Análise de Valores Extremos: Uma Introdução

M. Ivette Gomes
C.E.A.U.L. e D.E.I.O., F.C.U.L., Universidade de Lisboa,
Instituto de Investigação Científica Bento da Rocha Cabral

M. Isabel Fraga Alves
D.E.I.O., F.C.U.L. e C.E.A.U.L., Universidade de Lisboa

Cláudia Neves
C.E.A.U.L., Universidade de Lisboa,
DMat, Universidade de Aveiro

Edições SPE

Ficha Técnica:

Análise de Valores Extremos: Uma Introdução[1]

M. Ivette Gomes
C.E.A.U.L. e D.E.I.O., F.C.U.L., Universidade de Lisboa,
Instituto de Investigação Científica Bento da Rocha Cabral

M. Isabel Fraga Alves
D.E.I.O., F.C.U.L. e C.E.A.U.L., Universidade de Lisboa

Cláudia Neves
C.E.A.U.L., Universidade de Lisboa,
DMat, Universidade de Aveiro

Editora: Sociedade Portuguesa de Estatística

Capa: Carina Sousa

Impressão: Instituto Nacional de Estatística

Tiragem: 200 exemplares

ISBN: 978-972-8890-30-8

Depósito Legal: 366446/13

[1]Investigação parcialmente financiada pelos fundos nacionais da **FCT**—Fundação para a Ciência e a Tecnologia, projecto PEst-OE/MAT/UI0006/2011, EXTREMA, PTDC/MAT/101736/2008 e PTDC/MAT/112770/2009: EXTREMES IN SPACE.

Conteúdo

Prefácio

Neste texto procedemos em grande parte a uma compilação do material leccionado em cadeiras das áreas de *Estatísticas Ordinais*, de *Teoria de Valores Extremos*, de *Estatística de Extremos* e de *Modelação de Acontecimentos Raros*, ampliado com alguns desenvolvimentos recentes.

Trata-se de um manual de trabalho, ainda em fase embrionária, em que se procurou encontrar um compromisso entre o rigor teórico e uma abordagem intuitiva às áreas em estudo, disseminando técnicas simples, mas poderosas da área de *Estatística de Extremos*, que têm sido largamente utilizadas nos mais variados campos, entre os quais destacamos *Ciências Ambientais*, *Finanças* e *Seguros*.

Começamos por apresentar no Capítulo 2 alguma *Motivação* para a necessidade da *Teoria de Valores Extremos* (TVE), muito frequentemente denotada EVT, do ingês '*Extreme Value Theory*'. No Capítulo 3 avançamos com algumas *Técnicas Gráficas* usadas na análise preliminar de qualquer tipo de dados, tais como os QQ-plots e os PP-plots, e *Técnicas Gráficas* específicas da área de valores extremos, como os ME-plots e os W-plots. No Capítulo 4, através de alguns exemplos de aplicação a dados univariados, tentamos responder à pergunta *Porquê a Teoria de Valores Extremos*? Mas em EVT, e mais geralmente, em quase todos as áreas da *Estatística*, a ordenação de uma amostra aleatória univariada, como base para uma representação clara do conteúdo dessa amostra, é crucial. Tal justifica a consideração dos Capítulos 5 e 6, respectivamente sobre o *Comportamento Distribucional Exacto* e o *Comportamento Distribucional Assintótico* das estatísticas ordinais. Finalmente, nos Capítulo 7, 8 e 9, debruçamo-nos sobre *Estatística de Extremos*, área de grande utilidade em aplicações quando se pretende inferir na cauda de um modelo, estimando parâmetros de acontecimentos raros, como por exemplo quantis elevados ou períodos de retorno de níveis elevados. No Capítulo 7, abordamos as perspectivas paramétricas de inferência estatística em acontecimentos raros. O Capítulo 8 é dedicado a alguns métodos de inferência semi-paramétrica. Finalmente, no Capítulo 9, procedemos à análise de três casos de estudo.

O texto é, como convém, consideravelmente mais ambicioso do que será o

curso breve no XXI *Congresso Anual da Sociedade Portuguesa de Estatística*. Fica no entanto como elemento de referência para os interessados, enquanto num curso de algumas horas, mesmo intensivas e com a celeridade que uma audiência conhecedora impõe, apenas os tópicos mais relevantes podem ser abordados. Qualquer curso é um compromisso, procurando um equilíbrio pessoal (neste caso de uma trindade geracional) entre o que é reconhecidamente fundamental e imprescindível, e os gostos e interesses de quem o escreve. Assim, ficaram naturalmente de fora questões muito importantes mas que não serviam o nosso *puzzle*, tal como ficaram de fora questões que estão entre os nossos interesses directos de investigação, tais como como estimação de viés reduzido, utilização de metodologias de re-amostragem, como o bootstrap e o jackknife em *Estatística de Extremos*, entre outros, mas que certamente iriam desequilibrar a dinâmica do texto.

Não é decerto por ingratidão, mas em todo o rol de agradecimentos há esquecimentos. Por isso preferimos os *clichés*: às nossas famílias e aos nossos amigos, aos nossos mestres, aos nossos colegas, aos nossos alunos.

Agradecemos por outro lado à *Sociedade Portuguesa de Estatística* (SPE) e aos organizadores do XXI *Congresso Anual da* SPE esta honra que nos conferiram. Agradecemos também o apoio institucional do CEAUL — Centro de Estatística e Aplicações da Universidade de Lisboa. E, claro, as palavras mágicas e o logotipo: Esta investigação foi parcialmente subsidiada por **FCT** — Fundação para a Ciência e a Tecnologia, projectos PEst-OE/MAT/UI0006/2011, EXTREMA, PTDC/MAT/101736/2008 e PTDC/MAT/112770/2009: EXTREMES IN SPACE.

FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA

M. Ivette Gomes
M. Isabel Fraga Alves
Cláudia Neves

Capítulo 1

Introdução e Comentários Bibliográficos

Em *Teoria de Valores Extremos* (TVE), muito frequentemente denotada EVT, do ingês '*Extreme Value Theory*' a ordenação da amostra é primordial. Mais geralmente, e em quase todos as áreas da Estatística, a ordenação de uma amostra aleatória univariada, como base para uma representação clara do conteúdo dessa amostra, foi desde há muito considerada importante. Tal importância permitiu chegar ao patamar em que estamos hoje — uma vasta metodologia estatística e associada teoria distribucional relativas a amostras ordenadas — tal como se pode ver nos livros de Sarhan & Greenberg[1] (1962), David[2] (1981), Arnold & Balakrishnan[3] (1989), Reiss[4] (1989), Arnold *et al.*[5] (1992; 2008) e David & Nagaraja[6] (2003), sobre estatísticas ordinais (e.o.'s),

[1] Sarhan, A.E. & Greenberg, B.G. (1962). *Contributions to Order Statistics*. Wiley.

[2] David, H.A. (1981). *Order Statistics*, 2nd Ed., Wiley.

[3] Arnold, B.C. & Balakrishnan, N. (1989). *Relations, Bounds and Approximations for Order Statistics*. Springer-Verlag.

[4] Reiss, R.-D. (1989). *Approximate Distributions of Order Statistics*. Springer-Verlag.

[5] Arnold, B., Balakhrishna, N. & Nagaraja, H. N. (1992; 2008). *A First Course in Order Statistics*. 1st Ed., Wiley; 2nd Ed., SIAM.

[6] David, H.A. & Nagaraja, H.N. (2003). *Order Statistics*. 3rd. Ed., Wiley.

e nos livros de Leadbetter *et al.*[7] (1984), Galambos[8] (1987), Resnick[9] (1987) e Falk *et al.*[10] (1994; 2005; 2010), sobre e.o.'s extremais.

Temos ainda de referir o clássico livro de Gumbel[11] (1958; 2004), livro pioneiro em *Estatística de Extremos*, e os livros de Beirlant *et al.*[12] (1996), Tiago de Oliveira[13] (1997), Reiss & Thomas[14] (1997; 2007), Embrechts *et al.*[15] (1998), Kotz & Nadarajah[16] (2000), Coles[17] (2001), Beirlant *et al.*[18] (2004), Castillo *et al.*[19] (2004), de Haan & Ferreira[20] (2006), Resnick[21] (2007) e Markovich[22] (2007).

Existe, por um lado, um *interesse natural pela ordenação*:

- Os valores extremos são crucialmente importantes como expressão do pior ou do melhor que pode ser encontrado numa amostra (temperaturas mínimas, níveis máximos de barragens, tempos de vida mínimos

[7] Leadbetter, R., Lindgren, G. & Rootzén, H. (1984). *Extremes and Related Properties of Random Sequences and Processes*. Springer-Verlag.

[8] Galambos, J. (1987). *The Asymptotic Theory of Extreme Order Statistics*. Krieger.

[9] Resnick, S. (1987). *Extreme Values, Regular Variation and Point Processes*. Springer-Verlag.

[10] Falk, M., Hüsler, J. & Reiss, R.-D. (1994; 2005; 2010). *Laws of Small Numbers: Extremes and Rare Events*. Birkhäuser.

[11] Gumbel, E.J. (1958; 2004). *Statistics of Extremes*. Columbia University Press, Dover Publications, Inc., New York.

[12] Beirlant, J., Teugels, J.L. & Vynckier, P. (1996). *Practical Analysis of Extremes*. Leuven University Press.

[13] Tiago de Oliveira, J. (1997). *Statistical Analysis of Extremes*. Pendor.

[14] Reiss, R.-D. & Thomas, M. (1997; 2007). *Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields*. 2nd Ed.; 3rd. Ed., Birkhaüser Verlag, Berlin.

[15] Embrechts, P., Klüppelberg, C. & Mickosh, T. (1998). *Modelling Extremal Events for Insurance and Finance*. Springer Verlag.

[16] Kotz, S. & Nadarajah, S. (2000). *Extreme Value Distributions – Theory and Applications*. Imperial College Press, London.

[17] Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer.

[18] Beirlant, J., Goegebeur, Y., Segers, J. & Teugels, J. (2004). *Statistics of Extremes. Theory and Applications*. Wiley.

[19] Castillo E., Hadi A.S., Balakrishnan, N. & Sarabia, J.M. (2004). *Extreme Value and Related Models with Applications in Engineering and Science*. Wiley.

[20] de Haan, L. & Ferreira, A. (2006). *Extreme Value Theory: an Introduction*. Springer Science+Business Media, LLC, New York.

[21] Resnick, S. (2007). *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer Verlag.

[22] Markovich, N. (2007). *Nonparametric Analysis of Univariate Heavy-Tailed Data: Research and Practice*. Wiley.

em teoria da fiabilidade). Tais valores podem ainda dar-nos indicações sobre 'outliers', indicando influências estranhas ou erros no processo de colecção de dados.

- A amplitude da amostra é uma medida importante de escala.
- A mediana da amostra é uma medida importante de localização.
- A própria forma de um problema pode censurar um conjunto de dados, e somos muitas vezes forçados a trabalhar com um conjunto de e.o.'s superiores ou e.o.'s inferiores.

Tudo isto é *inevitável e natural*.

Alternativamente, um conjunto de observações pode ser *deliberadamente ordenado*, de forma a facilitar a análise estatística pretendida. Por exemplo, podemos estar interessados em:

- Estimadores centrados, de variância mínima, que sejam combinações lineares das e.o.'s.
- Métodos rápidos para estimação de parâmetros ou para testes de significância baseados em estatísticas sistemáticas, como a amplitude e a semi-amplitude.
- Eliminar valores extremos de forma a aumentar a robustez de um estimador, com a consideração de estimadores como a média *Winsorizada* e a *trimmed mean*.
- Usar a técnica de papel de probabilidade (PP) ou outras alternativas gráficas como o QQ-plot para validação de um modelo e estimação dos parâmetrosdesconhecidos, onde QQ é o acrónimo de *quantil versus quantil*.
- Usar a técnica dos resíduos ordenados em *Análise de Variância*.
- Metodologias de índole não paramétrica, onde a noção de ordem é fundamental.

De qualquer forma, e qualquer que seja a finalidade da ordenação dos dados, necessitamos obviamente do conhecimento prévio da teoria distribucional exacta, e frequentemente da teoria distribucional assintótica das e.o.'s.

No que se segue, vamos usar a notação usual X para uma variável aleatória (v.a.) genérica com função de distribuição (f.d.) F eventualmente dependente

de parâmetros desconhecidos de localização, $\lambda \in \mathbb{R}$, e de escala, $\delta \in \mathbb{R}^+$. Em situação univariada, a amostra original $(X_1, \ldots, X_n)$ é imediatamente ordenada ascendentemente e denotada $(X_{1:n} \leq \cdots \leq X_{n:n})$, com

$$X_{1:n} := \min_{1 \leq i \leq n} X_i \quad \text{e} \quad X_{n:n} := \max_{1 \leq i \leq n} X_i.$$

Usamos por vezes a notação $M_n^{(i)} := X_{n-i+1:n},\ 1 \leq i \leq n$, quando estamos interessados em e.o.'s superiores.

A notação Z é frequentemente usada para a v.a. standardizada, $Z = (X - \lambda)/\delta$, com a qual trabalhamos na maior parte das situações, por simplicidade e sem perda de generalidade, uma vez que $X_{i:n} = \lambda + \delta Z_{i:n}$, $1 \leq i \leq n$. Modelos de especial interesse no contexto das e.o.'s são o modelo Uniforme, o Exponencial e o Pareto. Usamos a notação óbvia para representar variáveis aleatórias (v.a.'s) provenientes desses modelos, U, E e P, respectivamente.

1.1 Tópicos a abordar

Neste livro, começamos por apresentar no Capítulo 2 alguma *Motivação* para a necessidade da EVT. No Capítulo 3 avançamos com algumas *Técnicas Gráficas* usadas na análise de valores extremos, tais como os QQ-plots, os PP-plots, os W-plots e os ME-plots. No Capítulo 4, através de alguns exemplos de aplicação a dados univariados nas áreas de ambiente, hidrologia, meteorologia e seguros, tentamos responder à pergunta *Porquê a Teoria de Valores Extremos*? No Capítulo 5 abordamos alguns resultados sobre a *Teoria Distribucional Exacta*, colocando-nos pois numa perspectiva probabilística. Referimos neste capítulo as distribuições exactas de e.o.'s, o cálculo dos seus momentos, a importância da sua estrutura markoviana, dando ainda algumas indicações sobre estatísticas sistemáticas e aproximações para os momentos. O Capítulo 6 incide sobre a *Teoria Distribucional Assintótica*, onde referimos as leis limite das e.o.'s centrais, extremais e intermédias, as leis limite estáveis para max e min–domínios de atracção, as condições necessárias e suficientes a impor nas caudas dos modelos subjacentes às amostras em estudo e os POT–domínios de atracção. São introduzidas neste capítulo a distribuição (geral) de valores extremos (GEV, do inglês '*general extreme value*' ou '*generalized extreme value*')

e a distribuição generalizada de Pareto (GP), bem como o índice de valores extremos (denotado EVI, do inglês '*extreme value index*') e a noção de peso de cauda, fortemente relacionada com a teoria das funções de variação regular (Bingham *et al.*[23], 1987). Finalmente, nos Capítulos 7, 8 e 9, debruçamo-nos sobre *Estatística de Extremos*, área de grande utilidade em aplicações quando se pretende inferir na cauda de um modelo, estimando parâmetros de acontecimentos raros, como por exemplo quantis elevados ou períodos de retorno de níveis elevados. No Capítulo 7, abordamos as perspectivas paramétricas de inferência estatística em acontecimentos raros e a escolha estatística de modelos extremais e de max-domínios de atracção, o chamado método dos máximos anuais (MMA), e entre outras, as metodologias POT (do inglês '*peaks over threshold*') e PORT (do inglês '*peaks over random threshold*'), muito úteis na inferência de acontecimentos extremos. O Capítulo 8 é dedicado a alguns métodos de inferência semi-paramétrica. No Capítulo 9, procedemos à análise de vários casos de estudo.

[23]Bingham, N., Goldie, C.M. & Teugels, J.L. (1987). *Regular Variation.* Cambridge Univ. Press, Cambridge.

Capítulo 2

Motivação

É perfeitamente natural perguntar qual o porquê da EVT. Para motivar o interesse por este tema, damos em seguida alguns exemplos recentes de grande relevância para a sociedade, e que envolvem esta teoria.

2.1 Katrina: Um desastre (não) natural?

Nova Orleães encontra-se situada abaixo do nível do mar, no meio de dois lagos, a Norte e a Este, e do rio Mississipi a sul. De acordo com as informações divulgadas pelas autoridades locais, a inundação provocada pelo Katrina deveu-se, sobretudo, a uma brecha de 60 metros num dique junto ao lago Pontchartrain.

Traduzimos de forma livre parte de uma notícia do *New York Times, Sept'05*, intitulada '*New Orleans After Hurricane Katrina: An Unnatural Disaster*?' Dizia o redator que o que teriam de fazer em seguida era construir um sistema de diques adequado, para o que necessitariam de engenheiros holandeses, capazes de desenhar essas estruturas. A primeira estrutura deveria ser uma barragem com pelo menos 40-50 pés de altura, construída ao longo do lago e de cada canal com ligação ao lago. Tratar-se-ia de um plano que custaria biliões, mas conseguir-se-ia assim que Nova Orleães NUNCA tornasse a en-

Figura 2.1: Nova Orleães após o furacão Katrina

frentar semelhante tragédia. E terminava esperando que se aprendesse a lição, de modo a não se ter uma repetição dentro dos próximos 20 anos.

Parece óbvio que não só este desastre, mas também cheias históricas, como a que aconteceu no Mar do Norte às primeiras horas da manhã do dia 1 de Fevereiro de 1953, podem servir de guia. Nesse dia, o nível das águas excedeu então os 5.6 metros acima do nível do mar, destruiu as defesas marítimas, tendo inundado áreas na Holanda, Inglaterra, Bélgica, Dinamarca e França e cerca de 2500 pessoas morreram. Como resultado, o governo holandês, constituiu uma comissão, designada '*Delta Committee*'. O governo decretou que os diques devem ser construídos com uma altura tal que

- **a probabilidade de uma inundação num determinado ano é de 1 em 10.000**.

Ora o período de observação dos dados é muitíssimo mais curto!... É então necessário proceder a uma ***extrapolação*** para além dos dados observados!! ...E a EVT consegue dar respostas fidedignas sobre a altura da referida barragem, entrando em linha de conta com aquilo a que chamamos período de retorno (conceito a ser definido mais adiante) de um acontecimento extremo, como o furacão Katrina.

Figura 2.2: A cheia no Mar do Norte a 1 de Fevereiro de 1953

2.2 Extremos no mercado financeiro

O Comité de Basileia sobre o controlo bancário formula normas e directrizes de supervisão e recomenda boas práticas para as instituições financeiras. Entre outras medidas de risco, essa regulamentação envolve a estimação de uma quantidade denominada *Value-at-Risk* (VaR), que não é mais do que um quantil extremo da distribuição de perdas e ganhos. Como poderá ser estimado o VaR a partir da série de retornos diários R_t (em percentagem), definidos por

$$R_t = 100 \log(P_t/P_{t-1}),$$

sendo P_t o preço de fecho no dia t? Para o PSI20, apresentamos na Figura 2.3 os valores P_t (*esquerda*) e R_t (*direita*).

A simulação histórica é muito pobre! E existem contribuições positivas da EVT, como se ilustra nas duas figuras seguintes, 2.4 (veja-se Araújo Santos[1], 2011) e 2.5. Um breve texto crítico sobre o cálculo do VaR feito através

[1] Araújo Santos, P. (2011). *Excesses, Durations and Forecasting Value-at-Risk*. PhD Thesis, University Lisbon.

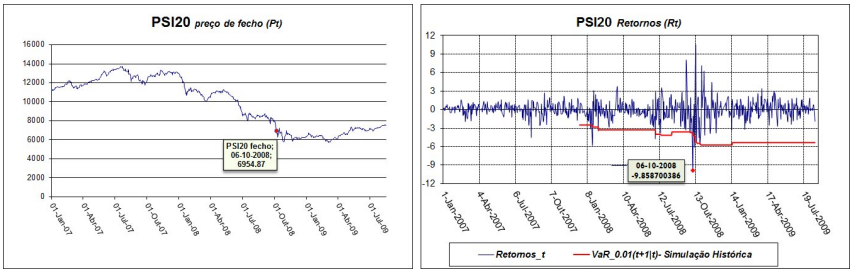


Figura 2.3: Preços de fecho (*esquerda*) e log-retornos diários (*direita*) do PSI20

das metodologias tradicionais (Normal-VaR) versus a EVT (EV-VaR), de que apresentamos a Figura 2.5, ilustrativa da comparação das duas metodologias, pode ser consultado em Aragonés *et al.*[2] (2000), um artigo introdutório em EVT.

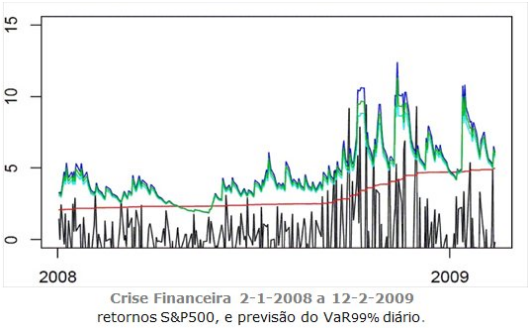


Figura 2.4: Previsão para o VaR99% diário do S&P500

As principais questões a ter em consideração são essencialmente as seguintes:

- Usualmente existem poucas observações na cauda da distribuição.
- São requeridas estimativas muito para além do máximo observado.
- Necessitamos de recorrer a modelos para a cauda baseados em resultados assintóticos.
- Será sensato usar esses modelos em todas as situações reais envolvendo acontecimentos raros?

[2] Aragonés, J., Blanco, C. & Dowd, K. (2000). The Learning Curve: Extreme Value Theory for VaR (`http://www.fea.com/resources/pdf/a_evt_1.pdf`Part 1 & `http://www.fea.com/resources/pdf/a_evt_2.pdf`Part 2).

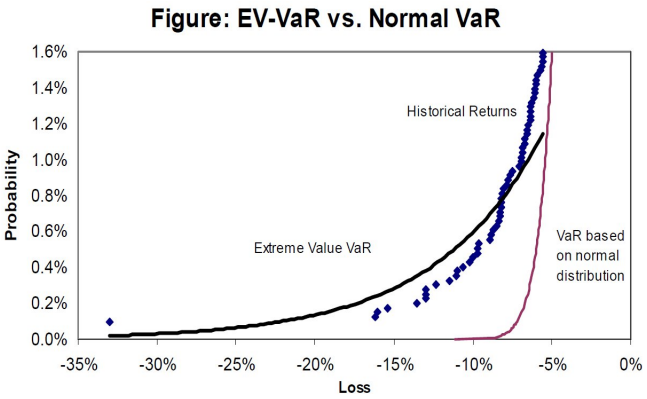


Figura 2.5: EV-VaR vs Normal-VaR

- É preciso não esquecer, parafraseando George Box (1919–2013), genro de Sir Ronald Fisher, que *'... all models are wrong but some models are useful'* (Box & Draper[3], 1987, p. 424).

Note-se desde já que as áreas de aplicação da EVT na análise de acontecimentos raros são tão diversas como o Ambiente, as Finanças, os Seguros, a Resistência de Materiais, o Desporto e a Sismologia, entre outras.

2.3 EVT: porque nem tudo é normal!

- De que altura deverá ser projectada uma barragem de aterro, de tal forma que o mar só atinja este nível uma vez em 1000 anos?
- Qual a probabilidade de rotura de determinado dique marítimo?
- Que ordem de grandeza poderá vir a atingir um '*crash*' bolsista amanhã?
- Qual a probabilidade de ser ultrapassada a melhor marca de 8.95m em salto em comprimento, dado o actual '*state of the art*'?

Muitas questões da vida real requerem a estimação sobre acontecimentos acerca dos quais os dados são inexistentes ou se existem são escassos — são

[3]Box, G.E.P. & Draper, N.R. (1987). *Empirical Model-Building and Response Surfaces.* Wiley.

os designados *acontecimentos extremos ou raros*. A EVT é um ramo probabilístico de suporte à Estatística que lida exactamente com tais situaçõoes, ajudando a descrever e a quantificar os ditos acontecimentos raros. Em particular, permite a estimação de probabilidades de acontecimentos que não contêm dados, ou como usualmente dizemos, permite *extrapolar para além da amostra*.

Quantidades relevantes são, entre outras, um *quantil extremo*, noção já atrás referida, e o *período de retorno*, que não é mais do que o intervalo de tempo médio entre ocorrências de um determinado valor extremo.

Na análise de dados clássica os extremos podem vir a ser rotulados de '*outliers*', chegando por vezes mesmo a ser ignorados no estudo, uma vez que se afastam do modelo 'ajustado'. Se o objectivo for inferir acerca de acontecimentos do dia-a-dia, realmente poderá ser irrelevante suprimir tais dados das pontas, mas se a questão fulcral residir em eventos que não ocorrem com muita frequência então dever-se-á aplicar o contexto EVT, dando relevância exactamente a esses valores extremos.

Existirá um padrão escondido subjacente a todo o tipo de eventos?

Se medirmos as alturas de muitas pessoas de um mesmo estrato homogéneo e as representarmos por um simples histograma, facilmente descobrimos uma mesma regra, a famosa curva de Gauss, por vezes também denominada distribuição em forma de sino, que não é mais do que a constatação de que o modelo Normal como que 'regula' a característica em causa. Surpreendentemente (*ou talvez não ...*) muitos dos dados da vida real seguem a distribuição Normal e suas congéneres.

Metodologias estatísticas mais comuns assentam no pressuposto de que os dados disponíveis correspondem a realizações independentes de v.a.'s provenientes de uma população com distribuição Normal. É o caso, por exemplo, do teste-t para comparação de valores médios. Outras abordagens usuais são essencialmente motivadas pela concepção mais ou menos consensual de que qualquer fenómeno que dê origem a um grande número de observações independentes, em que nenhuma delas é dominante, pode ser convenientemente modelado por uma distribuição Normal. A manifesta vantagem que daqui deriva é a da possibilidade de simplificar um elevado número de situações, decorrente de propriedades que surgem como apanágio da distribuição Normal,

nomeadamente a que resulta da aplicação do Teorema Limite Central (TLC) para somas ou a que deriva desta distribuição poder ser completamente especificada à custa dos seus momentos.

Contudo, quando nos focamos nos extremos, localizados nas *caudas das distribuições*, esta deixa de ser uma verdade irrefutável.

Contrariamente à condição de normalidade acima descrita, não é difícil deparar, no decurso da vida quotidiana, com situações em que uma única observação que se afasta da tendência central dos dados poderá, pela sua magnitude, ser comparável à acumulação de todas as outras não dominantes. É também neste sentido que as e.o.'s extremais têm protagonizado tão grande número de situações práticas em áreas tão diversas quanto Seguros, Finanças, Hidrologia, Biologia, Controlo de Qualidade, Telecomunicações ou Teletráfego, ao ponto de justificar o constante desenvolvimento da *Teoria de Valores Extremos*.

Por exemplo, no campo financeiro, e em particular nas distribuições associadas aos retornos, é habitual encontrar caudas mais pesadas do que as abordagens clássicas consideram. Isto quer basicamente dizer o seguinte: os acontecimentos extremos, embora improváveis por hipótese, são mais frequentes do que seria de esperar segundo o modelo gaussiano, um modelo com caudas leves, de tipo Exponencial.

Existem situações onde a abordagem EVT é primordial. A distribuição associada às maiores observações para aplicações a dados ou temperaturas anuais de pico, por exemplo. Por outro lado, a distribuição das menores observações é aplicada a problemas de resistência de materiais, onde o princípio do elo mais fraco impera, ou ainda a fenómenos como a duração da vida humana, com limite superior de suporte necessariamente finito.

Em *Estatística*, o *Teorema de Fisher-Tippett-Gnedenko* (o teorema fulcral dos tipos em valores extremos) é um resultado acerca da distribuição assintótica das e.o.'s extremais. O teorema dos tipos extremais desempenha um papel análogo ao tão famoso TLC para as médias (somas). Basicamente, estabelece que *o máximo amostral convenientemente normalizado converge para uma de* 3 *distribuições possíveis, a* Gumbel *a* Fréchet *ou a* Max-Weibull, a serem estudadas mais adiante, no Capítulo 7, mas abordadas em situações várias ao longo deste livro. Independentemente da forma do centro de distribuição, a *cauda assume formas sempre muito especiais* quando estamos suficiente-

mente longe na cauda. O crédito deste resultado é devido essencialmente a Gnedenko[4] (1943), embora versões anteriores tivessem sido estabelecidas por Fréchet[5] (1927) e Fisher & Tippett[6] (1928).

2.4 Estatísticos históricos na área de extremos

Começamos por referir Fisher e Tippet, na Figura 2.6, e em seguida Weibull, Gumbel e Fréchet, na Figura 2.7. Finalmente, na Figura 2.8, referimos von Mises (veja-se von Mises[7], 1936), outro dos pioneiros na área.

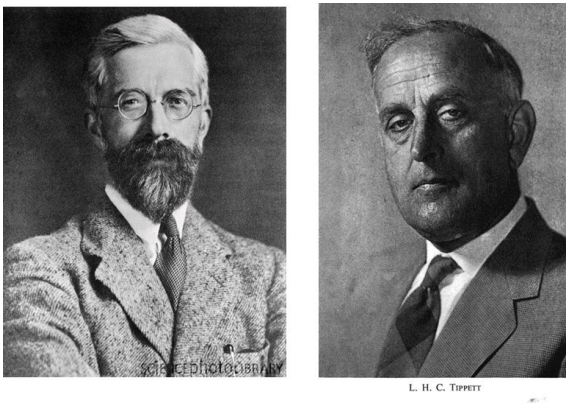


Figura 2.6: Sir Ronald Alymer Fisher (1890-1962) e Leonard Henry Caleb Tippett (1902-1985)

Referimos em seguida algumas das frases célebres de Emil Gumbel, um dos nome sonantes e pioneiros na área de *Estatística de Extremos*:

[4]Gnedenko, B.V. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *Annals of Mathematics* **44**:6, 423–453.

[5]Fréchet, M. (1927). Sur le loi de probabilité de l'écart maximum. *Ann. Société Polonaise de Mathématique* **6**, 93–116.

[6]Fisher, R.A. & Tippett, L.H.C. (1928). Limiting forms of the frequency distributions of the largest or smallest member of a sample. *Proceedings Cambridge Philosophical Society* **24**, 180–190.

[7]Mises, R. von (1936). La distribution de la plus grande de n valeurs. *Revue Math. Union Interbalcanique* **1**, 141–160. Reprinted in *Selected Papers of Richard von Mises*, Amer. Math. Soc. **2** (1954), 271–294.

Figura 2.7: Ernst Hjalmar Waloddi Weibull (1887-1979), Emil Julius Gumbel (1891-1966), e Maurice René Fréchet (1878-1973)

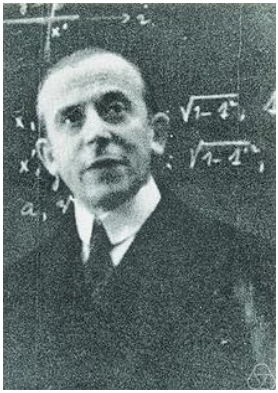


Figura 2.8: Richard Edler von Mises (1883-1953)

'It seems that the rivers know the theory. It only remains to convince the engineers of the validity of this analysis.'
'Il est impossible que l'improbable n'arrive jamais.'
'Il y aura toujours une valeur qui dépassera toutes les autres.'

Existem hoje em dia vários 'R-*Packages for Extreme Values*', tais como **evd**, **evdbayes**, **evir**, **ismev**, **extRemes**, **extremevalues**, **fExtremes**, **lmom**, **lmomRFA**, **lmomco**, **POT**, **SpatialExtremes**, alguns dos quais a serem usados neste livro.

Capítulo 3

Metodologias Gráficas para Análise Preliminar de Valores Extremos (APVE)

É óbvio que a *linearidade num gráfico* pode ser facilmente constatada por observação directa de uma *nuvem de pontos*, e quantificada em termos do *coeficiente de correlação*. A ideia subjacente aos PP-plots, ou equivalentemente aos actuais QQ-plots, existentes em quase todos os 'packages' estatísticos, e a estudar mais em pormenor nas secções 3.1 e 3.2, respectivamente, surgiu da necessidade de responder à pergunta:

Será que um determinado modelo probabilístico fornece um ajustamento sensato à distribuição subjacente aos dados em causa?

O método gráfico mais antigo para selecção de modelos é a técnica do *papel de probabilidade*, que introduzimos em seguida, e que pode ser visto com mais detalhe em Gomes *et al.*[1] (2010).

[1] Gomes, M.I., Figueiredo, F. & Barão, M.I. (2010). *Controlo Estatístico da Qualidade*. Edições INE.

3.1 Método Gráfico Clássico de Selecção de Modelos — Papel de Probabilidade (PP-plot)

A técnica do *papel de probabilidade* que, com modificações convenientes, pode ser usada para dados contínuos ou discretos, completos ou censurados, tem sido usada nas mais variadas formas, desde que Hazen[2] (1914) (veja-se também Hazen[3], 1930) sugeriu o princípio de linearização da f.d. Normal, num estudo de cheias, mas a sua principal aplicação tem sido na *obtenção de uma confirmação visual rápida do ajustamento de determinado modelo probabilístico, sugerido por exemplo pelo histograma, a dados* $(x_1, \ldots, x_n)$*, permitindo ainda a estimação grosseira de parâmetros.*

O *papel de probabilidade* é frequentemente usado quando os dados, $(x_1, \ldots, x_n)$, podem ser considerados observações independentes de uma v.a. X com f.d. do tipo $F((x-\lambda)/\delta)$, λ e δ parâmetros de localização e escala, respectivamente. Trata-se de um método de linearização da f.d.: face à amostra ordenada $(x_{1:n} \leq \cdots \leq x_{n:n})$, e para um modelo $F(x) = F((x-\lambda)/\delta)$, represente-se graficamente a nuvem de pontos:

$$(x_{i:n},\ y_i := F^{\leftarrow}(p_i)), \quad p_i := i/(n+1),\ 1 \leq i \leq n, \tag{3.1}$$

onde $F^{\leftarrow}$ denota a inversa generalizada de F, i.e.,

$$F^{\leftarrow}(x) := \inf\{y : F(y) \geq x\},\ 0 \leq x \leq 1. \tag{3.2}$$

Se o gráfico resultante mostrar que existe uma relação linear entre $x_{i:n}$ e y_i temos uma validação informal do modelo $F(\cdot)$, postulado. A intersecção com o eixo das abcissas e a inclinação da recta fornecem-nos então estimativas grosseiras de λ e δ.

Na realidade, admitindo que $F^{-1}(\cdot)$ existe, sendo $F^{-1}(x)$ o valor de y tal que $F(y) = x$, e escrevendo

$$p_i = F((x_{i:n} - \lambda)/\delta), \quad 1 \leq i \leq n,$$

[2] Hazen, A. (1914). Storage to be provided in impounding reservoirs for municipal water supply. *Trans. Amer. Soc. Civil Engrs.* **77**, 1539-1659.

[3] Hazen, A. (1930). *Flood Flows. A Study of Frequencies and Magnitudes.* Wiley.

tem-se

$$y_i = F^{-1}(p_i) = x_{i:n}/\delta - \lambda/\delta \iff x_{i:n} = \lambda + \delta y_i, \quad 1 \leq i \leq n,$$

i.e. existe uma relação linear entre $x_{i:n}$ e $y_i = F^{-1}(p_i)$, devendo p_i ser qualquer estimativa plausível de $F((X_{i:n} - \lambda)/\delta)$. Uma escolha possível para os valores de p_i, $1 \leq i \leq n$, as chamadas '*plotting positions*', foi dada em Weibull[4] (1939). Trata-se dos valores $p_i = i/(n+1)$, $1 \leq i \leq n$, já definidos em (3.1), os valores de $\mathbb{E}\big(F((X_{i:n} - \lambda)/\delta)\big)$, $\forall F(\cdot)$ absolutamente contínua, uma vez que então

$$F\Big(\frac{X_{i:n} - \lambda}{\delta}\Big) \stackrel{d}{=} B_{i,n-i+1},$$

onde $B_{p,q}$ denota uma v.a. Beta de parâmetros p e q, i.e., uma v.a. com f.d.p.,

$$f(z; p, q) = \frac{1}{B(p,q)}\, z^{p-1}(1-z)^{q-1}, \quad 0 \leq z \leq 1,$$

com $B(\cdot, \cdot)$ a função Beta completa.

Observação 3.1.1. *A função* Beta completa *é o integral,*

$$B(\alpha, \beta) := \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx = \frac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \quad \alpha, \beta \in \mathbb{R}^+, \tag{3.3}$$

com $\Gamma(\cdot)$ *a função* Gama (factorial) completa, *i.e.,*

$$\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1}e^{-x}dx, \quad \alpha \in \mathbb{R}^+. \tag{3.4}$$

A função Gama *é a extensão da função* factorial *a qualquer real positivo* (*na realidade, a função* Gama *pode mais geralmente ser definida para qualquer complexo* z*, cuja parte real seja positiva — para detalhes, veja-se por exemplo o* Capítulo 6 *de* Abramowitz & Stegun[5]*,* 1972). *Para valores de* $n \in \mathbb{N}_0$*, inteiro não negativo, temos*

$$\Gamma(n+1) = 1 \times 2 \times 3 \times \cdots \times (n-1) \times n = n!, \quad (0! \equiv 1).$$

Não podemos aqui deixar de referir uma relação de recorrência relativa à função Gama *frequentemente utilizada ao longo deste livro,*

$$\Gamma(\alpha+1) = \alpha\,\Gamma(\alpha), \quad \alpha > 0.$$

[4] Weibull, W. (1939). *A Statistical Theory of Strength of Materials.* Ing. Vet. A.K, Handl., 151, Genelstabens Litografiska Anstals Forlg Stocklholm, Sweden.

[5] Abramowitz, M. & Stegun, I.A. (1992). *Handbook of Mathematical Functions.* Dover, New York.

Valores particulares importantes são

$$\Gamma\left(\tfrac{1}{2}\right) = 2\int_0^\infty e^{-t^2}dt = \sqrt{\pi} = 1.77245\,38509\ldots, \quad \Gamma(\tfrac{3}{4}) = 1.22541\,67024\ \ldots.$$

Podemos ainda dizer que em modelo F absolutamente contínuo se tem

$$\mathbb{E}\left(F\Big(\frac{X_{i:n}-\lambda}{\delta}\Big)\right) = \mathbb{E}(B_{i,n-i+1}) = \mathbb{E}(U_{i:n}) = \frac{i}{n+1}, \quad 1 \le i \le n.$$

Para outras possíveis escolhas de *plotting positions* em papel de probabilidade veja-se, por exemplo, Barnett[6] (1975).

Tal como já foi referido, se o gráfico resultante mostrar que existe uma relação linear entre $x_{i:n}$ e $y_i = F^{-1}(i/(n+1)) =: Q(i/(n+1))$, com $Q(\cdot)$ a função quantil, temos uma validação informal da forma da distribuição $F(\cdot)$, postulada. A intersecção com o eixo das abcissas e a inclinação da recta fornecem-nos então estimativas grosseiras de λ e δ. Caso exista linearidade, a estimação dos parâmetros pode então ser feita através do módulo de regressão de qualquer *package* estatístico.

3.1.1 Referência histórica aos papéis de probabilidade

Como o processo pretendia ser um método visual rápido, tornava-se importante a facilidade da sua aplicação, tendo sido produzidos tipos especiais variados de *papel de probabilidade*, com uma escala funcional que mede convenientemente $F^{-1}(p)$, mas que é graduada em p. Torna-se então unicamente necessário representar graficamente $x_{i:n}$ versus p_i na(s) escala(s) transformada(s). Com a acessibilidade a algoritmos computacionais de cálculo de $F^{-1}(\cdot)$ para uma grande variedade de modelos, estes *papéis de probabilidade* têm hoje em dia apenas **interesse histórico**.

O exemplo mais vulgar, ainda muito usado em aplicações diversas, particularmente em áreas de *Hidrologia* e *Climatologia Estatística*, é o *papel de probabilidade Normal* (Chernof & Lieberman[7], 1954), acessível em qualquer papelaria do Reino Unido pelo menos até meados dos anos 80, e representado graficamente na Figura 3.1.

[6] Barnett, V. (1975). Probability plotting methods and order statistics. *Applied Statistics* **24**, 95–108.

[7] Chernoff, H. & Lieberman, G.J. (1954). Use of normal probability paper. *J. Amer. Statist. Assoc.* **49**, 778–785.

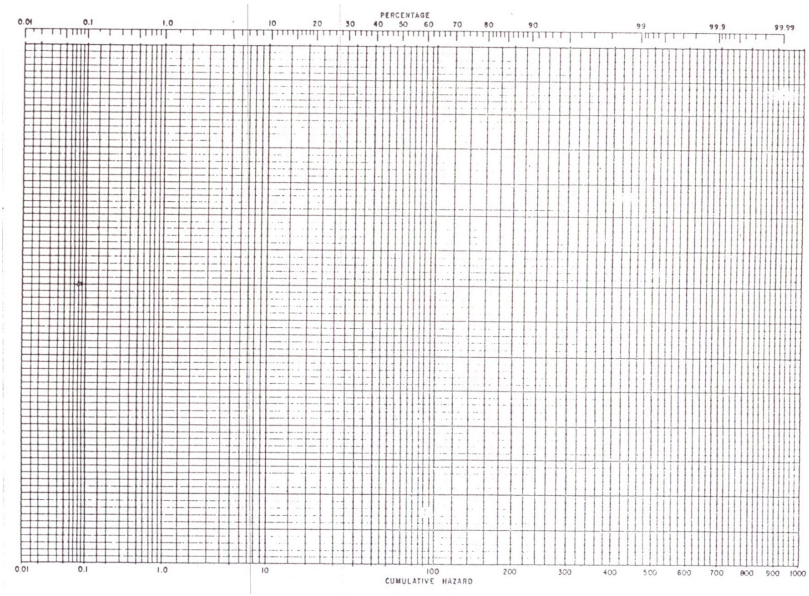


Figura 3.1: Papel de probabilidade Normal

No *papel de probabilidade Normal*, ilustrado na Figura 3.1, uma das escalas (neste caso, a das ordenadas) é uma escala aritmética, em que se marcam as observações ordenadas. A outra escala (a das abcissas) é uma escala probabilística, graduada em $\Phi^{-1}(p)$, com $\Phi(\cdot)$ a f.d. da Normal reduzida, $\mathcal{N}(0,1)$, mas em que se marca p (ou $100 \times p$). Esta escala funcional aparece ilustrada na Figura 3.2, e tem a seguinte tabela associada:

p	0.001	0.01	0.1	0.2	0.3	0.4	0.5	0.6	0.7	...
$\Phi^{-1}(p)$	-3.09	-2.33	-1.28	-0.84	-0.52	-0.25	0.0	0.25	0.52	...

Se as observações recolhidas forem efectivamente $\mathcal{N}(\lambda, \delta)$ teremos um gráfico do tipo do ilustrado na Figura 3.3.

Ao ajustarmos uma recta aos pontos marcados em papel de probabilidade Normal, obtemos facilmente estimativas de λ e δ, dadas por

$$\begin{aligned} \lambda^* &= \text{abcissa no ponto } 50\%, \\ \delta^* &= \tfrac{1}{2} \text{ (diferença entre as abcissas dos pontos } 84\% \text{ e } 16\%). \end{aligned}$$

Note-se no entanto que se para dados normais não usarmos *papel de probabi-*

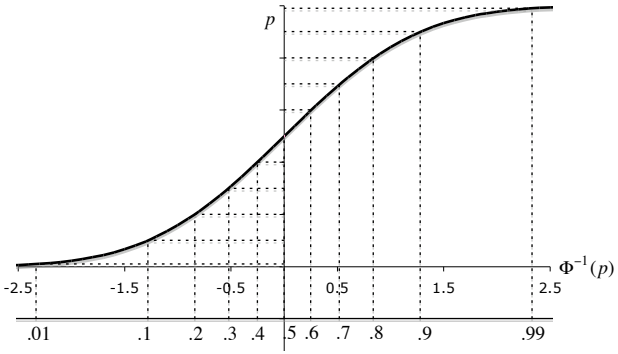


Figura 3.2: Escala funcional de um papel de probabilidade Normal

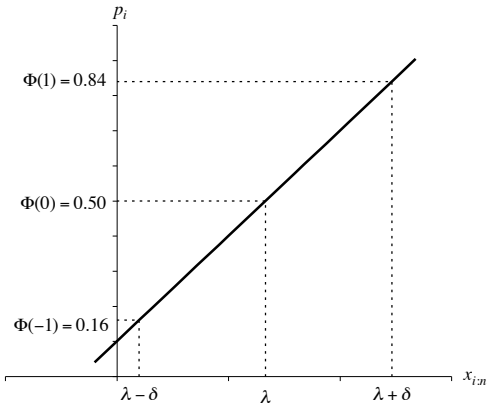


Figura 3.3: Gráfico em papel de probabilidade Normal

lidade, mas fizermos um gráfico de $\left(\Phi^{-1}(i/(n+1)), x_{i:n}\right)$, $1 \leq i \leq n$, com Φ a f.d. da $\mathcal{N}(0,1)$ (o actualmente chamado QQ-plot), como se tem

$$x_{i:n} = \lambda + \delta \ \Phi^{-1}\Big(\frac{i}{n+1}\Big),$$

obtemos as estimativas:

λ^{**} = intersecção com o eixo dos $x_{i:n}$ (ordenadas),

δ^{**} = inclinação da recta ajustada.

Exemplo 3.1.1. *Utilizando o* `package R`*, gerámos* 250 *observações de um modelo* $\mathcal{N}(3,1)$*. Temos para isso disponível a função* `rnorm`*. O gráfico da* Figura 3.4*, à esquerda, está associado às instruções:*

```
> x <- rnorm(250,3,1)
> x_in <- sort(x)
> n <- length(x)
> p_i <- (1:n)/(n+1)
> y_i <- qnorm(p_i)
> plot(y_i, x_in, col="blue", xlab=expression(y[i]), ylab=expression(x[i:n]),cex.lab=1.2)
> res1 <- lm(x_in ~ y_i)
> abline(res1, col="red",lty=1,lwd=2)
> legend(0,1,c("lmline"), col="red", lty=1,lwd=2, bty="n", cex=1)
```

A recta `abline` *está relacionada com a regressão linear, e o método de mínimos quadrados.*

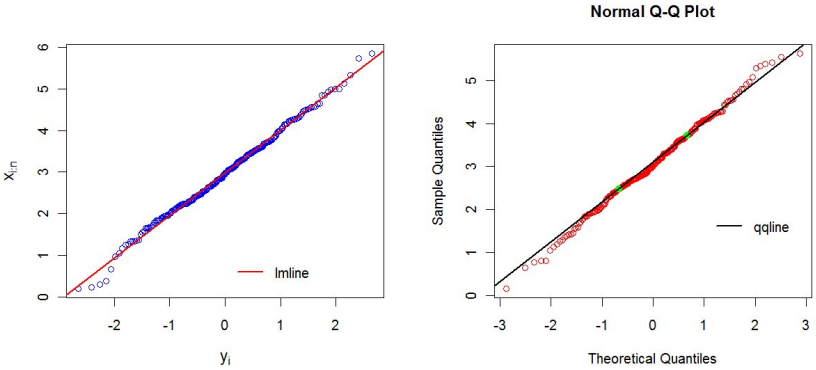


Figura 3.4: Papel de probabilidade (*esquerda*) e QQ-plot (*direita*) normais

Obtemos um gráfico semelhante, com a função `qqnorm`*, que permite o traçado do chamado* QQ-plot Normal*, ao qual ajustámos também a recta* `qqline`*. No caso do* QQ-plot Normal*, a função* `qqline` *permite o ajustamento de uma recta que passa pelo* 1^o *e* 3^o *quartis.*

O gráfico da Figura 3.4*, à direita, foi obtido através das instruções:*

```
> qqnorm(x, col="red")
> points(c(qnorm(0.25),qnorm(0.75)),
+ quantile(x,c(.25,.75)),col="green",cex=1.1,bg="green",pch=21)
```

```
> qqline(x, lwd=1.8)
> legend(1,2,c("qqline"), col="black", lty=1,lwd=2, bty="n", cex=1)
```

Foi ainda feita uma análise dos resíduos associados à regressão linear, que nos fornece estimativas para λ e δ. O 'output' foi o seguinte:

```
>  summary(res1)
Call: lm(formula = x_in ~ y_i)
Residuals:
 Min                    1Q            Median          3Q              Max
-0.725256   -0.037243   0.006295   0.043307   0.208526
Coefficients:
                      Estimate     Std. Error   t value   Pr(>|t|)
(Intercept)   2.983100=: lambda**   0.005704    523.0     <2e-16     ***
y_i           1.031064=: delta**    0.005814     77.3     <2e-16     ***
---
Signif. codes:  0'***' 0.001'**' 0.01'*' 0.05'.' 0.1' ' 1

Residual standard error: 0.09019 on 248 degrees of freedom
Multiple R-squared: 0.9922,Adjusted R-squared: 0.9921
F-statistic: 3.145e+04 on 1 and 248 DF,  p-value: < 2.2e-16
```

Outro exemplo simples é o *papel de probabilidade* Gumbel, um modelo muito usual em *Estatística de Extremos*, como veremos mais adiante.

Exemplo 3.1.2. *Se $F \equiv \Lambda$, a f.d.* Gumbel*, dada por:*

$$\Lambda(x;\lambda,\delta) = e^{-e^{-(x-\lambda)/\delta}},\ x \in \mathbb{R} \quad \Longrightarrow \quad x_{i:n} = \lambda + \delta(-\log(-\log(p_i))).$$

Consequentemente, o papel de probabilidade Gumbel *terá uma escala aritmética (onde marcamos as observações ordenadas ascendentemente, $x_{i:n}$, $1 \leq i \leq n$), versus uma escala duplamente logarítmica (onde marcamos as 'plotting positions', $p_i = i/(n+1),\ 1 \leq i \leq n$).*

Do ponto de vista conceptual, é óbvio que também podemos marcar $x_{i:n}$ versus $y_i = -\log(-\log(i/(n+1)))$ (ou $y_i = -\log(-\log(i/(n+1)))$ versus $x_{i:n}$), $1 \leq i \leq n$, num papel milimétrico usual, ou utilizar um QQ-plot*. É aliás isto que se faz, quando possuímos facilidades computacionais, como o package* `R`*, entre outros, a situação usual nos dias de hoje.*

A geração de observações ou números pseudo-aleatórios (NPA's) Gumbel *é simples. Procedemos pois à geração de* 250 NPA's Gumbel(0,1)*, colocados no vector* `gumb`*. O gráfico da* Figura 3.5 *(esquerda) foi obtido através dos comandos:*

```
> qqnorm(gumb, col="red")
> points(c(qnorm(0.25),qnorm(0.75)),
+ quantile(gumb,c(.25,.75)),col="green",cex=1.1,bg="green",pch=21)
> qqline(gumb, lwd=1.8)
> legend(1,0,c("qqline"), col="black", lty=1,lwd=2, bty="n", cex=1)
```

Trata-se pois de um QQ-plot Normal*, que fornece indicação imediata da não-normalidade dos dados. O traçado da nuvem de pontos* $\big(y_i = -\log(-\log(i/(n+1))), x_{i:n}\big)$*,* $1 \leq i \leq n$*, forneceu o gráfico da* Figura 3.5 (*direita*)*, e foi obtido de forma análoga ao que fizemos anteriormente para os dados de uma* $\mathcal{N}(3,1)$*.*

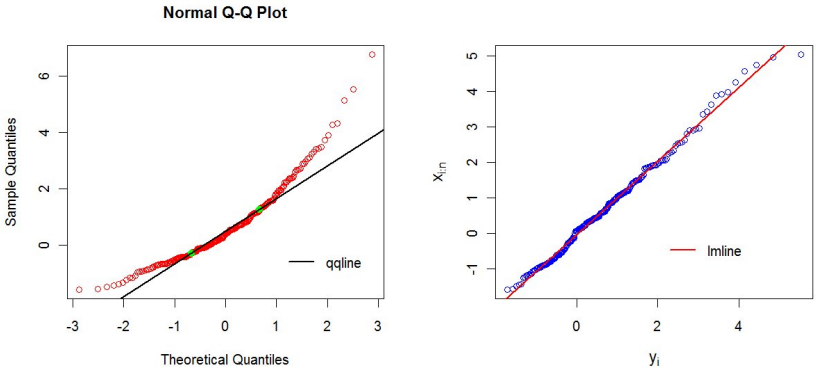


Figura 3.5: Dados Gumbel em '*papel de probabilidade*' Normal (*esquerda*) e Gumbel (*direita*)

A recta dos mínimos quadrados fornece-nos as estimativas,

$$\lambda^{**} = -0.102893, \qquad \delta^{**} = 0.993790.$$

Quando o gráfico em papel de probabilidade é nitidamente não linear, resultando consequentemente a rejeição do modelo postulado, $F(\cdot)$, podemos obter informação adicional a partir do gráfico (para mais detalhes veja-se Bury[8], 1975; Gomes *et al.*, 2010).

[8] Bury, K.V. (1975). *Statistical Models in Applied Science*. Wiley.

3.2 QQ-plots: outra perspectiva equivalente

3.2.1 QQ–plot: modelo Exponencial

Em *Estatística de Extremos*, e como veremos mais adiante nos Capítulos 6, 7–9, o modelo Exponencial de parâmetro $\lambda > 0$, denotado por $\mathcal{E}(\lambda)$, desempenha um papel bem mais importante do que o modelo Normal. A função de sobrevivência (ou cauda) é para a $\mathcal{E}(\lambda)$,

$$1 - F_\lambda(x) := \exp(-\lambda x),\ x > 0,$$

e para o caso da Exponencial standard ou reduzida,

$$1 - F_1(x) := \exp(-x),\ x > 0.$$

Será que a distribuição subjacente às observações $x_1, \ldots, x_n$ pertence a esta família $\mathcal{E}(\lambda)$? Para este modelo, temos a *função quantil*,

$$Q_\lambda(p) = F^{\leftarrow}(p) = -\frac{1}{\lambda}\log(1-p),\ p \in (0,1),$$

e para a $\mathcal{E}(1)$ temos pois a função quantil,

$$Q_1(p) = -\log(1-p),\ p \in (0,1).$$

Existe então uma relação linear

$$Q_\lambda(p) = \frac{1}{\lambda}Q_1(p) = \frac{1}{\lambda}(-\log(1-p)),\ p \in (0,1).$$

Dada uma amostra de observações $(x_1, \ldots, x_n)$, substitua-se $Q(p)$ pela contrapartida empírica $\widehat{Q}_n(p)$, e represente-se num sistema de eixos ortogonais a nuvem de pontos

$$(-\log(1-p), \widehat{Q}_n(p)),\ \text{para valores de}\ \ p \in (0,1).$$

Se o modelo Exponencial for bem ajustado, espera-se que a nuvem de pontos se distribua ao longo de uma recta. O declive dessa recta pode ser identificado com $1/\lambda$ e usado para obtenção de uma estimativa preliminar de λ. Note-se que a ordenada na origem deverá ser nula, já que $Q(0) = 0$.

De um modo geral, e denotando por $x_{i:n}$ o *i-ésimo* valor amostral, a nuvem de pontos

$$\widehat{Q}_n(p) = x_{i:n}, \text{ para } \frac{i-1}{n} < p \leq \frac{i}{n},$$

deve ser aproximadamente linear. Tal como mencionámos atrás, são possíveis várias escolhas de p (*plotting positions*), de entre as quais referimos:

$$p \in \left\{ \tfrac{1}{n}, \tfrac{2}{n}, \ldots, \tfrac{n-1}{n}, 1 \right\},$$

$$p \in \left\{ \tfrac{1-.5}{n}, \tfrac{2-.5}{n}, \ldots, \tfrac{n-1-.5}{n}, \tfrac{n-.5}{n} \right\},$$

$$p \in \left\{ \tfrac{1}{n+1}, \tfrac{2}{n+1}, \ldots, \tfrac{n-1}{n+1}, \tfrac{n}{n+1} \right\},$$

sendo usual considerar as *plotting positions* definidas em (3.1), i.e., $p_i = i/(n+1),\ 1 \leq i \leq n$, que, tal como já foi dito anteriormente, podem ser consideradas como o posicionamento dado pelo valor médio das e.o.'s associadas ao modelo Uniforme, $\mathcal{U}(0,1)$, uma vez que $F(X_{i:n}) \stackrel{d}{=} U_{i:n},$ para $i = 1, \ldots, n$, e se tem $U_{i:n} \frown \text{Beta}(i, n-i+1)$ de valor médio $\mathbb{E}[U_{i:n}] = i/(n+1) =: p_i$.

Voltemos ao caso Exponencial. A recta (declive=a; ordenada na origem=0) ajustada à nuvem de pontos pelo *método dos mínimos-quadrados*, é obtida pela minimização de

$$\sum_{i=1}^{n} \left(x_{i:n} + a \, \log(1-p_i)\right)^2 ,$$

vindo

$$\hat{a} = \frac{\sum_{i=1}^{n} x_{i:n} q_i}{\sum_{i=1}^{n} q_i^2}, \quad \text{com } q_i := -\log(1-p_i), \quad i = 1, \ldots, n\,.$$

Para um conjunto de 1000 NPA's, $\mathcal{E}(\lambda)$, com $\lambda = 2$, gerados no `R`, o gráfico é o apresentado na Figura 3.6. O declive da recta, com ordenada na origem nula, é de $0.5091 = 1/\hat{\lambda}$ e consequentemente, lembrando a relação teórica $Q_\lambda(p) = \frac{1}{\lambda}(-\log(1-p))$, temos uma estimativa preliminar dada por $\hat{\lambda} = 1.9643$.

Vejamos uma outra interpretação: A função que se está a aproximar ao marcar a nuvem de pontos $x_{i:n} \leftrightsquigarrow -\log(1-p_i),\ i = 1, \ldots, n$, é $x \mapsto -\log(1-F(x))$. Esta é exactamente a transformação que converte qualquer v.a. X, com f.d.

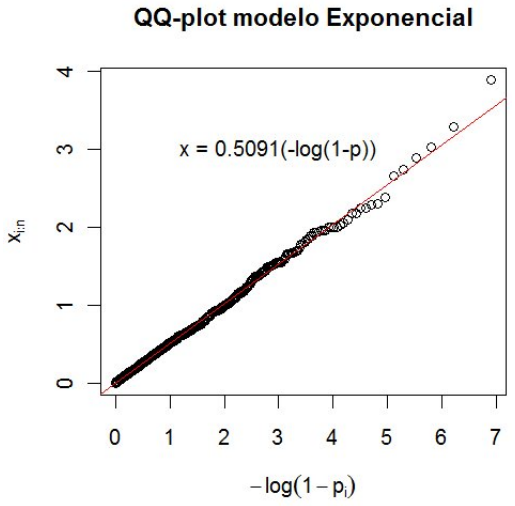


Figura 3.6: QQ-plot Exponencial

contínua F, na $\mathcal{E}(1)$. Realmente

$$\begin{aligned}\mathbb{P}[-\log(1-F(X)) \leq x] &= \mathbb{P}[X \leq Q(1-\exp(-x))] \\ &= F(Q(1-\exp(-x)) = 1-\exp(-x)\end{aligned}$$

ou seja,

$$-\log(1-F(X)) \frown \mathcal{E}(1).$$

Pensemos agora nas observações acima de um nível t (método POT, do inglês *peaks over threshold*). Na realidade, muitas vezes os dados só estão disponíveis acima de um nível t. Por exemplo, uma Resseguradora pode só receber informação acerca de pedidos de indemnização acima de um nível/franquia t, elevado. Abordemos o caso de X ser $\mathcal{E}(\lambda)$, e condicionemos no acontecimento $\{X > t\}$,

$$\mathbb{P}[X > x | X > t] = \frac{\mathbb{P}[X > x]}{\mathbb{P}[X > t]} = \exp(-\lambda(x-t)), \text{ para } x > t,$$

pelo que a correspondente *função quantil* é

$$Q(p) = t - \frac{1}{\lambda}\log(1-p),\ 0 < p < 1\,.$$

Então o QQ-plot tem ordenada na origem igual a t.

Como estimar um **quantil extremal**

$$q_p := Q(1-p), \text{ com } p \text{ pequeno ?}$$

Se pelo QQ-plot é sensato assumir o modelo Exponencial, então

$$\hat{q}_p = t - \frac{1}{\hat{\lambda}} \log(p).$$

Inversamente, uma **probabilidade de excedência pequena**

$$p \equiv p_x := \mathbb{P}[X > x | X > t]$$

pode ser estimada por

$$\hat{p}_x = \exp\left(-\hat{\lambda}(x-t)\right).$$

Estimação preliminar de λ: Poder-se-á estimar λ a partir do QQ-plot, através do método dos *mínimos quadrados*, ou alternativamente, considerar o estimador de *máxima verosimilhança* (ML, do inglês '*maximum likelihood*'),

$$\hat{\lambda} = 1/\big(\overline{x} - t\big).$$

3.2.2 QQ–plot: caso geral

No caso geral, seja Q_s a função quantil para o modelo standard de uma determinada família. De forma a aceitarmos um modelo como plausível para a população subjacente à amostra:

1. Deve existir uma relação linear entre os quantis teóricos $Q(p)$ e $Q_s(p)$.
2. Os quantis teóricos $Q(p)$, desconhecidos, devem ser substituídos pelos quantis empíricos $\widehat{Q}_n(p)$.
3. Devemos pois representar graficamente a nuvem de pontos

$$\left\{\left(Q_s(\tfrac{i}{n+1}), \widehat{Q}_n(\tfrac{i}{n+1})\right) = (Q_s(p_i), x_{i:n}) : i = 1, \ldots, n\right\}.$$

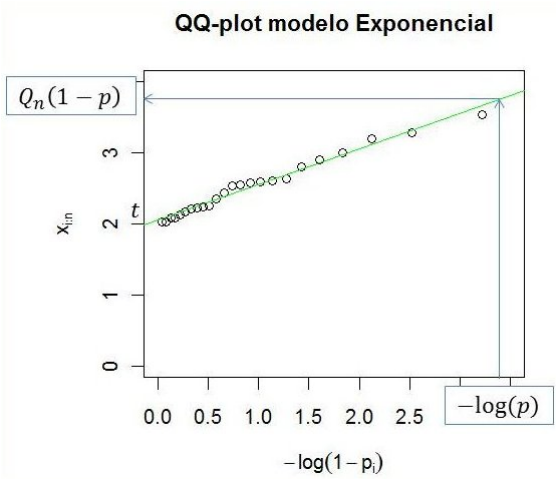


Figura 3.7: QQ-plot para dados $\mathcal{E}(\lambda)$ e estimação de quantis

4. Finalmente, devemos investigar a linearidade, levando a cabo uma regressão linear no QQ-plot, por exemplo.

Os quantis e períodos de retorno podem ser estimados através da aceitação de um relação linear no QQ-plot, $y = \hat{b} + \hat{a}x$, com

$$\overline{q} = \frac{1}{n}\sum_{i=1}^{n} Q_s(p_i), \quad \hat{a} = \frac{\sum_{i=1}^{n}(x_{i:n} - \overline{x})Q_s(p_i)}{\sum_{i=1}^{n}(Q_s(p_i) - \overline{q})^2} \quad \text{e} \quad \hat{b} = \overline{x} - \hat{a}\overline{q}\,.$$

Com F_s a f.d. reduzida e $Q_s = F_s^{\leftarrow}$ a inversa generalizada de F_s, $q_p = Q(1-p)$ e $p_x = \mathbb{P}[X > x]$, tem-se:

- **Estimação de quantis extremais:** $\hat{q}_p = \hat{b} + \hat{a}Q_s(1-p)$.
- **Probabilidades de excedência:** $\hat{p}_x = \overline{F}_s((x-\hat{b})/\hat{a})$, $\overline{F}_s := 1 - F_s$.

3.2.3 QQ–plots para modelos Normal e Log-Normal

Uma vez que os quantis da Normal, $\mathcal{N}(\mu, \sigma)$, se relacionam com os quantis da Normal *standard*, $\mathcal{N}(0,1)$, através de

$$Q(p) = \mu + \sigma\Phi^{-1}(p), \;\; \Phi^{-1} \text{ função quantil da } \mathcal{N}(0,1)\,,$$

as coordenadas do QQ-plot para este modelo são

$$\left(\Phi^{-1}(p_i), x_{i:n}\right), \ i = 1, \ldots, n\,.$$

Uma vez que a transformação logarítmica da Log-Normal é uma Normal, as coordenadas do QQ-plot para o modelo Log-Normal são obtidos pela transformação logarítmica dos dados

$$\left(\Phi^{-1}(p_i), \log x_{i:n}\right), \ i = 1, \ldots, n\,.$$

3.2.4 QQ–plot: Tabela de distribuições

A Tabela seguinte foi retirada de Beirlant *et al.* (2004):

Table 1.1 QQ-plot coordinates for some distributions.

Distribution	$F(x)$	Coordinates
Normal	$\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(u-\mu)^2}{2\sigma^2}\right) du$ $x \in \mathbb{R}; \mu \in \mathbb{R}, \sigma > 0$	$(\Phi^{-1}(p_{i,n}), x_{i,n})$
Log-normal	$\int_{0}^{x} \frac{1}{\sqrt{2\pi}\sigma u} \exp\left(-\frac{(\log u-\mu)^2}{2\sigma^2}\right) du$ $x > 0; \mu \in \mathbb{R}, \sigma > 0$	$(\Phi^{-1}(p_{i,n}), \log x_{i,n})$
Exponential	$1 - \exp(-\lambda x)$ $x > 0; \lambda > 0$	$(-\log(1 - p_{i,n}), x_{i,n})$
Pareto	$1 - x^{-\alpha}$ $x > 1; \alpha > 0$	$(-\log(1 - p_{i,n}), \log x_{i,n})$
Weibull	$1 - \exp(-\lambda x^{\tau})$ $x > 0; \lambda, \tau > 0$	$(\log(-\log(1 - p_{i,n})),$ $\log x_{i,n})$

3.3 QQ-plots e PP-plots: caso geral $F(\cdot|\theta)$

Até agora a maior parte dos modelos considerados permitiram a construção dos QQ-plots *sem qualquer conhecimento dos valores exactos dos parâmetros*. Aliás, *estimativas preliminares* desses valores puderam ser obtidas como resultado colateral do QQ-plot. Esta situação está essencialmente relacionada com

o caso de estarmos a lidar com modelos com *localização/escala*, para os quais a ordenada na origem representa a *localização*, estando o declive relacionado com a *escala*.

Mais geralmente, o *papel de probabilidade* usa-se quando $F(x_{i:n}, \underline{\theta})$, com $x_{i:n}$ a i-ésima estatística ordinal (e.o.) ascendente associada à amostra $(x_1, \ldots, x_n)$, e $\underline{\theta}$ vector de parâmetros desconhecidos, pode ser transformada numa relação linear, i.e. existem funções $g_i(\cdot),\ i = 1, 2, 3, 4$ tais que

$$g_1[F(x_{i:n}, \underline{\theta})] = g_2(\underline{\theta}) + g_3(\underline{\theta})\ g_4(x_{i:n}),$$

onde $F(x_{i:n}, \underline{\theta})$, desconhecido, é substituído por uma sua estimativa plausível, como por exemplo $p_i = i/(n+1)$, sempre que $F(\cdot)$ for absolutamente contínua (Chernoff & Lieberman[9], 1956). Têm-se pois gráficos do tipo do apresentado na Figura 3.8.

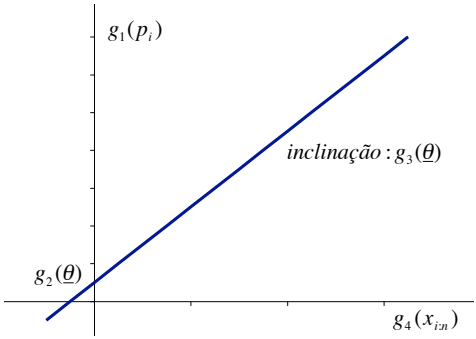


Figura 3.8: Aspecto possível de um papel de probabilidade genérico

Isto significa que o *papel de probabilidade* pode ser facilmente usado quando queremos testar informalmente a validade de uma população subjacente que, não dependendo apenas de parâmetros de localização e escala, pode ser transformada numa população com essas características, e o aspecto possível é o indicado na Figura 3.8. Vejamos alguns exemplos.

Exemplo 3.3.1. *Se estivermos interessados em validar informalmente um outro modelo muito comum em Estatística de Extremos, o modelo* Fréchet (*de*

[9] Chernoff, H. & Lieberman, G.J. (1956). The use of generalized probability paper for continuous distributions. *Ann. Math. Statist.* **27**, 806–818.

máximos) *com localização* $\lambda = 0$, *com f.d.*

$$F(x; 0, \delta, \alpha) = \exp(-(x/\delta)^{-\alpha}),\ x \geq 0,$$

temos

$$-\log(-\log(p_i)) = -\alpha \log \delta + \alpha \log x_{i:n},\ 1 \leq i \leq n,$$

i.e. um papel de probabilidade para esta população terá uma escala logarítmica (*onde marcamos* $x_{i:n}$), *sendo a outra escala duplamente logarítmica, a escala funcional* Gumbel.

Observação 3.3.1. *Na realidade o logaritmo de uma v.a.* Fréchet (*de máximos*) *tem distribuição* Gumbel (*de máximos*).

Exemplo 3.3.2. *Acontece algo semelhante para* X Log-normal, *com localização* $\theta = 0$, *pois então* $\log X$ *é* Normal, *tal como já foi referido. Um papel de probabilidade* Log-normal *terá uma escala logarítmica* (*onde marcamos* $x_{i:n}$), *sendo a outra escala a escala funcional* Normal. *Na situação actual, temos simplesmente de marcar a nuvem de pontos* $\left(\log x_{i:n}, \Phi^{-1}(i/(n+1))\right)$, $1 \leq i \leq n$, *ou, em* `R` *usar o* `qqnorm(y)` *com* $y = \log x$.

Neste caso geral, e mesmo que não seja viável proceder a uma transformação simples dos dados, os QQ-plots comparam os *dados ordenados* $X_{i:n}$ com os correspondentes *quantis da distribuição a ajustar*. Seja X uma v.a. com f.d. F_θ, com θ vector de parâmetros desconhecidos. Seja $(X_1, \ldots, X_n)$ uma amostra aleatória (a.a.) associada a X. Denotemos $\hat{\theta} \equiv \hat{\theta}(X_1, \ldots, X_n)$, um estimador consistente de θ. Então, podemos conceber um QQ-plot, em que se marcam os pontos $\left(F^{\leftarrow}_{\hat{\theta}}(p_i), X_{i:n}\right)$, $i = 1, \ldots, n$. O QQ-plot e o PP-plot são muito frequentemente definidos da mesma forma. Existe no entanto quem considere no PP-plot a marcação dos pontos $\left(F_{\hat{\theta}}(X_{i:n}), p_i\right)$, $i = 1, \ldots, n$.

3.4 W-plots: caso geral $F(\cdot|\theta)$

Tal como referido anteriormente, o princípio em que se baseia o PP-plot e o QQ-plot é a identificação

$$F_\theta(X_{i:n}) \stackrel{d}{=} U_{i:n}, \qquad i = 1, \ldots, n,$$

com $U_{i:n},\ i = 1, \ldots, n$ as e.o.'s associadas a uma a.a. da $\mathcal{U}(0,1)$. Então

$$-\log(1 - F_\theta(X_{i:n})) \stackrel{d}{=} E_{i:n}, \quad i = 1, \ldots, n,$$

com $E_{i:n}$ as e.o.'s associadas a uma a.a. de dimensão n da $\mathcal{E}(1)$.

O W-plot é outra representação gráfica, que decorre dos quantis da Exponencial. Marcamos

$$\big(-\log(1 - p_i), -\log(1 - F_{\hat{\theta}}(X_{i:n}))\big) \quad i = 1, \ldots, n\,,$$

e analisamos se a nuvem de pontos está razoavelmente perto da diagonal.

3.5 Função de excesso médio e ME-plot

Na prática actuarial, o condicionamento no acontecimento $\{X > t\}$ assume a maior importância, especialmente no *Resseguro*. Considere-se um tratado *excess-of-loss* com retenção t, para qualquer indemnização da carteira. O *Ressegurador* terá de pagar um montante aleatório $X - t$, *o excesso acima de* t, mas só se $X > t$. Tendo em vista o *cálculo do prémio*, o actuário pretende estabelecer uma *franquia* ou *nível* t, o chamado *t-threshold*, da metodologia POT, procedendo ao cálculo do *montante esperado a ser pago por cliente*, quando um dado nível t é escolhido.

Por exemplo, o actuário calcula a *função de excesso médio* (em inglês, '*mean excess function*'),

$$e(t) := \mathbb{E}\left[X - t | X > t\right], \tag{3.5}$$

assumindo que $\mathbb{E}[X] < \infty$.

3.5.1 ME-plots — *mean excess plots*

Na prática, a função de *excesso médio*, $e(\cdot)$, em (3.5), é substituída pela sua contrapartida empírica, $\hat{e}_n(\cdot)$, com base na amostra de dados observados $x_1, \ldots, x_n$, e definida por

$$\hat{e}_n(t) := \frac{\sum_{i=1}^n x_i I_{(t,+\infty)}(x_i)}{\sum_{i=1}^n I_{(t,+\infty)}(x_i)} - t\,, \text{ com } I_{(t,+\infty)}(x_i) := \begin{cases} 1 & x_i > t \\ 0 & \text{caso contrário.} \end{cases}$$

Usualmente, $\hat{e}_n$ é representada nos valores

$$t = x_{n-k:n},\ k = 1, \ldots, n-1,$$

ou seja utiliza-se a chamada metodologia PORT. Tem-se então,

$$\sum_{i=1}^{n} x_i I_{(t,+\infty)}(x_i) = \sum_{j=1}^{k} x_{n-j+1:n},$$

com $k \equiv \#\ x_i : x_i > t$ e as *estimativas dos excessos médios* dadas por

$$e_{k,n} := \hat{e}_n(x_{n-k:n}) = \frac{1}{k}\sum_{j=1}^{k} x_{n-j+1:n} - x_{n-k:n}.$$

3.5.2 Padrões das funções de excesso médio

Vamos em seguida investigar os padrões das *funções de excesso médio* associados a determinados modelos, i.e.

$$e(t) = \mathbb{E}\left[X - t | X > t\right] = \frac{\int_t^{x^F} (1 - F(u))du}{1 - F(t)}, \tag{3.6}$$

com $x^F := \sup\{x : F(x) < 1\} \leq \infty$, limite superior do suporte ou '*right endpoint*' de F.

Observação 3.5.1. *Note-se que, o numerador de* $e(t)$*, em* (3.6)*, é obtido pela inversão da ordem de integração* (Teorema de Fubini)

$$\begin{aligned}\int_t^{x^F} (x-t)dF(x) &= \int_t^{x^F} \left(\int_t^x du\right) dF(x) \\ &= \int_t^{x^F} \left(\int_u^{x^F} dF(x)\right) du = \int_t^{x^F} (1 - F(u))du.\end{aligned}$$

Mais uma vez a distribuição **Exponencial** desempenha um papel fulcral, devido por exemplo à ***propriedade de falta de memória da Exponencial***:

$$\begin{aligned}e(t) := \mathbb{E}[X - t | X > t] = \frac{\int_t^{x^F} (1 - F(u))du}{1 - F(t)} &= \frac{\int_t^{+\infty} e^{-\lambda u} du}{e^{-\lambda t}} \\ &= \frac{1}{\lambda}, \quad \forall t > 0,\end{aligned}$$

i.e., é *irrelevante o condicionamento em* $\{X > t\}$.

Genericamente, a ***forma*** *de* $e(\cdot)$, em (3.5) ou em (3.6), dá informação acerca de *caudas mais pesadas do que a da* Exponencial ou *mais leves que a da* Exponencial. Quando a distribuição de X tem cauda mais pesada do que a da Exponencial, a função de excesso médio $e(t)$ tende a exibir monotonia crescente para valores elevados de t. Na presença de caudas mais leves, a tendência desta função é no sentido decrescente.

3.5.3 Funções de excesso médio — modelo Weibull

O modelo Min-Weibull, ou simplesmente Weibull, tem função de sobrevivência,

$$\overline{F}(x) = 1 - F(x) = \exp(-\lambda x^{\tau}),\ x > 0,\ \tau > 0.$$

É relativamente fácil mostrar que

$$e(t) = \frac{t^{1-\tau}}{\lambda\tau}(1 + o(1))\,,$$

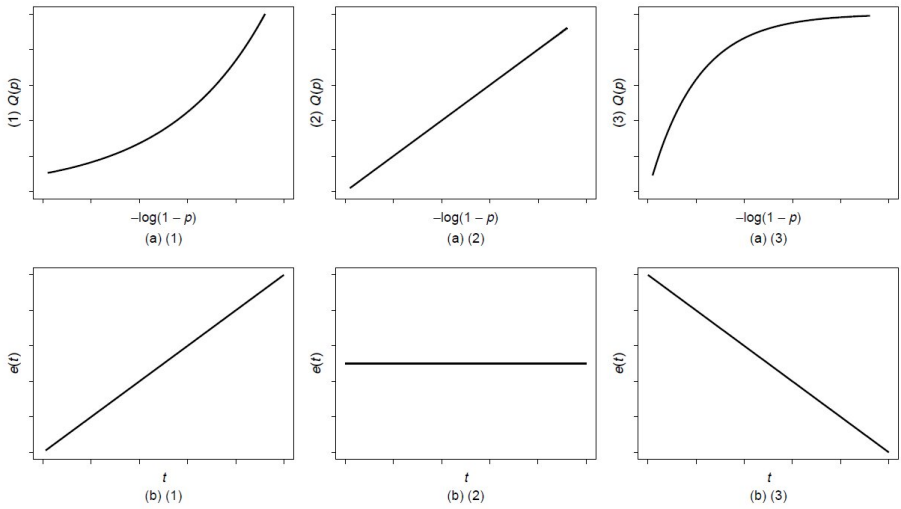
pelo que para grandes valores de t, $e(t)$ é crescente para $\tau < 1$, e decrescente para $\tau > 1$. Note-se que $\tau = 1$ **corresponde ao modelo Exponencial**, para o qual $e(t)$ é constante.

3.6 Caudas mais pesadas/leves (HTE/LTE) do que a Exponencial e caudas exponenciais

Na Figura 3.9 exibimos o aspecto dos QQ-plots e ME-plots para caudas HTE, do inglês '*heavier than exponential*', Exponencial e LTE, do inglês '*lighter than exponential*'.

3.7 Dados hidrológicos — parâmetros de interesse

O objectivo último na análise da frequência de cheias é a estimação do chamado 'T**-year flood discharge** (water level)', i.e., o nível das águas do caudal



(Beirlant *et al.*, 2004)

Figura 3.9: (a) QQ-plot Exponencial; (b) ME-plot para exemplos de distribuições com caudas tipo: (1) HTE, (2) Exponencial, e (3) LTE

do rio ultrapassado todos os T anos, em média.

Usualmente, toma-se para horizonte temporal $T = 100$, mas a estimação é levada a cabo tendo por base dados das descargas fluviais num período inferior.

Nos Países Baixos (Holanda e Bélgica), por exemplo, a exigência legal para a construção de diques exige que a sua altura é a que decorre de $T = 10^4$ para horizonte temporal de inundação, i.e., o nível ultrapassado apenas cada 10^4-anos, em média.

Outro fenómeno hidrológico em que tem interesse especial o estudo da cauda da distribuição é a intensidade da precipitação, dando atenção especial aos níveis de pluviosidade mais extremos.

3.7.1 Dados de máximos anuais

Muito frequentemente os dados disponíveis são de natureza periódica, em especial **Dados de Máximos Anuais**.

Alternativamente aos níveis T-anual (T–year water levels) a *análise de valores extremos* pode ser abordada em termos recíprocos através dos denominados **Períodos de Retorno**:

$$T(x) := \frac{1}{\mathbb{P}[Y > x]}\,,$$

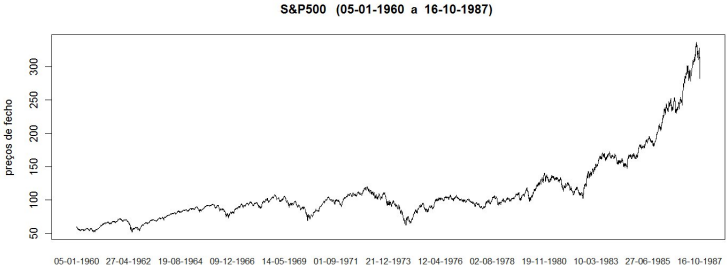
com Y a v.a. associada ao máximo periódico anual, por exemplo.

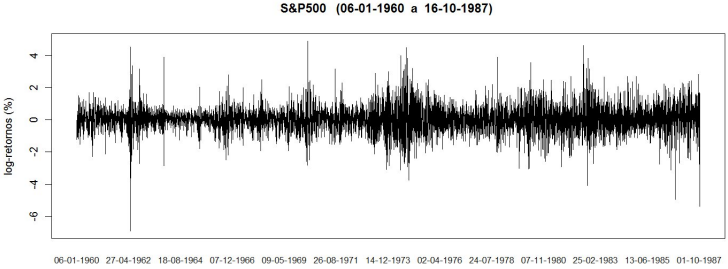
3.8 Dados financeiros

Séries temporais financeiras consistem em preços especulativos dos activos, como stocks, moeda estrangeira ou *commodities*. A gestão do risco num banco comercial destina-se a protecção contra os riscos de perda, devido a quedas nos preços dos activos financeiros, detidos ou emitidos pelo banco. Assim, as diferenças relativas de preços consecutivos, ou os 'log-retornos' são quantidades adequadas para ser investigados.

O VaR de um portfólio é essencialmente o nível abaixo do qual o portfólio futuro vai cair com apenas uma pequena probabilidade. O VaR é uma das importantes medidas de risco que têm sido utilizadas por investidores ou gestores de fundos para tentar avaliar ou prever o impacto de eventos desfavoráveis que podem ser piores do que o que foi observado durante o período para o qual estão disponíveis dados relevantes. O VaR é pois um *quantil extremal*.

Sem entrar ainda em detalhes de estimação de parâmetros, apresentamos em seguida os dados 'SP500.txt' (veja-se Beirlant *et al.*, 2004; `http://lstat.kuleuven.be/Wiley/`): Standard&Poors 500, com os preços de fecho (n=6985) e os log-retornos diários (n=6984), de 5 de Janeiro 1960 a 16 Outubro de 1987 (dia de mercado anterior ao *Big Crash Black Monday 19-Out-87*) (Beirlant et al., 2004).


S&P500 (05-01-1960 a 16-10-1987)
preços de fecho
50
100
150
200
250
300
05-01-1960
27-04-1962
19-08-1964
09-12-1966
14-05-1969
01-09-1971
21-12-1973
12-04-1976
02-08-1978
19-11-1980
10-03-1983
27-06-1985
16-10-1987


S&P500 (06-01-1960 a 16-10-1987)
log-retornos (%)
-6
-4
-2
0
2
4
06-01-1960
27-04-1962
18-08-1964
07-12-1966
09-05-1969
26-08-1971
14-12-1973
02-04-1976
24-07-1978
07-11-1980
25-02-1983
13-06-1985
01-10-1987