

Estatística Robusta aplicada no Mercado

Por Hindenburg Melão Jr.

<http://www.sigmasociety.com>

Alguns autores dizem que os primeiros rudimentos de Estatística surgiram em 3.500 a.C., quando os antigos chineses, egípcios, hebreus e mesopotâmios começaram a fazer censos populacionais, mas na verdade a Estatística é praticada muito antes, desde 3.500.000.000 a.C., quando os organismos primitivos começaram a desenvolver “critérios” para se estabelecer em ambientes que reunissem condições em que tivessem mais chances de sobrevivência. Isso já é Estatística. Quando um cachorro decide seguir por um caminho que ele estima ser mais seguro do que outro, como por exemplo atravessar um rio pela passarela mais larga, mais estável e mais sólida, ele está fazendo uma estimativa de qual caminho ele terá mais probabilidades de ter sucesso em sua travessia, pois para tomar sua decisão está operando num universo randômico em que imperam os conceitos estatísticos. Quando um matemático desenvolve um software para reconhecimento de imagens ou de voz, ele também usa Estatística. A diferença é que o cachorro apenas estima intuitivamente, enquanto o matemático calcula com o máximo rigor que suas ferramentas lhe permitem. Quando se constatou que o espelho parabólico do telescópio espacial Hubble havia sofrido uma deformação durante o processo de resfriamento, ficando com uma bolha de ar que comprometia a formação das imagens, os cientistas que trabalharam vários anos no projeto começaram a chorar, porque achavam que todo o trabalho havia sido em vão. No entanto, graças à Ciência da Computação conjugada à Estatística, foi possível reconstruir as imagens prejudicadas pela deformidade da lente, e o resultado final ficou tão bom que praticamente não se sente o efeito da distorção.

A Estatística é um dos campos da Matemática mais importantes para nossa vida cotidiana. Toda Ciência depende da Estatística para fundamentar empiricamente seus modelos teóricos, bem como delimitar até que ponto estes modelos são representativos dos fatos. Por meio da Estatística podemos verificar a eficiência de medicamentos, podemos prognosticar a produção agropecuária para determinado período, podemos projetar preços de ativos financeiros, podemos taxonomizar espécies animais e vegetais, entre muitas outras aplicações. É por isso que profissionais de praticamente todas as áreas tomam emprestadas várias ferramentas da Estatística para solucionar problemas de seu campo de atuação. Na Psicologia, por exemplo, tomamos emprestada uma enormidade de ferramentas. Na Física e na Astronomia, praticamente toda a parte experimental é determinada pela Estatística. Na Biologia, praticamente tudo que envolve números (Biometria) também envolve Estatística. Na Arqueologia, na Economia, na Geografia, na Política, na Educação, na Publicidade e em praticamente qualquer área, usa-se Estatística, graças ao vasto arsenal de poderosos instrumentos de apoio que ela nos oferece.

Contudo, existem ferramentas estatísticas que praticamente nunca são usadas, porque embora sejam muito eficientes, demandam muito poder de processamento, muita capacidade de armazenamento e muita memória de trabalho. Até a década de 1970 e mesmo no início dos anos 1980, era praticamente impossível aplicar algumas das melhores ferramentas estatísticas porque não havia computadores suficientemente velozes. Só grandes instituições científicas como o MIT e a NASA, bem como grandes universidades, como Harvard, Oxford e Moscou, é que dispunham de mainframes capazes de operar estas ferramentas. Mas nos dias de hoje, com R\$ 1.500,00 pode-se adquirir um PC caseiro 30.000 vezes mais veloz do que o computador (0,08 MHz) usado pela NASA, em 1969, para controlar a missão Apollo 11 que levou o homem à Lua.

Graças aos avanços na Computação, agora é possível fazer estudos de Análise Fatorial Hierárquica com distâncias de Minkovski sobre matrizes de 1000 x 100 em questão de segundos. Pode-se usar Redes Neurais para repetir milhares de iterações em modelagens de funções que exigem o ajuste simultâneo de vários parâmetros, em questão de poucos minutos. É possível plotar e ampliar pseudofractais com escala de fragmentação 10 ou 15 em poucos minutos. Pode-se, enfim, obter soluções numéricas para sistemas estocásticos e randômicos que, de outro modo, levariam séculos. E muito disto já está sendo feito. Porém, estas ainda não são as ferramentas mais sofisticadas e mais eficientes disponíveis. O que há de melhor é a Estatística Robusta. É sobre isso que vamos falar, abordando o tema de maneira simples e didática.

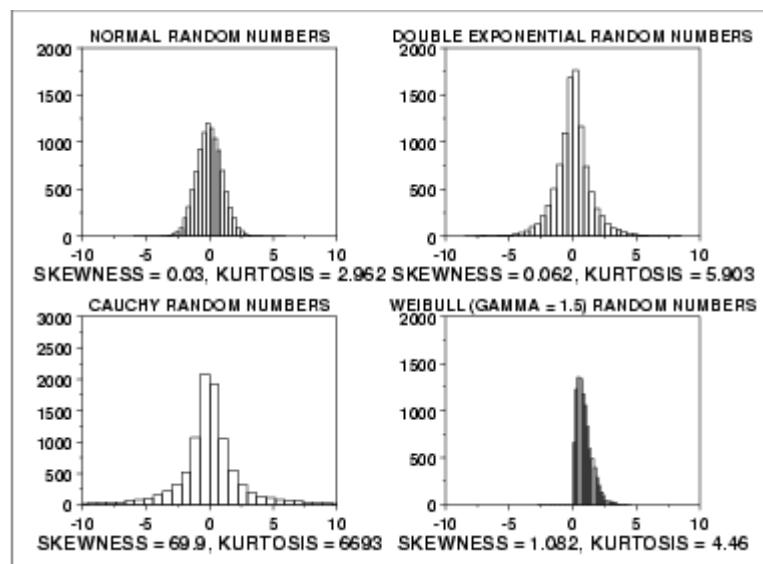
Tudo em Estatística gira em torno de dois conceitos básicos:

- 1 – Tendência central
- 2 – Dispersão

O estimador de tendência central mais simples e mais conhecido é a média aritmética. No âmbito da Estatística Básica, estuda-se também a média geométrica, a média harmônica, a mediana, a moda e a média ponderada. Todos estes estimadores têm o mesmo objetivo: identificar em torno de onde se distribuem os dados.

Após determinar a tendência central, é importante saber o quão dispersos estão os dados em torno desse ponto. Para isso usa-se o conceito de desvio-padrão ou variância (desvio elevado ao quadrado).

Assim, conhecendo a média e o desvio-padrão, pode-se calcular a probabilidade de atingir um valor situado à determinada distância da tendência central. A utilidade destas informações é



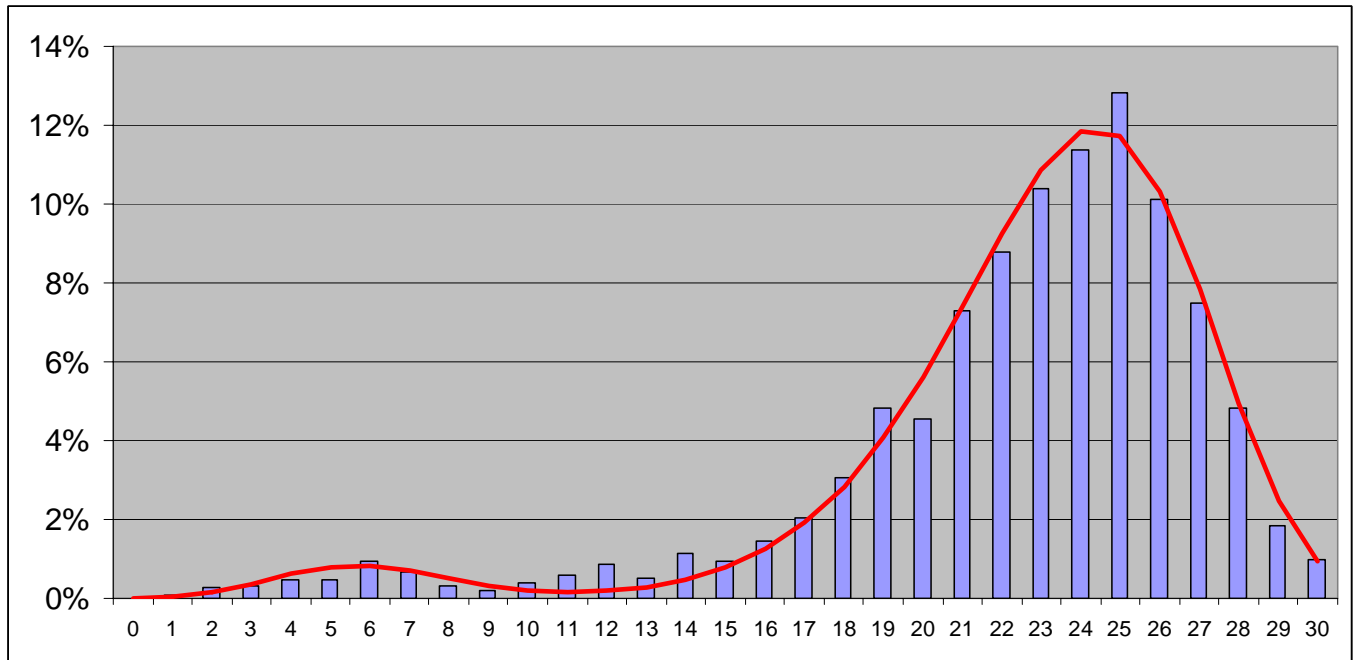
muito grande, porque com apenas 2 parâmetros, pode-se descrever um conjunto com milhares, milhões ou bilhões de dados, e saber quantos destes dados se encontram em cada intervalo. Mas isso só funciona bem quando a distribuição dos dados é gaussiana. Uma distribuição é gaussiana quando cumpre dois quesitos: simetria e mesocurtose. As 4 figuras ao lado exemplificam diferentes níveis de assimetria e de curtose. Quanto maior a diferença entre a metade esquerda e a direita da distribuição, maior é a assimetria. Quanto mais delgado for o miolo da distribuição e mais densa

forem as caudas, maior é a curtose. Quando a curtose é maior do que a de uma gaussiana, diz-se que a distribuição é leptocúrtica (lepto = leve, magro). Caso contrário, diz-se que é platicúrtica.

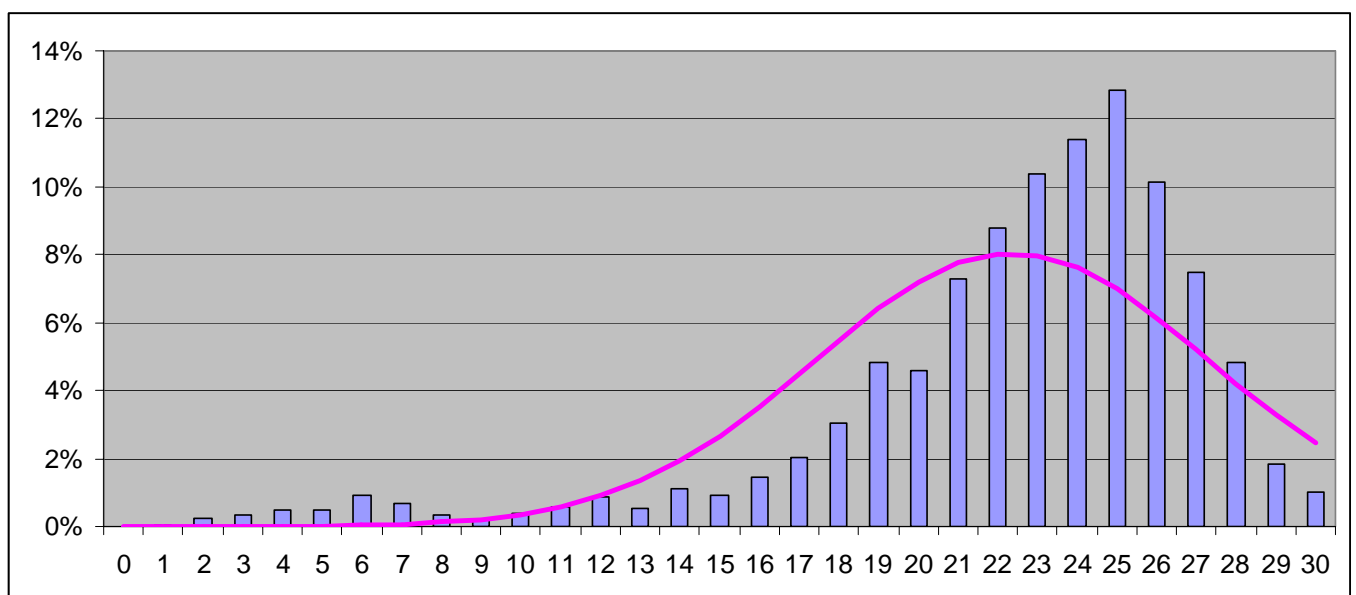
As distribuições do mundo real geralmente podem ser tratadas como se fossem gaussianas quando o grau de acurácia necessário é pequeno. Ao medir 20 vezes uma peça de automóvel, por exemplo, é suficiente adotar a hipótese de que a distribuição é gaussiana, já que o conjunto de dados é tão pequeno que seria inviável comparar a aderência dos dados às diferentes distribuições teóricas. Assim, basta usar uma gaussiana ou uma distribuição *t* de Student. Mas

quando há fartura de dados, com amostras reunindo centenas ou milhares de elementos, pode-se verificar qual função melhor se ajusta aos dados experimentais, em vez de simplesmente supor que a distribuição é gaussiana.

Um exemplo claro de não-normalidade pode ser observado em meu estudo do TIG-NV http://www.sigmasociety.com/artigos/norma_tig.pdf, em que uma distribuição de Weibull é usada para representar os níveis de habilidade e uma binomial para representar os acertos casuais.



O TIG é um teste cognitivo constituído por 30 itens de diferentes níveis de dificuldade e cada item tem 6 alternativas. Assim temos um miolo robusto a outliers e uma cauda sensível a outliers, produzindo uma representação muito mais fiel do que se se considerasse que a distribuição é gaussiana. Veja a gaussiana abaixo e compare à Weibull acima.

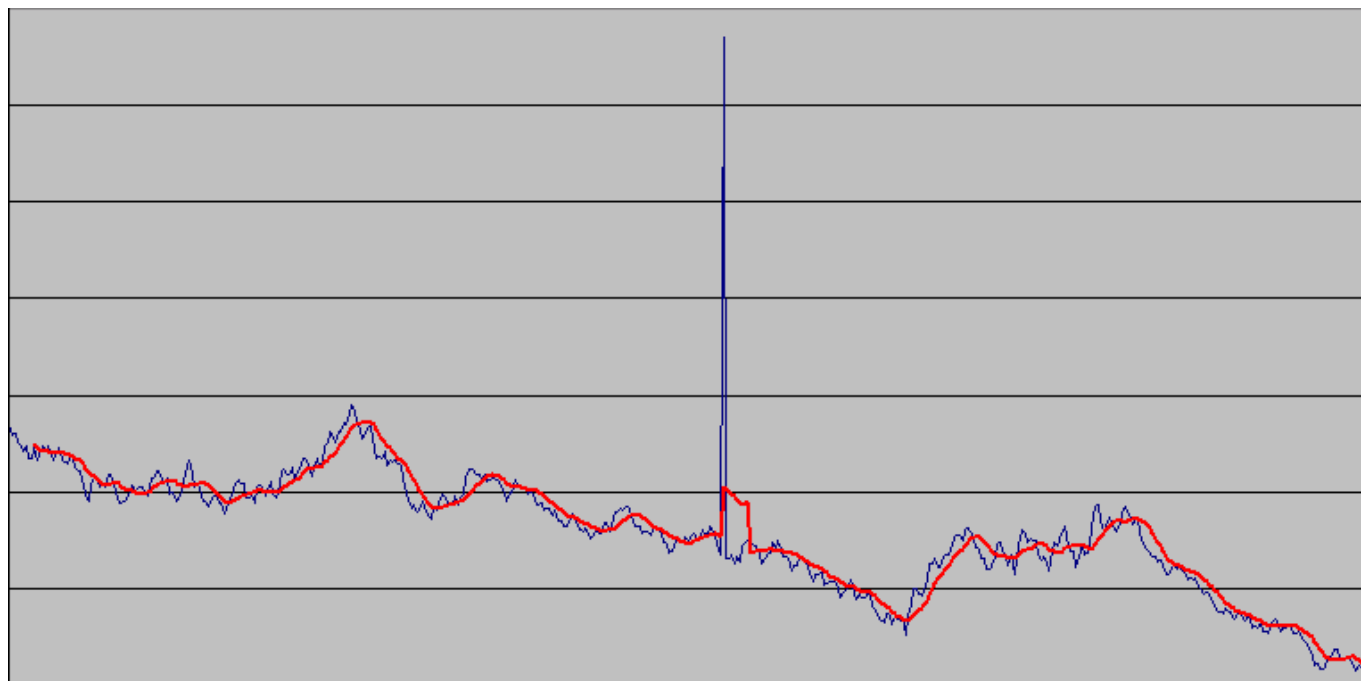


Usar uma Weibull mesclada a uma Binomial é uma solução apropriada neste caso, mas poderia não ser em circunstâncias diferentes. Aqui funciona bem por se tratar de um teste de múltipla escolha, em que a Weibull representa os níveis de habilidade + sorte das pessoas que

se esforçaram para resolver as questões e a Binomial representa os acertos casuais da massa de pessoas que “chutaram” todas ou quase todas as respostas. Portanto é um procedimento totalmente legítimo. Porém para o tratamento de caudas de distribuições de preços de ativos ou de volatilidade de ativos, seria preferível usar outras ferramentas, como bootstrap, por exemplo, para ajustar distribuições de Pareto às caudas.

Existe uma grande variedade de distribuições teóricas e algumas são muito mais versáteis do que outras, de modo que mesmo havendo dezenas de opções, poucas são úteis para ajustes de alta qualidade. A Weibull é uma das mais flexíveis, contudo existem casos em que uma Weibull ou uma Gumpel também não são apropriadas para representar os dados experimentais. Quando isso acontece, torna-se necessário utilizar Estatística Robusta e métodos meta-paramétricos. No mundo real, praticamente sempre isso acontece e basta dispor de uma amostra suficientemente grande, para que se possa constatar que os dados reais não seguem com rigor nenhuma distribuição teórica.

O principal objetivo da Estatística Robusta é atenuar o efeito de *outliers*, mas também tem outros papéis, como preservar a forma, a dispersão e a simetria dos dados reais. Antes de prosseguir, daremos uma breve explicação sobre o que é um “*outlier*”:



No gráfico acima, a linha azul representa os preços de um ativo e a linha vermelha representa a média aritmética móvel de 10 períodos. Podemos observar que na maior parte do tempo os preços oscilam com uma amplitude aproximadamente constante, mas, repentinamente, surge um valor isolado muito fora do “normal”, provocando um pico na linha azul e arrastando para cima a média móvel. Este pico permanece pouco tempo e depois retorna ao “normal”, ou seja, para perto do conjunto de dados. Este evento anômalo produz um deslocamento acentuado e **artificial** da média aritmética, da média geométrica, da média exponencial etc., causando uma distorção na posição da média em comparação à tendência central verdadeira. Desde o início do gráfico, podemos observar que a linha vermelha permanece a maior parte do tempo próxima à linha azul, sendo que cerca de metade dos preços ficam acima da média e outra metade fica abaixo. Isso indica que na maioria das vezes a média móvel está razoavelmente próxima à tendência central, de modo que os preços se distribuem em torno dela, aproximadamente metade dos preços acima da média e metade abaixo. Quando surge o *outlier*, a média dá um salto gigantesco e fica muito acima de todos os preços, quando o correto seria ficar acima de metade e abaixo de metade. A tendência central verdadeira

continua lá embaixo, mas a média foi violentamente deslocada. Uma das implicações disso é que os cálculos de dispersão deixam de funcionar e começam a fornecer indicações gravemente incorretas. Num caso destes, tanto IFR, quanto bandas de Bollinger, MACD, DI+/DI-, ADX e qualquer outro indicador são violentamente distorcidos e passam a dar indicações gravemente incorretas. E não apenas isso. O *outlier* também provoca erros em decisões baseadas em análise gráfica. Um exemplo didático deste efeito é com base num caso real, da ação LIGH3:



Um gráfico análogo a este (mesmo ativo, com níveis de Fibonacci nas mesmas posições e cobrindo mesmo período, usado candles etc.) foi postado num dos fóruns sobre investimentos mais tradicionais do Brasil, por um analista gráfico com MBA pela FGV, atua há 30 anos no mercado e ministra cursos sobre análise gráfica há mais de 10 anos. Antes de tudo, é importante esclarecer que embora o erro seja grave, não se pode atribuir a falha ao analista, mas ao sistema de análise que ele utiliza, por ser um sistema norteado por técnicas básicas. O analista apenas colocou os níveis onde a tradição manda colocar. Se cometeu algum erro, não foi técnico, mas conceitual. Ele deveria ter se questionado sobre a legitimidade de seguir o procedimento típico numa situação totalmente atípica. Há muitos detalhes neste gráfico que precisam ser comentados, como o uso de escala aritmética, numa situação em que uma escala logarítmica poderia ser mais apropriada, os preços foram indexados pelo real, que é um ativo de baixa estabilidade temporal e não é uma boa escolha para períodos que cobrem vários meses ou anos (seria preferível dolarizar os preços), entre outras coisas, mas estes são detalhes menores. O real problema no estudo é onde estão fixados os níveis. O nível 0, ou seja,

o topo, está em R\$ 48,20. Será que isso faz sentido? Qual a robustez da resistência em 48,20? Este preço é apenas a pontinha de um candle. Qual a força desta pontinha de candle? Aqui podemos fazer uma analogia muito didática: o que lhe ofereceria mais resistência para pendurar alguma coisa: uma agulha, um prego ou vários pregos? A resposta é óbvia: vários pregos. Quanto maior o diâmetro dos pregos e mais numerosos eles forem, tanto mais resistência ou suporte vão oferecer. Assim podemos entender que a falsa resistência em R\$ 48,20 é muito frágil, é como uma agulha bem fina. Isto é um exemplo de *outlier* em análise gráfica: um preço que está fora da distribuição dos preços, um preço que alcançou casualmente um nível muito alto, por alguma intervenção espúria, e assim como bateu em R\$ 48,20 poderia ter batido em R\$ 31,50 ou algo assim. Não é um valor que possa ser considerado como parte das informações relevantes para análise. Por outro lado, podemos observar um prego perto de R\$ 40,6 e vários pregos entre R\$ 36,7 e R\$ 40,5, oscilando em suas altitudes, e temos ainda outros pregos reforçando este nível entre outubro de 2001 e março de 2002. Isso indica uma resistência mais provável de ser alcançada e, além disso, se alcançada é mais provável que seja respeitada. O ponto isolado em R\$ 48,20 não é uma referência minimamente segura, além de ser improvável de alcançar, no caso de chegar lá não há motivos para supor que bata e volta. Outra falha é que neste gráfico temos um suporte muito frágil, com apenas 1 prego.

Vejamos um gráfico do mesmo ativo, porém desenhado de maneira mais apropriada, com escala logarítmica, preços dolarizados e níveis de Fibonacci fixados em valores mais robustos:

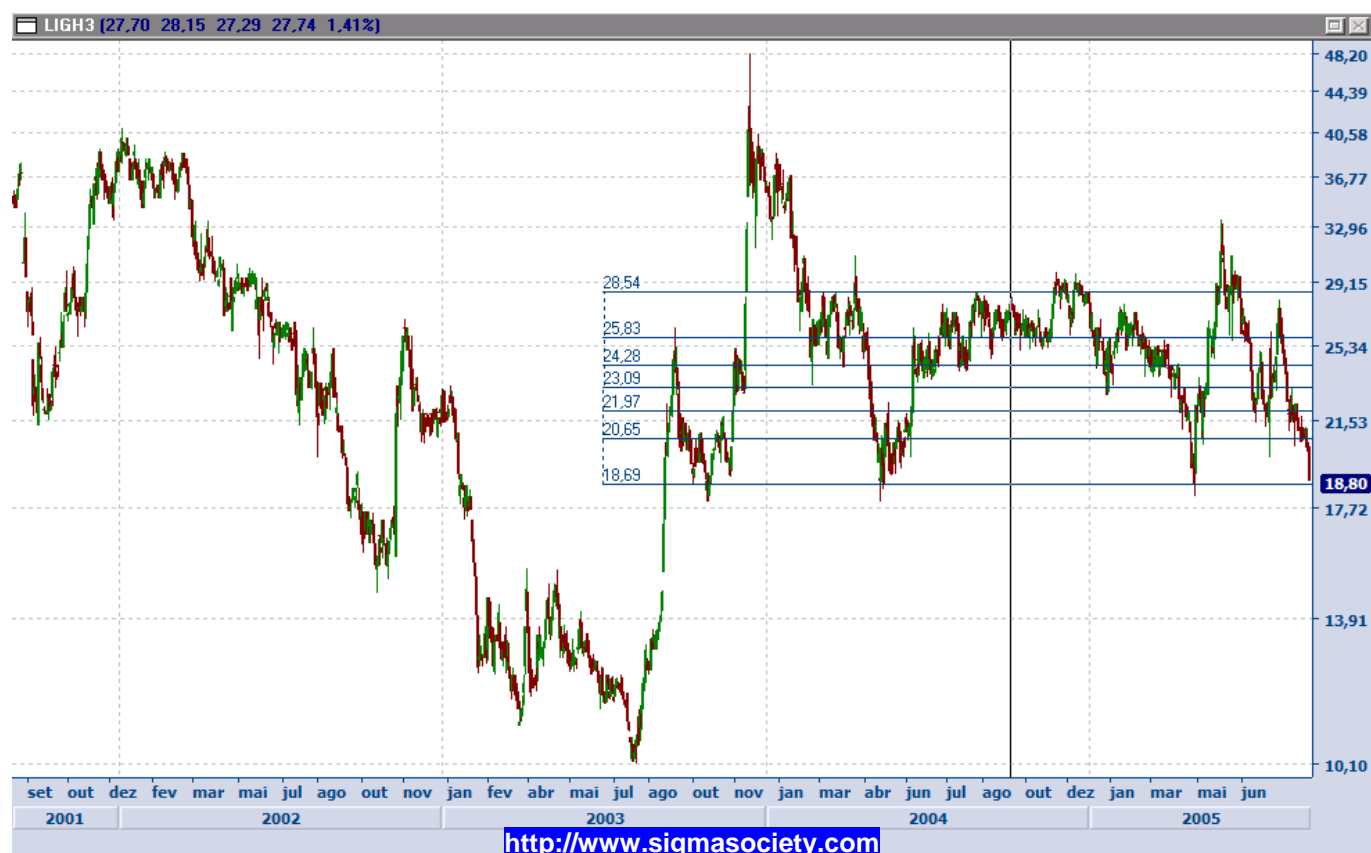


No gráfico incorreto tínhamos a base, ou seja, o nível 100%, situado perto de R\$ 10,00, considerando apenas o suporte de agosto de 2003 e negligenciando que há outro forte indicativo de suporte em abril de 2003. Isto fica muito mais fácil de enxergar com os preços dolarizados, como podemos ver acima, em que podemos observar também um suporte em fevereiro e outro em maio, portanto fortalecendo muito mais o suporte do que pegando a pontinha do candle de 6/8/2003. Agora vamos aproveitar esta oportunidade para desmistificar o uso indiscriminado dos níveis de Fibonacci e vejamos onde batem os demais níveis: podemos

constatar que não há nenhuma evidência de que os preços respeitam aos suportes ou às resistências coincidentes com qualquer dos níveis intermediários. Se alguns níveis coincidem com alguns picos ou vales, isso não acontece numa proporção muito maior do que em níveis aleatórios do tipo 60%, 28%, 17,2% e outros números "chutados". Mesmo desenhando os níveis corretamente, apenas os pontos 0 (US\$ 13.42) e 100% (US\$ 3.42) fazem sentido. O ponto 0 precisa ser "enterrado" um pouco abaixo dos picos, para "matar" ou atenuar *outliers*. Então vemos vários picos (vários pregos) entre dezembro de 2003 e janeiro de 2004 respeitando esta resistência e um pico recente, em junho de 2005. Fora isso, os demais níveis não são respeitados mais vezes do que quaisquer outros números escolhidos aleatoriamente. Claro que se pode ir colocando os níveis de Fibonacci em lugares diferentes, por tentativa e erro, até enxergar algo que faça sentido e encaixá-los de modo que pareça surgir um padrão, e em Ciência esse tipo de procedimento se chama "acoxambrão" ou "gambiarra" ou "adulteração". ☺ Com gambiarras, pode-se até deixar os níveis encaixados bonitinhos, "pra inglês ver", mas não oferece nenhuma garantia de que os preços futuros continuarão se encaixando nestes moldes.

Isso não quer dizer que Fibonacci não sirva para nada. Na verdade serve, mas precisa ser usado com bons critérios. Futuramente publicaremos um artigo sobre o grau de validade dos níveis de Fibonacci em comparação a níveis aleatórios, usando para isso coeficientes de compactação, que é a mesma técnica usada em 2003 para verificar o nível de significação estatística dos supostos códigos da Bíblia.

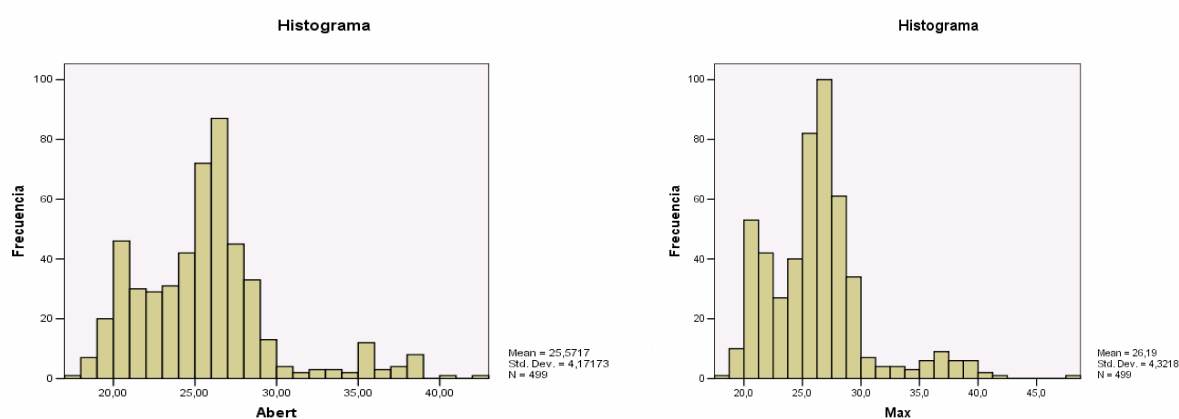
Vejamos mais um estudo da LIGH3, com os preços em Reais:

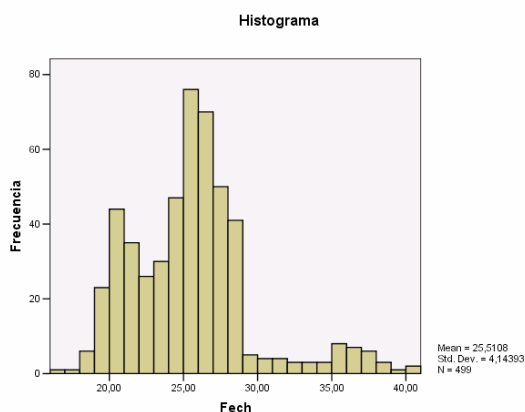
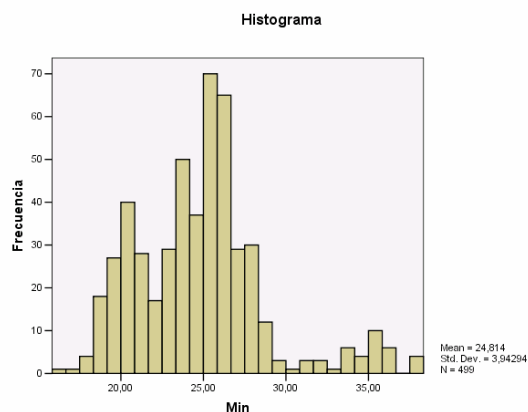


A resistência fica difícil de definir, mas o suporte fica perto de R\$ 18,70, com pontos de sustentação em outubro de 2003, maio de 2004 e maio de 2005. Novamente podemos ver que os níveis intermediários não fazem sentido.

Estes gráficos ajudam a visualizar o problema e a compreender com mais clareza em que consistem as falhas de usar Estatística Clássica, em lugar de Estatística Robusta. Mas isso não significa que estas análises gráficas sejam recomendáveis para operar no mercado. Um dos motivos disso é que ao estudar visualmente um gráfico, só se considera uma pequena fração da totalidade dos dados disponíveis (os topos e os fundos), ao passo que num estudo estatístico de boa qualidade se utilizam todos os dados intermediários (além dos topos e fundos), e não se tratam apenas dos dados recentes, mas também dos mais antigos, tudo devidamente balanceado em função do tempo, atribuindo maior peso aos dados mais atuais. O estudo gráfico pode ser útil em alguns aspectos, e tem a virtude de ser fácil de operar, mas é importante não se deixar iludir com uma enormidade de falsos padrões descontextualizados. O contexto em que surge um padrão faz uma diferença imensa. Uma estrela ou um violino muito abaixo da tendência central de Tukey não podem ser interpretados da mesma maneira que as mesmas figuras muito acima da tendência central de Tukey, porque ainda que a forma e o tamanho sejam praticamente iguais, se a posição for diferente e se o histórico for diferente, isso muda tudo. Ainda não escolhi o melhor termo para diferenciar “Análise Gráfica” de “Análise Estatística”. Já pensei em “Análise Geometria” e “Análise Algébrica”, mas não são os termos mais apropriados. Talvez o sejam apropriados os termos “Análise Visual” e “Análise Científica”. A “Análise Visual” é aquela em que se escolhe no “olhômetro” onde colocar níveis de Fibonacci, LTAs, LTBs, pivots etc. E a “Análise Científica” é aquela que usa a totalidade dos dados disponíveis para que se possa compreender o comportamento dos preços e assim prognosticar os preços futuros, usando heurísticas apropriadas, fractais e estatística robusta para abastecer redes neurais com boas redes de neurônios, axônios e dendrites artificiais.

Vejamos mais um pouco a situação da Ligh3. Na figura a seguir, estão representadas as distribuições dos preços de abertura, fechamento, máximo e mínimo dos últimos 500 dias (excluindo o último). É importante compreender porquê escolhemos este período de 500 dias, em vez de qualquer outro intervalo. Em primeiro lugar, porque observamos no gráfico dos preços que neste período houve suficiente homogeneidade na forma e na amplitude das oscilações. Sem esta característica, não faria sentido tentar determinar a distribuição de preços. Em segundo lugar, porque cerca de 250 dias representam 1 ano econômico, logo um período de 500 dias não apresenta distorções causadas por sazonalidades.





Usa-se mais freqüentemente os preços de fechamento como referência, por isso fizemos um estudo mais detalhado apenas destes preços, aplicando testes de ajuste de qualidade de Kolmogorov-Smirnov-Lilliefors e Shapiro-Wilk para determinar a normalidade, e testes de Kolmogorov-Smirnov para testar ajustes a outras funções além da Normal. Os resultados dos testes de normalidade foram estes:

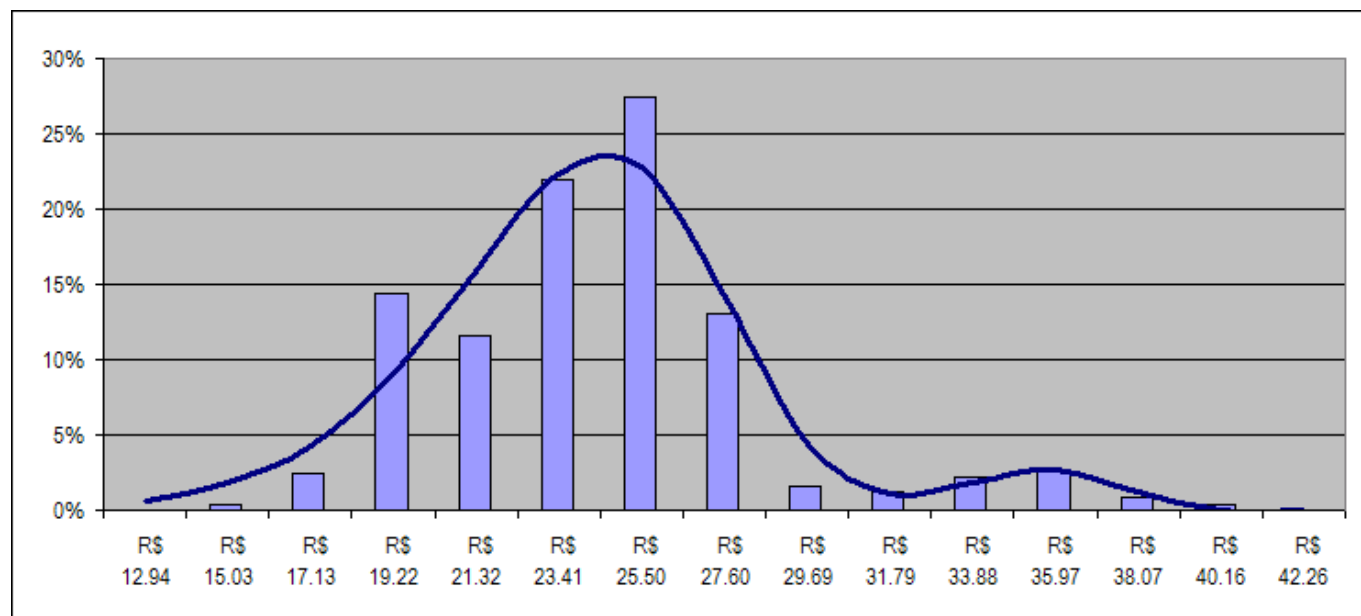
	Kolmogorov-Smirnov-Lilliefors			Shapiro-Wilk		
	Estatístico	Gl	Sig.	Estatístico	gl	Sig.
Abert	0,109	499	>0,001	0,918	499	>0,001
Max	0,126	499	>0,001	0,905	499	>0,001
Min	0,100	499	>0,001	0,932	499	>0,001
Fech	0,104	499	>0,001	0,925	499	>0,001

	N	Média	Desvio-padrão	Mínimo	Máximo	Percentis		
						25	50 (Mediana)	75
Fech	499	25,51	4,14	16,60	40,75	22,51	25,47	27,29
Abert	499	25,57	4,17	17,28	42,98	22,65	25,56	27,29
Max	499	26,19	4,32	18,70	48,20	23,19	26,15	27,74
Min	499	24,81	3,94	16,60	37,84	21,84	24,95	26,42

Também foi realizado um teste de normalidade usando cadeias de Markov no Método de Monte Carlo (MCMC), com 10.000 iterações, a fim de melhor calibrar a prova de Kolmogorov-Smirnov, e os resultados foram estes:

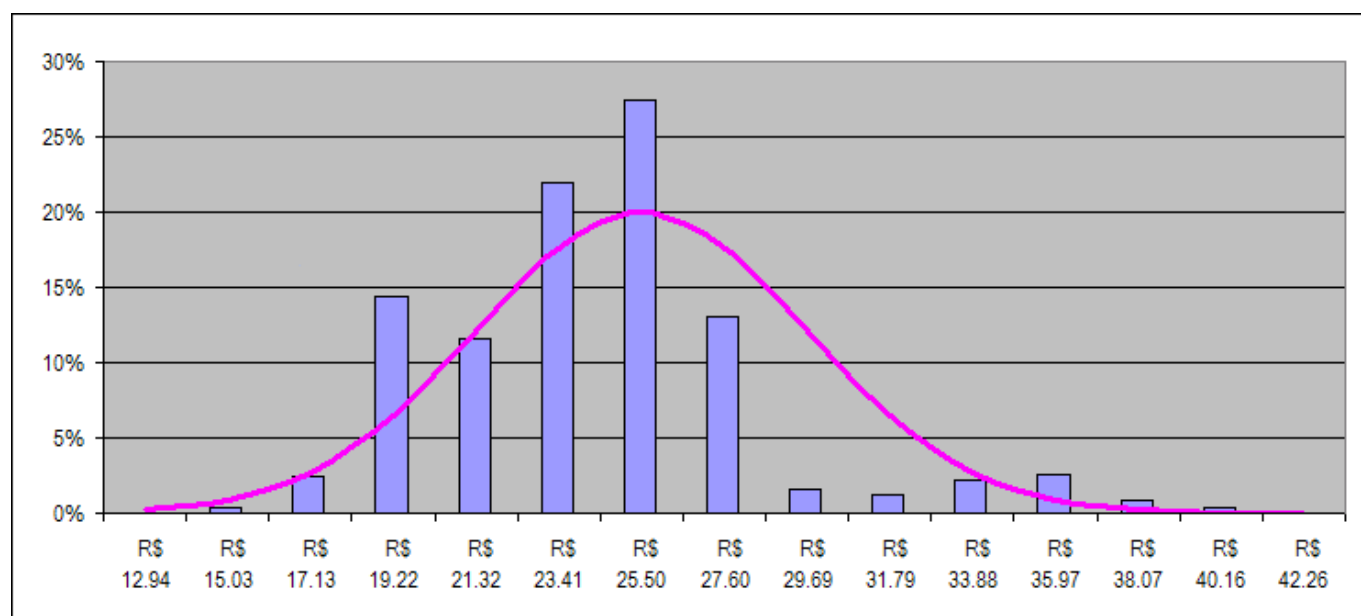
		Fech	Abert	Max	Min
N		499	499	499	499
Parâmetros normais (μ , σ)	Média	25,51	25,57	26,19	24,81
	Desvio-padrão	4,14	4,17	4,32	3,94
Diferenças extremas	Absoluta	0,104	0,109	0,126	0,100
	Positiva	0,104	0,109	0,126	0,100
	Negativa	-0,046	-0,046	-0,054	-0,046
Z de Kolmogorov-Smirnov		2,317	2,440	2,808	2,232
Sig. asintót. (bilateral)		0,000	0,000	0,000	0,000
Sig. Monte Carlo (bilateral)	Sig.	0,000	0,000	0,000	0,000
	Int de confiança 99%				
	Límite inferior	0,000	0,000	0,000	0,000
	Límite superior	0,000	0,000	0,000	0,001

Portanto nenhuma das amostras se distribui normalmente a um grau de significação 0,001, ou seja, há mais de 99,9% de chances de que nenhuma destas distribuições seja normal (a única que remotamente se assemelha é a dos preços mínimos). Após descartada a hipótese de normalidade, testamos as distribuições Lognormal, Gamma, Gumpel, Weibull, Laplace e Cauchy, usando Kolmogorov-Smirnov, e constatamos que a Weibull é a que apresenta melhor ajuste, contudo a disparidade ainda era grande, por isso foi feito um ajuste com duas Weibull mescladas, e isso produziu uma modelagem satisfatória, conforme a figura a seguir:



Assim podemos inferir que uma distribuição de Weibull é uma representação satisfatória para a distribuição dos preços da Ligh3 no período considerado, desde que seja descartada a alta ocorrida no final de 2003 e início de 2004. Se quisermos representar também esta alta, basta mesclar outra Weibull, tal como na figura acima. Para prognósticos de curto prazo, é suficiente usar o lado esquerdo da distribuição, já que sabemos que praticamente toda a cauda densa à direita foi causada pela alta em dez-2003 e fev-2004, tendo, portanto, menor influência nos preços mais recentes.

O benefício de usar este par de distribuições de Weibull é tão grande que mesmo visualmente podemos constatar que a aderência é muito melhor do que a uma gaussiana (figura abaixo).



Além disso, na Weibull podemos observar que a tendência central verdadeira se localiza muito próxima à média da distribuição, que é R\$ 25,00. No caso da gaussiana, a tendência central verdadeira também está perto de R\$ 25,00, no entanto a média e a moda estão perto de R\$ 25,50 (moda da amostra discretizada em estratos, obviamente).

Tudo isso pode ser feito visualmente, porém o “olhômetro” não é muito exato. Como podemos determinar a tendência central verdadeira com boa acurácia e nos valendo de um método objetivo? A resposta é: usando, por exemplo, tendência central bibalanceada de Tukey. Mas o que é tendência central de Tukey? Agora retornamos aos fundamentos: conforme já vimos, a média é uma estimativa básica de tendência central, e ela apresenta o problema que constatamos na página 4, em que a média foi deslocada da tendência central verdadeira por um *outlier*. A tendência central de Tukey tem a mesma finalidade que a média aritmética ou geométrica, porém é muito mais robusta, ou seja, é mais resistente aos puxões dos *outliers*, mantendo-se muito mais próxima à tendência central verdadeira onde se encontra a massa de dados. Na figura da página 4, quando ocorreu o *outlier* e a média deu um sobressalto, deixando os dados lá em baixo, a tendência central de Tukey permaneceu quase indiferente ao *outlier*, bem próxima aos dados reais, comportando-se da maneira desejada para possibilitar cálculos mais exatos. Existem outros estimadores, como tendência central da Hampel ou ondas de Andrews, todas muito mais robustas do que as médias. Além disso, existem estimadores robustos para medidas de dispersão (desvio de Stahel, de Ronchetti etc.), de assimetria (Skewness de Chambers, Kleiner, Tukey etc.) e até para regressão, correlação, análise fatorial etc. Os estimadores robustos são claramente melhores do que os procedimentos clássicos, e o fato de exigirem o uso de técnicas e conhecimentos mais avançados não constitui um problema operacional, porque existem procedimentos simples para atenuar *outliers*, como fizemos nos gráficos da LIGH3, “enterrando” mais fundo o nível 0, sem usar nenhuma Matemática para isso. Apenas usamos o “olhômetro”, e isso foi suficiente para obter um resultado semelhante ao que seria obtido usando estimadores robustos. Para cálculos de DI-/DI+ e ADX, por exemplo, também é fácil fazer estimativas visuais mais seguras do que usando os procedimentos tradicionais. E para quem quiser mais exatidão, existem softwares estatísticos dedicados que executam estas funções com apenas poucos cliques no mouse. Não é preciso conhecer nenhuma fórmula. Basta compreender o conceito. Entendendo para que serve a ferramenta e dispondo de um software estatístico que execute a função, pode-se aprimorar tremendamente a qualidade das análises. Quem tiver interesse em conhecer também a fórmula, para adaptá-la ou para uso “manual” (HP, Excel etc.), pode entrar em contato comigo pelo formulário para contatos de Sigma Society, em www.sigmasociety.com, solicitando a fórmula que deseja.

No caso da Ligh3, os estimadores robustos de tendência central indicam os seguintes valores:

	Estimador-M de Huber (a)	Bponderado de Tukey (b)	Estimador-M de Hampel (c)	Onda de Andrews (d)
Abert	25,33	24,99	25,12	24,985
Max	25,94	25,58	25,68	25,580
Min	24,64	24,30	24,45	24,302
Fech	25,30	25,00	25,09	24,997

a Constante de ponderação: 1,339.

b Constante de ponderação: 4,685.

c Constantes de ponderação: 1,700, 3,400 y 8,500.

d Constante de ponderação: $1,340 \cdot \pi$.

Para encerrar, convém citar um estudo feito por Luiz Álvares Rezende de Souza, em sua tese de mestrado em Administração pela FEA USP, em que ele propõe o uso de Estatística Robusta para o cálculo de *Value at Risk*. Trata-se de um dos poucos trabalhos no Brasil sobre aplicação

de Estatística Robusta no Mercado. São louváveis a iniciativa e os esforços de Luiz Álvares e de seu orientador, Prof. Dr. Marcos Eugênio da Silva, no sentido de tentar preencher esta gravíssima lacuna na literatura especializada, já que praticamente não há nada sobre o assunto em língua portuguesa, e mesmo em no idioma universal, o inglês, existe pouca bibliografia sobre o assunto. Não obstante meus sinceros reconhecimentos a estes méritos do autor e seu orientador, é preciso apontar e descrever alguns aprimoramentos que se fazem necessários nesta dissertação (algumas conclusões do autor, por exemplo). Sobre isso escreveremos outro artigo. Aqui nos limitaremos a apontar dois aprimoramentos:

1 – O autor usa provas de normalidade de Jarque-Bera, Bowman-Shenton e Doornik-Hansen, que são testes de Chi-quadrado, portanto de baixa sensibilidade à forma da distribuição. Nestas situações é muito mais recomendável usar o teste de ajuste de qualidade de Kolmogorov-Smirnov, que além de ser mais sensível à forma, é tanto melhor (em comparação a outros testes de aderência) se tanto menor for a amostra, portanto é particularmente recomendável para amostras pequenas. Além disso, Kolmogorov-Smirnov não é apenas um teste de normalidade, mas pode também ser usado para comparar qualquer par de funções teóricas, ou par de distribuições empíricas, ou uma função teórica e uma distribuição empírica, de modo que possibilitaria não apenas rejeitar o uso da gaussiana simétrica e mesocúrtica como também ofereceria uma farta gama de alternativas a serem estudadas, com especial destaque para as distribuições de Weibull e Gumpel, o que provavelmente possibilitaria alcançar alto grau de aderência aos dados empíricos até mesmo sem o uso de bootstrap para modelar as caudas. Se se constatar que a distribuição normal, a lognormal ou a de Cauchy apresentam boa aderência, então pode-se usar Anderson-Darling para obter uma medida mais acurada do ajuste e eleger qual das três é a melhor representação dos dados (conforme recomendação do NIST: <http://www.itl.nist.gov/div898/handbook/eda/section3/eda35g.htm>). Também é possível usar ferramentas mais robustas de ajuste, mas estas só fazem sentido se a amostra for suficientemente grande e houver necessidade de alta acurácia, para justificar o esforço, já que as fórmulas de recorrência deste gênero demandam muito tempo inclusive em processadores velozes.

2 – O autor demonstra preocupação em encontrar um ponto ótimo de corte para separar as caudas do miolo, e para isso tenta uma solução com o estimador de Hill, fazendo ajustes à base do “olhômetro” e lamentando-se por não haver uma solução “automática” e objetiva para o problema. Pois bem, vamos supor que tal tricotomização fosse uma solução aceitável, então bastaria usar nuvens dinâmicas e estratificar a amostra em 3 fatores. Isto proveria uma solução automática e objetiva, que atenderia aos propósitos do autor. Mas o fato é que solução almejada por Luiz Álvares não seria apropriada, já que haveria uma mudança súbita de robustez ao sair do miolo e entrar na cauda. Isso é artificial, porque não há uma mudança brusca nestas situações. A transição do miolo para a cauda é suave, e para que se possa dar uma boa representação aos dados experimentais é necessário que se adote um procedimento que produza esta transição suave. Resumindo: há duas maneiras melhores de solucionar este problema. A primeira seria usando Análise Fatorial Hierárquica e estratificar a distribuição em vários segmentos como “miolo central”, “miolo lateral esquerdo”, “pré-cauda esquerda”, “cauda média esquerda”, “cauda periférica esquerda” e “extremidade da cauda esquerda”, com equivalentes do lado direito. Esta solução é melhor do que cortar em 3 partes, contudo ainda seria granulosa. A melhor solução consiste num refinamento desta, em que, depois de fazer a Análise Fatorial, os níveis médios de robustez em cada estrato serviriam balanceados, plotados e em seguida seria feita uma regressão robusta para determinar os parâmetros de uma função contínua que definiria o grau de sensibilidade em função da distância à tendência central. Assim a solução definitiva seria dada por uma função cuja robustez variaria suavemente do centro para as extremidades das caudas, em vez de uma mudança brusca do miolo para cauda.

Estas e outras possibilidades de aperfeiçoamento não diminuem a importância do trabalho praticamente pioneiro de Luiz Álvares, e o estudo de sua tese é altamente recomendável como fonte de pesquisa a quem se interessa por *Value at Risk*, já que os procedimentos apresentados por Luiz Álvares, ainda que não sejam sempre os melhores possíveis, são seguramente muito superiores aos métodos tradicionais e possibilitam cálculos muito mais fidedignos. A quem tiver interesse em ler a tese completa de Luiz Álvares, pode baixá-la em: www.sigmasociety.com/var-dissertacao.pdf.

Vimos a importância da Estatística, fizemos uma comparação entre Estatística Clássica e Estatística Robusta e constatamos que o uso de Estatística Robusta é francamente superior. Vimos um exemplo teórico sobre o efeito de um *outlier* e em seguida estudamos um caso real, em que um *outlier* causou graves distorções numa análise gráfica. Constatamos que embora os estimadores robustos exijam matemática pesada para o uso formal, eles podem ser aproximadamente estimados com o famoso “olhômetro”, e até mesmo com “olhômetro” acabam funcionando melhor do que a Estatística Básica. E, por fim, comentamos brevemente um dos raros textos em português que trata de Estatística Robusta. Com isso podemos concluir que existem ferramentas melhores do que aquelas que são habitualmente utilizadas, vimos que no Brasil praticamente não há bibliografia sobre o assunto e vimos que usar estas técnicas pode ser mais fácil do que muitos imaginam.

Em nosso próximo artigo, trataremos das superstições sobre suporte e resistência e apresentaremos um estudo para desmistificar a crença de que quando um preço “rompe” um suporte ou uma resistência, isso indica maior probabilidade de continuar na mesma direção.