}.$$

Passemos agora ao cálculo de $\mathbb{E}[(\overline{X} - \mu)^2 (S^2 - \sigma^2)^2]$.

$$\mathbb{E}[(\overline{X} - \mu)^2 (S^2 - \sigma^2)^2] = \mathbb{E}[(\overline{X} - \mu)^2 (S^4 - 2\sigma^2 S^2 + \sigma^4)] =$$

$$= \mathbb{E}[(\overline{X} - \mu)^2 S^4] - 2\sigma^2\ (\frac{\mu_4}{n^2} + \frac{(n-3)\sigma^4}{n^2}) + \frac{\sigma^6}{n}$$

onde

$$\mathbb{E}[(\overline{X}-\mu)^2 S^4] = \frac{1}{n^2(n-1)^2}\,\mathbb{E}[(\sum_{i=1}^{n}(X_i-\mu))^2(\sum_{i=1}^{n}(\overline{X}_i-\overline{X})^2)^2] =$$

$$= \frac{1}{n^2(n-1)^2}\,\mathbb{E}[(\sum_{i=1}^{n}Y_i)^2(\sum_{j=1}^{n}Y_j^2 - \frac{1}{n}(\sum_{k=1}^{n}Y_k)^2)^2] =$$

$$= \frac{1}{n^2(n-1)^2}\,\mathbb{E}[(\sum_{i=1}^{n}Y_i)^2(\sum_{j=1}^{n}Y_j^2) - \frac{2}{n}(\sum_{i=1}^{n}Y_i)^4)\sum_{j=1}^{n}Y_j^2 + \frac{1}{n^2}(\sum_{i=1}^{n}Y_i)^6)] =$$

$$= \frac{1}{n^2(n-1)^2}[n\mu_6 + 3n(n-1)\sigma^2\mu_4 + 2n(n-1)\mu_3^2 + n(n-1)(n-2)\sigma^6 -$$

$$-\frac{2}{n}(n\mu_6 + 7n(n-1)\sigma^2\mu_4 + 4n(n-1)\mu_3^2 + 3n(n-1)(n-2)\sigma^6) +$$

$$+\frac{1}{n^2}(n\mu_6 + 15n(n-1)\sigma^2\mu_4 + 10n(n-1)\mu_3^2 + 15n(n-1)(n-2)\sigma^6)].$$

A 4ª parcela de $var(T)$, a menos de $0(\frac{1}{n^2})$ é, então, aproximada por

$$\frac{n}{4\sigma^6}[\frac{1}{n^2(n-1)^2}[n\mu_6 + 3n(n-1)\sigma^2\mu_4 + 2n(n-1)\mu_3^2 + n(n-1)(n-2)\sigma^6 -$$

$$-\frac{2}{n}(n\mu_6 + 7n(n-1)\sigma^2\mu_4 + 4n(n-1)\mu_3^2 + 3n(n-1)(n-2)\sigma^6) +$$

$$+\frac{1}{n^2}(n\mu_6 + 15n(n-1)\sigma^2\mu_4 + 10n(n-1)\mu_3^2 + 15n(n-1)(n-2)\sigma^6)] -$$

$$-\frac{2\sigma^2\mu_4}{n^2} + \frac{\sigma^6}{n}] \approx \left(\frac{3\gamma_2}{4n} + \frac{9}{4n} + \frac{\gamma_1^2}{2n} + \frac{1}{4}\right) - \frac{\gamma_2 - \frac{1}{2n}}{2n} - \frac{1}{4}$$

donde

$$var(T) \sim 1 + \frac{\gamma_1^2 - \gamma_2 + 7}{4n}.$$

Na derivação da expressão acima note-se que

$$(\sum_{i=1}^{n} Y_i)^2 = \sum_{i=1}^{n} Y_i^2 + \sum_{i \neq j} Y_i Y_j, \tag{5.1}$$

em que temos n termos Y_i^2 e $n(n-1)$ termos $Y_i Y_j$
e

$$(\sum_{i=1}^{n} Y_i^2)^2 = \sum_{i=1}^{n} Y_i^4 + \sum_{i \neq j} Y_i^2 Y_j^2, \tag{5.2}$$

com n termos Y_i^4 e $n(n-1)$ termos $Y_i^2 Y_j^2$.

Multiplicando (5.1) e (5.2)

$$(\sum_{i=1}^{n} Y_i)^2 (\sum_{j=1}^{n} Y_j^2)^2 = \sum_{i=1}^{n} Y_i^6 + 3\sum_{i \neq j} Y_i^4 Y_j^2 + 2\sum_{i \neq j} Y_i^3 Y_j^3 + \sum_{i \neq j \neq k} Y_i^2 Y_j^2 Y_k^2 +$$

+ termos em que aparece Y_i e que têm valor médio igual a zero.

Assim, temos n termos Y_i^6, $3n(n-1)$ termos $Y_i^4 Y_j^2$, $2n(n-1)$ termos $Y_i^3 Y_j^3$ e $n(n-1)(n-2)$ termos $Y_i^2 Y_j^2 Y_k^2$.

Sendo

$$(\sum_{i=1}^{n} Y_i)^4 = \sum_{i=1}^{n} Y_i^4 + 4\sum_{i \neq j} Y_i^3 Y_j + 3\sum_{i \neq j} Y_i^2 Y_j^2 + 6\sum_{i \neq j \neq k} Y_i^2 Y_j Y_k + \sum_{i \neq j \neq k \neq l} Y_i Y_j Y_k Y_l,$$

então

$$(\sum_{i=1}^{n} Y_i)^4 (\sum_{j=1}^{n} Y_j)^2 = \sum_{i=1}^{n} Y_i^6 + 7\sum_{i \neq j} Y_i^4 Y_j^2 + 4\sum_{i \neq j} Y_i^3 Y_j^3 + 3\sum_{i \neq j \neq k} Y_i^2 Y_j^2 Y_k^2 +$$

+ termos com Y_i e, portanto, com valor médio nulo.

Assim, temos n termos Y_i^6, $7n(n-1)$ termos $Y_i^4Y_j^2$, $4n(n-1)$ termos $Y_i^3Y_j^3$ e $3n(n-1)(n-2)$ termos $Y_i^2Y_j^2Y_k^2$.

Quanto a $(\sum_{i=1}^{n} Y_i)^6$,

$$(\sum_{i=1}^{n} Y_i)^6 = \sum_{i=1}^{n} Y_i^6 + 15\sum_{i\neq j} Y_i^4Y_j^2 + 10\sum_{i\neq j} Y_i^3Y_j^3 + 15\sum_{i\neq j\neq k} Y_i^2Y_j^2Y_k^2 +$$

+ termos em que aparece Y_i e que têm valor médio igual a zero, concluindo que, com valor médio não nulo, temos n termos Y_i^6, $15n(n-1)$ termos $Y_i^4Y_j^2$, $10n(n-1)$ termos $Y_i^3Y_j^3$ e $15n(n-1)(n-2)$ termos $Y_i^2Y_j^2Y_k^2$, ficando, deste modo, justificada a referida expressão.

Por outro lado,

$$\mathbb{E}(T) \simeq -\frac{\gamma_1}{2\sqrt{n}}$$

e

$$var(T) \simeq 1 + \frac{\gamma_1^2 - \gamma_2 + 7}{4n}$$

são aproximações (a menos de $O(n^{-\frac{3}{2}})$ e de $O(n^{-2})$, respectivamente), e outras aproximações são possíveis.

Por exemplo, desenvolvendo

$$T^2 = f_{(\overline{X},W)}(\overline{X}, W) = n\frac{\overline{X} - \mu^2}{W},$$

onde $W = S^2$, em torno de (μ, σ^2), tem-se

$$T^2 = f(\overline{X}, W) = f(\mu, \sigma^2) + f_k^{'}(\mu, \sigma^2)(\overline{X} - \mu) + f_T^{'}(\mu, \sigma^2)(S - \sigma^2) +$$

$$+ f_{\overline{X}^2}^{''}(\mu, \sigma^2)\frac{(\overline{X} - \mu)^2}{2!} + f_W^{''}(\mu, \sigma^2)\frac{(S - \sigma^2)^2}{2!} +$$

$$+ 2f_{(\overline{X},W)}^{''}(\mu, \sigma^2)\frac{(\overline{X} - \mu)(S - \sigma^2)}{2!} + ...$$

e de

$$f(\mu,\sigma^2)=\frac{n(\mu-\mu)^2}{\sigma^2}=0$$

$$f'_{\overline{X}}(\overline{X},W)=\frac{2n(\overline{X}-\mu)}{W} \qquad f'_{\overline{X}}(\mu,\sigma^2)=0$$

$$f''_{\overline{X}^2}(\overline{X},W)=\frac{2n}{T} \qquad f''_{\overline{X}^2}(\mu,\sigma^2)=\frac{2n}{\sigma^2}$$

$$f''_{\overline{X}W}(\overline{X},W)=\frac{-2n(\overline{X}-\mu)}{T^2} \qquad f''_{\overline{X}W}(\mu,\sigma^2)=0$$

$$f''_{W}(\overline{X},W)=\frac{-n(\overline{X}-\mu)^2}{T^2} \qquad f''_{W}(\mu,\sigma^2)=0$$

$$f''_{W^2}(\overline{X},W)=\frac{2n(\overline{X}-\mu)^2}{T^3} \qquad f''_{W^2}(\mu,\sigma^2)=0$$

vem

$$T^2\simeq\frac{2n}{\sigma^2}\,\frac{(\overline{X}-\mu)^2}{2}$$

e tomando valores médios

$$\mathbb{E}(T^2)\sim\frac{2n\sigma^2}{\sigma^2 2n}=1,$$

donde a aproximação

$$var(T)\sim 1-(-\frac{\gamma_1}{2\sqrt{n}})^2=1-\frac{\gamma_1^2}{4n}$$

mais fácil de deduzir, mas evidentemente muito pobre.

Para uma população parente $X\frown Gama(\alpha,\delta)$ com $\alpha\leq\frac{1}{2}$, obtém-se avaliação negativa para $var(T)$ para valores baixos de n.

Claro que podemos truncar a expansão em série nos termos de ordem

2, em vez dos termos de ordem 1; como

$$f'''_{\overline{X}^3}(\overline{X},W)=0 \qquad \Rightarrow \quad f'''_{\overline{X}^3}(\mu,\sigma^2)=0$$

$$f'''_{\overline{X}^2W}(\overline{X},W)=\frac{-2n}{T^2} \qquad \Rightarrow \quad f'''_{\overline{X}^2W}(\mu,\sigma^2)=\frac{-2n}{\sigma^4}$$

$$f'''_{\overline{X}W^2}(\overline{X},W)=\frac{4n(\overline{X}-\mu)}{T^2} \qquad \Rightarrow \quad f'''_{\overline{X}W^2}(\mu,\sigma^2)=0$$

$$f'''_{W^3}(\overline{X},W)=\frac{-6n(\overline{X}-\mu)^2}{T^4} \quad \Rightarrow \quad f'''_{W^3}(\mu,\sigma^2)=0$$

obtém-se a expansão

$$T^2 \sim \frac{2n(\overline{X}-\mu)^2}{2\sigma^2}-6\frac{n}{\sigma^4}\,\frac{(\overline{X}-\mu)^2(S^2-\sigma^2)}{6!}$$

Tomando valores médios

$$\mathbb{E}(T^{)} \sim 1-\frac{n}{\sigma^4}\,\mathbb{E}[(\overline{X}-\mu)^2(S^2-\sigma^2)]$$

Atrás calculámos

$$\mathbb{E}[(\overline{X}-\mu)^2(S^2-\sigma^2)]=\frac{\mu_4}{n^2}+\frac{n-3}{n^2}\sigma^4-\frac{\sigma^4}{n}$$

donde

$$\mathbb{E}(T^2) \sim 1-\frac{n}{\sigma^4}(\frac{\mu_4}{n^2}-\frac{3}{n^2}\sigma^4)=1-\frac{\gamma_2}{n}$$

e

$$var(T) \sim 1-\frac{\gamma_2}{n}-\frac{\gamma_1^2}{4n}.$$

A expressão que deduzimos,

$$var(T) \simeq 1+\frac{\gamma_1^2-\gamma_2+7}{4n},$$

que logo à partida parece mais conforme com o que se espera de uma variância, é, sem dúvida, de dedução mais trabalhosa.

Capítulo 6

Bem-estar e progresso académico dos alunos da FCUL provenientes de PALOP

6.1 Introdução

Os alunos de PALOP da FCUL têm por vezes uma progressão escolar lenta — mas estão longe de ser os únicos a quem tal acontece. Constituem, porém, uma população de dimensão modesta, e com características especiais que facilitavam uma parte do estudo que tínhamos planeado.

De facto, uma das vertentes da investigação que nos pareceu interessante foi avaliar a que ponto o recurso a um entrevistador que tivesse empatia com os entrevistados influenciava a disponibilidade dos respondentes, não só na entrevista directa como em posterior entrevista telefónica ou questionário escrito, um e/ou outro mascarados de operação de controle do trabalho do entrevistador.

Nessa fase, adiante descrita com mais detalhe, foi contratado um aluno finalista do DEIO, natural de Cabo Verde, dotado de qualidades de inteligência e memória, e dotes de observação, adequados ao que pretendíamos.

A fase seguinte foi, infelizmente, muito mais limitada por imperativos logísticos. Nela pretendia-se investigar a importância do uso de dados

administrativos no controle da fidedignidade das respostas em questões que podiam causar algum embaraço aos inquiridos, como o seu êxito escolar, ou o eventual atraso no pagamento das propinas. Os serviços académicos — e, mais uma vez, se agradece à Professora Doutora Carla Kulberg, do Conselho Directivo da FCUL, que autorizou, e à Dra. Aldina Vieira, que foi de uma disponibilidade ímpar em auxiliar-nos, apesar das suas múltiplas ocupações urgentes — forneceram-nos os dados cujo tratamento inicial apresentamos. Por outro lado, usámos os contactos telefónicos dos alunos para entrevistar 50 — com uma taxa de não-resposta de cerca de $\frac{1}{3}$, como esperávamos e nos convinha, e com não-resposta muito ocasional a alguma das questões postas.

Com base na análise dos dados, fizemos um estudo muito elementar sobre estratégias de imputação de dados em falta para reduzir o viés de estimadores.

Relatamos ainda algumas considerações que nos parecem importantes sobre a preparação de instrumentos de recolha de dados, na vertente não matemática, mas essenciais para o êxito do trabalho estatístico, pois se os dados não prestarem não há análise que os salve.

6.2 Elaboração de um Questionário

A elaboração de um questionário é uma etapa delicada na qual se deve ter em conta alguns princípios. Em termos muito gerais apresentamos um conjunto de princípios importantes a observar na elaboração de cada questão:

* Deve conter uma e uma só ideia;
* Deve ser simples: recorrer a palavras simples e a uma linguagem acessível;
* Deve ser curta e directa (evitando as negações e, sobretudo, as duplas-negações);
* Deve ser lida e entendida facilmente;
* Não deve sugerir uma resposta particular;
* Não deve incluir elementos de emotividade;
* Não deve contribuir para o surgimento de não-resposta o que leva a cuidados adicionais na formulação de "questões delicadas".

6.3 Inquérito

6.3.1 Introdução

Com o objectivo de estudar as condições em que os alunos africanos se encontram a fazer o seu curso superior nas universidades portuguesas procedemos a um estudo piloto envolvendo 107 alunos da Universidade de Lisboa, da Universidade Nova de Lisboa e da Universidade Técnica de Lisboa. Por outro lado, na perspectiva de investigar a qualidade dos dados em investigação social, decidimos comparar as informações obtidas via entrevista, e através de inquérito. Ainda no que respeita à entrevista, interessou-nos ter uma avaliação preliminar da interferência e inerente limitação à espontaneidade, no caso de ser visível pelo entrevistado um instrumento de registo, fosse este com suporte em papel ou suporte magnético.

Os 107 alunos eram originários de Cabo Verde (32), da Guiné (22), de Angola (30), de São Tomé e Príncipe (3) e de Moçambique (20). Apenas um dos alunos cursava pós-graduação. Embora, em geral, os estudantes africanos que estão em Portugal sejam bem aceites pela comunidade universitária, enfrentam problemas que acabam por ser negativos na sua vida estudantil e pessoal. Assim, o estudo incide sobre a qualidade de vida, a satisfação, a integração, o aproveitamento, a auto-estima, a fidelização, e as intenções no que se refere a regresso ao país de origem.

Nesta fase preparatória, interessava-nos investigar se a obtenção de dados resultaria melhor usando questionário ou entrevista. Para esta abordagem contribuiu um recém-licenciado do DEIO, antes de regressar a Cabo Verde, com fácil contacto com outros estudantes africanos,que foi contratado para a fase inicial deste trabalho devido às seguintes características: capacidade intelectual, comprovada por um progresso regular nos estudos; dotes de memória (testada); aparência física e contacto agradável, sem ser excessivamente descontraído; isenção, isto é, capacidade treinada de "opacidade" de opinião, por forma a minimizar situações de empatia com o entrevistado.

As primeiras 22 entrevistas foram consideradas da fase preparatória do entrevistador e, por isso, descartadas nos estudos comparativos posteriores; nessa fase preparatória, o entrevistador memorizou cada

um dos itens do inquérito e foi treinado com voluntários, que não faziam parte da população alvo, a conduzir a entrevista "de memória", memorizando as respostas que lhe eram dadas tão fidedignamente quanto possível e tendo o cuidado de passar ao registo escrito assim que deixava de estar na presença do entrevistado.

Das remanescentes 85, 35 foram feitas com apoio do inquérito presente no momento da entrevista e as restantes 50 usando o treino da memória do entrevistador. Em 15 casos foi solicitada ao entrevistado a autorização para gravar a entrevista, mas o registo foi feito de memória, sendo posteriormente o registo magnético usado para avaliar a fidedignidade dos dados obtidos com recurso a memorização. No caso de potenciais entrevistados manifestarem pouca vontade em colaborarem, alegando mesmo que já tinham sido amplamente interrogados sobre aquelas questões, havia alguma pressão emocional e apelo à solidariedade, referindo-se que com este trabalho estava a ganhar o dinheiro necessário para concluir os estudos e regressar ao seu país de origem, o que funcionou em todos os casos em que foi necessário recorrer a essa pressão. No que se refere a estes 85 alunos, 20 tinham sido antes (cerca de 1 mês) interrogados usando o questionário e outros 20 foram interrogados usando o questionário, cerca de 1 mês depois da entrevista; oito foram inquiridos antes e depois, cerca de um mês em cada caso, da entrevista.

6.3.2 Inquérito Usado no Estudo Piloto

Ao fazer este estudo um dos passos importantes foi a elaboração do inquérito utilizado. Malhotra e Birks (1999) recomendam que um inquérito deverá transformar a informação que procuramos num conjunto de questões específicas que leve os inquiridos a cooperar, a interessarem-se e a dar respostas correctas, completas e honestas, e minimizar os erros nas respostas. Na construção de um inquérito deverão ser tomados em consideração alguns aspectos importantes. Utilizar questões abertas ou de escolha múltipla, não usar palavras ambíguas, evitar alternativas implícitas, evitar generalizações e a ordem das questões são alguns exemplos dados por Malhotra e Birks (1999).
Passamos a apresentar o inquérito elaborado para este estudo, acompanhado de um breve preâmbulo em que se procura captar a colaboração dos abordados para obtenção dos dados.

6.3.2.1 Preâmbulo

Apesar de todas as medidas tomadas para evitar a discriminação, é evidente que há ainda um longo caminho a percorrer para que todos os alunos africanos a estudar em universidades portuguesas se sintam plenamente integrados, obtendo uma satisfação justa e perdurável da sua vivência do ensino superior.

Notamos que os estudantes africanos, apesar de bem aceites pela comunidade universitária, se sentem pouco à vontade, ocasionalmente contrafeitos, o que acaba por ter um efeito negativo sobre a sua vida estudantil e pessoal. Este inquérito é um primeiro instrumento para nos guiar nas correcções que podem minorar os problemas e, por isso, agradecemos a sua colaboração:

* respondendo ao inquérito escrito;
* respondendo, também, a entrevista pessoal, caso esta lhe seja proposta, mesmo que lhe pareça que as perguntas são as mesmas; de facto, numa segunda abordagem, pode já ter amadurecido as ideias sobre as questões que abordamos e dar-nos uma informação mais rica.

  Muito obrigado pela colaboração.

6.3.2.2 Inquérito

1. Onde vivia antes de iniciar os seus estudos superiores?
2. No ano de ingresso no ensino superior em que data começou a frequentar as aulas?
3. Qual a fonte do seu sustento (bolsa de estudo, rendimentos familiares, emprego)?
4. No caso de ter bolsa de estudo, qual é a entidade financiadora (fundação, governo português, governo do país de origem, outra)?
5. Onde habita (casa de família, residência, quarto alugado)?
6. Quanto gasta em alojamento?

7. Quanto gasta em vestuário?
8. Quanto tempo, em média, gasta para chegar à universidade?
9. Que transportes usa (comboio, barco, autocarro, metro, a pé)?
10. Em geral, come na(s) cantina(s) universitária(s)?
11. Qual o preço médio da sua refeição principal?
12. A relação quantidade/preço é boa?
13. Onde vai nas noites de sexta e/ou sábado (casa de amigos, discoteca, bar, cinema)?
14. Quantos amigos próximos tem em Portugal?
15. Destes amigos:
    (a) quantos são portugueses?
    (b) quantos frequentam a sua universidade?
16. Quando faz trabalhos em grupo, qual o número de elementos?
17. O grupo inclui quantos elementos africanos?
18. Nos seus contactos sociais, quanto tempo, em média, fala uma língua africana?
19. Participa nas actividades da Associação de Estudantes da sua universidade?
20. Pratica desporto universitário?
21. Se pratica desporto universitário, qual é a modalidade?
22. Se não pratica desporto universitário, pratica alguma modalidade fora da universidade?
23. Há quantos anos está na universidade?
24. Em quantas disciplinas obteve aprovação?
25. Em termos médios, que tipo de aluno se considera (suficiente, bom, muito bom?
26. Se tem bolsa de estudo, a sua atribuição influenciou a escolha do curso que frequenta?

27. O curso tem correspondido às suas expectativas?

28. Enquanto estudante está a ter sucesso?

29. Houve alguma situação em que se sentiu discriminado?

30. Sente-se bem fisicamente?

31. Tem tido oportunidades para desenvolver a sua personalidade?

32. Em geral, os professores da sua universidade mostram-se disponíveis para lhe tirar dúvidas?

33. Alguma vez algum dos seus professores o(a) ajudou na resolução de algum problema pessoal?

34. Alguma vez algum dos seus colegas o(a) ajudou na resolução de algum problema pessoal?

35. Recomendaria a sua universidade a pessoas do seu país que queiram vir estudar para Portugal?

36. A sua formação aqui em Portugal é importante para o desenvolvimento do seu país?

37. Gostaria de permanecer em Portugal depois de terminar o seu curso?

38. No caso de regressar ao país de origem, gostaria de voltar a Portugal para cursos de pós-graduação ou de especialização?

Os dados obtidos mostram que os inquiridos parecem mais disponíveis a responder com mais detalhe a um inquérito depois de terem sido entrevistados sobre as questões que ele aborda. Há indicações de que a entrevista feita por um entrevistador relativamente ao qual se possa criar alguma projecção favorece a quantidade de informação que se consegue de cada entrevistado.

6.4 Análise de Dados Administrativos

A análise que apresentamos de seguida é baseada na informação administrativa fornecida pela Divisão Académica da FCUL em Agosto de 2008.

O objectivo deste estudo é investigar as dificuldades académicas dos alunos de PALOP, inscritos na faculdade à data da obtenção da informação administrativa.

O estudo engloba 126 alunos dos quais 43 são provenientes de Angola, 60 de Cabo Verde, 9 da Guiné, 9 de Moçambique e 5 de São Tomé e Príncipe. Existem 95 alunos a frequentar uma Licenciatura, 19 em Mestrado ou Curso de Especialização e 12 em Doutoramento.

Codificámos o país de origem da seguinte forma:

A: Angola
CV: Cabo Verde
G: Guiné
M: Moçambique
ST: São Tomé e Príncipe.

A Tabela 6.1 apresenta alguns dados gerais.

Na Tabela 6.2 podemos ver a distribuição dos alunos por país de origem e por nível de ensino.

A Tabela 6.3 apresenta o número de alunos segundo o país de origem e o sexo. Existe um maior número de alunos do género masculino.

Tabela 6.1: Dados gerais

	A	CV	G	M	ST
Nº total de alunos	43	60	9	9	5
Total Feminino	18	32	1	4	2
Total Masculino	25	28	8	5	3
Média de idades	31.40	23.95	34.67	30.22	28

Tabela 6.2: País de origem/Nível de ensino

	Licenciatura	Mest/Esp	Doutoramento	Totais
A	33	1	9	43
CV	46	12	2	60
G	7	2	0	9
M	4	4	1	9
ST	5	0	0	5
Totais	95	19	12	126

Tabela 6.3: País de origem/Sexo

	Feminino	Masculino	Totais
A	18	25	43
CV	32	28	60
G	1	8	9
M	4	5	9
ST	2	3	5
Totais	57	69	126

Na Tabela 6.4 podemos observar como os alunos estão distribuídos pelos 3 níveis de ensino. Repare-se que são do género masculino a maioria dos alunos em Doutoramento, não havendo grande diferença nos outros níveis de ensino.

Tabela 6.4: Nível de ensino/Sexo

	Feminino	Masculino	Totais
Licenciatura	46	49	95
Mest/Esp	9	10	19
Doutoramento	2	10	12
Totais	57	69	126

As Tabelas 6.5 e 6.6 mostram como os alunos se distribuem segundo o país de origem e o nível de ensino relativamente à idade. A maioria dos alunos tem idades compreendidas entre os 17 e os 31 anos.

Tabela 6.5: Idade/País de origem

idade	A	CV	G	M	ST	Totais
17–24	14	39	1	3	1	58
24–31	10	17	2	3	3	35
31–38	9	3	2	1	1	16
38–45	6	1	4	2	0	13
45–52	2	0	0	0	0	2
52–59	1	0	0	0	0	1
59–66	1	0	0	0	0	1
Totais	43	60	9	9	5	126

Tabela 6.6: Idade/Nível de ensino

idade	Licenciatura	Mest/Esp	Doutoramento	Totais
17–24	51	7	0	58
24–31	27	7	1	35
31–38	9	4	3	16
38–45	6	1	6	13
45–52	1	0	1	2
52–59	0	0	1	1
59–66	1	0	0	1
Totais	95	19	12	126

O que apresentamos a seguir envolveu unicamente os alunos a frequentar uma Licenciatura. Do total de 95 alunos foram retirados 7 que, embora apresentando um total de ECTS superior a 180 e não sendo alunos da Licenciatura em Geologia, ainda não tinham terminado o curso, dado não terem sido aprovados em disciplinas consideradas obrigatórias no respectivo curso.

Note-se que, presentemente, na Faculdade, as Licenciaturas são de 3 anos, tendo os alunos que obter 180 ECTS (60 por ano), exceptuando a Licenciatura em Geologia que tem duração de 4 anos com um total de 240 ECTS.

Ao pretender saber qual a percentagem de curso realizada por cada aluno até à data do estudo, agruparam-se os alunos pelo número de anos

em que se inscreveram na Faculdade. A referida percentagem foi obtida através do quociente

$$\frac{y}{x}$$

sendo y o total de ECTS obtidos e $x = 60$ se $y = 1$, $x = 120$ se $y = 2$, $x = 180$ se $y \geq 3$, e $x = 240$ se $y \geq 4$ para os alunos da Licenciatura em Geologia. A Tabela 6.7 apresenta os resultados obtidos.

Tabela 6.7: Percentagem média de curso realizada

N.º de anos em que o aluno de inscreveu	1	2	≥ 3	≥ 4 (Geologia)
Percentagem média de curso realizada	18%	38%	57%	77%

O que apresentamos em seguida enquadra-se no problema das não-respostas. Dos 88 alunos a frequentar uma Licenciatura, e considerados na análise anterior, foram seleccionados 50, aleatoriamente, aos quais foi efectuada uma entrevista via telefone. As questões colocadas foram:

1. (a) Idade
   (b) Nacionalidade
   (c) Sexo
2. Quanto tempo, em média, demora, por dia, em transportes?
3. Qual o ano de ingresso nesta faculdade?
4. No ano lectivo de 07/08 em quantas disciplinas esteve inscrito?
5. No ano lectivo de 07/08 em quantas disciplinas obteve aprovação?
6. Indique o seu grau de satisfação com o seu sucesso escolar:
   (a) Muito fraco
   (b) Fraco
   (c) Suficiente/Mediano
   (d) Bom
   (e) Muito bom
7. Que percentagem do seu curso realizou até agora?

(a) Menos de 25%

(b) Entre 25% e 50%

(c) Entre 50% e 75%

(d) Mais de 75%

8. Já, em algum ano, teve redução de propinas?

9. Paga as propinas dentro do prazo?

10. Na generalidade como classifica a qualidade e o contacto com os seus professores?

    (a) Qualidade: de 1 a 5

    (b) Contacto: de 1 a 5

Dos 50 inquiridos obtivemos 33 respondentes a todas as questões, um não respondente unicamente a uma questão e os restantes 16 foram considerados não respondentes a todo o inquérito.

Quanto ao não respondente unicamente a uma questão (questão 5) foi feita a imputação do valor 5, assim determinado: considerando alunos com o mesmo número de anos em que se inscreveram na Faculdade e estando inscritos em igual número de disciplinas no ano lectivo 07/08, a média do número de disciplinas em que obtiveram aprovação no mesmo ano lectivo foi igual a 5; ao considerar alunos com o mesmo número de anos em que se inscreveram na Faculdade estando inscritos em 11, 12 ou 13 disciplinas no ano lectivo 07/08, a média do número de disciplinas em que obtiveram aprovação no mesmo ano lectivo foi igual a 8; dado que a resposta do referido aluno à questão 6 foi Suficiente/Mediano, poderíamos pensar no valor 8 como sendo o valor a imputar, mas, dado o facto de não ter respondido, levou-nos a decidir pelo valor 5.

Ao confrontarmos a informação administrativa fornecida, verificámos que o referido aluno tinha, de facto, sido aprovado em 5 disciplinas no ano lectivo passado, até à data do estudo.

Quanto à questão 10 sobre a qualidade e o contacto com os seus professores, a Tabela 6.8 apresenta os valores a imputar aos não respondentes. Foram consideradas duas técnicas de imputação:

∗ imputação "padrão" em que se substitui os valores em falta por valores obtidos por alunos que apresentam comportamento similar aos alunos não respondentes, tendo em conta a Licenciatura e o número de anos em que se inscreveram na Faculdade;

∗ imputação média por grupo (os alunos foram agrupados pelo número de anos em que se inscreveram na Faculdade.

Tabela 6.8: Valores a imputar na questão 10

Nº de anos	Licenciatura	AQ	BQ	AC	BC
1	Biologia	3	3	3	3
	Engenharia Geográfica	1	3	3	3
	Engenharia Informática	3	3	2	3
	Estatística Aplicada	4	3	3	3
	Geologia	4	3	4	3
	Matemática	1	3	3	3
	Química Tecnológica	4	3	2	3
2	Biologia	3	3	3	3
	Engenharia Geográfica	1	3	3	3
	Engenharia Informática	3	3	2	3
	Estatística Aplicada	5	3	4	3
	Geologia	4	3	4	3
3	Biologia	4	4	4	4
	Bioquímica	5	4	4	4
	Eng.Energia e Ambiente	3	4	1	4
	Engenharia Geográfica	1	4	3	4
	Engenharia Informática	4	4	3	4
	Estatística Aplicada	4	4	4	4
	Física	3	4	3	4
	Geologia	4	4	4	4
≥ 4	Biologia	4	4	4	4
	Bioquímica	5	4	5	4
	Engenharia Física	3	4	3	4
	Engenharia Informática	4	4	3	4
	Estatística Aplicada	5	4	5	4
	Geologia	3	4	4	4
	Matemática Aplicada	4	4	5	4
	Química Tecnológica	4	4	3	4

Na coluna AQ (AC), relativamente à Qualidade (Contacto), o

método de imputação é a imputação "padrão" e na coluna BQ (BC) é usada a imputação média por grupo.

Mais tarde foi seleccionada, aleatoriamente, uma nova amostra com 16 alunos entre os 38 não incluídos na amostra inicial de dimensão 50. Procedemos às entrevistas e obtivemos 11 respondentes e 5 não respondentes. Tinham sido imputados valores à questão 10 para cada um destes 16 alunos. Depois das entrevistas compararam-se os valores imputados com os valores reais.

Tabela 6.9: Percentagem de erro

Método de imputação	Qualidade	Contacto
Imp."padrão"	36%	82%
Imp. Média por grupo	45%	55%

Tabela 6.10: Erros de imputação para Qualidade

Método de imputação	Erro Médio	Erro Médio Absoluto	Erro Médio Quadrático
Imp."padrão"	-0.64	0.82	2.27
Imp. Média por grupo	-0.55	0.73	1.27

Tabela 6.11: Erros de imputação para Contacto

Método de imputação	Erro Médio	Erro Médio Absoluto	Erro Médio Quadrático
Imp."padrão"	-1.09	1.27	2.36
Imp. Média por grupo	-0.55	0.91	1.64

Na Tabela 6.9, relativamente à Qualidade, a percentagem de erro é menor com a imputação "padrão", embora se passe o contrário em relação ao Contacto. Comparando Qualidade e Contacto há uma maior diferença na percentagem de erro quando se utiliza a imputação "padrão". Estas diferenças podem ter explicação no facto de os alunos terem comportamentos bastante diferenciados no contacto com os seus professores.
Nas Tabelas 6.10 e 6.11 são apresentados os erros de imputação (erro médio, erro médio absoluto e erro médio quadrático). Repare-se, na Tabela 6.11, as diferenças entre os dois métodos de imputação.

Bibliografia

• Aleixo, S. M., Diamantino, M. F., e Pestana, D. D. (2008). Família parametrizada GLE. *Notas e Comunicações do CEAUL*, Nota nº15.

• Aleixo, S., Brilhante, F., Diamantino, F., Mendonça, S., and Pestana, D. (2007). Non-response and sample size, *Bulletin of the International Statistical Institute LVI* (electronic publication).

• Barndorff-Nielsen, O., and Cox, D. R. (1989). *Asymptotic Techniques for Use in Statistics*, Chapman and Hall, London.

• Barnett, V. (2002). *Sample Surveys — Principles and Methods*, 3rd ed., Arnold, London.

• Benjamin, Y. (1983). Is the t test really conservative when the parent distribution is long-tailed?, *J. Amer. Statist. Assoc.* **78**, 645–654.

• Bishop, Y. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*, MIT Press, Cambridge, MA.

• Brilhante, M. F., Pestana, D. D., e Rocha, J. (1996). Inferência sobre o parâmetro de localização de uma população exponencial, II: Studentização interna. In *A Estatística a Decifrar o Mundo*, 57–63, Salamandra, Lisboa.

• Brilhante, M. F., Pestana, D., Rocha, J., e Velosa, S. (2001). *Inferência Estatística sobre Localização e Escala*, Sociedade Portuguesa de Estatística, Ponta Delgada.

• Casella, G., and Berger, R. L. (2002). *Statistical Inference*, 2nd ed., Duxbury, Pacific Grove.

• Chandra, T. K. (1999). *A First Course in Asymptotic Theory of Statistics*, Narosa, New Delhi.

• David, H. A. (1981). *Order Statistics*, Wiley, New York.

• Devroye, L. (1986). *Non-Uniform Random Variate Generation*, Springer-Verlag, New York.

• Diamantino, F. (1998). Análise de escala em pequenas amostras. *Notas e Comunicações do CEAUL*, Nota nº9.

• Diamantino, F., e Pestana, D. (1997). Perturbações da gaussiana — sua influência em estatísticas studentizadas. In *A Estatística a Decifrar o Mundo*, 65–71, Salamandra, Lisboa.

• Efron, B. (1968). Student's t-test under symmetry conditions. *J. Amer. Statist. Assoc.* **64**, 1278–1302.

• Erdös, P., and Rényi, A. (1959). On a central limit theorem for samples from a finite population, *Publ. Math. Instit. Hungar. Acad. Sci.* **4**, 49–61.

• Feller, W. (1971). *An Introduction to Probability Theory and its Applications*, vol.II, Wiley, New York.

• Fraser, D.A.S. (1957). *Nonparametric Methods in Statistics*, Wiley, New York.

• Gomes, P. (1998), *Tópicos de Sondagens*, Sociedade Portuguesa de Estatística, Tomar.

• Hotelling, H. (1961). The behavior of some standard statistical tests under nonstandard conditions. *Proc. 4th Berkeley Symp. Mathematical Statistics and Probability*, **I**, 319–359, California University Press, Berkeley.

• Iglewicz, B. (1981). Robust scale estimators and confidence intervals for location. *In* D. Hoaglin, F. Mosteller and J. Tukey (eds.), *Understanding Robust and Exploratory Data Analysis*, Wiley, New York.

• Johnson, N. L., Kotz, S., and Balakrishnan, N. (1994). *Continuous Univariate Distributions*, vol. 1, 2nd ed., Wiley, New York.

• Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995). *Continuous Univariate Distributions*, vol. 2, 2nd ed., Wiley, New York.

• Kish, L. (1965). *Survey Sampling*, Wiley, New York.

• Kovalenko, I. N. (1965). On a class of limit distributions for rarefied flows of homogeneous events, *Lit. Mat. Sbornik* **5**, 569–573. (*Selected*

Transl. Math. Statist. and Prob. **9**, Providence, Rhode Island, 1971, 75–81.)

• Kozubowski, T. J. (1994). Representation and properties of geometric stable laws, *Approximation, Probability, and Related Fields*, Plenum, New York. 321–337.

• Kundu, D., and Gupta, R. D. (2007). A convenient way of generating gamma random variables using generalized exponential distribution, *Comp. Stat. & Data Analysis*, **51**, 2796–2802.

• Law, A. M., and Kelton, W. D. (1982). *Simulation, Modeling and Analysis*, Academic Press, New York.

• Lehman, E. L., and Casella, G. (2003). *Theory of Point Estimation*, Springer, New York.

• Logan, B. F., Mallows, C. L., Rice, S. O., and Shepp, L. A. (1973). Limit distributions of self-normalized sums. *Ann. Probability* **1**, 788–809.

• Malhotra, N., and Birks, D. (1999), *Marketing Research. An Applied Approach*, Prentice Hall, New York.

• Malva, M. (2006). *Distribuições Conjugadas e Aproximações*, Tese de Doutoramento, Universidade de Lisboa.

• Montgomery, D. C. (1997). *Design and Analysis of Experiments*, 4th ed., Wiley, New York.

• Mosteller, F., and Tukey, J. W. (1977). *Data Analysis and Regression — a Second Course in Statistics*, Addison-Wesley, Reading, Mass.

• Pearson, E. S. (1941–42). The probability integral of the range in samples of n observations from a normal population. *Biometrika* **32**, 301.

• Penny, K. I., and Chesney, T. (2008). A comparison of missing value imputation methods for classifying patient outcome following trauma injury. *Proc. of the ITI 2008 30th Int. Conf. on Information Technology Interfaces*, June 23–26, Cavtat, Croatia.

• Perlo, V. (1933). On the distribution of 'Student's' ratio for samples of three drawn from the rectangular distribution. *Biometrika* **25**, 203–204.

• Pestana, D., e Rocha, J. (1993). Análise de escala — modelo exponencial. In *A Estatística e o Futuro e o Futuro da Estatística*, 295–303, Salamandra, Lisboa.

• Pestana, D., e Rocha, J. (1995). Studentização interna com distribuição parente simétrica. In *Bom Senso e Sensibilidade, Traves Mestras da Estatística*, 377–381, Salamandra, Lisboa.

• Pestana, D. D., Sequeira, F., and Velosa, S. F. (2001). Parseval's relation and self-reciprocal characteristic functions, *Rev. Estat./Statist. Rev. — 23rd European Meeting of Statisticians, Contributed Papers* II, 315–316.

• Pestana, D. D., e Velosa, S. F. (2008). *Introdução à Probabilidade e à Estatística*, Vol. 1, 3ª edição. Lisboa: Fundação Calouste Gulbenkian.

• Rao, P. S. R. S. (2000), *Sampling Methodologies with Applications*, Chapman & Hall, New York.

• Rényi, A. (1956). A characterization of the Poisson process, *MTA Mat. Kut. Int. Közl.* **1**, 519–527. (Reeditado em *Selected Papers of Alfréd Rényi*, P. Turán, ed, Akadémiai Kiadó, Budapest, 1976, vol. I, p. 622–628.)

• Rocha, J. M. (1995). *Localização e Escala em Situações não Clássicas*, Tese de Doutoramento, Universidade dos Açores.

• Ross, S. M. (1997). *Simulation. Statistical Modeling and Decision Science*, 2nd ed., McGraw-Hill, San Diego.

• Särndal, C. E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*, Sringer-Verlag, New York.

• Shohat, J. A., and Tamarkin, J. D. (1943). *The Problem of Moments*, American Mathematical Society, Providence, Rhode Island.

• Stuart, A., Ord, K. (1992). *Kendall´s Advanced Theory of Statistics*, vol. 1, 5th ed., Charles Griffin & Company Limited, London.

• Student (1908). On the probable error of the mean. *Biometrika* **6**, 1–25. (Reeditado em Pearson and Wishart, 1958).

• Thiele, (1931). Theory of Observations, *Ann. Math. Statist.*, **2**, 165–308.

• Tukey, J. W. (1977). *Exploratory Data Analysis*, Addison-Wesley, Reading, Mass.

• van Belle, G. (2008). *Statistical Rules of Thumb*, 2nd ed., Wiley, New York.

Apêndice