ESTIMAÇÃO ROBUSTA NO MODELO DE POSIÇÃO

OSCAR BUSTOS

INSTITUTO DE MATEMÁTICA PURA E APLICADA
Rua Luiz de Camões, 68
20.060 - Rio de Janeiro - RJ

Prefácio

A intenção das presente notas é que elas possam servir de ajuda dos interessados na busca de técnicas de estimação aplicáveis nos modelos onde as hipóteses habituais de "normalidade" se satisfaçam só aproximadamente. Somente se analisa o caso de estimação de um parâmetro de posição univariado.

O rigor matemático foi excluído totalmente pensando num prossível leitor que conheça o essencial da Probabilidade e Estatística, segundo como se estuda nas nossas Universidades nos cursos de Formação na Estatística.

Na verdade, foi uma primeira intenção fazer um estudo do modelo mais amplamente usado: o modelo de regressão. Mais posteriormente, se julgo melhor fazer um estudo detalhado de um modelo mais simples, assim futuramente o leitor poderá sozinho e quiçá com uma base mais firme continuar o estudo dos trabalhos que tratam da Estimação Robusta nos modelos mais complexos.

Muito desse julgamento foi baseado na experiência que o autor teve de conversas com alunos e colegas, especialmente aquelas de um curso de verão na U.S.P. O autor agradece a valiosa colaboração brindada pelos alunos desse curso.

Finalmente o autor agradece a seus amigos e colegas do IMPA: Ana Maria Lima de Farias, Rosely Moraes Garcia, Nelson Ithiro Tanaka e Nuno Duarte da Çosta Bittencourt os que não somente fizerão a tradução do texto originariamente em espanhol, mais

também apontarão críticas e sugestões que ajudaram a melhorar estas notas.

ÍNDICE

CAPÍTULO I

INTRODUÇÃO ÀS TÉCNICAS ROBUSTAS DE ESTIMAÇÃO: SUA ORIGEM E NECESSIDADES

1.1 - Conceito Geral de Robustez

Em diversas atividades interessa estudar certos fenômenos ou procedimentos que ao serem observados em diversas ocasiões, apresentam uma certa variabilidade em seus resultados. Se se deseja apreender o geral ou essencial deste fenômeno, terá que se saber extrair essa essência desta massa de resultados obtidos em observações particulares. A Estatística, como sabemos, é uma ferramenta importantíssima neste processo de passar do particular ao geral.(inferência). De fato, poderíamos dizer, como em CRAMER [6], que o objetivo principal da Estatística é extrair inferências válidas de um conjunto de dados. Deveríamos acrescentar que a Estatística também nos ensina, através da Teoria de Amostragem e Planejamento de Experimentos, como fazer para que os dados obtidos nos dêem informação útil sobre o fenômeno em estudo.

Da nossa parte, vamos supor que o conjunto de dados obtido é "bom", no sentido de que se teve cuidado de seguir as prescrições a que nos referíamos no final do parágrafo anterior. Também vamos supor que esses dados são numéricos. Com o objetivo de formalizar e sistematizar a análise desses dados, vamos ajustar a eles um modelo matemático.

Isto é bem fácil de dizer e entender mas é muito difí-

cil de levar à prática de modo que estejamos seguros de que o modelo sobre o qual iremos trabalhar se ajusta corretamente aos dados. Existe uma produção crescente de trabalhos referentes à maneira de proceder na construção de modelos matemáticos. Podemos encontrá-los em diversos livros e publicações geralmente pertencentes à área de Análise de Dados (por exemplo: TUKEY [50], MOSTELLER E TUKEY [40], DACHS [8]). Para uma visão geral e, digamos, filosófica deste tema, se poderia consultar BOX [5]. Mas por melhor que procedamos, por mais cuidados que tomemos, sempre chegaremos a um modelo matemático que, precisamente por seu caráter abstrato, constitui somente uma descrição aproximada do fenômeno físico que se deseja estudar. Mais ainda, para que a construção deste modelo seja possível é necessário fazer suposições sobre o fenômeno, suposições que são difíceis de testar, ou que são feitas como uma primeira aproximação ou porque foram feitas em estudos "parecidos" ou, o que é pior, se deseja forçar a realidade para que se justifique a aplicação de uma certa técnica... De todas as maneiras, a partir do modelo que julgamos (subjetivamente em maior ou menor medida) como o mais adequado, deduzimos certas técnicas de inferência.

Como nunca podemos ter a certeza de que valem as suposições que nos conduziram à construção dessas técnicas, é natural que busquemos técnicas ou procedimentos que sejam mais ou menos resistentes diante de desvios das suposições feitas. Desde um certo tempo, se começou a chamar a essas técnicas ou procedimentos "robustas". Mas ainda não se chegou a uma formalização mate

mática deste conceito que seja suficientemente geral ou que goze da aceitação dos estatísticos. O melhor será, então, aceitar o significado que dão KENDALL E BUCKLAND [33]:

> "Robustez ... um procedimento estatístico é chamado robusto se não é muito sensível a desvios das suposições sobre as quais se baseia".

como vemos, se poderia analizar o conceito anterior para técnicas pertencentes a diversas áreas da Estatística. De fato, na literatura encontramos trabalhos referentes a testes de hipóteses robustos", "planejamentos de experimentos robustos", "robustez de modelos", "estimadores robustos", etc. Nosso estudo ficará limitado a uma pequena parte do tema de "estimadores robustos". Só estudaremos técnicas de estimação robusta válidas para o modelo de posição. A definição desse modelo será vista mais adiante. Não obstante, o que for visto deve servir como base para analisar a estimação robusta nomodelo possivelmente mais usado nas aplicações: o modelo linear ou o de regressão linear. Para destacar a importância de prosseguir essa análise nos referimos brevemente nessas notas a tal modelo.

Recordemos, então, sua definição: dizemos que uma sucessão de variáveis aleatórias (observações) $Y_1,\ldots,Y_n$ satisfaz um modelo linear geral se

$$(1.1.1)\quad Y_i = \theta_1 X_{i1} + \ldots + \theta_p X_{ip} + U_i \quad , \quad 1 \leq i \leq n$$

onde todas as X_{ij} são constantes conhecidas, os θ_j são parâ-

metros desconhecidos a estimar e as U_i são variáveis aleatórias.

Um caso particular importante do modelo anterior e do modelo deposição é o "modelo de medição", obtido de (1.1.1) fazendo $p = 1$, $X_{i1} = 1$, isto é:

$$(1.1.2) \qquad Y_i = \mu + U_i \qquad 1 \leq i \leq n$$

onde μ é o parâmetro a estimar.(parâmetro de medição). Vamos considerar também que as variáveis U_i são "erros de observação".

1.2 - Breve resenha histórica

Nesta seção não veremos mais que um resumo da seção 1 - Capítulo I - de JAMES e BUSTOS [32]. Consideremos o modelo de medição. É fácil ver que, na formulação (1.1.2), se queremos construir algum estimador "razoável" de μ, devemos fazer alguma suposição sobre a forma da distribuição do vetor aleatório $(U_1, \ldots, U_n)$, vetor dos "erros de observação". A suposição de que as Y_i's (observações) se relacionam com as U_i's segundo (1.1.2) e a suposição sobre a distribuição destas últimas determinam a hipótese sobre a distribuição das Y_i's.

Poderíamos dizer, como em MOSTELLER E TUKEY [40], que a história da Inferência Estatística se reduz a uma certa mistura de otimismo e ceticismo acerca da hipótese de que as observações se comportam segundo uma certa distribuição.

Nos primeiros cursos de Estatística constuma-se dar

uma atenção tão preponderante a uma certa distribuição do tipo contínuo que se poderia ter daí a impressão de que quase sempre as observações se comportam segundo essa distribuição. Essa distribuição recordemos, é a distribuição normal que está definida pela função de densidade (normal ou "gaussiana"):

$$(1.2.1) \quad X \mapsto \varphi(X;\mu,\sigma) =: \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2\right\}$$

Denotaremos sua função de distribuição por

$$(1.2.2) \quad \Phi(X;\mu,\sigma) = \int_{-\infty}^{X} \varphi(t;\mu,\sigma)\, dt$$

simplificaremos a notação colocando $\varphi(X) = \varphi(X;0,1)$ e $\Phi(X) = \Phi(X;0,1)$.

Em nosso (modelo de medição), o que se faz habitualmente é supor que as variáveis aleatórias $U_1,\ldots,U_n$ são independentes, identicamente distribuidas (i.i.d) com distribuição comum $N(0,\sigma^2)$ (normal com média 0 e variância σ^2). Esta hipótese implica supor que as $Y_1,\ldots,Y_n$ são i.i.d. com distribuição comum $N(\mu,\sigma^2)$.

Pois bem, como essa distribuição chegou a ter esta posição tão destacada ? com respeito a isso, é interessante ler a quem sugeriu seu uso: GAUSS. Em "Theoria Motus" (traduzida para o Inglês e com um apêndice do tradutor em DAVIS [9]), GAUSS se interessava pela determinação da órbita mais provável que descreveria um certo corpo celeste. Em uma seção desta obra, ele

explicita esse estudo e realiza considerações de índole geral mas seguramente influenciado pelo problema que tinha em mente. Aplicando suas idéias ao caso do modelo de medição se chega a uma equação que deve ser satisfeita por μ e pela distribuição das U_is. Como é impossível conhecer uma dessas grandezas sem conhecer a outra, Gauss preferiu considerar como axioma a"excelência" da média amostral (tal como é geralmente reconhecida) e deduzir, a partir deste axioma, a forma que devia ter a função de distribuição dos erros para que valesse a equação já estabelecida.

Devemos destacar que Gauss tirou sua conclusão guiado principalmente por considerações que seguramente eram razoáveis dentro do problema que estudava: medições astrônomicas, mas cuja generalização a outras questões é discutível. Mesmo assim, dentro da astronomia, NEWCOMB em 1825 observava que era mais realista considerar que os erros seguiam uma distribuição que admitisse a existência de observações efetuadas com diversos graus de precisão.

Por outra parte, notemos que a média amostral foi considerada "boa" por GAUSS de maneira axiomática. Bem sabemos que um axioma ou um sistema de axiomas pode servir para analizar adequadamente um fenômeno mas não outro.

Na verdade, podemos ler em REY [44] que, há vários séculos antes de Cristo, já se usava outros estimadores, diferentes da média amostral, para se estimar parâmetros de posição. De todas as maneiras, devemos concordar que este estimador é fácil de ser calculado. Tal parece ter sido uma das principais razões

para seu destacado papel na história da Estatística e que se prolonga até os nossos dias (mais que o desejável).

A tal ponto chegou a aceitação cega de que as observações se distribuem segundo a distribuição normal que muitas vezes, observações que evidentemente não se ajustavam a esse modelo foram consideradas errôneas e então descartadas. Outros estatísticos, mais cuidadosos, não descartavam assim tão facilmente tais observações "anormais" ou "outliers". Sem dúvida, seus estudos constituem as primeiras páginas da história da "estimação robusta". Quem estiver interessado em se aprofundar um pouco mais nos estudos de tais trabalhos pode consultar STIGLER [46], a série de artigos escritos por HARTER em "Int. Statist. Rev.", HUBER [28], HAMPEL [20], etc.

Além das dificuldades computacionais que seguramente desestimularam a mais de um no esforço de buscar técnicas de estimação alternativas para a média amostral, havia também o fato de se ter pouco conhecimento do comportamento catastrófico de tal estimador quando a distribuição verdadeira está ligeiramente afastada do normal.

Pouco antes de 1960, o surgimento de computadores mais velozes e o desenvolvimento das chamadas "técnicas não paramétricas ou de distribuição livre" começaram a provocar uma mudança de atitude de um crescente número de estatísticos.

Para não nos extendermos demasiado, mencionaremos apenas alguns dos trabalhos relevantes publicados desde 1960 sobre robus-

tez em estimação:

- HUBER [25]. Podemos considerá-lo como o início de uma busca de estimadores robustos segundo um ponto de vista formal. Sugeriu como medida adequada da robustez para estimadores assintoticamente normais, o supremo da variância assintótica quando a distribuição das observações se move em uma vizinhança "conveniente" da distribuição tomada a priori como modelo. Introduziu uma classe de estimadores (M-estimadores) que veremos detalhadamente mais adiante).

- HUBER [26]. Provou a consistência e normalidades assintótica dos M-estimadores sob condições muito gerais. Tem servido base para analisar o comportamento assintótico dessa classe de estimadores em modelos como o (1.1.1), em séries temporais, etc.

- HAMPEL [18]. Definiu formalmente o conceito de robustez qualitativa: um estimador é qualitativamente robusto se sua distribuição varia pouco quando a distribuição das observações varia pouco. Nessa formalização usou trabalhos anteriores de VON MISES e de PROHOROV que analisam a relação existente entre a teoria assintótica de estimação e as situações práticas nas quais a amostra tem tamanho finito. Definiu também outros conceitos que são utilizados para formalizar o que se quer dizer com a expressão estimador robusto. Estes conceitos são, curva de influência e ponto de ruptura. Voltaremos a falar neles mais adiante.

- ANDREWS E OUTROS [1]. Usando técnicas de simulação, estudaram o comportamento de 65 estimadores de um parâmetro de posição, sobre amostras finitas. Esses estimadores haviam sido sugeridos em sua maioria como alternativas, diante da falta de robustez da média amostral (que foi também incluida).

- HUBER [29]. Colocou o problema de robustificar o estimador de mínimos quadrados no modelo linear. Deu algoritmos para calcular os M-estimadores que extendeu para este modelo.

- YOHAI [52]. Estudou o comportamento assintótico dos M-estimadores de $\theta_1,\ldots,\theta_p$ em (1.1.1).

- MARONNA [37]. Definiu e analisou o comportamento de M-estimadores para a média e matriz de dispersão no caso multivariado.

- MARTIN [38]. Expôs o que se tinha feito para estimar robustamente os parâmetros no modelo autoregressivo.

Existem muitos outros trabalhos de igual importância aos mencionados que devem ser consultados no caso de se querer estudar com mais detalhe o presente tema. Uma lista bastante completa pode ser encontrada em REY [44], HUBER [30], LAUNER E WILKINSON (ed) [35], JAMES E BUSTOS [32], etc.

Ao autor dessas notas, pareceu interessante terminar esta brevíssima introdução histórica mencionando apenas um aspecto um tanto polêmico levantado em BOX [5] e que lança luzes so

bre o uso que se poderia fazer, e que tem sido feito em alguns casos particulares, dos métodos de robustez para ajustar um modelo matemático ao fenômeno real em estudo.

Sem dúvida alguna, uma das atividades mais transcendentes nas quais um estatístico pode ajudar a um investigador em disciplinas "aplicadas" é a de cooperar na construção de modelos que descrevam um certo fenômeno da maneira mais fiel e simples possível. Esta construção será convenientemente realizada mediante um processo por etapas. A partir de um modelo simples, introduzir modificações que sejam sugeridas ou pelas observações ou por um melhor conhecimento teórico do fenômeno. A necessidade de tais modificações deve evidenciar-se através de diversas técnicas de testes de modelos, entre as quais uma das mais importantes é a de análise de resíduos. Se chega assima a um modelo que poderá ser mais satisfatório que o inicial. Agora, podem ocorrer discrepâncias que sejam impossíveis de detectar por meio de tais técnicas. Diante desta situação, muitos estatísticos aconselham a usar métodos de inferência resistentes frente a essas discrepâncias. BOX [5] pensa que é melhor continuar se esforçando na busca de um modelo mais adequado. De todas as maneiras, deve-se destacar que o desenvolvimento de técnicas robustas não tem por que ser visto como um caminho contraposto ao sugerido pelo Professor BOX e sim, deve ser encarado como um meio valioso para percorrê-lo.

De fato, HOGG [24] parece adotar esse ponto de vista ao fazer as seguintes recomendações em uma seção do trabalho re-

cém citado, intitulada "O uso da robustez hoje em dia".
Nas operações é conveniente realizar as seguintes etapas:

a) Efetuar a análise habitual, usando alguma das técnicas clássicas;

b) Usar depois um procedimento robusto;

c) Se os resultados de a) e b) coincidem, realizar o informe de síntese estatística habitual;

d) Se os resultados de a) e b) não coincidem, estudar o problema globalmente, tantò os dados como o modelo.

1.3 - C por que da robustez

Vamos recordar o que a Estatística clássica nos aconselha fazer para estimar μ em (1.1.2) e os θ_j em (1.1.1). Veremos logo o que sucede com esses estimadores se não são satisfeitas algumas das hipóteses em que se baseiam. Finalmente, mostraremos que a violação de tais hipóteses constitue mais a regra, que a exceção, nos casos práticos.

Consideremos o modelo de medição. Isto é, suponhamos que $Y_1,\ldots,Y_n$ são variáveis aleatórias tais que

$$(1.3.1)\quad Y_i = \mu + U_i \qquad 1\leq i\leq n$$

onde μ é um parâmetro desconhecido a estimar e $\underline{U} = (U_1,\ldots,U_n)^T$ é o vetor dos "erros de observação". (Nestas notas, os vetores de dimensão n serão pensados como matrizes $n \times 1$. Também, se A é uma matriz, denotaremos por A^T a matriz transposta de A). As

hipóteses habitualmente feitas são (BICKEL E DOKSUM [4]):

(i) a distribuição de U é independente de μ;

(ii) o valor do erro cometido em uma observação não afeta o erro cometido em outras observações ($U_1,\ldots,U_n$ são independentes);

(iii) a distribuição do erro em uma observação é a mesma que nas outras observações ($U_1,\ldots,U_n$ são identicamente distribuídas);

(iv) a distribuição comum dos erros está dada por uma densidade f que é simétrica em torno da origem e que, ou é totalmente conhecida, ou é conhecida a menos de um fator de escala (σ) que costuma ser estimado simultaneamente com μ.

Se pode provar que, sob estas hipóteses, $Y_1,\ldots,Y_n$ são variáveis aleatórias independentes identicamente distribuídas (v.a.i.i.d) com distribuição comum dada pela densidade

$$f_\mu(y) = f(y-\mu).$$

Sob estas hipóteses, um dos métodos de estimação favoritos da Estatística clássica e o de máximo verossimilhança. Vamos recordá-lo: seja $L:\mathbb{R}\times\mathbb{R}^n \to \mathbb{R}$ a função de verossimilhança, isto é:

$$L(m,(y_1,\ldots,y_n)) = f(y_1-m)\ \ldots\ f(y_n-m).$$

Se define como estimador de máxima verossimilhança a $\hat{\mu} = \hat{\mu}(y_1,\ldots,y_n)$ tal que

$$L(\hat{\mu},(y_1,\ldots,y_n)) \geq L(m,(y_1,\ldots,y_n)) \quad \forall\ m$$

Não é difícil ver que, sob condições de regularidade para f, $\hat{\mu}$ é equivalentemente definido por

$$\sum_{i=1}^{n} \Psi (y_i - \hat{\mu}) = o$$

com $\Psi(y) = - f'(y)/f(y)$, sendo f' a derivada de f.

Se agora acrescentamos a hipótese de "normalidade" (talvez seja mais exato dizer de "gaussianidade") dos erros, isto é, se

$$(v) \quad f(x) = \varphi(x;o,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2\right),$$

obtém-se que

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} = \bar{y} \quad ,$$

ou seja, $\hat{\mu}$ é a média amostral.

As propriedades ótimas que apresenta esse estimador sob as hipóteses (i) a (v) são estudadas nos cursos habituais de estatística. Recordemos a mais importante: $\bar{y}$ é um estimador ENVUMV para μ, seja σ conhecido ou não (ENVUMV significa estimados uniformemente de mínima variância entre os não-viciados).

O modelo de medição não é mais que um caso particular do modelo linear geral (1.1.1). Consideremos, então o modelo (1.1.1). Isto é, suponhamos que $Y_1,\ldots,Y_n$ são variáveis aleatórias tais que

$$(1.3.2) \quad Y_i = \theta_1 X_{i1} + \ldots + \theta_p X_{ip} + U_i \qquad 1 \leq i \leq n$$

condição essa que pode ser expressa de maneira mais compacta usando notação matricial:

$$(1.3.3) \qquad \underline{Y} = \tilde{X}\,\theta + \underline{U}$$

onde $\tilde{X}$ é uma matriz nxp cuja i-ésima linha é $\underline{X}_i^T = (X_{i1},\ldots,X_{ip})$, $\underline{Y} = (Y_1,\ldots,Y_n)^T$ e $\underline{U} = (U_1,\ldots,U_n)^T$.

As hipóteses usualmente feitas são: (H1) $U_1,\ldots,U_n$ são independentes identicamente distribuidas com distribuição comum F;

(H2) $E_F\, U_i = 0$

(H3) $\text{Var}\, U_i = \sigma_F^2 < +\infty$

Classicamente, o estimador de θ mais usado é o $\hat{\theta}_{MQ} = \hat{\theta}_{MQ}(y_1,\ldots,y_n)$ que minimiza a função

$$v \mapsto \sum_{i=1}^{n} (y_i - \underline{x}_i^T\, \underline{v})^2$$

ou equivalentemente, $\hat{\theta}_{QM}$ é o que satisfaz

$$(1.3.4) \qquad \sum_{i=1}^{n} (y_i - \underline{x}_i^T\, \hat{\underline{\theta}}_{MQ})\, \underline{x}_i = \underline{0} \quad ,$$

equação que na forma matricial é escrita como

$$\tilde{X}^T\, \tilde{X}\, \hat{\underline{\theta}}_{MQ} = \tilde{X}^T\, \underline{Y}$$

onde $\underline{y} = (y_1,\ldots,y_n)^T$. Suponhamos, para simplificar, que $\tilde{X}^T\tilde{X}$ é não singular. Um resultado importante sobre as propriedades de $\hat{\underline{\theta}}_{MQ}$ (estimador de mínimos quadrados) sob as condições

(H1) a (H3) é:

Teorema de Gauss-Markov: Sejam $w \in \mathbb{R}^n$, $\xi = \underline{w}^T \tilde{X} \underline{\theta}$,

$\mathcal{L} = \{\underline{a}^T \underline{Y} : a \in \mathbb{R}^n, E(\underline{a}^T \underline{Y}) = \xi\}$ (quer dizer, $\mathcal{L}$ é a classe de todos os estimadores lineares não viciados de ξ).

Então:

(i) $\hat{\xi}_{MQ} = \underline{w}^T \tilde{X} \hat{\theta}_{MQ}$ está em $\mathcal{L}$.

(ii) $\xi^* \in \mathcal{L} \Rightarrow \text{Var}(\hat{\xi}_{MQ}) \leq \text{Var}(\xi^*)$.

Muitas vezes, e levados talvez pelas suposições feitas para o modelo de posição, costuma-se acrescentar às hipóteses (H1) a (H3) o seguinte hipótese

(H4) $F(x) = \Phi(x;o,\sigma)$

Sob as hipóteses (H1) a (H4) obtem-se, então, o seguinte resultado (ver, por exemplo, BICKEL e DOKSUM [4]):

Se $\underline{d} \in \mathbb{R}^p$ então $\underline{d}^T \hat{\theta}_{MQ}$ é o estimador não viciado em $\underline{d}^T\theta$ de mínima variância entre os estimadores não viciados de $\underline{d}^T\theta$.

Vamos analisar agora o que se passa com $\hat{\mu}$ e $\hat{\theta}_{MQ}$ se não se cumpre alguma das hipóteses (i) a (iv) para $\hat{\mu}$, ou (H1) a (H4) para $\hat{\theta}_{MQ}$.

Exemplo 1.3.1: Suponhamos que em (1.3.1) os valores observados de $Y_1,\ldots,Y_n$ são:

$y_1 = 2.422$, $y_2 = 0.130$, $y_3 = 2.232$, $y_4 = 1.700$, $y_5 = 1.903$

$y_6 = 0.725$, $y_7 = 2.031$, $y_8 = 0.515$, $y_9 = -0.684$, $y_{10} = 2.788$

Então $\hat{\mu} = 1.376$. Se, por alguma razão, o dado lido não tivesse sido $y_9 = -0.684$ e sim $y_9 = -68.4$, teríamos obtido $\hat{\mu} = -5.395$. Os valores de $y_1,\ldots,y_{10}$ acima foram extraídos de uma tabela de "números aleatórios normais com média 2 e variância 1", publicada em DIXON e MASSEY [11]; assim $\hat{\mu}$ não deveria estar muito afastado de 2, o que acontece com $\hat{\mu} = 1.376$. Mas bastou que uma única observação não seguisse a mesmo lei que as outras, para que o valor estimado $\hat{\mu}$ de μ não tivesse nenhuma relevância. Vemos assim, que a média amostral é muito sensível diante de desvios das suposições (i) a (v). Ou em outras palavras, talvez mais significativas: a média amostral dá igual importância a todas as observações, minimizando a informação fornecida por todo o conjunto de dados.

Exemplo 1.3.2: Analisemos agora o comportamento de $\hat{\theta}_{MQ}$. Para simplificar, vamos ficar com o modelo (1.1.1) sujeito à restrição $p = 1$; isto é, suponhamos que $Y_1,\ldots,Y_n$ são variáveis aleatórias tais que

$$Y_i = \theta x_i + U_i \qquad 1 \leq i \leq n$$

e que os valores conhecidos das x_i's e os observados das Y_i's são:

i	x_i	y_i
1	1.0	0.96
2	1.2	0.74
3	1.4	3.16
4	1.6	0.48
5	1.8	0.08
6	2.0	1.3
7	2.2	0.81
8	2.4	2.5
9	2.6	1.54
10	10.0	4.04

A partir de (1.3.4) é fácil ver que

$$\hat{\theta}_{MQ} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \approx 0.47 \ .$$

Na figura 1 temos a reta $y = \hat{\theta}_{MQ} X$ e os pontos (x_i, y_i) ajustados por ela.

Se agora trocamos $y_6 = 1.3$ por $y_6 = 13.0$ obtemos $\hat{\theta}_{MQ} \approx 0.65$, aumentando a inclinação da reta de ajuste por um fator 1.4. A figura 2 mostra o novo conjunto de pontos com sua reta de ajuste.

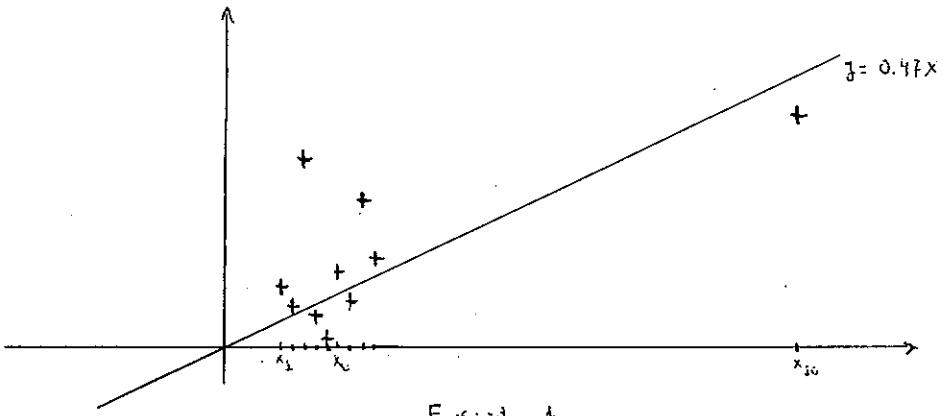

Figure 1

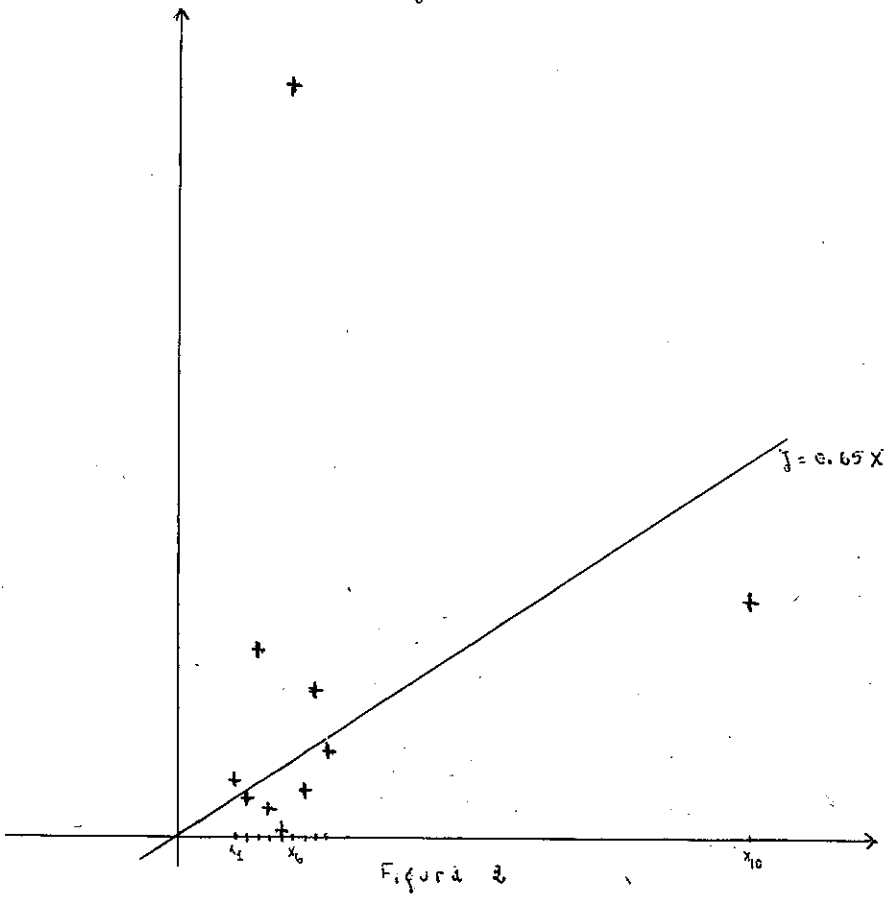

Figure 2

Vamos usar agora este conjunto de dados para mostrar um fenômeno que torna o estudo do robustez para o modelo de regressão mais necessário mas também mais difícil que nocaso do modelo de medição. Este fenômeno é o do "influência de x's grandes no ajuste por mínimos quadrados".

Suponhamos que $y_{10} = 7.46$, em lugar de $y_{10} = 4.04$. Então $\hat{\theta}_{MQ} = 0.73$ (ver figura 3). Uma mudança por um fator 1.8 na observação correspondente a $x = 10.0$ provocou uma mudança por um fator 1.6 na inclinação da reta, enquanto que para $x = 2.0$ um fator 10 provocou uma alteração menor.

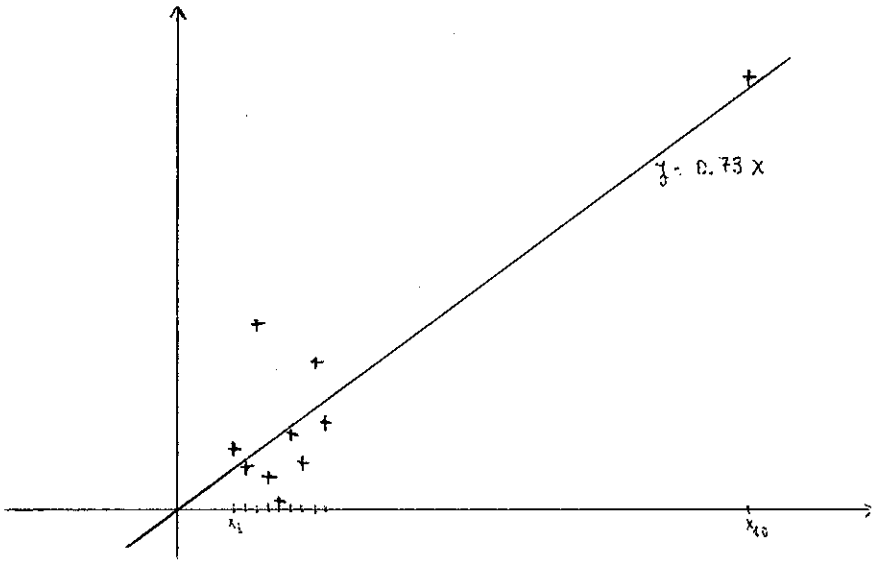

Figura 3

Comparemos agora as figuras 2 e 3. Inspecionando os resíduos $r_i^2 = :(y_i - \hat{\theta}_{MQ}\, x_i)^2$ no caso mostrado na figura 2 temos:

$r_1^2 = 0.096 \quad r_2^2 = 0.002 \quad r_3^2 = 5.063 \quad r_4^2 = 0.314 \quad r_5^2 = 1.188$

$r_6^2 = 136.89 \quad r_7^2 = 0.384 \quad r_8^2 = 0.884 \quad r_9^2 = 0.023 \quad r_{10}^2 = 6.052$

donde se manifesta o caracter de "outlier" de y_6. Para o caso da figura 3 temos:

$r_1^2 = 0.053 \quad r_2^2 = 0.018 \quad r_3^2 = 4.571 \quad r_4^2 = 0.473 \quad r_5^2 = 1.523$

$r_6^2 = 0.026 \quad r_7^2 = 0.634 \quad r_8^2 = 0.560 \quad r_9^2 = 0.128 \quad r_{10}^2 = 0.026$

A inspeção desses resíduos não serviria para detectar a modificação que houve em y_{10} nem, o que é pior, uma mudança maior ainda (veja o que passa com $y_{10} = 0.0$).

Os exemplos tratados anteriormente exibem o que se costuma chamar "outliers grosseiros", isto é, observações que se afastam da massa de dados em forma notável. Vimos a instabilidade dos métodos clássicos frente a esse tipo de outlier.

Além de tais observações notáveis, existem na prática numerosos exemplos de desvios não tão fáceis de notar e que, conforme veremos, afetam sensivelmente o rendimento das técnicas clássicas.

Para fixar idéias, tornemos a considerar o modelo de medição. Já no ano de 1825, NEWCOMB levantou a idéia de que era mais realista supor, em lugar de (v), que f era um "mistura" de normais com distintas variâncias. A literatura estatística poste

terior mostra exaustivamente a adequação de tal suposição. TUKEY [49] analisou com detalhes o que sucede com a média amostral se substituimos (v) por:

(v') $f(x) = (1-\varepsilon)\,\varphi(x;0,\sigma) + \varepsilon\,\varphi(x;0,\tau)$

($\varphi(.,.,.)$ como em (1.2.1))

com $0 \leq \varepsilon < 1/2$. Com isto, estamos supondo que uma proporção $(1-\varepsilon)$ dos erros tem precisão σ e proporção restante, precisão τ.

Como sabemos, um dos critérios mais usados para comparar o rendimento entre diferentes estimadores é o da eficiência (exata ou assintótica); isto é: entre vários estimadores de μ, se escolhe aquele que tenha variância (exata ou assintótica) mínima. Em diversos trabalhos (por exemplo, REY [44], JAMES e BUSTOS [32]) se mostra os valores da variância assintótica da média amostral e da mediana amostral sob as hipóteses (i) a (v), com $\sigma=1$ e sob as hipóteses (i) a (v') com $\sigma=1$ e ε,τ como se indica na seguinte tabela, extraída destes trabalhos:

ε	τ	Média Amostral	Mediana Amostral
0	1	1.000	1.571
0.002	3	1.014	1.575
0.03	3	1.226	1.632
0.1006	3	1.8047	1.8047
0.2	3	2.600	2.091
0.1	7	5.800	1.879

Em primeiro lugar, notemos que tomar $\varepsilon=0$ em (v') é o mesmo que supor (v). Vejamos agora o que nos diz a tabela passada. A mediana amostral é uns 36% menos eficiente que a média amostral sob a hipótese (v) (ou $\varepsilon=0$). ($0.36 \approx 1/1.571$) mas basta apenas um pouco mais de 10% de contaminação, com observações só 3 vezes menos precisas que sob normalidade, para que essa diferença disapareça; a situação se inverte para $\varepsilon=0.2$: a média amostral é 20% menos eficiente que a mediana amostral. A perda de eficiência da média frente à mediana é mais notável se as observações contaminadas são menos precisas, como podemos deduzir da última linha da tabela. Assim, o critério que nos levaria a escolher a média em lugar da mediana sob (v) nos levaria a escolha contrária sob (v') com ε e τ próximos aos valores que têm em (v). Não queremos com isto dizer que devemos usar a mediana em lugar da média para estimar μ. Só destacamos a necessidade de contar com técnicas de estimação de μ que sejam robustas, pelo menos no que se refere à eficiência assintótica; mais precisamente, técnicas que percam pouca eficiência em relação à média amostral sob normalidade, mas que sejam mais eficientes que ela sob contaminações, variando o menos possível. A urgência de tais técnicas se manifesta mais notavelmente se levarmos em conta a dificuldade em detectar a presença de contaminações que, com veremos mais adiante, se encontram com frequência nos problemas aplicados. De fato, REY [44] analisou qual seria a quantidade mínima de observações que se deveria fazer para testar, ao nível de 0.95, se estas observações viriam de uma distribuição estritamente normal

ou de uma normal contaminada. Em termos mais precisos: seja f a densidade da distribuição comum das U_i's .

H_o: $f(x) = \varphi(x;o,\sigma_\varepsilon)$

H_1: $f(x) = (1-\varepsilon)\,\varphi(x;0,1) + \varepsilon\,\varphi(x;0,3)$

onde $\sigma_\varepsilon^2 = (1-\varepsilon) + 9\varepsilon$ é a variância sob H_1.

Seja $\lambda(X_1,..,X_n)$ a estatística do teste da razão de verossimilhança para testar H_o conta H_1. Seja λ_o tal que $P(\lambda(U_1,...,U_n) \geq \lambda_o) \leq 0.05$. De REY [44] extraímos a seguinte tabela para o n mínimo que satisfaz a última desigualdade:

$\varepsilon = 0.1006$	$\varepsilon = 0.2436$
137	83

Também neste artigo de REY se mostram os resultados de um estudo semelhante ao que vimos mas trocando em (v'), $\varphi(x;o,\tau)$ por $\varphi(x;\mu_1,\tau)$ com $\mu_1 > o$, isto é, admitindo que os erros poderiam ser assimétricos (a hipótese (iv) poderia não se cumprir). Estes resultados reforçam ainda mais a necessidade de se mudar a técnica da média amostral, se se deseja um estimador robusto de μ em (1.3.1).

A dificuldade para se testar algumas das hipóteses que sustentam o uso de técnicas clássicas, como a de mínimos quadrados, já fez com que investigadores experientes e cuidadosos inicialmente acreditassem nelas e depois, com uma análise mais atenta, che

gassem à conclusão de que tais hipóteses eram insustentáveis. A este respeito, seria conveniente que o leitor se informasse sobre uma experiência realizada por PIERCE em 1873 e revisada 60 anos mais tarde por WILSON e HILFERTY (ver, por exemplo, MOSTELLER e TUKEY [50], ou JAMES e BUSTOS [32]).

É interessante destacar que a necessidade de técnicas robustas não passou desapercebida a vários estatísticos notáveis. Assim, várias publicações dispersas mostram os esforços realizados na busca de tais técnicas, mas limitados a um determinado problema de aplicação. No entanto, parece que as técnicas robustas encontradas até agora para o modelo de medição, levando em conta considerações de índice geral, como em HUBER [25], ANDREWS e OUTROS [1], levam vantagem sobre aquelas sugeridas por gente experiente diante da única visão dos dados. A este respeito, aconselhamos a leitura do estudo realizado por RELLES e ROGERS [43].

Para terminar esta seção, vejamos o que nos diz HAMPEL [20] e [22] sobre a frequência com que, nos casos práticos, se apresentam sérios desvios às suposições habituais, que invalidam quase totalmente o uso das técnicas clássicas. Poder-se-ia destacar como principais fontes de tais desvios a: (i) arredondamento e agrupamento de dados; (ii) ocorrência de "erros grosseiros" como leituras equivocadas em um instrumento, colocação errada da vírgula decimal em algum dado que se copia, aos que se acrescenta atualmente a perfuração errada de cartões; (iii) observação de uma variável com distribuição diferente das outras; (iv) o modelo

subjacente foi concebido só como uma aproximação da realidade que se pretende estudar. A frequência com que tais erros sé apresentam depende naturalmente da qualidade dos dados. Mas é interessante destacar alguns valores. HAMPEL, fundamentando sua opinião em diversos autores, diz que habitualmente os dados, em aplicações à engenharia, apresentam em torno de uns 10% de erros grosseiros;; em dados de diversas atividades industriais esta frequência vai desde 1% até 10%, havendo casos de 20%; ainda em dados provenientes de medições cuidadosas em experiências físicas ou astronômicas é possível detectar frequentemente tal tipo de erro, até o ponto de que uma quantidade deles numa proporção entre 5% e 10% parece constituir mais a regra que a exceção. Quanto a forçar a "normalidade" nos erros de observação, parece que foi uma prática frequente em Geodésica e Astronomia. HAMPEL conclui citando as seguintes palavras de um estatístico que, parece, teve grande experiência no manejo de dados: "A normalidade é um mito; nunca houve nem haverá uma distribuição normal (de observações)" (GEARY [14]).

1.4 - Síntese do Capítulo I

Um trabalho comum em Estatística é o da estimação de certos parâmetros que ajustam os dados a um modelo hipotético

com o qual se pretende descrever um certo fenômeno. As técnicas clássicas são, em geral, muito sensíveis diante de desvios de tais hipóteses, ou seja, são "não-robustas".

Há vários anos e com maior intensidade a partir do desenvolvimento tecnológico crescente de computação, muitos estatísticos se têm dedicados a buscar métodos robustos de estimação e começar a construir uma teoria em torno deles. Nestas notas estudaremos o essencial do que tem sido feito a respeito do Modelo de Posição.

Na seção 1.2 vimos como surgiu a hipótese tão difundida de que os erros de observação se distribuem segundo a lei normal ou Gaussiana. Destacamos ali também que, apesar das dificuldades, sobretudo computacionais, que se apresentavam no uso de outras técnicas diferentes da de mínimos quadrados (indiscutível se o modelo de erros gaussianos é adequado), houve quem tentou usar técnicas robustas em seus trabalhos antes de 1960. Em torno deste ano, começa o desenvolvimento de trabalhos sobre teoria e prática da estimação robusta.

Na seção 1.3 recordamos algumas propriedades do estimador de máxima verossimilhança para o modelo de medição supondo normalidade e do estimador de mínimos quadrados para o modelo linear geral.

CAPÍTULO II

REVISÃO DE ALGUNS CONCEITOS DE INFERÊNCIAS PARAMÉTRICA

2.1 - O modelo da Inferência Paramétrica

Seja $\underline{Y} = (Y_1,\ldots,Y_n)^T$ o vetor de observações de um certo fenômeno aleatório. O ponto de partida da Inferência Paramétrica é supor que a distribuição de $\underline{Y}$ é conhecida, a menos de um parâmetro $\underline{\theta}$ que está em um certo conjunto de $\mathbb{R}^p$. Precisamente: suponhamos que a distribuição de $\underline{Y}$ é uma certa probabilidade $P_{\underline{\theta}}$ que pertence a uma família $\mathcal{P} = \{P_{\underline{\theta}} : \underline{\theta} \in \Theta\}$ de probabilidades sobre $\mathbb{R}^n$, donde $\Theta \subset \mathbb{R}^p$. Seja $q: \Theta \to \mathbb{R}^S$ $(S \geq 1)$ uma função. O problema consiste em estimar $q(\underline{\theta})$ por meio de estimadores $T = T(\underline{Y}) = T(Y_1,\ldots,Y_n)$ que são variáveis aleatórias dependentes das observações.

Exemplo 2.1.1: Modelo linear geral

Consideremos o modelo (1.3.2) (ou sua formulação matricial (1.3.3)) com as hipóteses (H1), (H2), (H3) e (H4). Então $\mathcal{P} = \{P_{(\theta,\sigma)} : (\theta,\sigma) \in \mathbb{R}^p \times (0,\infty)\}$ sendo $P_{(\theta,\sigma)}$ a distribuição normal multivariada cujo vetor de médias é $\tilde{X}\underline{\theta}$ e cuja matriz de covariância é $\sigma^2 I$. Em outras palavras, $P_{(\underline{\theta},\sigma)}$ está definida pela densidade.

$$(2.1.1)\quad \underline{y} = (y_1,\ldots,y_n) \mapsto (2\pi\sigma^2)^{-n/2} \exp\{-\frac{1}{2\sigma^2}\Sigma(y_i - \tilde{X}\underline{\theta})^2\}$$

Assim, então, a densidade de $\underline{Y}$ sob $P_{(\underline{\theta},\sigma)}$ será dada por (2.1.1).

Exemplo 2.1.2. Modelo de Medição

Consideremos o modelo (1.3.1) com as hipóteses (i), (ii), (iii), (iv) e (v). Então $\mathcal{P} = \{P_{(\mu,\sigma)}:(\mu,\sigma) \in \mathbb{R} \times (0,\infty)\}$ sendo o $P_{(\mu,\sigma)}$ a distribuição normal multivariada de média $(\mu,\ldots,\mu)^T$ e matriz de covariância $\sigma^2 I$.

Nos livros-texto de Estatística temos aprendido diversas propriedades de estimadores "razoáveis" de $q(\underline{\theta})$. Também aprendemos diversos critérios para selecionar esses estimadores. Um dos mais difundidos é aquele baseado no erro quadrático médio. Vejamos sinteticamente sua definição e recordemos algumas das considerações de importância em sua aplicação à classificação de estimadores. Seguiremos nesta parte a BICKEL e DOKSUM [4].

2.2 - Erro quadrático médio - ENVUMV

Definição 2.2.1: Chama-se erro médio quadrático (EMQ) de T a

$$R(\underline{\theta},T) = E_{P_{\underline{\theta}}} (|T - q(\underline{\theta})|^2)$$

definido $\forall\ \underline{\theta} \in \Theta$ (usamos a seguinte notação: se $\underline{v} = (v_1,\ldots,v_s)^T \in \mathbb{R}^s$, então $|\underline{v}|^2 = \underline{v}^T\underline{v} = \sum_{i=1}^{s} v_i^2$). É fácil ver que o EMQ está determinado pela variância de T e por sua média. Se $R(\underline{\theta},T) < +\infty$, então:

$$(2.2.1)\quad R(\underline{\theta},T) = Var(\underline{\theta},T) + B^2(\underline{\theta},T)$$

onde $B(\underline{\theta},T) = E_{P_{\underline{\theta}}} (T - q(\underline{\theta}))$ é o vício de T para estimar $q(\underline{\theta})$.

Como o $\underline{\theta}$ que determina a distribuição de $\underline{Y}$ poderia ser qualquer ponto de Θ, dados dois estimadores S e T parece natural preferir o uso de T ao de S se $R(\underline{\theta},T) \leq R(\underline{\theta},S)$ $\forall\, \underline{\theta}$ com desigualdade estrita para algum $\underline{\theta}_o$. Em tal caso se diz que T é melhor que S e que S é inadmissível.

Exemplo 2.2.1: Modelo de Medição

Suponhamos estar na situação do exemplo 2.1.2. Seja $\hat{\mu}$ o estimador média amostral. ($\hat{\mu} = (1/n) \sum_{i=1}^{n} Y_i$). Um cálculo direto prova que, se $q(\mu,\sigma) = \mu$, então $B((\mu,\sigma),\hat{\mu}) = o$ e que $R((\mu,\sigma),\hat{\mu}) = Var((\mu,\sigma),\hat{\mu}) = \sigma^2/n$, $\forall (\mu,\sigma) \in \Theta = \mathbb{R} \times (o,\infty)$. em particular, vemos que $\hat{\mu}$ é não viciado para estimar μ.

Suponhamos agora que $S = a\hat{\mu}$ com $o<a<1$ fixo. Então

$$B((\mu,\sigma),S) = (a-1)\mu$$
$$R((\mu,\sigma),S) = a^2 \frac{\sigma^2}{n} + (a-1)^2 \mu^2 .$$

Logo, se μ está "próximo" de 0, então será melhor usar S em vez de $\hat{\mu}$.

O exemplo que acabamos de ver nos mostra a dificuldade na escolha de um estimador se consideramos todos os estimadores possíveis. Mais ainda, é bem sabido que é impossível encontrar um estimador ótimo no sentido de que seja o de menor EMQ entre todos os estimadores. Daí segue a necessidade de se restringir a classe dos estimadores a considerar. Uma classe razoavelmente ampla é a dos T tais que $E_{p_{\underline{\theta}}}(T - q(\underline{\theta})) = 0$ $\forall\, \underline{\theta}$, isto é, a classe dos estimadores não viciados. No entanto, podem existir estimado

res viciados (conforme exemplo 2.2.1) que tenham EMQ menor que algum não viciado. De todas as maneiras, considerar somente a classe dos estimadores não viciados tem algumas vantagens, dignas de se levar em conta: em primeiro lugar nos asseguramos de não sobre - nem sub-estimar; em segundo lugar, trabalhando nessa fica fora de consideração alguns estimadores absurdos como as constantes; finalmente, em muitos casos é possível encontrar dentro desta classe um estimador T que tenha EMQ mínimo entre os não viciados, isto é

$$\text{Var}(\underline{\theta},T) \leq \text{Var}(\underline{\theta},S) \quad \forall\, \underline{\theta}$$

onde S é qualquer estimador não viciado. A tal T chamamos ENVUMV. Não recordaremos nestas notas os importantes resultados relacionados com os ENVUMV, tais como os teoremas de RAO BLACKWELL, LEHMANN-SCHEFFE, etc. e que estão detalhadamente expostos em muitos textos. Certamente existem dificuldades para se seguir rigidamente este critério de busca de estimadores ótimos entre os não-viciados. Com efeito: pode suceder que não existam estimadores não viciados; ou que exista um ENVUMV absurdo sob o ponto de vista estatístico; também, como vimos no exemplo 2.2.1, podem existir estimadores S e T razoáveis sugeridos por outros critérios que não sejam comparáveis se usamos o critério baseado no EMQ, pois, para certos valores do parâmetro, S é melhor que T e para outros, T é melhor que S. Frequentemente, o cálculo do EMQ é bastante difícil de ser feito. No entanto, quando o tamanho da amostra é "grande", comportamento de muitos estimadores é tal que as comparações a critérios de escolha são

mais compreensíveis e simples, ao menos estatisticamente. A dificuldade no estudo de amostras grandes (teoria assintótica) está na ferramenta matemática necessária para sua formalização. Por isto, não entraremos em muitos detalhes na exposição desses formalismos, deixando seu estudo para a inquietação e necessidade do leitor, que pode recorrer a uma abundante e excelente bibliografia como a citada, por exemplo, em BICKEL e DOKSUM [4].

2.3 - Algumas formalizações sobre a teoria assintótica

Sejam Ω o espaço amostral; Θ um subconjunto de $\mathbb{R}^p$; $\mathcal{P} = \{P_{\underline{\theta}} \ \underline{\theta} \in \Theta\}$ uma família de probabilidades sobre Ω; $Y_1, Y_2, \ldots$ uma sequência de variáveis aleatórias definidas sobre Ω; $q : \Theta \to \mathbb{R}^s$ uma função; para cada $n=1,2,\ldots$, T_n um estimador de $q(\underline{\theta})$ baseado em $Y_1, \ldots, Y_n$.

Definição 2.3.1: Se diz que a sequência (T_n) é consistente para estimar $q(\underline{\theta})$ se

$$P_{\underline{\theta}}(|T_n - q(\underline{\theta})| > \varepsilon) \to 0, \quad n \to \infty, \quad \forall\, \varepsilon > 0 \text{ e } \forall\, \underline{\theta}$$

(consistência fraca)

ou se

$$P_{\underline{\theta}}(|T_n - q(\underline{\theta})| \to 0) = 1 \quad \forall\, \underline{\theta}$$

(consistência forte)

Definição 2.3.2: Se (T_n) converge, fraca ou fortemente, a $q_0(\underline{\theta})$ $\forall\, \underline{\theta}$, então se chama vício assintótico de (T_n) a $q(\underline{\theta}) - q_0(\underline{\theta})$

Definição 2.3.3: Se diz que (T_n) é assintoticamente normal se existem $q_o(\underline{\theta})$ e $\sigma_o(\underline{\theta})$ tais que

$$\sqrt{n}\ (T_n - q_o(\underline{\theta})) \to N(0,\ \sigma_o^2\ (\underline{\theta}))\quad (D),\ n\to\infty,\quad \forall\ \underline{\theta}$$

o que é equivalente a

$$P_{\underline{\theta}}(\sqrt{n}\,(T_n - q_o(\underline{\theta}))/\sigma_o(\underline{\theta}) \le x) \to \Phi\,(x),\ n\to\infty,\quad \forall\ x,\quad \forall\ \underline{\theta}\ .$$

A $\sigma_o^2(\underline{\theta})$ se chama variância assintótica de (T_n) sob $\underline{\theta}$.

Os conceitos dados acima são usados para comparar o rendimento de diversas sequências de estimadores de $q(\underline{\theta})$, mediante a eficiência assintótica relativa. Em termos mais precisos:

Definição 2.3.4: Sejam $(T_n^{(1)})$ e $(T_n^{(2)})$ duas sucessões de estimadores de $q(\underline{\theta})$. Se existe

$$(2.3.1)\quad EFA((T_n^{(1)}),(T_n^{(2)}),\underline{\theta}) =: \lim \frac{R(\underline{\theta},T_n^{(2)})}{R(\underline{\theta},T_n^{(1)})}$$

então esse limite é chamado eficiência assintótica de $(T_n^{(1)})$ relativa a $(T_n^{(2)})$ sob $\underline{\theta}$.

Vimos que, dados dois estimadores de $q(\underline{\theta})$, digamos T e S, ambos baseados em uma amostra de tamanho fixo, era razoável usar T em lugar de S se $R(\underline{\theta},T) \le R(\underline{\theta},S)$. $\forall\ \underline{\theta}$. De modo análogo, será preferível usar $(T_n^{(2)})$ em lugar de $(T_n^{(1)})$ se

$EFA\ ((T_n^{(1)}),\ (T_n^{(2)}),\ \underline{\theta}) < 1\quad \forall\ \underline{\theta}$.

Notemos que, por (2.2.1), se $T_n^{(1)}$ e $T_n^{(2)}$ são não viciados para estimar $q(\underline{\theta})$ $\forall$ n, então (2.3.1) é o mesmo que

$$(2.3.2)\quad EFA((T_n^{(1)}),(T_n^{(2)}),\underline{\theta}) = \lim_{n\to\infty} \frac{Var(\underline{\theta},T_n^{(2)})}{Var(\underline{\theta},T_n^{(1)})}$$

Esta consideração conduz a uma ligeira modificação na definição 2.3.4 para o caso de estimadores assintoticamente normais com vício assintótico nulo.

Isto é:

Definição 2.3.5: Sejam $(T_n^{(1)})$ e $(T_n^{(2)})$ duas sequências assintoticamente normais, ambas com vício assintótico nulo e variâncias assintóticas $\sigma_1(\underline{\theta})$ e $\sigma_2(\underline{\theta})$ respectivamente. Chama-se eficiência assintótica de $(T_n^{(1)})$ relativamente a $(T_n^{(2)})$ sob $\underline{\theta}$ a

$$(2.3.3)\quad EFA((T_n^{(1)}),(T_n^{(2)}),\underline{\theta}) = \frac{\sigma_2^2(\underline{\theta})}{\sigma_1^2(\underline{\theta})}$$

Nota: Devemos destacar que, ainda sob as hipóteses da definição 2.3.5, o limite do membro direito de (2.3.1) não seria necessariamente que existir, nem tampouco $\lim_{n\to\infty} E_{\underline{\theta}}(T_n^{(i)} - q(\underline{\theta})) = 0$ ou $\lim Var(\underline{\theta},\sqrt{n}(T_n^{(i)}-q(\underline{\theta})) = \sigma_i^2(\underline{\theta})$.

Por isso, a definição 2.3.5 é uma modificação da definição 2.3.4 que será de utilidade, como veremos mais adiante, para comparar o rendimento de diversos estimadores sob o ponto de vista da teoria assintótica. Claro está que, se $T_n^{(i)}$ é não viciado $\forall n$ e

$$n\ Var\ (\underline{\theta},\ T_n^{(i)}) \to \sigma_i^2(\underline{\theta}) \qquad \forall\ \underline{\theta}$$

qualquer que seja $i = 1,2$, então (2.3.2) e (2.3.3) coincidem.

2.4 - Estimador de Máxima Verossimilhança (EMV)

As propriedades assintóticas dos ENVUMV também estão

estudadas com certos detalhes, por exemplo, em BICKEL e DOKSUM [4]. Esta resultam ser semelhantes às de outro estimador consagrado pelo uso em Estatística, sobretudo a partir dos trabalhos de Fisher: o estimador de máxima verossimilhança (EMV), que já tivemos oportunidade de estudar na seção 1.3 para os modelos de medição e regressão. Vejamos sua definição para o modelo mais geral proposto nesta seção.

Definição 2.4.1: Seja $n \geq 1$ fixo; chama-se função de verossimilhança a $L: \Theta \times \mathbb{R}^n \to \mathbb{R}$ definida por

$$L(\underline{\theta}, (y_1, \ldots, y_n)^T) = p((y_1, \ldots, y_n)^T, \underline{\theta})$$

onde, para cada $\underline{\theta} \in \Theta$ a função $(y_1, \ldots, y_n)^T \to p((y_1, \ldots, y_n)^T, \underline{\theta})$ é uma densidade de $P_{\underline{\theta}}$ (supomos que $P_{\underline{\theta}}$ está definida por uma densidade).

Para cada $\underline{y} = (y_2, \ldots, y_n)^T \in \mathbb{R}^n$ se supõe que existe $\hat{\theta}_{EMV}(\underline{y}) \in \Theta$ tal que

$$L(\hat{\theta}_{EMV}(\underline{y}), \underline{y}) \geq L(\underline{\theta}, \underline{y}) \quad \forall \quad \underline{\theta} \in \Theta .$$

Chama-se estimador de máxima verossimilhança de $q(\underline{\theta})$ a $T_n(y_1, \ldots, y_n) = q(\hat{\theta}_{EMV}(y_1, \ldots, y_n))$. (A $\hat{\theta}_{EMV} = \hat{\theta}_{EMV}(y_1, \ldots, y_n)$ se chama estimador de máximo verossimilhança de $\underline{\theta}$) .

2.5 - Estimadores assintoticamente normal eficientes (ANE)

Os estimadores de máximo verossimilhança, na maioria dos casos, têm propriedades assintóticas notavelmente mais fortes

que as de consistência e/ou normalidade. Na verdade, tipicamente são assintoticamente normas e eficientes (ANE) (BICKEL e DOKSUM [4]) Recordemos essa definição: consideremos o modelo do início da seção 2.3. Suponhamos que, sob $P_{\underline{\theta}}$, $Y_1, Y_2, \ldots$ constituem uma sucessão de variáveis aleatórias independentes e identicamente distribuidas (i.i.d.). E também que $P_{\underline{\theta}}$ está dada por uma densidade $y \to p(y,\underline{\theta})$ definida sobre $\mathbb{R}$ (isto é: $P_{\underline{\theta}}(Y \in B) = \int_B p(y,\underline{\theta})dy \quad \forall \ B \in \mathcal{B}_1$, onde $\mathcal{B}_1$ representa a família de todos os borelianos de $\mathbb{R}$ e Y é qualquer das Y_i's). Finalmente, suponhamos que Θ é um subconjunto aberto de $\mathbb{R}$. Sob certas condições de regularidade (BICKEL e DOKSUM [4]) tem-se que

Desigualdade de Rao-Cramer: Seja $n \geq 1$. Para cada θ

$$q_n(\theta) = E_{P_\theta} T_n$$

Se $\mathrm{Var}(\theta, T_n) < +\infty \quad \forall \ \theta$, então q_n é diferenciável e

$$\mathrm{Var}\ (\theta, T_n) \geq \frac{(q_n'(\theta))^2}{nI_1(\theta)}$$

onde

$$I_1(\theta) = E_{P_\theta}\{(\frac{\partial \ \mathrm{Log}\ p(X,\theta)}{\partial \theta})^2\}$$

é o número de informação de Fisher (que supomos positivo).

A partir deste resultado, é fácil provar que

Proposição 2.5.1: Suponhamos que

(i) $q_n'(\theta) \to q'(\theta), \quad n \to \infty \quad \forall \ \theta \in \Theta$

(ii) (T_n) é assintoticamente normal com variância assintótica $\sigma^2(\theta)$.

(iii) $n \operatorname{Var}(\theta, T_n) \to \sigma^2(\theta)$, $n \to \infty$, $\forall\, \theta \in \Theta$

Então

$$(2.5.1) \quad \sigma^2(\theta) \geq \frac{(q'(\theta))^2}{I_1(\theta)} \qquad \forall\, \theta \in \Theta$$

Se nesta última desigualdade se cumpre o "=" $\forall\, \theta$, então se diz que (T_n) é assintoticamente normal eficiente. Notemos que se T_n é não viciado para estimar $q(\theta)$ $\forall$ n, $(q_n(\theta) = q(\theta), \forall\, n, \forall\, \theta)$ então (i) se cumpre obviamente.

Para fixar um pouco este conceito, vejamos que no modelo de medição sob normalidade (Exemplo 2.2.1) a sucessão

$$(\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i) \quad \text{é} \quad \text{ANE.}$$

Exemplo 2.5.1 :

Sejam Ω o espaço amostral; $\sigma_o > 0$; para cada $\mu \in \mathbb{R}$ seja P_μ a probabilidade sobre Ω dada por

$$P_\mu(Y_1 \leq y) = \Phi(y; \mu, \sigma_o)$$

Logo, P_μ está dada pela densidade $y \mapsto \varphi(y; \mu, \sigma_o)$.

Sob P_μ, suponhamos que $Y_1, Y_2, \ldots$ é uma sucessão i.i.d. (isto é análogo a dizer que, para cada n, $Y_1, \ldots, Y_n$ é uma amostra de tamanho n de $N(\mu, \sigma_o^2)$), $q: \mathbb{R} \to \mathbb{R}$ é definida por $q(\mu) = \mu$; finalmente, para cada $n = 1, 2, \ldots$ $\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i$. Como sabemos, $\hat{\mu}_n$ é não viciado para estimar μ. Pelo Teorema central do Limite temos que, sob P_μ,

$$\sqrt{n}\,(\hat{\mu}_n - \mu) \to N(0,\sigma_o^2) \qquad (D).$$

Ademais, $\mathrm{Var}(\mu,\hat{\mu}_n) = \sigma_o^2/n$; logo, $n\,\mathrm{Var}(\mu,\hat{\mu}_n) = \sigma_o^2$ $\forall$ $n=1,2,\ldots$ Por outro lado,

$$\log \varphi\,(y;\mu,\sigma) = -\frac{1}{2}\left(\frac{x-\mu}{\sigma_o}\right)^2 - \log\sqrt{2\pi}\ ,$$

de onde é fácil ver que $I_1(\theta) = 1/\sigma_o^2$. Logo, a igualdade se cumpre em (2.5.1).

Para finalizar essa breve recordação de certos conceitos clássicos da Inferência Paramétrica, resumamos as principais propriedades dos ENVUMV e dos EMV (mais detalhes podem ser vistos em BICKEL e DOKSUM [4]). Do ponto de vista assintótico, os ENVUMV e os EMV dão essencialmente a mesma resposta. Logo, a decisão de usar um ou outro depende da facilidade do cálculo dos mesmos. Parece que na prática, a existência de EMV é mais frequente que a de ENVUMV e também na maioria dos casos são mais fáceis de calcular. Se nos guiamos por estas propriedades, deveríamos preferir os EMV. Mas destaquemos uma vez mais que tanto um como o outro procedimento de estimação são muito sensíveis ao não cumprimento da hipótese de que as observações se ajustam a uma das distribuições da família $\mathcal{P}$. Por outro lado, quase nunca se sabe quão grande deve ser o tamanho da amostra (o número n de Y's) para que seja lícito descrever o comportamento de T_n pelo comportamento assintótico de (T_n).

Mais adiante estudaremos novas técnicas de estimação que são robustas no sentido visto no Capítulo I. O estudo analí

tico de suas propriedades para amostras de tamanho finito é complicado e muito difícil de realizar, a não ser por meio dos chamados "procedimentos de simulação" ou "Montecarlo", que nem sempre conduzem a resultados confiáveis. Por isso, se dedica um considerável esforço no estudo do comportamento assintótico de tais estimadores, o que também torna possível a comporação entre estas novas técnicas e as clássicas.

CAPÍTULO III

O MODELO DE POSIÇÃO

3.1 - Definição do Modelo de Posição

Nestas Notas trabalharemos com o seguinte modelo, chamado "modelo de posição"e que é um pouco mais geral que o já considerado modelo de medição. Este modelo já foi estudado em diversos trabalhos tanto do ponto de vista paramétrico, não paramétrico e de Robustez como HUBER [25] ANDREWS E OUTROS [1],etc.

Seja F uma função distribuição simétrica em torno de zero, isto é:

$$(3.1.1) \quad F(y) = 1 - \lim_{t \to o} F(-y-t) \quad \forall\, y \in R$$

E no caso de F ser contínua (3.1.1) se reduz a

$$(3.1.2) \quad F(y) = 1 - F(-y) \quad \forall\, y \in R$$

Se F admite uma densidade f de (3.1.2) conclui-se que

$$f(y) = f(-y) \quad \forall\, y \in R$$

Para cada $\mu \in R$ seja F_μ a distribuição sobre R definida por:

$$(3.1.3) \qquad F_\mu(y) = F(y-\mu)$$

Seja $\underline{Y} = (Y_1,\ldots,Y_n)^T$ um vetor aleatório (vetor de observações) que supomos ser uma amostra de tamanho n de alguma F_μ com μ desconhecida a ser estimada (recordemos que isto significa que $Y_1,\ldots,Y_n$ são i.i.d. com distribuição comum F_μ).

Por último, suporemos que como resultado de um experimento tivessemos obtido $Y_1 = y_1, \ldots, Y_n = y_n$ ($y_1, \ldots, y_n$ números reais).

Se não damos a forma de F explicitamente, contentando-nos em supor somente algumas propriedades sobre ela: simetria, unimodalidade, etc.; estaremos no ponto de partida da inferência não-paramétrica. Se damos a forma de F explicitamente, por exemplo supormos $F(\chi) = \Phi(\chi;0,1)$, estaremos sobre as hipóteses da inferência paramétrica. Também aqui é possível supor que F depende de parâmetros que se pode querer ou não estimar (parâmetros nuissance), por exemplo $F(\chi) = \Phi(\chi,0,\sigma)$ com $\sigma > o$ desconhecida).

O ponto de partida da estimação Robusta é supor que F é só parcialmente conhecida. Mais precisamente: que F está em uma visinhança de uma distribuição F_o (distribuição hipotética) que é totalmente conhecida ou conhecida exceto parâmetros "nuissance". O caso que tem recebido maior atenção é aquele que supõe.

(3.1.4) $$F = (1-\varepsilon)F_o + \varepsilon H$$

Com $o \leq \varepsilon \leq 1$ conhecido e H uma distribuição simétrica desconhecida. Neste capítulo daremos especial atenção ao caso que $F_o(\chi) = \Phi(\chi)$. Vários trabalhos tem considerado o caso em que H é não simétrica, entre eles: HUBER [25], JAECKEL [31] porém este caso não estrará em consideração pois além da complicação matemática do seu tratamento, não está bem esclarecido o significado do ponto de vista estatístico do parâmetro μ que

se quer estimar.

Em nosso estudo consideraremos casos de contaminação on de as observações não seguem a suposta distribuição Φ mas uma de caudas mais pesadas como uma Student com poucos graus de liberdade, uma Cauchy, uma normal contaminada ou uma exponencial dupla. Recordemos suas definições.

3.2 - Distribuições das Observações

Distribuição t-Student

Se diz que F é uma distribuição t-Student com m graus de liberdade se sua densidade é dada por:

$$(3.2.1) \quad ST(y;m) = \frac{1}{\sqrt{m}\,\beta(1/2, \frac{m}{2})} (1 + \frac{y2}{m})^{-\frac{m+1}{2}} \quad \forall\, y \in R$$

onde: $\beta(a,b) = \Gamma(a)\,\Gamma(b) / \Gamma(a+b)$

Distribuição de Cauchy

É um caso especial da anterior para m=1. Logo F segue uma distribuição de Cauchy se é dada pela densidade:

$$C(y) = \frac{1}{\pi(1+y^2)}$$

Distribuição exponencial dupla

F é uma distribuição exponencial dupla se e dada pela densidade

$$DE(y) = \frac{1}{2}\, e^{-|y|}$$

Distribuição normal contaminada

Quando F e como em (3.1.4) com $F_o(\chi) = \Phi(\chi)$ um caso

particular de interesse e aquele em que

$H(\chi) = \Phi(\chi;0;\tau)$ com $\tau > 1$. É fácil verificar que em tal caso F é definida pela densidade

$$CN(\chi;\varepsilon,\tau) = (1-\varepsilon)\,\varphi(\chi;0,1) + \varepsilon\,\varphi(\chi;0,\tau).$$

Distribuição logística

F é a distribuição cuja a densidade é:

$$L(y) = e^{-y}/(1+e^{-y})^2$$

Na figura 4 vemos um gráfico das densidades anteriores

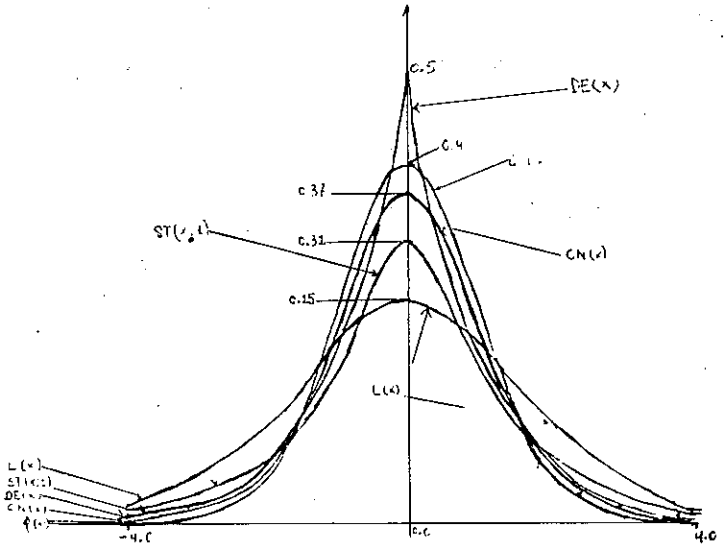

Figura 4

3.3 - Estimadores

Destaquemos uma vez mais o propósito principal deste estudo: Dado um modelo de posição e a hipótese comum da inferência paramétrica clássica: $F(\chi) = \Phi(\chi)$, queremos encontrar técnicas de estimação alternativas a sugerida pelo método habitual: o método da máxima verossimilhança. Fazemos assim pois este método nos conduz a tomar como estimador de μ a média amostral como vimos na secção 1.3; estimador que tem um comportamento muito ruim quando a F (distribuição das observações) não é Φ e sim alguma outra distribuição nas proximidades de Φ; o que também já vimos na secção 1.3.

Estimador média ponderada

Sejam $W_1,\ldots,W_n$ números reais positivos tais que $\sum_{i=1}^{n} W_i = 1$. Chama-se a média ponderada das observações $y_1,\ldots,y_n$ com pesos $W_1,\ldots,W_n$ ao estimador:

$$\hat{\mu}_W = \hat{\mu}_W(y_1,\ldots,y_n) = \sum_{i=1}^{n} W_i\, y_i$$

Este estimador foi sugerido com o intuito de ponderar (pesar) as observações de tal forma que aquelas mais afastadas da massa de dados recebam um peso menor. Desta maneira se obtem um estimador resistente ao efeito dos chamados "outliers selvagens".

Na verdade vários estimadores de μ que veremos podem ser pensados como uma média ponderada se admitirmos que os pesos W_i podem depender das observações o que parece natural.

Um dos estimadores que já tivemos ocasião de ver na seção 1.3, a mediana amostral, constitui um caso particular de uma fmília de estimadores usada já no século passado, ainda que de forma muito limitada, chamada da família dos L-estimadores.

3.3.1 - L-estimadores ou combinações lineares das estatísticas de ordem

Recordemos a definição das estatísticas de ordem

Seja $R_o^n = \{(Z_1,\ldots,Z_n)^T \in R^n : Z_1 \leq Z_2 \leq \ldots \leq Z_n\}$,

$O_n : R^n \to R_o^n$ a função definida por

$$O_n((y_1,\ldots,y_n)^T) = (y_{(1)},\ldots,y_{(n)})^T$$

sendo $(y_{(1)},\ldots,y_{(n)})^T$ o vetor $(y_1,\ldots,y_n)^T$ ordenado de modo que $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$

por exemplo:

$$(y_1,y_2,y_3,y_4)^T = (-1,-3,4,2)^T \text{ então:}$$

$$(y_{(1)},y_{(2)},y_{(3)},y_{(4)})^T = (-3,-1,2,4)^T$$

Chama-se estatística de ordem do vetor aleatório

$\underline{Y} = (Y_1,\ldots,Y_n)^T$ a $\underline{Y}(\cdot) = O_n(\underline{Y}) = (Y_{(1)},\ldots,Y_{(n)})^T$

A estatística definida por $\underline{Y} = (Y_1,\ldots,Y_n)^T \mapsto Y_{(i)}$ se chama a i'ésima estatística de ordem de $\underline{Y}$.

Definição 3.3.1: Sejam $a_1,\ldots,a_n$ números reais tais que

$\sum_{i=1}^{n} a_i = 1$ chama-se L-estimador induzido por $a_1,\ldots,a_n$ baseado em $y_1,\ldots,y_n$ a

$$(3.3.1)\quad L_n = L_n(y_1,\ldots,y_n) = \sum_{i=1}^{n} a_i\ y_{(i)}$$

Segundo a maneira que se derivam os ai's resultam diferentes tipos de L-estimadores. Os mais tradicionais são:

Mediana Amostral: chama-se mediana de $y_1,\ldots,y_n$ a;

$$(3.3.2)\quad MED=MED(y_1,\ldots,y_n) = \begin{cases} y_{(k)} & \text{se } n=2k-1 \\ \dfrac{y_{(k)} + y_{(k+1)}}{2} & \text{se } n=2k. \end{cases}$$

é fácil comprovar que (3.3.2)

é o mesmo que:

$$MED = MED\ (y_1,\ldots,y_n) = (y_{(M)} + y_{(L)})/2$$

Sendo $M = [(n+1)/2]$, $L = [(n+2)/2]$

([.t.] significa a parte inteira de t).

Que MED é um L estimador se pode ver se em (3.3.2) se faz:

$$\left.\begin{array}{l} a_k = 1 \\ a_i = 0 \quad i \neq k \end{array}\right\} \quad \text{Se} \quad n = 2k-1$$

$$\left.\begin{array}{l} a_k = 1/2 \\ a_{k+1} = 1/2 \\ a_i = 0 \quad i \neq k \quad e \quad i \neq k+1 \end{array}\right\} \qquad \text{Se} \quad n = 2k$$

Média α truncada: Seja $o<\alpha< 1/2$. Chama-se média α-truncada de $y_1,\ldots,y_n$ ao L estimador obtido de (3.3.1) fazendo-se:

$$a_i = \frac{1}{(n-2[n\alpha])} \qquad \text{Si} \quad i = ([n\alpha]+1,\ [n\alpha]+2,\ldots,n-[n\alpha])$$

$$a_i = o \qquad \text{caso contrário}$$

isto é, ao estimador:

$$(3.3.3)\ \alpha\, T = \alpha T(y_1,\ldots,y_n) = \frac{1}{n-2[n\alpha]} \sum_{i=[n\alpha]+1}^{n-[n\alpha]} y_{(i)}$$

os valores de α que se usam com maior frequência são os α tais que $0.05\leq\alpha\leq 0.15$.

Notemos que este estimador consiste em truncar, tirar, uma proporção α dos valores maiores e uma proporção igual dos menores e tomar a média dos valores restantes.

Outro L-estimador com o qual se tem trabalhado já se faz vários anos e a:

Média α-winsorizada: Seja $o<\alpha<1/2$.
chama-se média α-winzorizada de $y_1,\ldots,y_n$ ao estimador obtido de (3.3.1) fazendo:

$$a_i = 1/n \quad \text{se} \quad [n\alpha]+2 \leq i \leq n-[n\alpha]-1$$

$$a_i = \frac{1+[n\alpha]}{n} \quad \text{se} \quad i = [n\alpha]+1 \quad \text{ou} \quad i = n-[n\alpha]$$

$a_i = o$ caso contrário

isto é ao estimador.

$$\alpha_{W,T} = \alpha_{W,T}(y_1,\ldots,y_n) = \frac{1}{n}\left(\sum_{i=[n\alpha]+1}^{n-[n\alpha]} y_{(i)} + \right.$$

$$\left. + \alpha(y_{([n\alpha]+1)} + y_{(n-[n\alpha])})\right)$$

Vários outros L-estimadores tem sido sugeridos ultimamente, se poderia ver ANDREWS e OUTROS [1] para a definição de alguns deles. Entre eles se destaca um estudado por Gastwirth e Rubin chamado:

TRI-MÉDIA: Chama-se estimador tri-média (ou M-estimador de Gastwirth-Rubin) baseado em $y_1,\ldots,y_n$ ao:

$$T_{GR} = T_{GR}(y_1,\ldots,y_n) = 1/4\ y_{(q-)} + 1/2\ MED + 1/4\ y_{(q+)}$$

onde $q- = [n/4]+1$, $q+ = n - [n/4]$ e MED a mediana de $y_1,\ldots,y_n$

Já tinhamos visto na secção 2.4, a definição do estimador de máxima verossimilhança de $q(\theta)$ (Definição 2.4.1) sob o modelo proposto no início da secção 2.1. Não obstante, vejamos o que obteremos ao aplicar aquela definição para estimar μ sob o modelo de posição.

3.3.2. Estimador de Maxima verossimilhança - função "score"

Suponhamos que a função de distribuição F admite uma densidade f. Chama-se estimador de máxima verossimilhança baseado em $y_1,\ldots,y_n$ à $\hat{\mu}_{EMV}$ que satisfaz:

$$L(\hat{\mu}_{EMV};\ y_1,\ldots,y_n) \geq L(m;y_1,\ldots,y_n) \quad \forall\ m \in \mathbb{R}\ ,$$

onde $L(m;y_1,\ldots,y_n) = f(y_1-m)\ldots f(y_n-m) \quad \forall \quad m.$

É fácil ver que $\hat{\mu}_{EMV}$ está definido também por:

$$(3.3.4)\quad \sum_{i=1}^{n} \rho(y_i-\hat{\mu}_{EMV}) \leq \sum_{i=1}^{n} \rho(y_i-m) \quad \forall \quad m$$

onde $\rho(t) = -\log f(t)$ (log é função logaritmica natural).

Sob certas condições de regularidade sob f, $\hat{\mu}_{EMV}$ pode ser definido equivalentemente por:

$$(3.3.5)\quad \sum_{i=1}^{n} \Psi(y_i-\hat{\mu}_{EMV}) = 0\ .$$

Sendo $\Psi(t) = \dfrac{d\rho(t)}{dt} = -\dfrac{\frac{d}{dt} f(t)}{f(t)}$.

Alguns autores (por exemplo MARTIN [39] chamam a essa função Ψ "score function" associada a f. Possivelmente a razão pelo qual essa função tem recebido um nome próprio é que ela dá uma idéia de como são as "caudas" da distribuição F. Com efeito, $\Psi(t)$ mede a razão relativa de decrescimento da função densidade f.

Na tabela seguinte temos quais são as funções "scores" das funções definidas na secção 3.2 e na figura 5 seus gráficos aproximados nos dão uma idéia mais eloquente das diferenças entre as caudas dessas distribuições.

Densidade	Função Score
$\varphi(t)$	$\Psi_N(t) = t$
$ST(t;m) = \frac{1}{\sqrt{m}\ \beta(\frac{1}{2},\frac{m}{2})}(1+\frac{t^2}{m})^{-\frac{m+1}{2}}$	$\Psi_{ST(m)}(t) = \frac{m+1}{m}\ \frac{t}{(1+\frac{t^2}{m})}$
$C(t) = \frac{1}{\pi(1+t^2)}$	$\Psi_C(t) = \frac{2t}{1+t^2}$
$DE(t) = \frac{1}{2}\ e^{-\lvert t\rvert}$	$\Psi_{DE}(t) = \text{sinal } t$
$CN(t;\varepsilon,\tau) = (1-\varepsilon)\varphi(t)+\frac{\varepsilon}{\tau}\ \varphi(\frac{t}{\tau})$	$\Psi_{CN(\varepsilon,\tau)}(t) =$ $= \frac{(1-\varepsilon)t\varphi(t) + \frac{\varepsilon}{\tau}\ \frac{t}{\tau}\ \varphi(\frac{t}{\tau})}{(1-\varepsilon)\varphi(t) + \frac{\varepsilon}{\tau}\ \varphi(\frac{t}{\tau})}$
$L(t) = \frac{e^{-t}}{(1+e^{-t})^2}$	$\Psi_L(t) = \frac{1-e^{-t}}{1+e^{-t}} = tgh(\frac{t}{2})$

Os gráficos aproximados das funções scores são:

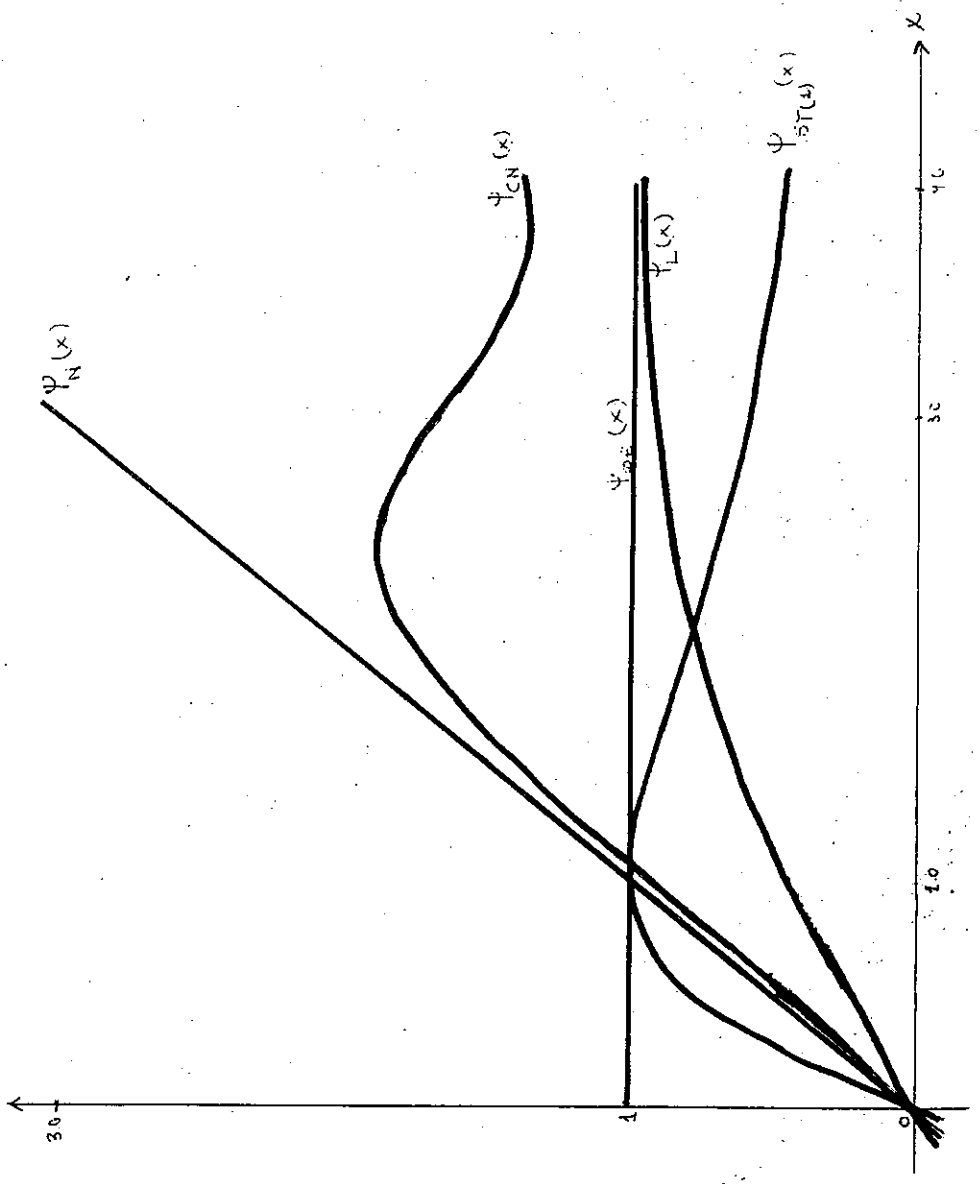

Figura 5

A partir de (3.3.5) e da tabela anterior podemos calcular $\hat{\mu}_{EMV}$ sob a hipótese de que as observações se distribuem segundo F. Assim por exemplo teremos:

$\hat{\mu}_{EMV}$ sob $F(x) = \Phi(x)$

$$\sum_{i=1}^{n} (y_i - \hat{\mu}_{EMV}) = 0$$

de onde resulta

$$\hat{\mu}_{EMV} = \frac{1}{n} \sum_{i=1}^{n} y_i \ .$$

confirmando assim o que diziamos no início desta secção 3.3.

$\hat{\mu}_{EMV}$ sob F exponencial dupla

Pode-se provar que neste caso $\hat{\mu}_{EMV}$ coincide com MED. Estimador que como já vimos não está influenciado por observações muito afastadas da massa de dados e é portanto, robusto nesse sentido. Assim então, se em um problema temos decidido usar como distribuição hipotética a dada pela exponencial dupla, a técnica clássica nos dá um estimador robusto. O mesmo não ocorre se a distribuição hipotética é a normal.

A partir das conclusões expostas em diversos trabalhos (por exemplo: HUBER [25], HAMPEL [21], ANDREWS e OUTROS [1]) têm-se chegado à conclusão de que uma alternativa robusta muito conveniente quando se quer evitar as catastróficas consequências de usar a média amostral quando as observações se desviam pouco da "normalidade", á usar uma classe de estimadores similares ao

de máxima verossimilhança. Esta classe é a dos:

3.3.3 M-Estimadores

Antes de ver sua definição precisa vejamos brevemente de onde surgem. Como vimos a média amostral coincide com $\hat{\mu}_{EMV}$ sob $F(x) = \Phi(x)$ que por (3.3.4) está dado por

$$(3.3.6) \quad \sum_{i=1}^{n} (y_i - \hat{\mu}_{EMV})^2 \leq \sum_{i=1}^{n} (y_i - m)^2 \quad \forall \ m \ ,$$

de onde se vê que os resíduos $r_i = (y_i - \hat{\mu}_{EMV})$ grandes (y_i muito afastado do valor central da massa de dados) têm uma influência excessiva na determinação do $\hat{\mu}_{EMV}$ que resolve (3.3.6). Com esta idéia em mente é natural considerar estimadores de μ que sejam solução de (3.3.4) porém com ρ escolhida de forma conveniente e não necessariamente como $-\log f$. A conveniência de escolher uma determinada ρ depende de diversas considerações: algumas bem nítidas comoveremos mais adiante, outras baseadas quiçá na forma de certas funções "score", porém todas elas com a idéia de minimizar a importância dos resíduos grandes. De forma precisa temos que

Definição 3.3.2. Seja $\rho : R \to R$ uma função não negativa. Chama-se M-estimador de μ definido por ρ baseado em $y_1, \ldots, y_n$ a um $T_n = T_n(y_1, \ldots, y_n)$ tal que

$$(3.3.7) \quad \sum_{i=1}^{n} \rho(y_i - T_n) \leq \sum_{i=1}^{n} \rho(y_i - m) \quad \forall\, m \in R$$

Em princípio, ρ poderia ser qualquer, porém se algumas propriedades sobre ρ não forem requeridas um T_n que satisfaça (3.3.7) poderá existir apenas para conjuntos $\{y_1, \ldots, y_n\}$ que muito di-

ficilmente apareceram na prática. Daí, em geral, se pede que ρ seja simétrica, convexa e $\rho(t) \to \infty$ quando $|t| \to \infty$. Na verdade, em trabalhos de índole mais prática que teórica, se tem trabalhado com M-estimadores dados por ρ com propriedades de regularidade suficientes para que T_n possa ser definido equivalentemente por:

$$(3.3.8) \qquad \sum_{i=1}^{n} \Psi(y_i - T_n) = 0$$

onde: $\Psi(t) = \frac{d}{dt}\rho(t)$.

Por isso alguns autores preferem definir os M-estimadores através de funções Ψ, isto é, usando(3.3.8). Aos M-estimadores que resultam de usar as Ψ da tabela das funções "score" se acrescentaram várias outras definidas por outras Ψ. Algumas das mais usuais são:

<u>M-estimador com Ψ do tipo Huber de parâmetro k</u>

baseado em $y_1,\ldots,y_n$ é o $\hat{\mu}_{H,k} = T_n$ definido por (3.3.8) com $\Psi = \Psi_{H,k}$ dada por:

(3.3.9) $\Psi_{H,k}(t) = \min(|t|,k)$ sinal (t) onde k é um parâmetro cuja determinação veremos mais a frente.

<u>M-estimador com Ψ do tipo Hampel de parâmetros A,B,C.</u>

baseado em $y_1,y_2,\ldots,y_n$ e o $\hat{\mu}_{HA,A,B,C} = T_n$ definido por (3.3.8) com $\Psi = \Psi_{HA,A,B,C}$ dada por

$$(3.3.10)\ \Psi_{HA,A,B,C}(t) = \begin{cases} t, & \text{se } 0 \le |t| \le A \\ A \text{ sinal } (t), & \text{se } A \le |t| \le B \\ A\,\frac{C-|t|}{C-B} \text{ sinal}(t), & \text{se } B \le |t| \le C \\ 0\ , & \text{se } C \le |t| \end{cases}$$

M-estimador com Ψ do tipo Seno (ou de Andrews) de parâmetro k

baseado em $y_1,\ldots,y_n$ é o $\hat{\mu}_{A,k} = T_n$ definido por (3.3.8) com $\Psi = \Psi_{A,k}$ dado por:

$$(3.3.11)\quad \Psi_{A,k}(t) = \begin{cases} \operatorname{Sen}(\frac{t}{k}) & \text{se} \quad |t| \leq k\pi \\ 0 & \text{caso contrário} \end{cases}$$

M-estimador com Ψ do tipo "biquadrada" (ou de Tukey) de parâmetro k

baseado em $y_1,\ldots,y_n$ e o $\hat{\mu}_{B,k} = T_n$ definido por (3.3.8) com $\Psi = \Psi_{B,k}$ dada por:

$$(3.3.12)\quad \Psi_{B,k}(t) = \begin{cases} t(1-(\frac{t}{k})^2)^2 & \text{se} \quad |t| < k \\ 0 & \text{C.C.} \end{cases}$$

Como no caso do M-estimador com Ψ de Huber, mais adiante nos referiremos a determinação dos parâmetros de (3.3.11) e (3.3.12).

Por último, na figura 6 podemos ver os gráficos destas funções Ψ.

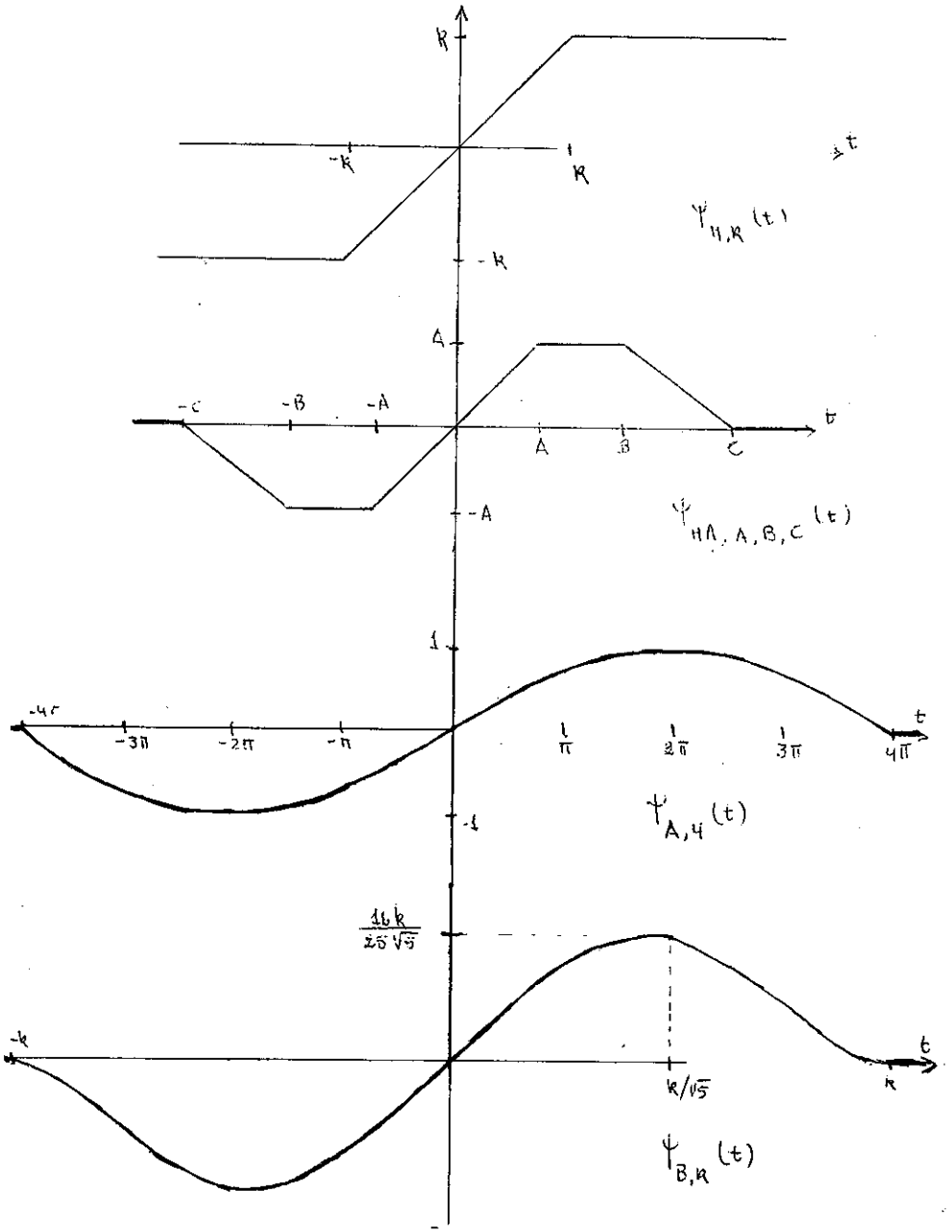
k
-k
1
k
t
$\psi_{4,k}(t)$
-k
A
-C
-B
-A
A
B
C
t
-A
$\psi_{4A,A,B,C}(t)$
1
-4π
-3π
-2π
-π
1/π
1/2π
1/3π
4π
t
-1
$\psi_{A,4}(t)$
16k/25√5
-k
k/√5
k
t
$\psi_{B,k}(t)$

Figura 6

Para terminar esta secção veremos uma classe de estimadores baseados em resultados de uma teoria estatística começada a ser desenvolvida antes de 1960:. A dos testes não-paramétricos para amostras emparelhadas baseados em postos. Esta família e a dos:

3.3.4 - R-estimadores

Antes da sua definição recordemos brevemente o sugerido pela teoria citada anteriormente. Seja G uma distribuição simétrica. Suponhamos que $U_1,\ldots,U_n$ é uma amostra de tamanho n de G e $V_1,\ldots,V_n$ uma amostra de tamanho n de G_μ (onde $G_\mu(t) = G(t-\mu)$), sendo os valores observados: $u_1 = U_1,\ldots,u_n = U_n$ $v_1 = V_1,\ldots,v_n = V_n$. Se deseja testar a hipóteses de que $u_1,\ldots,u_n$ e $v_1,\ldots,v_n$ vem da mesma distribuição G. Em termos precisos: testar $H_o:\mu=o$ contra $K:\mu \neq 0$. A teoria de testes não paramétricos recomenda-se proceder assim: (ver, por exemplo, BICKEL e DOKSUM [4], HAJEK e SIDAK [17], LEHMANN [36]). Sejam $J:(0,1) \to \mathbb{R}$ uma função não decrescente tal que $J(1-t) = -J(t)$; sejam também $a_n :\{1,2,\ldots,2n\} \to \mathbb{R}$ uma função definida por $a_n(k) = J(\frac{k}{2n+1})$

e finalmente

$$S_n = \frac{1}{n} \sum_{i=1}^{n} a_n(R_i)$$

onde R_i é o posto de $U_i = u_i$ na amostra combinada $\{u_1,\ldots,u_n, v_1,\ldots,v_n\}$.

Se a hipótese H_o está correta então S_n deve estar perto de zero, logo o teste não paramétrico baseado na estatística S_n rejeita H_o para valores grandes de S_n. A razoabilidade

deste procedimento não é difícil de ver-se e o leitor será beneficiado tratando de lê-la em alguns dos textos citados anteriormente.

Vejamos agora a conexão do dito acima e a teoria de estimação de μ no modelo de posição.

Se $Y_1,\ldots,Y_n$ é uma amostra de tamanho n de F_μ então $U_1 = Y_1 - \mu,\ldots,U_n = Y_n-\mu$ é uma amostra de F e por ser F simétrica em torno de zero resultará que $V_1 = -(Y_1-\mu),\ldots,\ldots,V_n = -(Y_n-\mu)$ é também uma amostra de F, logo:

$$S_n(\mu) = S_n(\mu; y_1,\ldots,y_n) = \frac{1}{n}\sum_{i=1}^{n} a_n(R_i)$$

estará perto de zero; sendo $S_n: R \times R^n \to R$ a função definida por:

(3.3.14) $S_n(m; X_1,\ldots,X_n) = \frac{1}{n}\sum_{i=1}^{n} a_n(R_i)$.

onde: R_i é o posto de X_i-m no conjunto:

$\{X_1-m,\ldots,X_n-m,-(X_1-m),\ldots,-(X_n-m)\}$; e a_n e J como anteriormente definidos.

Como $S_n(\mu)$ estará perto de zero quando μ é o parâmetro a ser estimado se pensa que um estimador razoável de μ baseado em $y_1,\ldots,y_n$ é um T_n tal que

$$S_n(T_n) = S_n(T_n;\ y_1,\ldots,y_n) = 0$$

Resumindo:

Definição 3.3.3: Seja $J:(0,1) \to R$ uma função não decrescente tal que $J(1-t) = -J(t)$, a_n como em (3.3.13). S_n co-

mo em (3.3.14). Chama-se R-estimador de μ definido por J baseado em $y_1,\ldots,y_n$ a um $\hat{\mu}_{R,J}$ tal que:

$$(3.3.15)\quad S_n(\hat{\mu}_{R,J};\ y_1,\ldots,y_n) = 0$$

Uma conta fácil mostra que o posto R_i de X_i-m em $\{X_1-m,\ldots,X_n-m,-(X_1-m),\ldots,-(X_n-m)\}$ é o mesmo que o posto de X_i em $\{X_1,\ldots,X_n,\ 2m-X_1,\ldots,\ 2m-X_n\}$

Por isto $\hat{\mu}_{R,J}$ está também definido por:

$$(3.3.16)\quad S_n^*(\hat{\mu}_{R,J};\ y_1,\ldots,y_n) = 0$$

onde $S_n^*: R\times R^n \to R$ é a função definida por:

$$(3.3.17)\quad S_n^*(m;\ x_1,\ldots,x_n) = \frac{1}{n}\sum_{i=1}^{n} a_n\ (R_i^*)$$

onde R_i^* é o posto de x_i no conjunto $\{X_1,\ldots,X_n,2m-X_1,\ldots,2m-X_n\}$

Utilizando a nomenclatura usual da teoria de testes não-paramétricos aos "a_n" chamamos "scores" e a J função geratriz dos scores.

Diversos R-estimadores podem ser obtidos por meio de diferentes funções J. Também aqui, como para os M-estimadores, as funções J podem ser definidas (ainda que desnecessário) baseadas na distribuição F, mais precisamente através da sua densidade f, por meio de:

$$(3.3.18)\quad J(t) = -\frac{f'(F^{-1}(t))}{f(F^{-1}(t))} \qquad 0<t<1$$

obtemos assim, por exemplo:

<u>R-estimador baseado em "scores normais"</u> é o definido pelas fórmulas (3.3.15) ou (3.3.16) com J como em (3.3.18) e $F=\Phi$, isto é:

$$J(t) = \Phi^{-1}(t) \quad o<t<1$$

R-estimador baseado em "scores mediana" como o anterior com $f = DE$ (exponencial dupla); neste caso:

$$J(t) = \begin{cases} -1 & \text{se } o<t<1/2 \\ 1 & \text{se } 1/2 <t<1 \end{cases}$$

R-estimador baseado nos "scores de Wilcoxon" ou estimador de "Hodges-Lehmann" como antes com $f = L$ (Logistica), neste caso: $J(t) = 2t-1$. Muitos autores também chamam assim o estimador definido por $J(t) = t-1/2$

3.3.5 - Outros estimadores

A parte dos já vistos se tem estudado muitos outros estimadores de μ no modelo de posição.

Como dissemos, em ANDREWS e OUTROS [1] estudaram-se mais de 65 do ponto de vista das amostras finitas e alguns deles em AZENCOT E OUTROS [2] LAUNER E WILKINSON [35] e em outros trabalhos também do ponto de vista assintótico. Uns dos que estão recebendo cada vez mais atenção são os do tipo Pitman (ver, por exemplo, LAUNER E WILKINSON[35] e MARTIN [39]). Nós só nos dedicaremos a analisar com certo detalhe os que já vimos até agora e em especial os M-estimadores. Fazemos assim devido ao fato de serem eles os mais estudados para o modelo linear geral.

3.4 - Estimadores invariantes e equivariantes

Frequentemente é desejável que os estimadores com que

se está trabalhando tenham certas propriedades de invariância e equivariância sob translações e mudança de escala nos dados originais. Essas trocas podem ser feitas para que os números a trabalhar permaneçam dentro de certos limites, como também para reduzir-se os erros de arredondamento.

Definição 3.4.1 - Se diz que $T:R^n \to R$ é invariante sobre translações se:

$$T(y_1+C,\ldots,y_n+C) = T(y_1,\ldots,y_n)$$

$$\forall\ (y_1,\ldots,y_n)^T \in R^n \text{ e } C \in R$$

Se diz que T é equivariante sobre translações se:

$$T(y_1+C,\ldots,y_n+C) = T(y_1,\ldots,y_n) + C$$

$$\forall\ (y_1,\ldots,y_n)^T \in R^n \text{ e } C \in R$$

Se diz que T é invariante sobre troca de escala se:

$$T(ay_1,\ldots,ay_n) = T(y_1,\ldots,y_n) \quad \forall\ (y_1,\ldots,y_n)^T \in R^n \text{ e } a > 0$$

Se diz que T é equivariante sob mudança de escala se:

$$T(ay_1,\ldots,ay_n) = aT(y_1,\ldots,y_n) \quad \forall (y_1,\ldots,y_n)^T \in R^n$$

Finalmente, se $T_n = T(Y_1,\ldots,Y_n)$ é um estimador baseado nas observações $Y_1,\ldots,Y_n$; então se diz que T_n é invariante (ou equivariante), sob translações (ou mudança de escala) se assim o for a função T.

3.4.1 - Equivariância sob translações e mudança de escala dos L e R estimadores

Aplicando-se as definições 3.3.1 e 3.4.1 é imediato deduzir que os L-estimadores são equivariantes sob translações e mudança de escala das observações. Propriedade análoga tem os R-estimadores, como se pode ver através da aplicação direta das definições 3.3.3 e 3.4.1.

3.4.2 - Equivariância sob translações dos M-estimadores

Também aqui uma aplicação simples de (3.3.8) conduz a ver que os M estimadores são equivariantes sob translação das observações.

3.4.3 - Equivariância sob mudança de escala dos M-estimadores

As coisas aqui não são tão simples. Necessitamos modificar a definição dos M-estimadores dada por (3.3.8) a fim de obter a equivariância dos mesmos.

Antes de mais nada uma observação sobre notação. Chamaremos de $\sigma(Z)$ ao parâmetro de escala da variável aleatória Z que satisfaz:

(3.4.1) $\sigma(Z-b) = \sigma(Z) \quad \forall\, b \in R$

(3.4.2) $\sigma(aZ) = a\sigma(Z) \quad \forall\, a > o$

Mais adiante tornaremos mais preciso este conceito de parâmetro de escala. Consideremos a situação habitual: (início da Secção 3.1) sejam: Y uma, qualquer, das variáveis $Y_1,\ldots,Y_n$; $\Psi:R \to R$ e $T:R^n \to R$ funções e $T_n = T(y_1,\ldots,y_n)$ tais que:

$$(3.4.3)\quad \sum_{i=1}^{n} \Psi\left(\frac{y_i - T(y_1,\ldots,y_n)}{\sigma(Y-\mu)}\right) = \sum_{i=1}^{n} \Psi\left(\frac{y_i - T(y_1,\ldots,y_n)}{\sigma(Y)}\right) = 0$$

quaisquer que sejam os valores observados $y_1,\ldots,y_n$ de $Y_1,\ldots,Y_n$; finalmente seja $a>0$; então:

$$\sum_{i=1}^{n} \Psi\left(\frac{ay_i - aT(y_1,\ldots,y_n)}{\sigma(aY)}\right) = 0$$

Isto nos diz que se trocamos (3.3.8) por (3.4.3) e colocando $T_n = T(y_1,\ldots,y_n)$, então o estimador T_n que resulta é invariante por mudança de escala. Ainda subsistem alguns problemas: 1) Que definir "naturalmente" como parâmetro de escala ou de dispersão de uma variável aleatória Z ? ; 2) do ponto de vista prático o $\sigma(Y)$ que aparece em (3.4.3), será desconhecido a maioria das vezes, como estimá-lo "razoavelmente" e antes de tudo "robusto" ? Pois não teria sentido estimar μ robustamente e não proceder da mesma maneira com o parâmetro de escala.

Se lessemos a literatura existente veríamos que não há acordo a respeito das respostas a nenhuma das perguntas anteriores.

Com respeito a questão 1) é comum definir como parâmetro de escala de uma variável Z, que tem momento de 2ª ordem finito a:

$$\sigma_1(Z) = DS(Z) = \sqrt{VAR(Z)}$$

sendo DS(Z) o desvio padrão de Z (e VAR(Z)é a variância de Z). Sucede muitas vezes que Z não tem momento de 2ª ordem finito,

por isto e por outros que não vem ao caso, se usa também como parâmetro de escala a:

$$\sigma_2(Z) = MAD(Z) = MEDM(|Z - MEDM(Z)|)$$

Sendo MAD a mediana dos desvios absolutos e MEDM a mediana:

$$MEDM(Z^*) = F^{-1}_{Z^*}(1/2)$$

qualquer que seja a variável aleatória Z^*, F_{Z^*} sua função de distribuição e:

$$F^{-1}_{Z^*}(t) = \inf\{y : F(y) \geq t\} \quad \forall \quad o<t<1$$

Notemos que $\sigma_2(Z)$ estará sempre bem definido. Também o estará o seguinte parâmetro de escala que tem sido usado com frequência

$$\sigma_3(Z) = F^{-1}_Z(3/4) - F^{-1}_Z(1/4)$$

É fácil de comprovar que os parâmetros de escala $\sigma_i(Z)$ definidos anteriormente satisfazem as condições (3.4.1) e (3.4.2).

Vejamos agora a questão 2. Notemos em primeiro lugar que os $\sigma_i(Z)$ já definidos dependem não dos valores de Z mas da sua função distribuição, ponemos assim $\sigma_i(F_Z) =: \sigma_i(Z)$ e consideramos $\sigma_i(G)$ definido para toda função distribuição G por:

$$\sigma_1(G) = \int (y-E(G))^2 \, dG(y)$$

$$E(G) = \int y \, dG(y)$$

(3.4.4)

$$\sigma_2(G) = \text{MEDM}(|G_g|)$$

$$\sigma_3(G) = G^{-1}(3/4) - G^{-1}(1/4)$$

Sempre que existam os elementos envolvidos no membro direito.

Em (3.4.4) usamos a seguinte notação

(3.4.5) $g = \text{MEDM}(G) = G^{-1}(1/2)$

$$G_g(y) = G(y-g) \quad \forall \quad y \ .$$

$$|G_g|(y) = \begin{cases} 0 & \text{se} \quad y < 0. \\ G_g(y) = \lim\limits_{t \to o+} G_g(y-t) & \text{se} \quad y \geq 0 \end{cases}$$

É natural agora tomar como estimador de $\sigma_i(Y)$ em (3.4.3) a $s_i = \sigma_i(F_n)$ sendo F_n a distribuição empírica de $\{y_1,\ldots,y_n\}$. Se adotam então os seguintes estimadores de escala.

$$s_1 = s_1(y_1,\ldots,y_n) = \sqrt{1/n \sum_{i=1}^{n} (y_i-\bar{y})^2}$$

$$\bar{y} =: \frac{1}{n} \sum_{i=1}^{n} y_i$$

$$s_2 = s_2(y_1,\ldots,y_n) = \text{MED}\{|y_1-T_o|,\ldots,|y_n-T_o|\}$$

$$T_o = \text{MED}(y_1,\ldots,y_n).$$

(ver(3.3.2) a definição de MED)

$$s_3 = s_3(y_1,\ldots,y_n) = Q_3 - Q_1 \ .$$

Onde Q_1 = primeiro quartil amostral.

Q_3 = terceiro quartil amostral

Voltemos agora a questão de definir os M-estimadores de modo que resultem equivariantes sob mudanças de escala. Uma vez que escolhido o $S_n=s_i$ que julguemos conveniente, preferindo sempre s_2 como é aconselhado do ponto de vista da robustez, aplicamos o seguinte:

Definição 3.4.2. Seja $\Psi : R \to R$ uma função, chama-se M-estimador de μ definido por Ψ baseado em $y_1,\ldots,y_n$ com escala estimada por $S_n=s_i(y_1,\ldots,y_n)$ a um $T_n=T_n(y_1,\ldots,y_n)$ tal que:

$$(3.4.6) \qquad \sum_{j=1}^{n} \Psi\left(\frac{y_j-T_n}{S_n}\right) = 0$$

É, agora, muito fácil de ver que o estimador T_n é equivariante sob translações e mudança de escala. Existe também outro procedimento em uso na prática da estimação robusta para estimar a escala dos resíduos. Este procedimento foi sugerido por HUBER [25], e então passou a ser conhecido por "proposta 2 de HUBER", que consiste em estimar μ e a escala dos resíduos simultaneamente, resolvendo simultaneamente (3.4.6) e:

$$\sum_{j=1}^{n} \chi\left(\left(\frac{y_i-T_n}{S_n}\right)^2\right) = 0$$

onde $\chi:[0,+\infty) \to R$ é uma função a ser escolhida adequadamente.

Mais detalhes podem ser visto no trabalho de HUBER já citado e em ANDREWS e OUTROS [1].

CAPÍTULO IV

CÁLCULO DOS ESTIMADORES

4.1 - Cálculo dos L-estimadores

Como vimos basta aplicar diretamente a fórmula(3.3.1). Em forma de algoritmo:

L1 - Ordenar $\{y_1,\ldots,y_n\}$

L2 - Calcular os a_i

L3 - Fazer $L_n = \sum_{i=1}^{n} a_i\, y_i$ (os y_i's já estão ordenados de modo que $y_1 \leq \ldots \leq y_n$).

4.2 - Cálculo dos M-estimadores

Existem vários algoritmos para o cálculo destes estimadores. Alguns deles tem sido elaborados para certas Ψ particulares com o objetivo de ganhar uma maior eficiência tanto no tempo de computação como em exatidão e precisão do resultado. Em particular, para os M-estimadores com Ψ do tipo HUMBER e Ψ do tipo HAMPEL, bem como de muitos outros M-estimadores, o leitor interessado poderia consultar ANDREWS e OUTROS [1]. Em JAMES E BUSTOS [32] se descreve um algoritmo que poderia ser aplicado para resolver (3.4.6) qualquer que seja Ψ, obtendo pelo menos uma resposta, ainda que como já dissemos algoritmo especiais para cada Ψ são mais eficientes em cada caso.

Para não estendermos demasiadamente e ter ao menos um método razoável para calcular esses estimadores repassaremos bre

vemente esse algoritmo chamado IWLS (Iterated-Weighted-Least-Square).

Nosso problema então é: dado $y_1,\ldots,y_n$ e $\Psi:R\to R$ encontrar T_n tal que:

$$\sum_{i=1}^{n} \Psi\left(\frac{y_j-T_n}{S_n}\right) = 0 \tag{4.2.1}$$

onde $S_n = MED\{|y_i-T_o|,\ldots,|y_n-T_o|\}$

$T_o = MED\{y_1,\ldots,y_n\}$

A idéia do algoritmo IWLS é a seguinte: Suponhamos que T_n é uma solução de (4.2.1), e que $T_n \neq y_j \ \forall \ j=1,\ldots,n$ (não há perda de generalidade nesta suposição pois sempre trabalharemos com Ψ tal que $\Psi(o)=o$). Então:

$$\sum_{j=1}^{n} \frac{\Psi\left(\frac{y_j-T_n}{S_n}\right)}{\frac{y_j-T_n}{S_n}} \left(\frac{y_j-T_n}{S_n}\right) = 0$$

Para cada $j=1,\ldots,n$ seja:

$$W_j = \frac{\Psi\left(\frac{y_j-T_n}{S_n}\right)}{\frac{y_j-T_n}{S_n}} \tag{4.2.2}$$

Logo: $$\sum_{j=1}^{n} W_j \left(\frac{y_j-T_n}{S_n}\right) = 0$$

De onde resulta

$$T_n = \frac{\sum_{j=1}^{n} W_j\, y_j}{\sum_{j=1}^{n} W_j} \tag{4.2.3}$$

Reciprocamente, se T_n satisfaz (4.2.3) com W_j como em (4.2.2) então T_n é uma solução de (4.2.1). A fórmula (4.2.3) também nos permite obter T_n por meio de um processo iterativo (notemos que T_n aparece num dos membros de (4.2.3)). Este processo é o que constitue o algoritmo IWLS que em forma sintética o descreveremos assim:

M1 - Ordenar $y_1,\ldots,y_n$ de modo que $y_1 \leq \ldots \leq y_n$

M2 - Calcular $T_o = \text{MED}\{y_1,\ldots,y_n\}$

M3 - Calcular $S_n = \text{MED}\{|y_1-T_o|,\ldots,|y_n-T_o|\}$

M4 - Seja $m=o$; $T(o)=T_o$

M5 - Para cada $j=1,2,\ldots,n$ calcular

$$W_j = \frac{\Psi\left(\frac{y_j - T(m)}{S_n}\right)}{\frac{y_j - T(m)}{S_n}}$$

M6 - Definir

$$T(m+1) = \frac{\sum_{j=1}^{n} W_j\, y_j}{\sum_{j=1}^{n} W_j}$$

M7 - Se $|T(m+1) - T(m)| < \varepsilon\, |T(m+1)|$ com ε pré fixado (na maioria dos casos bastaria fazer $\varepsilon = 0.001$) fazer $T_n = T(m+1)$ e parar

M8 - Se $|T(m+1) - T(m)| \geq \varepsilon\, |T(m+1)|$ e $m+1 < ISTOP$ (número máximo de iterações permitidas, poderia ser ISTOP=20), fazer $m=m+1$ e voltar a M5.

M9 - Se $|T(m+1)-T(m)| \geq \varepsilon\, |T(m+1)|$ e $m+1 = ISTOP$, parar fazendo

notar que o algoritmo não convergiu, mas fazer $T_n=T(m+1)$ de qualquer maneira.

4.3 - Cálculo dos R-estimadores

Também aqui o algoritmo IWLS nos permite calcular um $\hat{\mu}_{RJ}$ solução de (3.3.17). Vejamos isto com detalhe. Nosso problema consiste em:

dados $y_1,\ldots,y_n$ e $J:(0,1)\to R$ encontrar T_n tal que:

$$S_n^*(T_n;y_1,\ldots,y_n) = 0$$

o que é o mesmo que:

$$\sum_{j=1}^{n} a_n(R_j^*) = 0 \qquad (4.3.1)$$

onde $a_n(k) = J(\frac{k}{2n+1})$ para $k=1,\ldots,2n$ e R_j^* é o posto de y_j em $\{y_1,\ldots,y_n,\ 2T_n-y_1,\ldots,2T_n-y_n\}$.

Suponhamos que T_n é uma solução de (4.3.1) e que $T_n \neq y_i$ $\forall 1\leq j\leq n$. De (4.3.1) obteremos:

$$\sum_{j=1}^{n} W_j^*(y_j-T_n) = 0 \qquad (4.3.2)$$

sendo

$$W_j^* = \frac{a_n(R_j^*)}{y_j - T_n} \qquad j=1,\ldots,n$$

De (4.3.2) teremos:

$$T_n = \frac{\sum_{j=1}^{n} W_j^* y_j}{\sum_{j=1}^{n} W_j^*} \qquad (4.3.3)$$

Somo já vimos para os M-estimadores, se T_n satisfaz (4.3.3) e $T_n \neq y_j$ $\forall$ $j=1,\ldots,n$; então T_n é uma solução de

(4.3.1). Também o algoritmo de cálculo pode ser descrito assim:

R1 - Ordenar $y_1,\ldots,y_n$ de modo que $y_1 \leq \ldots \leq y_n$

R2 - Calcular $T_o = \text{MED}\{y_1,\ldots,y_n\}$

R3 - $m=0$, $T(o) = T_o$

R4 - Seja $Z_j = y_j$

$Z_{n+j} = 2T(m) - y_j$ $\Big\}$ p/ $j=1,\ldots,n$

R5 - Ordenar $Z_1,\ldots,Z_{2n}$ de modo que $Z_1 \leq \ldots \leq Z_{2n}$

R6 - Para cada $j=1,\ldots,n$ calcular

$R_j^* =$ posto de y_j em $\{Z_1,\ldots,Z_{2n}\}$

R7 - Para cada $j=1,\ldots,n$ calcular

$$W_j^* = \frac{a_n(R_j^*)}{y_j - T(m)}$$

R8 - Definir $T(m+1) = \dfrac{\sum_{j=1}^{n} W_j^* y_j}{\sum_{j=1}^{n} W_j^*}$

R9 - Se $|T(m+1) - T(m)| < \varepsilon\,|T(m+1)|$ com $\varepsilon>0$ pré-fixado fazer $T_n = T(m+1)$ e parar

R10- Se $|T(m+1)-T(m)| \geq \varepsilon\,|T(m+1)|$ e $m+1 < ISTOP$ fazer $m=m+1$ e voltar para R4

R11- Se $|T(m+1)-T(m)| \geq \varepsilon\,|T(m+1)|$ e $m+1=ISTOP$, parar fazendo notar que o algoritmo não convergiu e tomar $T_n=T(m+1)$.

É necessário advertir que esses algoritmos são um tanto ingênuos e que para implement-alos eficientemente será necessário atender a certos detalhes. O leitor que está interessado

poderia consultar DUTTER [12] e também MARTIN [39] para uma exposição mais simples.

CAPÍTULO V

MEDIDAS DE SENSIBILIDADE DOS ESTIMADORES PARA AMOSTRAS FINITAS

Neste capítulo nos propomos estudar o seguinte:

Sejam $y_1,\ldots,y_n$ as observações, comose afetam os diversos estimadores de μ já definidos, quando j dessas n observações são "ruins" (não se ajustam ao modelo) e as restantes são "boas" ?

5.1 - Definição de "amostra típica"

Como queremos extrair conclusões de validade mais geral possível vamos trabalhar com uma "amostra típica" de F, isto é, com uma amostra tal que se tomamos um grande número de vezes amostras de tamanho n de F então a "amostra típica" ou "aproximações" aparecem várias vezes. Precisamente:

Definição 5.1.1: Seja F uma distribuição. Diremos que $\chi_1,\ldots,\chi_n$ é uma amostra típica de tamanho m de F se:

$$(5.1.1) \qquad \chi_1 = E\, X_{(i)} \qquad 1\leq i\leq m \ ,$$

onde $X_{(1)},\ldots,X_{(m)}$ são as estatísticas de ordem de uma amostra $X_1,\ldots,X_m$ de tamanho m de F.

Pode-se ver, por exemplo em BICKEL E DOKSUM [4] que (5.1.1) é o mesmo que:

$$\chi_j = \frac{m!}{(j-1)!(m-j)!} \int_{-\infty}^{+\infty} \chi\, F(\chi)^{j-1}(1-F(\chi))^{m-j}\, dF(\chi)$$

ou se F está definida por uma densidade f:

$$\chi_j = \frac{m!}{(j-1)!(m-j)!} \int_{-\infty}^{+\infty} \chi\, f(\chi)F(\chi)^{j-1}(1-F(\chi))^{m-j}\, d\chi$$

Os valores de certos χ_j para determinados F tem sido tabelados em diversos trabalhos. Quando $F = \Phi$, os extraíremos do excelente trabalho de TIETJEN E OUTROS [48] onde estão tabulados os χ_j correspondentes a n, para n=2 até n=50 com uma precisão de 10 casas decimais.

Exemplo 5.1.1: Sejam: n=15, $F = \Phi$. Então

χ_1	= -1.74	χ_9	= 0.16
χ_2	= -1.25	χ_{10}	= 0.34
χ_3	= -0.95	χ_{11}	= 0.52
χ_4	= -0.71	χ_{12}	= 0.71
χ_5	= -0.52	χ_{13}	= 0.95
χ_6	= -0.34	χ_{14}	= 1.25
χ_7	= -0.16	χ_{15}	= 1.74

$\chi_8 = 0$

5.2 - Curva de Sensibilidade

Suponhamos agora que temos efetuado n-1 observações. Para ver como uma nova observação "y" influi no estimador T_n parece bastante natural analizar a variação de

$$T_n(y_1,\ldots,y_{n-1},y) - T_n(y_1,\ldots,y_{n-1})$$

de onde se tira a importância do conceito de curva de sensibilidade, cuja definição formal é:

Definição 5.2.1: Sejam T_{n-1} e T_n estimadores de μ baseados

em amostras de tamanho $n-1$ e n respectivamente (ambos são estimadores definidos pela mesma regra; com a única diferença relativa aos tamanhos amostrais); sejam $y_1,\ldots,y_{n-1}$ números reais. Chama-se curva de sensibilidade de T_n com respeito a $T_{n-1}(y_1,\ldots,y_{n-1})$ à função $SC_{n-1}:R \to R$ definida por:

$$(5.2.1)\quad SC_{n-1}(y) = n(T_n(y_1,\ldots,y_{n-1},y) - T_{n-1}(y_1,\ldots,y_{n-1})).$$

Como todos os estimadores de μ que estudaremos são invariantes com respeito a translação, não há perda de generalidade em supor que o parâmetro μ a estimar (ver (3.1.3)) é $\mu=0$. Recordemos também que sempre supomos que a distribuição "hipotética" das observações é Φ e que a "verdadeira" é $F_\mu=F$.

A partir de (5.2.1) pode-se ver que os estimadores "robustos" no sentido de que sejam resistentes à influência de uma única observação separada da massa de dados, são os que têm SC_{n-1} limitada.

Vejamos agora o aspecto de SC_n para alguns dos estimadores definidos no Capítulo III. Uma lista mais completa pode ser vista em ANDREWS E OUTROS [1]. Sejam $n=16$, $y_i=x_i$ $1\le i\le 15$ onde os x_i são como no exemplo 5.1.1.

1) Curva de Sensibilidade para a média amostral

Uma aplicação direta de (5.2.1) nos leva a:

$$(5.2.2)\qquad SC_{15}(y) = y$$

2) Curva de sensibilidade para a mediana amostral

$$(5.2.3)\quad SC_{15}(y) = \begin{cases} -1.28 & \text{se } y \leq -0.16 \\ 8y & \text{se } -0.16 \leq y \leq 0.16 \\ 1.28 & \text{se } y \geq 0.16 \end{cases}$$

3) Curva de sensibilidade para a média 0.1 truncada

$$(5.2.4)\quad SC_{15}(y) = \begin{cases} (16/14)(-1.74) & \text{se } y \leq -1.74 \\ (16/14)y & \text{se } -1.74 \leq 1.74 \\ (16/14)(1.74) & \text{se } y \geq 1.74 \end{cases}$$

A seguinte figura nos mostra os gráficos das curvas (5.2.2),(5.2.3) e (5.2.4). É possível ver que os M-estimadores têm curvas de sensibilidade com forma similar a da função Ψ (somente diferem por um fator).

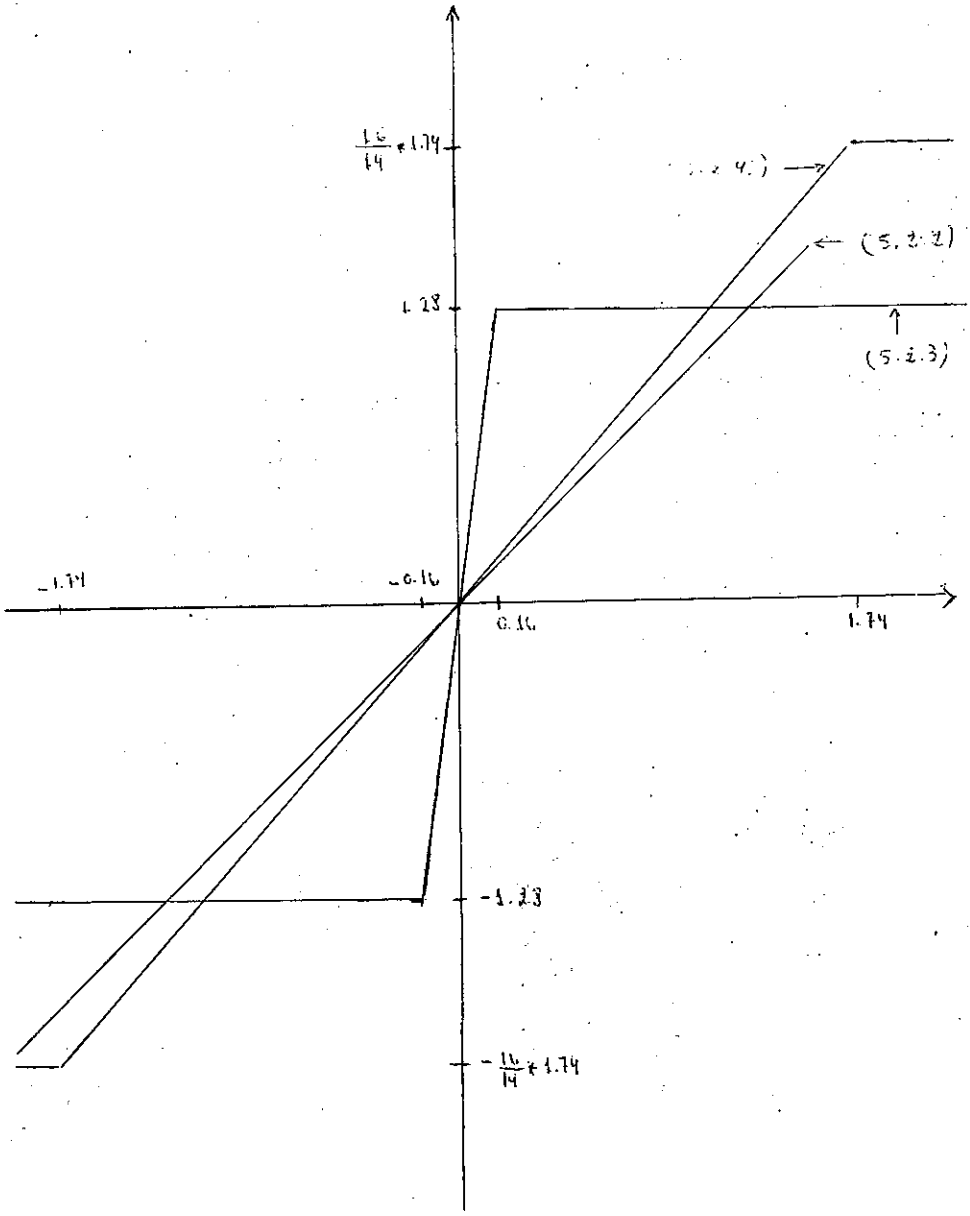

Figura 7

5.3 - Ponto de ruptura não assintótico

Com este conceito formalizaremos a noção de proporção máxima de pontos "ruins" que se pode admitir numa amostra a partir da qual o estimador T_n não nos dá qualquer informação útil sobre o parâmetro a estimar.

O conceito que veremos aqui é o definido em ANDREWS E OUTROS [1]. É bastante limitado desde que só é aplicável ao problema de estimação de μ no modelo de posição e se considera somente "outliers selvagens" (pontos muitos afastados da massa de dados). Não obstante, nos dá uma idéia de como poderia se proceder em outros problemas e serve também como critério de comparação do rendimento de diversos estimadores.

Antes de tudo, devemos decidir quando um estimador T_n de μ-baseado numa amostra $y_1,\ldots,y_n$ não nos dá informação relevante sobre μ: Uma decisão natural é a de considerar o afastamento de T_n de μ, supondo μ conhecido. Só para fixar as idéias, diremos que $T_n = T_n(y_1,\ldots,y_n)$ não dá nenhuma informação relevante sobre μ se:

$$|T_n(y_1,\ldots,y_n) - \mu| \geq 3 \tag{5.3.1}$$

De (5.3.1) e pelo fato de que a totalidade estimadores que estamos estudando são equivariantes sob translações podemos supor que o parâmetro verdadeiro é $\mu=0$.

Assim como na definição 5.1.1 formalizamos o conceito de "amostra típica" de uma distribuição, necessitamos agora co-

locarmos de acordo sobre o que vamos entender por "amostra típica de tamanho n com j observações ruins". Tendo em conta o que dissemos antes da definição 5.1.1, também de que só vamos nos interessar aqui por "outliers selvagens" e o que se faz em ANDREWS E OUTROS [1] chegamos a:

Definição 5.3.1: Seja $0 \leq j \leq n$ inteiro. Diremos que $y_1, \ldots, y_n$ é amostra típica de tamanho n com j observações ruins se: $y_1, \ldots, y_{n-j}$ é amostra típica de tamanho n-j de Φ (ver Definição 5.1.1) e $y_{n-j+1} = 100, \ldots, y_n = 100\ j$.

Agora sim estamos em condições de definir "ponto de ruptura não assintótico de um estimador".

Definição 5.3.2: Seja T_n um estimador de μ ($\mu=0$). Chamaremos ponto de ruptura não assintótico de T_n a:

$$\sigma_n^*(T_n) = \frac{j_o}{n}\ 100\ ,$$

onde $j_o = \max\{0 \leq j \leq n : |T_n(y_1, \ldots, y_n)| \leq 3$, sendo $y_1, \ldots, y_n$ a amostra típica de tamanho n com j observações ruins$\}$

Obviamente, do ponto de vista da robustez preferiremos trabalhar com um estimador que tenha o mais alto ponto de ruptura possível.

A seguinte tabela, extraída de JAMES E BUSTOS[32], nos mostra os pontos de ruptura de vários estimadores para n=5, 20 e 40. Uma vez mais destacaremos a falta de "robustez" da média amostral em contraste com o excelente rendimento dos M-estimadores.

Estimador \ n =	5	20	40
Média amostral	0.	0.	2.5
Mediana amostral	40.	45.	47.5
Média 0.1-truncada	0.	10.	10.
M-estimador com Ψ de Huber e k=1.345	20.	30.	32.5
M-estimador com Ψ de Hampel e A=2.5,B=4.5,C=9.5	40.	45.	47.5
M-estimador com Ψ "bicuadrada" e k=4.685	40.	45.	47.5
R-estimador de Hodges-Lehmann	20.	25.	27.5

Tabela de δ_n^*

CAPÍTULO VI

ESTIMADORES DEFINIDOS POR FUNCIONAIS

6.1 - Definição de estimadores através de funcionais

HAMPEL [18] introduziu uma forma de definir os estimadores de μ no modelo de posição que mostrou ser bastante frutífera pois permitiu formalizar um aspecto importante da robustez do qual fizemos referência na seção 1.2 (robustez qualitativa). Também facilitou o estudo da teoria assintótica dos estimadores ligando trabalhos teóricos na área de Análise funcional com a robustez.

Para darmos conta da idéia de HAMPEL [18] necessitamos antes de mais nada ver como é possível identificar um conjunto de observações $y_1,\ldots,y_n$ com uma distribuição ou probabilidade definida sobre R. Naturalmente, podemos fazer isto através do conceito de "distribuição empírica".

Definição 6.1.1: Sejam $y_1,\ldots,y_n$ números reais. Denominamos distribuição empírica de $y_1,\ldots,y_n$ a probabilidade $\mu[y_1,\ldots,y_n]$ sobre R definida por

$$\mu[y_1,\ldots,y_n](B) = \frac{1}{n}\sum_{i=1}^{n} I_B(y_i)$$

onde B é um subconjunto boreliano de R e I_B é a função indicadora de B (em outras palavras $\mu[y_1,\ldots,y_n](B)$ é a proporção de pontos do conjunto $\{y_1,\ldots,y_n\}$ que está em B).

É de interesse destacar o gráfico de $F_n = F_{\mu[y_1,\ldots,y_n]}$ função de distribuição da probabilidade $\mu[y_1,\ldots,y_n]$

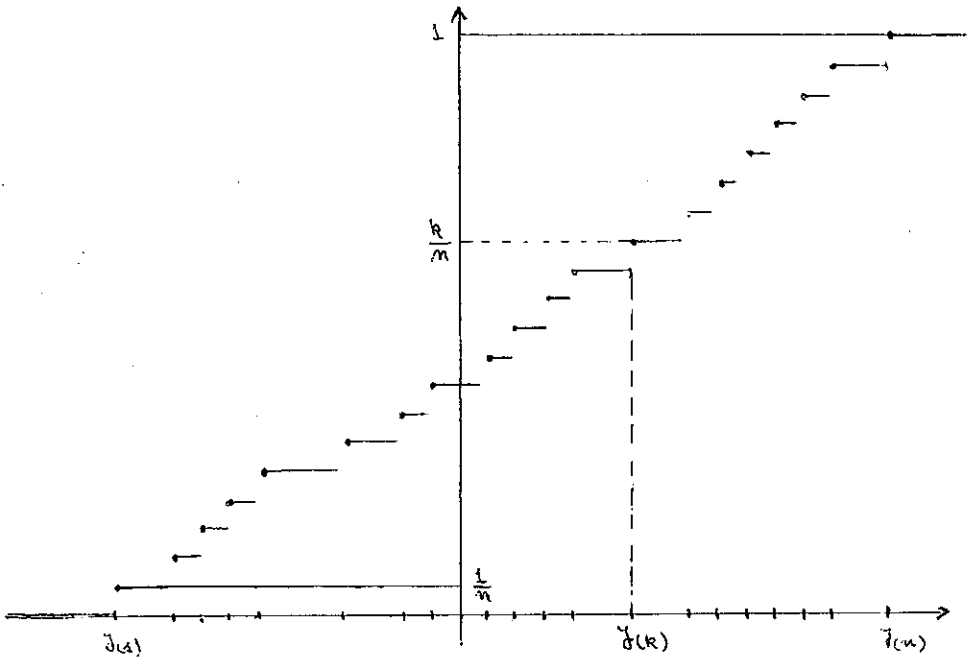

Figura 8

Seja agora $Z(R)$ o conjunto de todas as probabilidades sobre R. Para cada n seja $\mathfrak{F}_n$ o conjunto de todas as distribuições empíricas associadas com amostras de tamanho n(logo $\mathfrak{F}_n \subset Z(R)$)

$$\mathfrak{F}_n = \{\mu[y_1,\ldots,y_n] \ : \ y_1,\ldots,y_n \text{ sejam números reais}\}$$

Notemos agora que todos os estimadores de μ estudados tem a propriedade de não alterar seu valor se permutamos os índi-

ces de $\{y_1,\ldots,y_n\}$, isto é, todos os estimadores T_n satisfazem

$$T_n(y_1,\ldots,y_n) = T_n(y_{\pi(1)},\ldots,y_{\pi(n)})$$

qualquer que seja a permutação π definida sobre $\{1,2,\ldots,n\}$. Assim vemos quese valor depende não tanto dos pontos $y_1,\ldots,y_n$ mas exatamente do conjunto $\{y_1,\ldots y_n\}$ o mesmo se sucede com a distribuição empírica de $\{y_1,\ldots,y_n\}$; é imediato comprovar que:

$$\mu[y_1,\ldots,y_n] = \mu[y_{\pi(1)},\ldots,y_{\pi(n)}]$$

Desta maneira pensa-se naturalmente em definir os estimadores não por meio de funções definidas sobre R^n e sim por funções definidas sobre subconjuntos de $Z(R)$ (funcionais).

Definição 6.1.2. Diremos que um estimador T_n está definido por uma funcional T definida em $Z(R)$ (em símbolos: $T_n=T|_{\mathfrak{Z}_n}$) se existe uma função T definida em um subconjunto DT de $Z(R)$ tal que:

$$(6.1.1)\quad T_n(y_1,\ldots,y_n) = T(\mu[y_1,\ldots,y_n])$$

sempre que os membros de (6.1.1) tenham sentido (isto é: $(y_1,\ldots,y_n)$ está no domínio de T_n e $\mu[y_1,\ldots,y_n] \in DT$).

Firmemos esta definição vendo como algum estimadores vistos no capítulo III podem ser definidos por funcionais após serem feitas ligeiras modificações.

Média amostral definida por funcional

Seja $DT = \{G:G$ é uma distribuição sobre R com momento de primeira ordem finito $(\int |x| dg(x) < \infty)\}$; para cada

$G \in DT$ seja

(6.1.2) $\quad T(G) = \int x\, dG(x) = E_G(X)$

Então é fácil ver que

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i = T(\mu[y_1,\ldots,y_n])$$

Mediana amostral definida por funcional

Em (3.3.2) definirmos $MED(y_1,\ldots,y_n)$ = mediana de $y_1,\ldots,y_n$ em (3.4.5) definimos $MEDM(G)$ = mediana da distribuição G. Pode-se comprovar facilmente que se n é ímpar:

(6.1.3) $\quad MED(y_1,\ldots,y_n) = MEDM(F_{\mu[y_1,\ldots,y_n]})$

Se n é par (6.1.3) não está correta, porém as vantagens de se ter os estimadores definidos por funcionais faz que se veja conveniente modificar a definição da mediana amostral adotando a (6.1.3) qualquer que seja n. Em termos precisos: Chamaremos de mediana amostral (modificada) o estimador MEDM definido sobre R^n por:

$$MEDM(y_1,\ldots,y_n) = MEDM(F_{\mu[y_1,\ldots,y_n]})$$

Média α-truncada definida por funcional

Seja $0<\alpha<1/2$. Seja α-TM a função definida sobre $Z(R)$ por

(6.1.4) $\quad \alpha\text{-TM}(G) = \frac{1}{1-2\alpha} \int_{\alpha}^{1-\alpha} G^{-1}(t)\, dt$

É fácil ver que se $n\alpha=[n\alpha]$ ($n\alpha$ é inteiro) então

(6.1.5) $\quad \alpha_T(y_1,\ldots,y_n) = \alpha\text{-TM}\ (F_{\mu[y_1,\ldots,y_n]})$

(ver (3.3.3) para a definição de α_T)

A igualdade (6.1.5) nos conduz a seguinte definição:

Chamaremos de média α-truncada (modificada) o estimador α-TM definido sobre R^n por

$$\alpha\text{-TM}(y_1,\ldots,y_n) = \alpha\text{-TM}(F_{\mu[y_1,\ldots,y_n]})$$

Uma aplicação direta de (6.1.4) conduz a

$$\alpha\text{-TM}(y_1,\ldots,y_n) = \frac{1}{n(1-2\alpha)} \sum_{k=[n\alpha]+2}^{n-[n\alpha]-1} y_{(k)} + \frac{1-n\alpha+[n\alpha]}{n(1-2\alpha)} \left(y_{([n\alpha]+1)} + y_{(n-[n\alpha])}\right)$$

de onde, além da igualdade (6.1.5) para $n\alpha=[n\alpha]$, pode-se ver que α_T e α-TM diferem cada vez menos para n grande. Em verdade, o comportamento assintótico dos estimadores α_T e α-TM é o mesmo; o mesmo se sucede entre MED e MEDM

M-estimador definido por funcional

Seja $\Psi:R\to R$ uma função seja T_Ψ uma função definida em $Z(R)$ tal que:

$$(6.1.6)\quad E_G(\Psi(X-T_\Psi(G))) = \int \Psi(y-T_\Psi(G))\, dG(y) = 0$$

(para G em um subconjunto de $Z(R)$ onde tenha sentido a integral em (6.1.6) e a equação tenha solução). É imediato ver que se T_n está definida por T_Ψ então T_n satisfaz (3.3.8). Daí pode-se considerar que um M-estimador definido por Ψ vem dado por uma funcional como a T_Ψ. Se estamos considerando a escala devemos tomar $T_{\Psi,\sigma}$ definido de maneira tal que:

(6.1.7) $$E_G\Psi\left(\frac{X-T_{\Psi,\sigma}(G)}{\sigma(X)}\right) = \int \Psi\left(\frac{y-T_{\Psi,\sigma}(G)}{\sigma(G)}\right) dG(y) = 0$$

onde $\sigma(G) = \sigma(X)$ é um parâmetro de escala de G. Assim então pode-se ver que (T_n, S_n) definidos em (3.4.6) e $(T_{\Psi,\sigma_i}(\mu[y_1,\ldots,y_n]), \sigma_i(\mu[y_1,\ldots,y_n]))$ tendem a um comportamento muito similar, com a igualdade para $n \to \infty$.

R-estimador definido por funcional

Sejam $Z_1,\ldots,Z_N$ N números reais ($Z_i \neq Z_j$ se $i \neq j$) Um cálculo simples prova que

(6.1.8) Posto de Z_i em $\{Z_1,\ldots,Z_N\} = NF_{\mu[Z_1,\ldots,Z_N]}(Z_i)$

Seja agora S_n^* como em (3.3.17) tendo em conta a igualdade (6.1.8) é fácil provar que

(6.1.9) $$S_n^*(m;x_1,\ldots,x_n) = \frac{1}{n}\sum_{i=1}^{n} a_n(R_i^*) =$$

$$= \frac{1}{n}\sum_{i=1}^{n} J\left(\frac{nF_n(y_i)+1-F_n(2m-y_i-0))}{2n+1}\right) =$$

$$= \frac{1}{n}\sum_{i=1}^{n} J\left(\frac{F_n(y_i)+1-F_n(2m-y_i-0)}{2+\frac{1}{n}}\right) \cong$$

$$\cong \int J\left(\frac{F_n(y)+1-F_n(2m-y)}{2}\right) dF_n(y)$$

onde $F_n = F_{\mu[y_1,\ldots,y_n]}$; $\cong$ significa aproximadamente igual e usamos a notação $h(x-0) = \lim_{t\to 0^+} h(x-t)$ quaisquer que sejam h (função) e x (ponto do interior do domínio de h).

O raciocínio utilizado em (6.1.9) nos leva a tomar

como funcional, em $Z(R)$, definidora de $\hat{\mu}_{R,J}$ (ver (3.3.16)) a T_J dada por

$$\int J\left(\frac{G(y)+1-G(2T_J(G)-y)}{2}\right) dG(y) = 0 \qquad (6.1.10)$$

6.2 - Função de influência dos estimadores definidos por funcionais - "GES" ou sensibilidade a erros grosseiros

Quando os estimadores vem dados por funcionais definidas sobre subconjuntos de $Z(R)$, o conceito de curva de sensibilidade (Definição 5.2.1) pode ser estudado de uma maneira global.

Este conceito foi definido por HAMPEL [18] (também aparece em HAMPEL [19] e [21]) denominado função de influência.

Definição 6.2.1. Seja T uma função definida em um subconjunto DT de $Z(R)$; $G \in DT$ tal que

$$f(t) = f(t;T,G,y) = T((1-t)G+t\,\mu[y])$$

Ela está definida para cada $y \in R$ e cada t em um intervalo da forma $[o,\delta) \subset [0,1)$ Chama-se função de influência de T em G à função definida em $y \in R$ por

$$IC(y) = IC(y;T,G) = \lim_{t\to o^+} \frac{f(t) - f(0)}{t} \qquad (6.2.1)$$

$$= \lim_{t\to o} \frac{T((1-t)G + t\mu[y]) - T(G)}{t}$$

Vemos assim que este número nos dá uma idéia da velocidade com que troca o valor de T quando o modelo G é contaminado por uma distribuição com massa 1 em "y". Do ponto de vista da robustez preferiremos então usar aqueles estimadores definidos

por T com IC mais próximos possíveis a zero para todo ponto y. A conexão entre curva de sensibilidade e função de influência pode ser vista por exemplo em HUBER [30] ou JAMES E BUSTOS [32]. Lá

(6.2.2) $IC(y;T,\mu[y_1,\ldots,y_{n-1},y]) \cong \frac{n-1}{n} SC_{n-1}(y)$

onde SC_{n-1} é como em (5.2.1) com T_n definido pela funcional T. De (6.2.2) deduzimos então que muitas propriedades de SC_{n-1} (cotação por exemplo) podem ser estudadas através de IC e vice-versa.

O cálculo de IC não é simples na maioria dos casos. Tudo consiste em manejar as "coisas" habilmente a fim de calcular a derivada lateral à direita no ponto zero da função f, tal como se pede em (6.2.1).

Requere-se também certas propriedades de regularidade tanto da distribuição G (simetria, continuidade, etc.) como das funções Ψ e J para todos os estimadores M e R. Em relação a distribuição, como dissemos no início da seção 3.3 vai nos interessar principalmente o caso $G = \Phi$. Com relação as condições de regularidade requeridas para Ψ elas são satisfeitas pelas Ψ's do tipo HUBER, HAMPEL, "seno" e "biquadrada" (recordar as fórmulas (3.3.9) a (3.3.12)). Analogamente pelas J que definem os estimadores R baseados em "escores normais", "escores mediana" e "escores de WILCOXON".

De nossa parte limitaremos a sintetizar toda a informação que fizemos referência na seguinte lista de funções de influência de diversos estimadores em $G = \Phi$:

Funções de influência em Φ de:

Média amostral

(6.2.3) $\quad IC(y) = y$

Mediana amostral

(6.2.4) $IC(y) = \dfrac{\text{sinal}(y)}{2\,\varphi(o)}$

Média α-truncada (α-TM)

$$(6.2.5)\quad IC(y) = \begin{cases} \Phi^{-1}(\alpha)/(1-2\alpha) & \text{se } y<\Phi^{-1}(\alpha) \\ y/(1-2\alpha) & \text{se } \Phi^{-1}(\alpha)\leq y\leq \Phi^{-1}(1-\alpha) \\ \Phi^{-1}(1-\alpha)/(1-2\alpha) & \text{se } y> \Phi^{-1}(1-\alpha) \end{cases}$$

M-estimador com $\Psi = \Psi_{H,k}$ e escala conhecida (ver (3.3.9))

$$(6.2.6)\quad IC(y) = \frac{\Psi_{H,k}(y)}{2\Phi(k)-1}$$

M-estimador com $\Psi=\Psi_{HA,A,B,C}$ e escala conhecida (ver (3.3.10))

$$(6.2.7)\quad IC(y) = \frac{\Psi_{HA,A,B,C}(y)}{2\Phi(A)-1\ -(\frac{2A}{C-B})\,(\Phi(C)-\Phi(B))}$$

M-estimador com $\Psi=\Psi_{B,k}$ e escala conhecida (ver (3.3.12))

$$(6.2.8)\quad IC(y) = \frac{\Psi_{B,k}(y)}{\sqrt{2k^3}\ \int_0^1 v^2(1-v^2)^2\,\varphi(kv)dv}$$

R-estimador baseado em "escores normais"

(6.2.9) $\quad IC(y) = y$

R-estimador baseado em "escores de WILCOXON" (o estimador de HODGES-LEHMANN).

$$(6.2.10)\quad IC(y) = \frac{\sqrt{2}}{\varphi(o)}\,(\Phi(y) - 1/2)$$

Obviamente a influência máxima que pode ter um "outlier" sobre um estimador será dada pelo supremo da função de influência, valor que HAMPEL [18] definiu como GES (sensibilidade aos erros grosseiros).

Em fórmulas:

$$GES(T,G) = \sup_{y\in R} |IC(y;T,G)|$$

De onde deduz-se a seguinte tabela

Estimador	GES(·,Φ)
Média amostral	$+\infty$
Mediana amostral (MEDM)	$1/(2\varphi(0)) \cong 1.253$
Média α-truncada (α-TM)	$\Phi^{-1}(1-\alpha)/(1-2\alpha)$
M-estimador com (esc.conhecida) $\Psi=\Psi_{H,k}$	$k/(2\,\Phi(k)-1)$
M-estimador com (esc.conhecida) $\Psi=\Psi_{HA,A,B,C}$	$A/(2\Phi(A)-1-2(A/(C-B))(\Phi(C)-\Phi(B)))$
M-estimador com (esc.conhecida) $\Psi=\Psi_{B,k}$	$8/(25\sqrt{5}k^2 \int_0^1 v^2(1-v^2)^2\,\varphi(kv)dv)$
R-estimador baseado em "escores normais"	$+\infty$
R-estimador baseado em "escores de Wilcoxon"	$1/\sqrt{2\,\varphi(o)} \cong 1.120$

Finalmente, na figura 9 podemos ver os gráficos das curvas de influência (6.2.3) a (6.2.10).

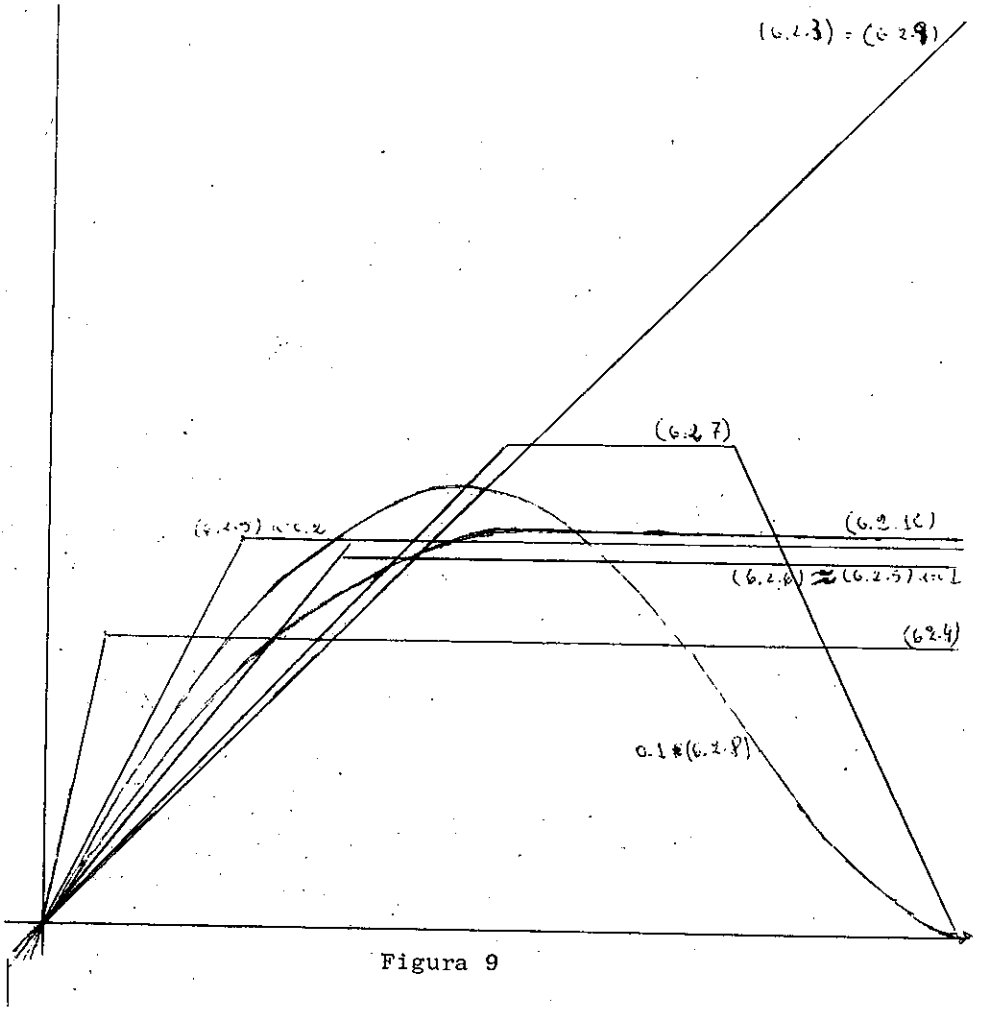

Figura 9

6.3 - Comportamento assintótico de estimadores definidos por funcionais: consistência e normalidade assintóticas

Podemos encontrar a teoria assintótica dos estimadores robustos em diversas publicações tais como: HUBER [25], [26] ANDREWS E OUTROS [1] JAECKEL [31], AZENCOT E OUTROS[2], KLEIN E YOHAI[34] MARTIN [39], etc. Em geral é simples em suas idéias essenciais, porém muito complicada e técnica na formalização matemática de tais idéias como também é no que se refere a detalhes sobre a validez da aplicação dessas idéias a diversos casos.

De nossa parte faremos uma breve resenha dessas idéias e exporemos muito sucintamente os detalhes referentes ao caso de nosso interesse (exposto no princípio da seção 3.3).

Suponhamos que (T_n) é uma sequência de estimadores cada um deles definido por uma mesma funcional T (definição 6.1.2)

Por outro lado, na Teoria de Probabilidade prova-se que (ver, por exemplo BILLINGSLEY [3], PARTHASARATHY [41]):

"Seja P uma probabilidade sobre um espaço amostral $\Omega; Y_1, Y_2, \ldots$ uma sequência iid definida sobre Ω com distribuição G. Então:

$$\mu[Y_1, \ldots, Y_n] \to G \quad (D), n \to \infty \text{ "}$$

Logo se a função T definida em Z(R) é contínua em relação a convergência em distribuição para a distribuição G, então teremos:

(6.3.1) $T_n = T(\mu[Y_1,\ldots,Y_n]) \to T(G)$ (D)

Diz-se (T_n) é fracamente consistente para estimar $T(G)$ (recordar Definição 2.3.1).

A continuidade de T em muitos casos ou não é certa (caso $T(G) = E_G(X)$, (fórmula (6.1.2))) ou é difícil de provar diretamente. Por tais razões, em geral as técnicas para provar consistência são desenhadas especialmente para cada caso ou para uma determinada variedade de casos.

Também é possível ver que tipicamente os estimadores de posição definidos por funcionais, além de (6.3.1) para G com certas propriedades de regularidade, satisfazem

$$\sqrt{n}\,(T_n - T(G)) \to N(0, VA(T,G))$$

e que frequentemente (ver por exemplo PROHOROV [42], VON MISES [51])

(6.3.2) $$VA(T,G) = \int IC^2(y;T,G)\, dG(y)$$

Porém, como já dissemos ao tratar da consistência, demonstrações que usem diretamente (6.3.2) não são viáveis salvo segundo condições demasiadamente restritivas. De todas as maneiras, em geral pode-se dizer que se (6.3.2) tem sentido, então pode-se provar sua validez. A fim de expor com maior precisão os detalhes referentes aos casos de nosso interesse consideremos o seguinte modelo.

Seja $\mathfrak{F}$ o conjunto de funções de densidades definidas

sobre R formado por:

(6.3.3) $\mathfrak{F} = \{\varphi, ST(\cdot;m), DE, CN(\cdot;\varepsilon,\tau), L{:}\, m\geq 3,\ 0<\varepsilon<1/2,\ \tau>1\}$

Seja Ω o espaço amostral, $\mathcal{P}$ uma família de probabilidades sobre Ω; $Y_1, Y_2, \ldots$ uma sequência de variáveis aleatórias tal que para cada $P \in \mathcal{P}$ $Y_1, Y_2, \ldots$ é iid com função de distribuição comum $F = F_p$ dada por uma das densidades de $\mathfrak{F}$ (isto é: para cada $P \in \mathcal{P}$ existe uma única $f_p \in \mathfrak{F}$ tal que:

$$P(Y_i \leq y) = F(y) = \int_{-\infty}^{y} f_p(t)dt \qquad \forall\, y, \forall i \ )$$

As Y_i's representam as observações. As vezes diremos que estamos segundo o modelo "puro" ou gaussiano se $f_p = \varphi$.

Como sempre trabalharemos com estimadores equivariantes segundo mudança de escala (no caso de supor a escala desconhecida) e equivariantes segundo translações (sempre), não há nenhuma restrição no caso de nosso interesse, se nos limitamos a considerar o modelo recém definido.

Comportamento assintótico da Média amostral

Usando a lei dos grandes números ve-se que

$$\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i \to E_p(Y_1) = \int y\, f_p(y)dy \qquad \forall\, P \in \mathcal{P}$$

Usando o teorema central do limite prova-se que

(6.3.4) $VA(\text{Média amostral}, f_o) = Var_p(Y_1) = \int y^2 f_p(y)dy \ \forall\ P \in \mathcal{P}$

Comportamento assintótico da Mediana amostral (MEDM)

Em HUBER [30] prova-se que o funcional MEDM definido

em (3.4.5) é contínua em toda G tal que G seja contínua e estritamente crescente em $G^{-1}(1/2)$ logo:

$$MEDM(Y_1,\ldots,Y_n) \to MEDM(F_p) = 0 \quad (P) \quad \forall\ P \in \mathcal{P}$$

Em ANDREWS E OUTROS [1] calcula-se a função de influência de MEDM e aplicando (6.3.4) chega-se a

$$(6.3.5)\quad VA(MEDM, f_p) = \frac{1}{4f_p^2(MEDM(F_p))} = \frac{1}{4f_p^2(o)} \quad \forall\ P \in \mathcal{P}$$

Comportamento assintótico da Média α-truncada(α-TM)

Em HUBER [30] estão provadas as propriedades assintóticas dos L-estimadores em geral e com muito detalhe em AZENCOT E OUTROS [2]. Dalí pode-se deduzir que α-TM é contínua em toda G tal que G é contínua em $G^{-1}(\alpha)$ e $G^{-1}(1-\alpha)$. No nosso caso tem-se

$$\alpha\text{-TM}(Y_1,\ldots,Y_n) \to \alpha\text{-TM}(f_p) = o \ (P),\ \forall\ P \in \mathcal{P}$$

Procedendo analogamente que para MEDM teremos (ver ANDREWS E OUTROS [1]).

$$(6.3.6)\quad VA(\alpha\text{-TM}, f_p) = \frac{1}{(1-2\alpha)^2}\left\{\alpha(F_p^{-1}(\alpha))^2 + \alpha(F_p^{-1}(1-\alpha))^2 + \int_{F_p^{-1}(\alpha)}^{F_p^{-1}(1-\alpha)} y^2 f_p(y)dy\right\}$$

Comportamento assintótico de M-estimadores

Em HUBER [30] e com maior detalhe em HUBER [26] pro-

va-se que se Ψ é cotada e não decrescente e, mais ainda, dada a distribuição G a função $t \to \int \Psi(y-t)\, dG(y)$ tem um único zero em $t = T_\Psi(G)$ então a funcional T_Ψ defininida em (6.1.6) é contínua em G, logo.

$$(6.3.7)\quad T_\Psi(Y_1,\ldots,Y_n) \to T_\Psi(f_p) = o \quad \forall\ P \in \mathcal{P}$$

Também HUBER [30] mediante um raciocínio heurístico que se pode fazer rigoroso com propriedades adequadas de regularidade sobre Ψ calcula

$$(6.3.8)\quad IC(y;T_\Psi,G) = \frac{\Psi(y-T_\Psi(G))}{-\int \Psi'(y-T_\Psi(G))dG(y)}$$

Usando (6.3.7), (6.3.8) e inspirando-se em (6.3.2) HUBER prova que

$$(6.3.9)\quad VA(T_\Psi,f_p) = \frac{E_p\Psi^2(X)}{(E_p\Psi'(X))^2} = \frac{\int \Psi^2(y)f_p(y)d_y}{(\int \Psi'(y)f_p(y)d_y)^2} \quad \forall\ P \in \mathcal{P}$$

para Ψ não decrescente (HUBER [25]) e para Ψ geral supondo (6.3.7) em HUBER [26]).

Temos então que para $\Psi = \Psi_{H,k}$ (ver (3.3.9)) foram provadas (6.3.7) e (6.3.8). Para Ψ como em (3.3.10). (3.3.11) ou (3.3.12) também chamadas "redescendentes" os argumentos de HUBER não são aplicáveis. Felizmente as fórmulas (6.3.7) e (6.3.9) são válidas para estas últimas Ψ's porém os argumentos são diferentes e estão baseados no algoritmo que resolve a equação (3.3.8); a dificuldade consiste no fato que para estas Ψ(3.3.8) não tem solução única. Os detalhes podem ser estudados em KLEIN E YOHAI [34].

No caso de estimação da escala, vimos que a fórmula (6.1.6) deve ser trocada por (6.1.7) e se o estimador S_n que se usa em (3.4.6) satisfaz $S_n \to \sigma(F_p)$ segundo P então (6.3.7) segue valendo (com $T_{\Psi,\sigma}$ em lugar de T_Ψ), enquanto que (6.3.9) deve ser trocada por

$$(6.3.10)\quad VA(T_{\Psi,\sigma}, f_p) = \frac{E_p(\Psi^2(\frac{Y}{\sigma(Y)}))}{(E_p(\Psi'(\frac{Y}{\sigma(Y)})))^2}\,\sigma^2(Y) =$$

$$= \frac{\int \Psi^2(\frac{y}{\sigma(f_p)})f_p(y)dy}{(\int \Psi'(\frac{y}{\sigma(f_p)}f_p(y)dy)^2}\,\sigma^2(f_p) \qquad \forall\, p \in \mathcal{P}$$

Os detalhes da demonstração de (6.3.10) podem ser encontrados nas publicações citadas anteriormente.

Comportamento assintótico dos estimadores R

Seja $J:(0,1) \to R$ tal que $J(t) = -J(1-t)$ J é contínua e não decrescente. Consideremos o funcional T_J definido em (6.1.10) e para cada n=1,2,...

$$T_n = T_n(y_1,\ldots,y_n) = T_J(\mu[y_1,\ldots,y_n]),\ (y_1,\ldots,y_n)^T \in R^n$$

Em AZENCOT E OUTROS [2] prova-se que

$$(6.3.11)\quad T_n(Y_1,\ldots,Y_n) \to T_J(F_p) = 0\ (P) \quad \forall\, p \in \mathcal{P}$$

Em ANDREWS E OUTROS [1] e HUBER [30] deduz-se a função de influência de T_J em G segundo a suposição de diferenciabili-

dade de T_J. Essa fórmula nos prova que:

$$IC(y;T_J,F_p) = \frac{J(F_p(y))}{\int J'(F_p(y))f_p^2(y)dy} \quad \forall\ p\in\mathcal{P}$$

Aplicando esta fórmula em (6.3.2) e considerando para os detalhes que se supomos válida (6.3.11) então o estudo de T_J pode ser levado ao de uma T_Ψ como em (6.1.6) (ver MARTIN [39]), teremos que

$$VA(\text{R-estimador baseado em escores normais}, F_p) =$$

$$= \frac{\int(\Phi^{-1}(F_p(y)))^2\, f_p(y)dy}{\left[\int \frac{(f_p(y))^2}{\Psi(\Phi^{-1}(F_p(y)))}\,dy\right]^2} \quad \forall\ p\in\mathcal{P}$$

(6.3.12) $VA(\text{R-estimador de HODGES-LEHMANN}, F_p) =$

$$= \frac{1}{12\left[\int f_p^2(t)dt\right]^2} \quad \forall\ p\in\mathcal{P}$$

Em JAMES E BUSTOS [32] estão calculadas explicitamente as fórmulas (6.3.4),(6.3.5),(6.3.6),(6.3.9),(6.3.10) e (6.3.12) para várias Ψ segundo o modelo "puro" ($f_p=\varphi$) e segundo o modelo "contaminado" com contaminações do tipo $f_p = CN(\cdot;\xi,\tau)$

6.4 - Eficiência assintótica: Calibração das constantes nos M-estimadores

Continuamos usando nesta seção o modelo da seção anterior. Para cada $f\in\mathcal{F}$ e cada $\mu\in R$ sejam

$$f(y,\mu) = f(y-\mu) \quad \forall\ y \in R$$

$$F_\mu(y) = \int_{-\infty}^{y} f(t,\mu)dt \quad \forall\ y \in R$$

Logo $F_\mu(y) = F_o(y-\mu) \quad \forall\ y \in R, \ \forall\ \mu$

Conforme vimos na seção anterior todos os estimadores que nos interessam satisfazem:

(6.4.1) $$T(F_o) = 0 \quad \forall\ f \in \mathfrak{F}$$

De (6.4.1) e da equivariância segundo translações deduzimos que todos nossos estimadores cumprem

$$T(F_\mu) = \mu \quad \forall\ \mu \in R, \ \forall\ f \in \mathfrak{F}$$

propriedade que é conhecida como "consistência Fischer" (ver HAMPEL [21], HUBER [30])

Em HUBER [30] prova-se o seguinte: "seja T um funcional que satisfaz a propriedade da consistência Fischer; $\{T_n\}$ uma sequência de estimadores de μ cada um deles definido por T. Então $\{T_n\}$ é ANE (recordar Seção 2.5) Se e somente se

$$IC(y;T,F_\mu) = \frac{1}{I_1(F_\mu)} \frac{\partial}{\partial\mu} \log f(y,\mu) \quad \forall\ y,\mu$$

onde $$I_1(F_\mu) = \int \left(\frac{\partial}{\partial\mu} \log f(y,\mu)\right)^2 dF_\mu(y)$$

é suposto ser um número real positivo".

Deste resultado podemos concluir a validez da seguinte lista:

Estimadores ANE segundo $F_o=\Phi$: Média amostral e estimador-R baseado em "escores normais".

Estimadores ANE segundo F_o=distribuição logística ($f(y,0)=L(y)$

Estimador-M com $\Psi=\Psi_L$ (função escore de L), que não é outra coisa senão o EMV segundo a distribuição logística também é ANE o estimador-R baseado em"escores de Wilcoxon" ou estimador de HODGES-LEHMANN.

Calibração de constantes nos M-estimadores

Vemos que nas definições dos M-estimadores com $\Psi=\Psi_{H,k}$, $\Psi_{A,k}$, $\Psi_{B,k}$ dadas por (3.3.9),(3.3.11) e (3.3.12) as constantes k não estão determinadas. O critério em uso para determiná-las baseia-se nc seguinte raciocínio:

O modelo "puro" ou hipotético é o gaussiano; segundo este modelo sabemos que a média amostral é ANE. Por outro lado, usando (6.3.9) pode-se ver que (ver também JAMES E BUSTOS [32]).

$$(6.4.2)\quad VA(\text{M-estimador com } \Psi=\Psi_{H,k};\Phi) = \frac{2k^2-2k^2\Phi(k)-2(-k\varphi(k)+\Phi(k)-\frac{1}{2})}{(2\Phi(k)-1)^2}$$

$$(6.4.3)\quad VA(\text{M-estimador com } \Psi=\Psi_{B,k};\Phi) = \frac{\int_o^1 v^2(1-v^2)^4\,\varphi(kv)dv}{2k^3\left[\int_o^1 v^2(1-v^2)^2\,\varphi(kv)dv\right]^2}$$

Uma conta mostra que ambas VA são maiores que 1 para $k<\infty$, logo os M-estimadores com $\Psi=\Psi_{H,k}$ e $\Psi=\Psi_{B,k}$ serão assintoticamente menos eficientes que a Média amostral segundo o modelo puro. No entanto, vimos já que estes M-estimadores tem propriedades de robustez muito mais interessantes não importando que valor tenha

$k<\infty$. Parece natural então escolher k de modo que em (6.4.2) e (6.4.3) VA $\approx$ 1.053 com o qual a perda em eficiência seria de 5% aproximadamente. Procedendo assim chegamos a conclusão de adotar:

(6.4.4) $k = 1.345$ para $\Psi = \Psi$ de Huber

$k = 4.685$ para $\Psi = \Psi$ "biquadrada"

No caso de estimar também a escala pode se ver que se adotamos como estimador de escala

$$s_n = s_n(y_1,\ldots,y_n) = \frac{\text{MED}(|y_1-T_o|,\ldots,|y_n-T_o|)}{0.6745}$$

então $s_n \to 1$ segundo o modelo puro e assim as constantes serão como em (6.4.4).

Finalmente de JAMES E BUSTOS [32] podemos extrair a seguinte tabela que mostra quão rapidamente se perde a propriedade da eficiência da Média amostral ante ligeiras contaminações no modelo, sendo em troca mais estável o comportamento dos M-estimadores com $\Psi = \Psi_{H,1.345}$ e $\Psi = \Psi_{B,4.685}$ (sobretudo este último). Notemos também a estabilidade da Mediana amostral e do estimador de Hodges-Lehmann.

ESTIMADOR	f=φ	f = CN(·;ε,τ)					
		ε = 0,1			ε = 0,2		
		τ=3.	τ=7.	τ=11.	τ=3.	τ=7.	τ=11.
Média amostral	1,000	1,800	5,800	13,000	2,600	10.600	25,000
MEDM	1,571	1,803	1,879	1,901	2,091	2,288	2,346
0,1-TM	1,060	1,296	1,419	1,460	1,629	2,123	2,384
M-estim.$\Psi=\Psi_{H,1.345}$	1,053	1,296	1,417	1,455	1,623	2,002	2,138
M-estim.$\Psi=\Psi_{B,4.685}$	1,053	1,272	1,249	1,218	1,592	1,594	1,508
Estim.de Hodges-Lehmann	1,047	1,311	1,458	1,506	1,651	2,079	2,234

Variâncias assintóticas segundo f

A título de exercício convida-se o leitor a calibrar o valor de k para $\Psi = \Psi_{A,k}$ (ver 3.3.11)).

CAPÍTULO VII

O USO DE MÉTRICAS NO ESPAÇO DAS DISTRIBUIÇÕES PARA ANALISAR ROBUSTEZ

7.1 - Métricas no conjunto das distribuições

Ao considerar o modelo de posição vimos que o ponto de partida da Inferência Paramétrica clássica é supor que F é uma certa distribuição F_o (na maioria das vezes F_o é suposto ser Φ). Suponhamos agora que (T_n) é uma sequência de estimadores tal que para cada n, T_n está baseado em n observações, isto é, T_n é uma variável aleatória da forma

$$(7.1.1) \qquad T_n = t_n(Y_1,\ldots,Y_n)$$

onde $t_n:R^n \to R$

Veremos aqui como formalizou HAMPEL [18] ou [19] o conceito de "robustez qualitativa". Segundo vimos na Seção 1.2 este conceito aplica-se a sequências (T_n) tais que se a "verdadeira" F não é F_o mas sim está "perto" de F_o, então as distribuições de T_n segundo F e F_o também estarão "perto", $\forall n$. Antes de tudo: O que entendemos por distribuição de T_n segundo F?

Definição 7.1.1. Sejam: Ω o espaço amostral; P uma probabilidade sobre Ω, G uma distribuição (ou probabilidade) sobre R; $Y_1,\ldots,Y_n$ uma amostra de tamanho n de G (isto é, $Y_1,\ldots,Y_n$ são variáveis aleatórias definidas sobre Ω independentes e identicamente distribuídas com distribuição comum G); T_n uma esta-

tística baseada em $Y_1,\ldots,Y_n$ (T_n é como em (7.1.1)) chama-se distribuição de T_n segundo G a probabilidade $\mathcal{L}_G(T_n)$ sobre R definida por.

$$\mathcal{L}_G(T_n)(B) = P(t_n(Y_1,\ldots,Y_n) \in B) = P((Y_1,\ldots,Y_n) \in t_n^{-1}(B))$$

Usando a notação da definição anterior diremos que (T_n) é robusta em F_o se quando F está"próximo" de F_o, então $\mathcal{L}_{F_o}(T_n)$ está "próximo" de $\mathcal{L}_F(T_n)$ para todo n. Porém o que significa "próximo" formalmente falando ? Obviamente necessitamos uma noção de distância entre distribuições ou para dizer numa linguagem mais matemática: entre pontos do conjunto Z(R).

Temos uma noção de distância entre pontos da reta, ou do plano ou do espaço; porém agora necessitamos transladar esta noção a pontos de um conjunto abstrato. Na Matemática já a bastante tempo trabalha-se com distâncias entre pontos de um conjunto abstrato e logrou-se construir uma teoria bastante geral que permite realizar analogias entre situações na reta, plano ou espaço e situações em conjuntos mais gerais. Esta teoria constitui uma parte importantíssima da Anaíise Matemática e é conhecida pelo nome de Teoria dos Espaços Métricos. Trata do seguinte:

Definição 7.1.2 - Seja M um conjunto não vazio. Chama-se métrica ou distância entre pontos de M uma função $d:MXM \to R$ tal que:

(D1) $d(A,B) \geq o \quad \forall A \in M, \ \forall B \in M$

(D2) $d(A,B)=o$ se e só se $A=B$

(D3) $d(A,B)=d(B,A) \quad \forall A \in M \quad \forall B \in M$

(D4) $d(A,C) \leq d(A,B)+d(B,C) \quad \forall A \in M, \ \forall B \in M, \ \forall C \in M$

Esta definição abstrai as propriedades essenciais da noção de distância entre pontos da reta, plano ou espaço. Com efeito

(D1) diz que a distância entre dois pontos é um número não negativo

(D2) diz que a distância entre dois pontos é zero se e só se os pontos são o mesmo

(D3) diz que a distância entre A e B é a mesma que entre B e A

(D4) é chamada desigualdade triangular e justifica seu nome de maneira primordial e evidente (ver Figura 10) de que para ir de A a C é mais rápido ir diretamente que ir de A e B e depois de B a C.

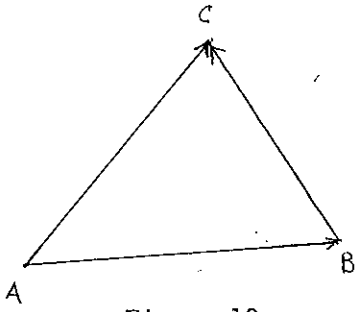

Figura 10

Não vamos nos aprofundar mais nesta frutífera parte da Matemática. Apenas para fixar idéias recomendamos ao leitor verificar que as seguintes funções definem métricas no plano:

$R^2 = \{(a,b): a \in R, b \in R\}$

$d_1((a,b),(a_1,b_1)) = \sqrt{(a-a_1)^2 + (b-b_1)^2}$

$d_2((a,b),(a_1,b_1)) = \text{Sup}\ (|a-a_1|, |b-b_1|)$

$d_3((a,b),(a_1,b_1)) = |a-a_1| + |b-b_1|$

$$d_4((a,b),(a_1,b_1)) = \begin{cases} 1 & \text{se } a \neq a_1, \text{ ou } b \neq b_1 \\ 0 & \text{se } a = a_1, \text{ e } b = b_1 \end{cases}$$

Da definição 7.1.2 vemos que várias funções poderiam ser definidas sobre $Z(R) \times Z(R)$ com valores em R e que constituiam distâncias entre distribuições. Porém, qual é a adequada do ponto de vista estatístico? Aqui não há muito acordo, pois depende de vários fatores do problema concreto em estudo. Em geral adota-se como adequadas aquelas mais usadas na Teoria de Probabilidade e entre elas a chamada métrica de LEVY.

Definição 7.1.3 - Chama-se distância LEVY entre distribuições sobre R a função $d_L: Z(R) \times Z(R) \to R$ definida por:

$$(7.1.2) \quad d_L(G,G_1) = \inf\{\varepsilon : G_1(x-\varepsilon)-\varepsilon \leq G(x) \leq G_1(x+\varepsilon)+\varepsilon, \ \forall\, x\}$$

Para ter uma idéia gráfica do que significa esta distância observemos a seguinte:

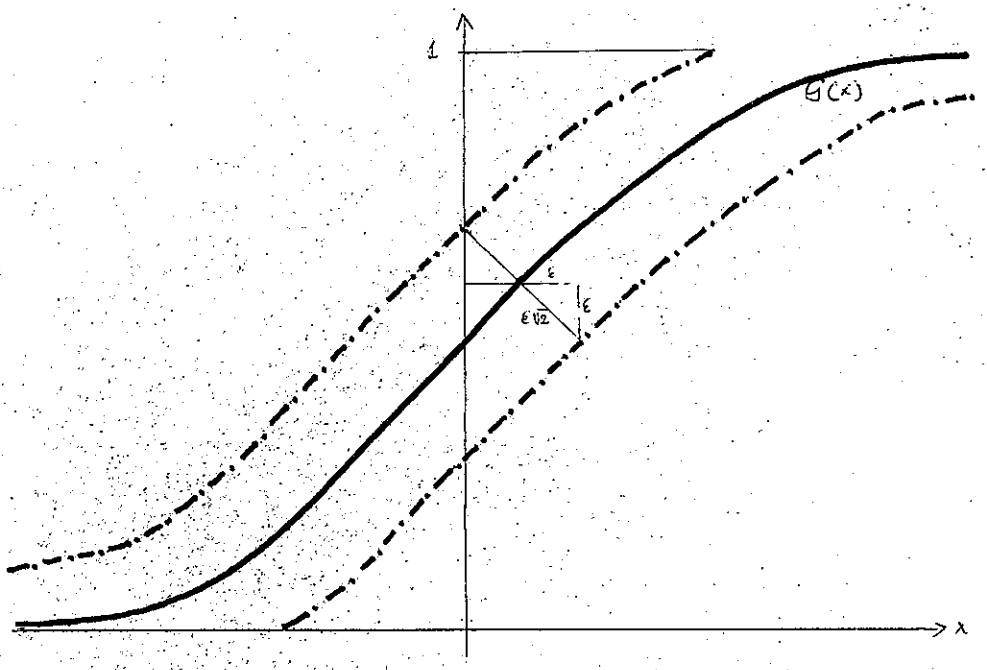

Figura 11

Não é difícil comprovar que $d_L(G,G_1) \leq \epsilon$ se e só se o gráfico de G_1 jaz inteiramente na zona sombreada.

Tampouco é muito complicado verificar que efetivamente a função definida por (7.1.3) constitui uma distância, ou seja, satisfaz (D1) a (D4) da definição 7.1.2.

Seguramente a importância de tal distância na Teoria de Probabilidade e na de Estatística deve-se ao fato da convergência em distribuição coincidir com a convergência na distância LEVY; isto é, vale o seguinte resultado (ver, por exemplo BILLINGSLEY [3]).

Sejam $G, G_1, G_2, \ldots$ distribuições sobre R. $G_n \to G(D)$ ($G_n(y) \to G(y)$ $\forall$ y onde G é contínua) se e só se $d_L(G_n, G) \to o$"

Agora sim estamos em condições de fazer rigoroso o conceito de robustez qualitativa (HAMPEL [18] e [19]).

7.2 - Robustez qualitativa

Definição 7.2.1 - Seja Ω o espaço amostral $Y_1, Y_2, \ldots$ uma sequência de variáveis aleatórias definidas sobre Ω; (T_n) é uma sequência de estimadores (estatísticos) tal que para cada n, T_n depende de $Y_1, \ldots, Y_n$. Diz-se que (T_n) é qualitativamente robusta em F_o se para cada $e > o$ existe $d > o$ tal que

$$d_L(F_o, F) < d \Rightarrow d_L(\mathcal{L}_{F_o}(T_n), \mathcal{L}_F(T_n)) < e \quad \forall\ n = 1, 2, \ldots$$

Como vemos, a definição anterior pode ser aplicada para uma sequência de estimadores de μ no Modelo de posição qualquer. Porém segundo vimos na seção 6.1 os estimadores que nos interessam estão definidos por meio de funcionais definidas em $Z(R)$. Em tais casos, HAMPEL [19] proporciona uma utilíssima caracterização de sequências de estimadores qualitativamente robustas; é a que está precisada no seguinte.

Teorema: Seja T um funcional definido em $Z(R)$, para cada $n = 1, 2, \ldots$ $T_n = T|_{\mathcal{F}_n}$ (recordar definição 6.1.2) T é contínua em todo o seu domínio DT (com relação a convergência em distribuição) se e só se para toda $G \in DT$ cumpre-se: (i) (T_n) é fracamente consistente para estimar $T(G)$; (ii) (T_n) é qualitativamente robusta em G.

Em vários trabalhos analisa-se a robustez qualitativa dos estimadores estudados até agora (por exemplo AZENCOT E OUTROS [2], HUBER [30], HAMPEL [19]) lamentavelmente algumas demonstrações são complicadas e exigem condições de regularidade muito restritivas (por exemplo monotonia na Ψ que define os M-estimadores), em outros casos as afirmações são conjecturas justificadas com argumentos heurísticos. Não obstante julga-se de interesse a seguinte tabela extraída de HAMPEL [21] e completada com alguns resultados de HUBER [30].

Estimador	Condições sobre F_o	Robustez qualitativa em F_o
Média amostral		não
MEDM	Contínua e estritamente crescente em $F_o^{-1}(1/2)$	sim
α-TM	Contínua em $F_o^{-1}(\alpha)$ e $F_o^{-1}(1-\alpha)$	sim
M com Ψ-monotona	$\int \Psi(y-t)dF_o(y)=0$ se e só se $t=T_\Psi(F_o)$	sim
R com escores normais		sim
Hodges-Lehmann		sim

Um pouco a margem do tema que estamos considerando é interessante destacar o comportamento do R-estimador com escores normais.

Vemos na tabela acima que este estimador é qualitativamente robusto em $F_o = \Phi$ e segundo vimos na seção 6.4 é também ANE;

não obstante, sua sensibilidade frente a erros grosseiros (GES) é muito grande $(+\infty)$,
como observamos na tabela ao final da seção 6.2; daqui HUBER[30] conclue que não seria conveniente recomendar tal R-estimador para usos práticos. A moral: Ainda que um estimador seja qualitativamente robusto não necessariamente deve-se usá-lo sem uma análise detalhada do problema concreto a resolver. O que acabamos de dizer vale para qualquer regra de estimação: não usar nada de forma cega.

Em verdade, HAMPEL [19] trabalhou não com a métrica de LEVY mas com outra métrica denominada de PROHOROV cujo uso tem sido muito difundido na Teoria de Probabilidade pois não apenas se aplica em distribuições sobre R mas sobre qualquer R^k e ainda em espaços mais gerais (ver BILLINGSLEY [3]). A desvantagem que apresenta do ponto de vista prático é a dificuldade de calculá-lo numericamente na maioria dos casos. Por outro lado, ambas as métricas (de LEVY e PROHOROV) são equivalentes sobre Z(R) (equivalentes no sentido de que $d_L(G_n,G) \to o$ se e só se $d_p(G_n,G) \to o$, sendo d_p a métrica de PROHOROV).

7.3 - Pontos de ruptura assintóticas

Analogamente ao que vimos na seção 5.3 trata-se aqui de formalizar a seguinte noção: Sejam F_o a distribuição hipotética das observações e (T_n) uma sequência de estimadores de μ (tal que para cada n, T_n está baseado em n observações) Admitamos que a verdadeira distribuição das observações (F) não seja F_o; dizemos que ε^* é o ponto de ruptura de (T_n) em F_o, quando

(T_n) nos dá informação útil sobre o parâmetro a estimar se e só se a distância de F e F_o é menor ou igual que ε^*. Naturalmente consideramos robustos com relação a esta propriedade os estimadores com maior ε^*. De acordo com a maneira como precisemos matematicamente as palavras "informação útil sobre o parâmetro a estimar" e a qual "distância" usamos para medir o "afastamento" entre F_o e F, teremos diversas definições (matemáticas) de ponto de ruptura. As que são sugeridas até agora (HAMPEL [19], HUBER [30], MARTIN [39]) apresentam vantagens e desvantagens em cada caso particular. Pensa-se assim que dar estas definições pode oferecer utilidade para quem trabalha em problemas de aplicações de índole diversa.

Ponto de ruptura HAMPEL [19] é o ε_1^* definido por

$$\varepsilon_1^* = \text{Sup}\{\varepsilon \le 1 : \exists a(\varepsilon), b(\varepsilon) \text{ em } R \text{ satisfazendo}$$
$$d_L(F_o, F) \le \varepsilon \Rightarrow \mathcal{L}_F(T_n([a(\varepsilon), b(\varepsilon)])) \to 1 \}$$

Ponto de ruptura com relação ao vício assintótico (HUBER [30])
Suponhamos que $T_n = T|_{\mathcal{F}_n}$ onde T é uma funcional definida em $Z(R)$. Em tal caso define-se com o nome recém sublinhado a:

(7.3.1) $\varepsilon_2^* = \text{Sup}\{\varepsilon \le 1 : b(\varepsilon) < b(1)\}$

onde $b(\varepsilon) = \text{Sup}\{|T(F) - T(F_o)| : d_L(F, F_o) \le \varepsilon\} \vee 0 \le \varepsilon \le 1$

Ponto de ruptura com relação a variância assintótica (HUBER [30])
Suponhamos que $T_n = T|_{\mathcal{F}_n}$ e que (T_n) é assintoticamente normal segundo F com variância assintótica $VA(T,F)$, para toda F em DT. Define-se então:

(7.3.2) $\varepsilon_3^* = \text{Sup}\{\varepsilon \leq 1: v(\varepsilon) < v(1)\}$

onde $v(\varepsilon) = \text{Sup}\{VA(T,F): d_L(F,F_o) \leq \varepsilon\}$

O mesmo HUBER [30] critica as definições (7.3.1) e (7.3.2) e propõe outras alternativas mais difíceis de manejar porém que na maioria dos casos de interesse prático coincidem com as últimas. Maiores detalhes o leitor interessado pode encontrar no citado trabalho de Huber e também em JAMES E BUSTOS [32]. Devemos ter em conta que muitas vezes não estaremos interessados em usar a métrica de LEVY ou em considerar todas as possíveis distribuições diferentes de F_o. Assim, por exemplo, HAMPEL [19] em lugar de usar a métrica de LEVY usa a de PROHOROV. Também em lugar de $b(\varepsilon)$ (ou $v(\varepsilon)$) poderia-se usar

$$b^*(\varepsilon) = \text{Sup}\{|T(F)-T(F_o)|: d_L(F,F_o) \leq \varepsilon, F \in Q\}$$

sendo Q, por exemplo:

$$Q = \{F = (1-\delta)F_o + \delta H : H \in \mathcal{H}\}$$

com $\delta > o$ fixado (em torno de δ-contaminação) e $\mathcal{H}$ um certo subconjunto de Z(R). Para finalizar esta seção vejamos a seguinte tabela que mostra os valores de ε_2^* de alguns dos estimadores que estamos estudando para o cado $F_o = \Phi$ (ver HAMPEL [21], HUBER [30], MARTIN [39]).

ESTIMADOR	ε_2^*
Média amostral	0
Mediana amostral (MEDM)	0,5
Média α-truncada (α-TM)	α
M-estimador com Ψ monótona e impar	0,5
R-estimador com escores normais	0,24
Estimador de Hodges-Lehmann	$1-1/\sqrt{2}\approx 0,29$

Tabela de ε_2^*=ponto de ruptura em relação ao vício assintótico

CAPÍTULO VIII

OUTROS CONCEITOS DE ROBUSTEZ: MINIMAX, SENSIBILIDADE LOCAL A DESLOCAMENTOS E PONTO DE REJEIÇÃO

O primeiro destes conceitos (robustez minimax) foi desenvolvido por Huber (HUBER [25],[30]) e serviu como base teórica para propor um M-estimador de μ com $\Psi = \Psi_{H,k}$ quando a distribuição hipotética das observações é Φ e se admite que a verdadeira seja $(1-\varepsilon)\,\Phi + \varepsilon\,H$ com $0 \leq \varepsilon < 1$ e H é uma certa distribuição desconhecida. Apesar de sua importância teórica ser destacável, nestas notas não veremos mais que uma introdução "a vôo de pássaro" principalmente devido a complexidade matemática implicada em seu desenvolvimento rigoroso e detalhado. O leitor interessado pode consultar os trabalhos de Huber recém citados e também MARTIN [39].

Quanto aos outros dois conceitos: sensibilidade local a deslocamentos (LSS) e ponto de rejeição foram definidos por Hampel (HAMPEL [19],[21]). Trata-se de explorar certas propriedades matemáticas da curva de influência de um estimador, traduzindo-as em termos suscetíveis de uma interpretação estatística. Não tem tido muito difusão quiçá pelo fato desta interpretação parecer um tanto forçada.

8.1 - Robustez minimax

Desde já há vários anos se tem utilizado no desenvolvimento da teoria estatística certos elementos da teoria dos jogos.

A base para tal utilização se deve ao fato de se pensar que a resolução de um problema de Inferência Estatística é como o resultado de um jogo entre dois jogadores: a Natureza (jogador A) e o Estatístoco (jogador B). O conjunto das possíveis jogadas ou estratégias de A constituem os estados possíveis da Natureza, denotemos este dito conjunto por E(A), o correspondente a B está formado pelas diferentes técnicas de estimação que o Estatístico está disposto a usar, seja este conjunto E(B). Como num jogo, o resultado pode ser medido por meio de uma função de perda $L:E(A) \times E(B) \to R$ ($L(a,b)$ é o resultado de A jogar com a estratégia "a" e B com a "b") interpretada como "maldade" da estimação quando o estatístico usa um método $b \in E(B)$ e a Natureza se encontra no estado $a \in E(A)$. Esta "maldade" poderia ser medida por vários critérios eficiência, vício, sensibilidade a erros grosseiros, etc.

O raciocínio do parágrafo anterior que poderia ser aplicado e tem sido aplicado a diversos problemas de inferência (principalmente pelos "bayesianos") foi precisado por HUBER[25] para o problema de encontrar um M-estimador ótimo(no sentido de eficiência assintótica) do parâmetro μ no modelo de posição quando a verdadeira distribuição das observações é suposta ser $(1-\varepsilon)\Phi + \varepsilon H$. Com efeito. Sejam

(8.1.1) $E(A) = \{F=(1-\varepsilon)\ \Phi+\varepsilon H : H$ é uma distribuição simétrica$\}$

com $o<\varepsilon<1$ suposto conhecido (também se tem estudado um pouco o caso ε desconhecido);

(8.1.2) $E(B) = \{T:R \to R : T$ com certas propriedades$\}$

(não destacamos aqui estas "certas propriedades" a fim de não entrar em detalhes matemáticos)

(8.1.3) $L:E(A)XE(B) \to R$ definida por:

$$L(F,\Psi) = VA(T_\Psi,F) = \frac{E_F\Psi^2(X)}{(E_F\Psi'(X))^2}$$

(recordemos (6.3.9))

Sigamos agora tomando emprestado raciocínios da Teoria dos Jogos (ver MARTIN [39]). Suponhamos que para uma $\Psi \in E(B)$ dada, a Natureza opõe a "pior" (para o estatístico) distribuição possível que há em E(A), isto é $\tilde{F}(\Psi) \in E(A)$ tal que

$$L(\tilde{F}(\Psi),\Psi) = \max_{F\in E(B)} L(F,\Psi)$$

É natural agora buscar uma $\Psi_o \in E(B)$ tal que

(8.1.4) $L(\tilde{F}(\Psi_o),\Psi_o) = \min_{\Psi\in E(B)} L(\tilde{F}(\Psi_o),\Psi) =$

$$\min_{\Psi\in E(B)} \max_{F\in E(A)} L(F,\Psi)$$

Por tal razão se uma Ψ_o como esta existe diz-se que T_Ψ é um M-estimador robusto minimax. É muito difícil obter um estimador deste tipo,(neste caso e em outros analogos) calculando diretamente o último membro de (8.1.4). Pelo que se sabe da Teoria dos Jogos, será suficiente obter um "ponto de sela" da função L; isto é uma $(F_o,\Psi_o) \in E(A)XE(B)$ tal que

$$\min_{\Psi\in E(B)} L(F_o,\Psi) = L(F_o,\Psi_o) = \max_{F\in E(A)} L(F,\Psi_o)$$

HUBER [25] desenvolveu uma teoria de robustez minimax mais geral da qual obtemos que segundo (8.1.1),(8.1.2) e (8.1.3) existe um (único) "ponto de sela" de $L(F,\Psi)$ dado por $F_o=(1-\varepsilon)\Phi+\varepsilon H_o$ com H_o determinada pela densidade

$$h_o(t) = \begin{cases} 0 & |t| \le k_2 \\ \frac{1-\varepsilon}{\sqrt{2\pi}} \left[e^{-k|t| + \frac{k^2}{2}} - e^{-\frac{t^2}{2}} \right], & |t| > k \end{cases}$$

com k dependente de ε, de onde conclue-se que F_o está definida pela densidade

$$f_o(t) = \begin{cases} (1-\varepsilon)\varphi(t) & , |t| \le k \\ \frac{1-\varepsilon}{\sqrt{2\pi}} e^{k^2/2} \; e^{-k|t|} & , |t| > k \end{cases}$$

(F_o é como uma normal na parte central e com "caudas" como uma exponencial dupla); Ψ_o não é outra senão a função "escore" associada a F_o (ver Seção 3.3.2), ou seja

$$\Psi_o(t) = -\frac{f_o'(t)}{f_o(t)} = \begin{cases} k \; \text{sinal}(t) & , |t| > k \\ t & , |t| \le k \end{cases}$$

que não é outra senão $\Psi_{H,k}$. Os valores de k dependem de ε porém como na prática a proporção de contaminação é desconhecida escolhe o valor de k segundo vimos na seção 6.4. Não obstante é de interesse observar a seguinte tabela de valores de k em função de ε (ver HUBER [25] MARTIN [39]).

ε	0,01	0,05	0,1
$k = k(\varepsilon)$	1,95	1,40	1,14

Notemos de passagem que conforme as propriedades dos estimadores de máxima verossimilhança T_{Ψ} é ANE segundo F_o

8.2 - Sensibilidade local a deslocamentos (LSS)

Só veremos sua definição e seu valor segundo o modelo "puro" para alguns dos estimadores estudados aqui (valor extraído de HAMPEL [21]).

Definição 8.2.1 - Seja T um funcional em $Z(R)$, $G \in Z(R)$. Chama-se sensibilidade local a deslocamentos (LSS) de T em G a:

$$\lambda^*(T,G) = \text{Sup}\left\{\frac{|IC(y;T,G)-IC(x;T,G)|}{|y-x|} : y \neq x\right\}$$

Segundo Hampel, este valor (pode ser $+\infty$) é de interesse se estudamos o efeito produzido sobre o estimador T por arredondamento e/ou agrupamento de dados: $\lambda^*(T,G)$ deve ser o mais pequeno possível, do ponto de vista da robustez para o efeito recém citado. Vejamos agora o valor de $\lambda^*(T,\Phi)$ para alguns de nossos já familiares T.

Estimador	$\lambda^*(\cdot,\Phi)$
Média amostral	1,00
Mediana amostral	$+\infty$
0,1-TM	1,25
M-estimador com $\Psi=\Psi_{H,1.5}$	1,15
R-estimador com escores normais	1,00
Estimador de Hodges Lehmann	1,41

Vemos assim que um estimador robusto segundo um aspecto pode não se-lo segundo outro. Daí a importância de ver qual o aspecto que mais interessa no problema de aplicação que se está resolvendo.

8.3 - Ponto de rejeição (ρ^*)

Sua definição precisa pode ser encontrada em HAMPEL [19],[21] e também em JAMES E BUSTOS [32]. Para ter uma idéia gráfica deste conceito observemos a seguinte figura.

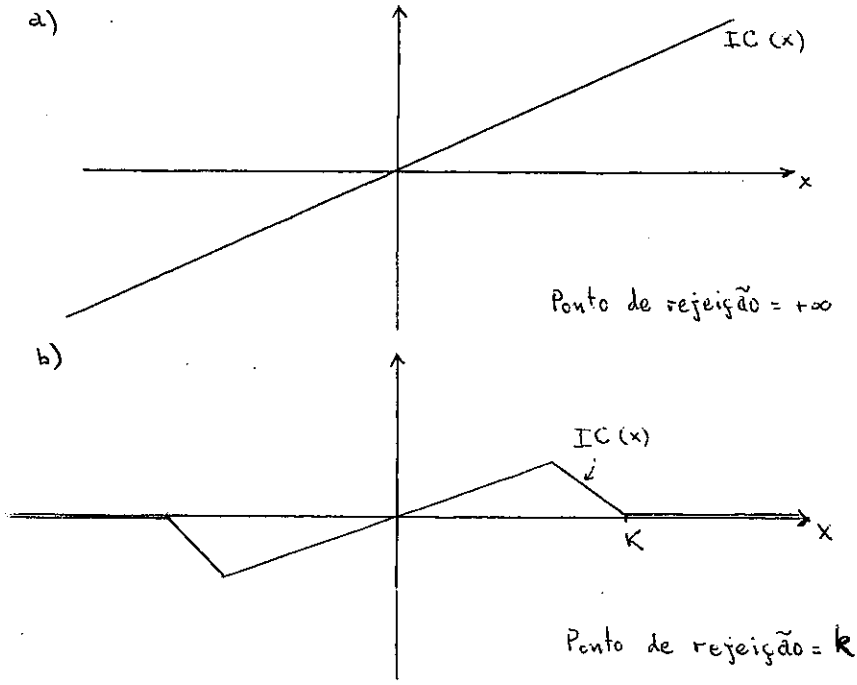

Figura 12

No primeiro caso $\rho^* = +\infty$, no segundo $\rho^* = k$.

Serve como uma regra de rejeição de "outliers": observações com valores maiores que ρ^* não são tomadas em conta pelo estimador cuja função de influência tem um gráfico como na figura 12.6)

CAPÍTULO IX

CONSTRUÇÃO DE INTERVALOS DE CONFIANÇA

9.1 - Intervalos de confiança para estimadores equivariantes por translações

Como sabemos, na Inferência Paramétrica se está interessado, mais que obter uma estimação pontual do valor do parâmetro, obter um intervalo ou região de confiança do qual saibamos que com certa probabilidade cobre o verdadeiro valor.

Recordemos brevemente o essencial do que sabemos sobre este tema (maiores detalhes podem ser encontrados, por exemplo em BICKEL E DOKSUM [4]). Aplicamo-lo ao problema de estimar μ no modelo de posição.

Suponhamos ademais que a distribuição hipotética ou nominal é $F=F_o$ (F_o será Φ na maioria das vezes).

Definição 9.1.1 - Seja $o<\gamma<1$. Diz-se que as estatísticas $T_{n_1}=T_{n_1}(Y_1,\ldots,Y_n)$, $T_{n_2}=T_{n_2}(Y_1,\ldots,Y_n)$ formam um intervalo de confiança para μ com nível de confiança $1-\gamma$ se:

$$F_\mu(T_{n_1} \leq \mu \leq T_{n_2}) \geq 1-\gamma \qquad \forall\, \mu \in R$$

(por abuso de notação F_μ denota aqui a probabilidade induzida pela função de distribuição F_μ (ver(3.1.3)). Seja $T_n=T_n(Y_1,\ldots,Y_n)$ um estimador de μ.
Suponhamos que existe uma distribuição H tal que:

(9.1.1) $\qquad H=\mathcal{L}_{F_\mu}(\sqrt{n}\,(T_n-\mu)) \qquad \forall\, \mu\in R$

Se H é conhecida é fácil ver que

(9.1.2) $T_{n_1} = T_n - H^{-1}(1-\frac{\gamma}{2})/\sqrt{n}$, $T_{n_2} = T_n - H^{-1}(\gamma/2)/\sqrt{n}$

definem um intervalo de confiança para μ com nível $1-\gamma$. Também é imediato provar que se T_n é equivariante segundo translações (recordar Definição 3.4.1). Então

(9.1.3) $\mathcal{L}_{F_o}(\sqrt{n}\ T_n) = \mathcal{L}_{F_\mu}(\sqrt{n}\ (T_n-\mu))$ $\forall\ \mu\in R$

por conseguinte para aplicar (9.1.1) e (9.1.2) basta conhecer $\mathcal{L}_{F_o}(T_n)$

9.2 - Intervalo de confiança induzido pela Média amostral segundo $F_o = \Phi$

Temos estudado nos primeiros cursos de estatística que se $Y_1,\ldots,Y_n$ é uma amostra de Φ então $T_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$ tem distribuição $N(0,\frac{1}{n})$ de onde $\sqrt{n}\ T_n$ tem distribuição $N(0,1)$. Aplicando (9.1.1),(9.1.2) e (9.1.3) deduzimos que

(9.2.1) $$T_{n_1} = \frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{\Phi^{-1}(1-\frac{\gamma}{2})}{\sqrt{n}}$$

$$T_{n_2} = \frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{\Phi^{-1}(\frac{\gamma}{2})}{\sqrt{n}}$$

define um intervalo de confiança para μ de nível $1-\gamma$. Outro conceito associado ao de nível de confiança de uma região ou intervalo de confiança é a imprecisão da mesma, isto é, no caso de intervalo de confiança $[T_{n_1}, T_{n_2}]$ será $d = T_{n_2} - T_{n_1}$. Como vemos esta magnitude nos dá uma medida da exatidão com que estimamos μ. Quanto menor seja d melhor. Porém ao diminuir d também

o nível de confiança diminuirá, o que não é desejável. Por esta razão é conveniente buscar intervalos de confiança com nível prefixado e imprecisão menor possível. Notemos que em geral d é uma variável aleatória o que complicará a análise. Uma forma de eliminar esta dificuldade é trabalhar com $E_\mu d = E_{F_\mu}(T_{n_2} - T_{n_1})$ o que tampouco é fácil. Porém se T_{n_1} e T_{n_2} são como em (9.1.2) temos que

$$d = T_{n_2} - T_{n_1} = \frac{H^{-1}(1-\frac{\gamma}{2}) - H^{-1}(\frac{\gamma}{2})}{\sqrt{n}}$$

que não é aleatória. Considerando (9.2.1) tem-se que:

Comprimento do intervalo de confiança induzido pela Média amostral segundo $F_o = \Phi$ = $\dfrac{\Phi^{-1}(1-\frac{\gamma}{2}) - \Phi^{-1}(\frac{\gamma}{2})}{\sqrt{n}}$

9.3 - Robustez de validez e de eficiência

O que fizemos até agora é o que se faz em inferência estatística clássica; isto é supondo que a verdadeira distribuição das observações ($F_\mu(\cdot) = F(\cdot - \mu)$ com $F = F_o$ (conhecida) e mais precisamente ainda, com $F_o = \Phi$. O que se passa do ponto de vista da robustez? Isto é, o que sucede se F está em uma "vizinhança" $\mathfrak{F}$ da hipotética F_o? A totalidade das respostas parciais que se tem dado a esta pergunta consideram $\mathfrak{F}$ como em (6.3.3.) ou com ligeiras modificações; isto é, na "vizinhança" de Φ. Para analisar a robustez dos intervalos de confiança tem-se definido os seguintes conceitos (HAMPEL [20], MARTIN [39]).

Robustez da validez. Diz-se que um intervalo de confiança para

μ, $[T_{n_1}, T_{n_2}]$ tem a propriedade de robustez da validez em F_o em relação a $\mathcal{F}$ se a função definida sobre $\mathcal{F}$ por

$F \to F_\mu$ ($T_{n_1} \leq \mu \leq T_{n_2}$) "não varia muito" ($F_\mu(\cdot) = F(\cdot - \mu)$).

Robustez de eficiência. Diz-se que $[T_{n_1}, T_{n_2}]$ tem tal propriedade se $F \mapsto E_{F_\mu}(T_{n_2} - T_{n_1})$ "não varia muito". Pode se ver em diversas publicações (MARTIN [39]) que os intervalos de confiança indicados pela Inferência Paramétrica clássica tais como (9.2.1) no caso de escala conhecida ($F_o = \Phi$) ou os t-intervalos

$$T_{n_1} = \frac{1}{n} \sum_{i=1}^{n} Y_i - \frac{S_n t(n-1,\ 1-\frac{\gamma}{2})}{\sqrt{n}}$$

$$T_{n_2} = \frac{1}{n} \sum_{i=1}^{n} Y_i + \frac{S_n t(n-1,\ 1-\frac{\gamma}{2})}{\sqrt{n}}$$

com $$S_n = \left(\frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2\right)^{1/2}$$

(9.3.1) $$1 - \frac{\gamma}{2} = \int_{-\infty}^{t(n-1,\ 1-\gamma/2)} ST(u;n-1)du \quad \text{(ver (3.2.1))}$$

no caso $F_o(\cdot) = \Phi(\frac{\cdot}{\sigma})$ com σ desconhecido, tem a primeira propriedade mas não a segunda. Procurando uma alternativa que seja robusta em amos sentidos pensou-se em estudar o uso de estimadores robustos de μ como os que temos estudado até agora.

Para fixar idéia consideramos apenas o realizado com M-estimadores. Quem está interessado no que sepode fazer com outros estimadores, alguma indicação pode ser encontrada em JAMES E BUSTOS [32] e em REY [44].

9.4 - Intervalos de confiança induzidos por M-estimadores

Seja (T_n) uma sequência de estimadores com T_n baseado em n observações e todos definidos por um mesmo funcional T_Ψ em $Z(R)$ que satisfaz (6.1.6).

Como T_n é equivariante segundo translações tem-se que:

$$\mathcal{L}_F(\sqrt{n}\ T_n) = \mathcal{L}_{F_\mu}(\sqrt{n}(T_n-\mu)) \quad \forall\ \mu \in R$$

Porém se queremos agora aplicar o que fariamos antes com a média amostral esbarramos com a dificuldade de não conhecer $\mathcal{L}_F(\sqrt{n}\,T_n)$. Pelo que vimos na seção 6.3 teremos

$$\mathcal{L}_F(\sqrt{n}\,T_n) \to N(0, VA(T_\Psi, F)) \ (\mathbb{N}), n \to \infty$$

$$(9.4.1) \quad VA(T_\Psi, F) = \frac{E_F \Psi^2(X)}{(E_F \Psi'(X))^2}$$

segundo vimos em (6.3.9)

Assim para n "grande" podemos tomar como aproximadamente certo.

$$(9.4.2) \qquad \mathcal{L}_F(\sqrt{n}\ T_n)(\chi) = \Phi(\chi; 0,\ VA(T_\Psi, F)) \quad \forall\ \chi$$

de onde

$$\mathcal{L}_F\left(\frac{\sqrt{n}\ T_n}{\sqrt{VA(T_\Psi, F)}}\right)(\chi) = \Phi(\chi) \quad \forall\ \chi$$

É natural então tomar como intervalo de confiança para μ de nível $(1-\gamma)$ a:

$$(9.4.3) \quad T_{n_1} = T_n - \frac{\Phi^{-1}(1-\frac{\gamma}{2})\sqrt{VA(T_\Psi\ F)}}{\sqrt{n}}$$

$$T_{n_2} = T_n - \frac{\Phi^{-1}(\frac{\gamma}{2})\sqrt{VA(T_\psi,F)}}{\sqrt{n}}$$

A dificuldade agora está no fato de não podermos calcular $VA(T_\psi,F)$ pois de F apenas sabemos que está em $\mathfrak{F}$. Esta dificuldade é resolvida estimando $VA(T_\psi,F)$ pelo que nos sugere (9.4.1), ou seja por

$$VA(T_\psi,F)_n^* = \frac{\frac{1}{n}\sum_{i=1}^{n}\psi^2(Y_i-T_n)}{\left(\frac{1}{n}\sum_{i=1}^{n}\psi'(Y_i-T_n)\right)^2}$$

Definitivamente, um intervalo de confiança para μ de nível $1-\gamma$ é $[T_{n_1},T_{n_2}]$ com T_{n_1} e T_{n_2} dados em (9.4.3) com $VA(T_\psi,F)$ substituido por $VA(T_\psi,F)_n^*$. Se (T_n) está definida por $T_{\psi,\sigma}$ que satisfaz (6.1.7), raciocinando como antes, tendo em conta (6.3.10) pensa-se que um intervalo de confiança para μ com nível $(1-\gamma)$ razoável é o dado por:

$$T_{n_1} = T_n - \frac{\Phi^{-1}(1-\frac{\gamma}{2})\,\sigma_n^*}{\sqrt{n}}$$

$$T_{n_2} = T_n - \frac{\Phi^{-1}(\frac{\gamma}{2})\,\sigma_n^*}{\sqrt{n}}$$

com σ_n^* dado por:

$$(9.4.4)\quad \sigma_n^{*2} = \frac{\frac{1}{n}\sum_{i=1}^{n}\psi^2\left(\frac{Y_i-T_n}{S_n}\right)}{\left(\frac{1}{n}\sum_{i=1}^{n}\psi'\left(\frac{Y_i-T_n}{S_n}\right)\right)^2}\,S_n^2$$

sendo S_n um estimador de $\sigma(F)$ tal como vimos na definição 3.4.2.

Agora a igualdade (9.4.2) ou sua similar para $T_{\Psi,\sigma}$ é apenas aproximadamente certa para n "grande" segundo o que podemos ler em vários trabalhos: HUBER [27], GROSS [15] e [16], SHORACK [45], HOGG [24], MARTIN [39] parece mais adequado aproximar $\mathcal{L}_F\left(\frac{\sqrt{n}\ T_n}{\sqrt{VA(T_{\Psi,\sigma},F)}}\right)$ por uma t-Student em lugar da Φ com o número de graus de liberdade menor ou igual que n-1. Concluindo: Segundo MARTIN [39] parece que um intervalo de confiança para μ com nível $1-\gamma$ suficientemente aceitável do ponto de vista de robustez da validez e eficiência é dado por:

$$T_{n_1} = T_n - \frac{t(v,1-\frac{\gamma}{2})\ \sigma_n^*}{\sqrt{n}}$$

$$T_{n_2} = T_n + \frac{t(v,1-\frac{\gamma}{2})\ \sigma_n^*}{\sqrt{n}}$$

onde σ_n^* é igual a expressão (9.4.4), $t(v,1-\frac{\gamma}{2})$ está definida de forma similar a (9.3.1) e

$$v \approx 0{,}75\ (n-1) \quad \text{se} \quad 10 \leq n \leq 20$$

$$v \approx 0{,}9\ (n-1) \quad \text{se} \quad n > 20$$

CAPÍTULO X

ANÁLISE DE EXEMPLOS COM DADOS "REAIS

Existem na literatura vários exemplos de aplicação de técnicas estatísticas clássicas e robustas a dados reais. Tais exemplos têm sido trabalhados através de diversos artigos por estatísticos de renome. Será conveniente analisar algum de tais estudos nestas notas.

"Os dados de CUSHNY e PEEBLES"

Segundo podemos ver em HAMPEL [20], CUSHNY e PEEBLES [7] analisaram os diferentes efeitos no aumento das horas de sono de dois isómeros opticos de uma chapa suporífera: a dextro : e a laevo-hyoscyamine hydrobromide (D-HH e L-HH). Foram medidos os tempos de sono de 10 pacientes sem efeitos de nenhum remédio, depois do tratamento (1) com D-HH e depois do tratamento (2) com L-HH. O aumento médio das horas de sono está na tabela abaixo. A conclusão a que chegaram é que (2) era de valor suporífero e que (1) não era.

Paciente	(1)	(2)	(2)-(1)
1	+ 0.7	+ 1.9	+ 1.2
2	- 1.6	+ 0.8	+ 2.4
3	- 0.2	+ 1.1	+ 1.3
4	- 1.2	+ 0.1	+ 1.3
5	- 0.1	- 0.1	0
6	+ 3.4	+ 4.4	+ 1.0
7	+ 3.7	+ 5.5	+ 1.8
8	+ 0.8	+ 1.6	+ 0.8
9	0	+ 4.6	+ 4.6
10	+ 2.0	+ 3.4	+ 1.4

Este dados foram analisados sob diversos pontos de vista em várias publicações, entre elas: STUDENT [47] (de onde obtivemos a tabela acima), FISHER [13], HAMPEL [20], MARTIN [39]. HAMPEL destaca que a partir do trabalho de FISHER numerosos livros-texto têm apresentado estes números como exemplos de dados distribuidos segundo uma normal univariado.

Só fazemos aqui uma simples olhada nos resultados obtidos, porém, sem detalhe nenhum. Seguimos assim a apresentação do HAMPEL[20] onde também são mostrados outros estimadores e os intervalos de confiança para alguns deles.

Por nossa parte fazendo uso das subrotinas do JAMES E BUSTOS[32] ontemos a seguinte tabela:

Media amostral	1,58
Mediana amostral	1,30
α_T ($\alpha = 0.1$)	1,40
M-estimador, Ψ de Huber e $k = 1.345$	1,42
M-estimador, Ψ de Hambel com: $A = 2.5$, $B = 4.5$, $C = 9.5$	1,29
M-estimador, Ψ "biquadrada" e $k = 4.685$	1,42
R-estimador de Hodges-Lehmann	1,30

O primeiro a salientar é o afastamento da média amostral do resto dos estimadores. Isto faz pensar que possivelmente em os dados exista um "outlier" com valor positivo grande. Com efeito, na tabela primeira, a diferença entre (2) e (1) para o paciente 9 é bem maior que para os restantes. Nesta fase caberia esperar que o pesquisador continuasse o seu trabalho tratando de avisar o porque de tal diferença e não como aconteceu na realidade que devido ao uso da média amostral, somente, tal diferença ficou escondida.

Na verdade, tal parece que o uso das técnicas de robustez está sendo cada vez mais de uma maior atenção no Análise de Dados. Possivelmente assim as técnicas aqui brevemente apresentadas possam vir a ocupar seu sitio certo na Estatística: não como "o que se deve fazer em lugar das técnicas clássicas" mas assim como uma complementação necessária para uma análise mais realista.

REFERÊNCIAS

[1] ANDREWS, D.F. E OUTROS (1972). Robust Estimates of Location: Survey and Advances, Princeton University Press, Princeton, New Jersey.

[2] AZENCOT, R. E OUTROS (1977). Théorie de la robustesse et estimation d'un paramètre. Em: Astérisque, 43-44, Societé Mathématique de France. Paris.

[3] BILLINGSLEY,P. (1968). Convergence of Probability Measures, Wiley, New York.

[4] BICKEL, P.J. E DOKSUM, K.A. (1977). Mathematical Statistics Holden-Day Inc., San Francisco

[5] BOX, G.E.P. (1979). Robustness in the strategy of scientific model building. Em: Robustness in Statistics, Launer e Wilkinson (eds), Academic Press, New York.

[6] CRAMER,H. (1946). Mathematical Methods of Statistics. Princeton, University Press, Princeton, New Jersey.

[7] CUSHNY Y PEEBLES (1905). The action of optical isomers.II. Hyoscines. Journal of Physiology, 32, 501-510.

[8] DACHS, J.N.W. (1978). Analise de dados e regressão. IMECC, UNICAMP, Campinas.

[9] DAVIS,C.H. (tradutor)(1973). Gauss,K.F.: Theory of the Motion of the Heavenly Bodies Moving about the Sun in Conic Section. Dover Publications, Inc., New York.

[10] DENBY,L.E MALLOWS,C.L. (1977). Two diagnostic displays for robust regression analysis. Technometrics, 19 1-13.

[11] DIXON,W.J. E MASSEY, J.Jr. (1969). Introduction to Statistical Analysis, Mc Graw-Hill, Kogakusha,Tokio.

[12] DUTTER,R.(1977). Numerical solution of robust regression problems: computational aspects, a comparison. Journal of Statistical Computation and Simulation, 5, 207-238.

[13] FISHER, R.A. (1925). Statistical Methods for Research Workers, Oliver & Boyd, Edinburgh.

[14] GEARY,R.C. (1947). Testing for normality. Biometrika, 34, 309-242.

[15] GROSS, A.M. (1976). Confidence interval robustness with long tailed symmetric distributions. Journal of American Statistical Association, 71, 409-416

[16] _____(1977). Confidence intervals for bisquare regression estimates. Journal of American Statistical Association, 72, 341-354.

[17] HÁJEK,J.E SIDAK Z. (1967). Theory of Rank Tests. Academic Press, New York.

[18] HAMPEL,F.R. (1968). Contributions to the theory of robust estimation. Tese de doutorado. University of California. Berkeley.

[19] _____(1971). A general qualitative definition of robustness. Annals of Mathematical Statistics, 42, 1887-1896.

[20] _____(1973). Robust estimation: a condensed partial survey. Z. Wahrscheinlichkeits-theorie und Verw. Gebiete, 27, 87-104.

[21] _____(1974). The influence curve and its role in robust estimation, Journal of American Statistical Association, 69, 383-393.

[22] _____(1977). Modern trends in the theory of robustness. Report Nº 13. Fachgruppe für Statistik. ETH. Zürich.

[23] HARTER,H.L. The method of least squares and some alternatives. Int. Statist. Rev., 42 (1974) 147-174, 235-264, 282, 43 (1975) 1-44, 125-190, 273-278, 269-272, 44. (1976), 113-159.

[24] HOGG,R.V. (1979). Statistical robustness: one view of its use in applications today. The American Statistician, 33, 108-115.

[25] HUBER,P.J.(1964). Robust estimation of a location parameter Annals of Mathematical Statistics, 35, 73-101.

[26] _____(1967). The behaviour of maximum likelihood estimates under nonstandard conditions. Proc. Fifth Berkeley Symp. Math. Statist. Prob., 1, 221-223.

[27] _____(1970). Studentizing robust estimates. Em: Non parametric Techniques in Statistical Inference, PURI, M.L.(ed), Cambridge University Press,Cambridge.

[28] _____(1973). Robust regression: asymptotics, conjectures and Monte Carlo. Annals of Statistics, 1, 799-821.

[29] _____(1972). Robust statistics: a review. Annals of Mathematical Statistics, 43, 1041-1067.

[30] _____(1977). Robust Statistical Procedures. Society of Industrial and Applied Mathematics. Philadelphia, Pennsylvania.

[31] JAECKEL,L.A.(1971). Robust estimates of locations: symmetry and asymmetric contamination, Annals of Mathematical Statistic, 42, 1020-1034.

[32] JAMES,K.L. E BUSTOS, O.H.(1980). Procedimentos Robustos, 4º Simpósio Nacional de Probabilidade e Estatística, 21 a 25 de julho de 1980, Rio de Janeiro.

[33] KENDALL,M.G. E BUCKLAND,W.R.(1971). A Dictionary of Statistical Terms. Oliver and Boyd, Edinburgh.

[34] KLEIN,R. E YOHAI,V.J.(1979). Asymptotic behavior of iterative M-estimators for location. Boletim da Sociedade Brasileira de Matemática, 10, 27-42.

[35] LAUNER,R.L. E WILKINSON, G.N. (1979) (eds.). Robustness in Statistics. Academic Press, New York.

[36] LEHMANN,E.L.(1975). Non parametrics: Statistics Methods Based on Ranks. Holden-Day, San Francisco

[37] MARONNA,R.A. (1976). Robust M-estimators of multivariate location and scatter. Annals of Statistics, 4, 51-67.

[38] MARTIN, R.D. (1979). Robust estimation of autoregressive Models. Em: Directions in Time Series, Brillinger, D.R. e Tiao, G.C. (eds.), Proceedings of the IMS Special Topics, Meeting on Time Series Analysis, 1 a 3 de Maio de 1978, Iowa State University, Ames.

[39] _____(1980). Robust estimation. Notas de Aula. V ELAM. Mar del Plata.

[40] MOSTELLER, F.E. E TUKEY, J.W.(1977). Data Analysis and Regression. Addison-Wesley. Reading, Massachusetts.

[41] PARTHASARATHY,K.R.(1968). Probability Measures on Metric Spaces. Academic Press. New York.

[42] PROHOROV,Y.V.(1956). Convergence of random processes and limit theorems in probability theory. Theor. Probab. Appl.; 1, 157-214.

[43] RELLES,D.A. E ROGERS,W.H.(1977). Statisticians are fairly robust estimators of location. Journal of American Statistical Association, 72, 107-111.

[44] REY, W.J.J.(1978). Robust Statistical Methods. Lectures Notes in Mathematics Nº 690. Springer Verlag. Heidelberg.

[45] SHORACK,G.R.(1976). Robust studentization of location estimates. Statistica Neerlandica, 30, 119-141.

[46] STIGLER,S.M.(1973). Simon Newcomb, Percy Daniell and the History of Robust Estimation: 1885-1920. Journal of American Statistical Association, 68, 872-879.

[47] "STUDENT"(1908). The probable error of a mean. Biometrika, 6, 1-25.

[48] TIETJEN,G.L., KAHANER,D.K. E BECKMAN,R.J.(1977). Variances and covariances of the normal order statistics for sample sizes 2 to 50. Selected Tables in Mathematical Statistics, V, 1-73. IMS.

[49] TUKEY,J.W.(1960). A survey of sampling from contaminated distributions. Em: Contributions to Probability and Statistics, Olkim, (ed.), Stanford University Press. Stanford.

[50] _____(1977). Exploratory Data Analysis. Addison-Wesley. Reading. Massachussets.

[51] VON MISES,R.(1947). On the asymptotic distribution of differentiable statistical functions. Annals of Mathematical Statistics, 18, 309-348.

[52] YOHAI,V.J.(1974). Robust estimation in the linear model. Annals of Statistics, 2, 562-567.