REDECA
REVISTA ELETRÔNICA DO DEPARTAMENTO DE CIÊNCIAS CONTÁBEIS & DEPARTAMENTO DE ATUÁRIA E MÉTODOS QUANTITATIVOS DA FEA

Detecção de Fraudes, Anomalias e Erros em Análise de Dados Contábeis: Um Estudo com Base em *Outlier*s

Cledson D. Oliveira[1]
Adhemar A. De Caroli[2]
Amaury S. Amaral[3]
Omar L. Vilca[4]

ABSTRACT

The studies of identification of anomalies in accounting registers it has been an object of research in the Auditing field. There may be a situation in which it has not been identified the biggest anomalies by the usual process of statistical sampling, or it can still take place only to partial detection of the cases, since statistical sampling depend on the concentration of the different weights or sampling errors in the set of data.
The approaches based on computation resulted from studies of distance data aiming at the identification of the most distant elements of a standard established through the measurement of the equidistant rays, which we will call in this article of outliers. We present comparative techniques of detection of these anomalies for studies of data mining.

Keywords: Outliers detection; Data mining; Data distance.

RESUMO

Os estudos de identificação de anomalias em registros contábeis tem sido objeto de pesquisas no campo de Auditoria. Pode ocorrer a situação em que as maiores anomalias não tenham sido identificadas pelo processo usual de amostragem estatística, ou ainda em que haja somente a detecção parcial dos casos, pois os levantamentos manuais dependem da concentração dos pesos divergentes ou dos erros amostrais no conjunto de dados.
Os enfoques baseados em computação resultaram dos estudos de afastamentos de dados uns dos outros, objetivando a identificação de elementos mais distantes de um padrão, estabelecida por meio da medição dos raios equidistantes, chamados neste artigo de *outliers*. Apresentamos técnicas comparativas de detecções dessas anomalias por estudos de mineração de dados. Trabalhamos com os métodos quantil-quantil, hampel, boxplot, distribuição t de Student, e distribuição Qui-quadrado. Para a detecção de *outliers*, comparamos os testes: Grubbs, Dixon, e Generalizado ESD.

Palavras-chave: Detecção de *Outliers*; Mineração de Dados; Distancia de Dados.

1 Especialista MBA - Universidade Federal do Rio de Janeiro (UFRJ) - Rio de Janeiro - RJ – Brasil cledson@mircontabil.com.br

2 Professor - Departamento de Contabilidade - Pontifícia Universidade Católica de São Paulo (PUC-SP) - São Paulo - SP – Brasil adhemar@pucsp.br

3 Professor - Departamento de Contabilidade - Pontifícia Universidade Católica de São Paulo (PUC-SP) – São Paulo - SP e Mestre em Ciência da Computação _ Universidade Federal do ABC (UFABC) - Santo André - SP – Brasil asamaral@pucsp.br

4 Pós-Graduação em Ciência da Computação _ Universidade Federal do ABC (UFABC) - Santo André - São Paulo - SP – Brasil omar.vilca@ufabc.com.br

1. INTRODUÇÃO

As técnicas de *Outliers* em Análises de Dados Contábeis trazem resultados adequados para que um analista ou auditor possa detectar, de forma genérica, indícios de erros, fraudes ou anomalias. Objetiva-se a apresentar uma técnica de estudo para Ciência Contábil, vinculando-se a contas contábeis e se identificando possíveis anomalias, erros e até mesmo fraudes.

Ocorre que quando se opera com uma massa de dados muito grande, como é o caso de uma conta de clientes de um banco, a verificação dos dados na pesquisa torna-se tarefa árdua e demanda certo período de tempo, ainda que o processo realize-se por amostragem. Muitas vezes, tais amostragens são constituídas por vieses ou podem estar influenciadas pela concentração de determinados números. Nessa vertente, a amostra utilizada para verificação mascarou o estudo e não foi suficiente para uma conclusão acerca do número analisado, ou ainda, da conta contábil pesquisada. Este estudo objetiva trazer uma ferramenta que poderá auxiliar o analista ou o contador nesse estudo, contribuindo para o aperfeiçoamento de sua análise.

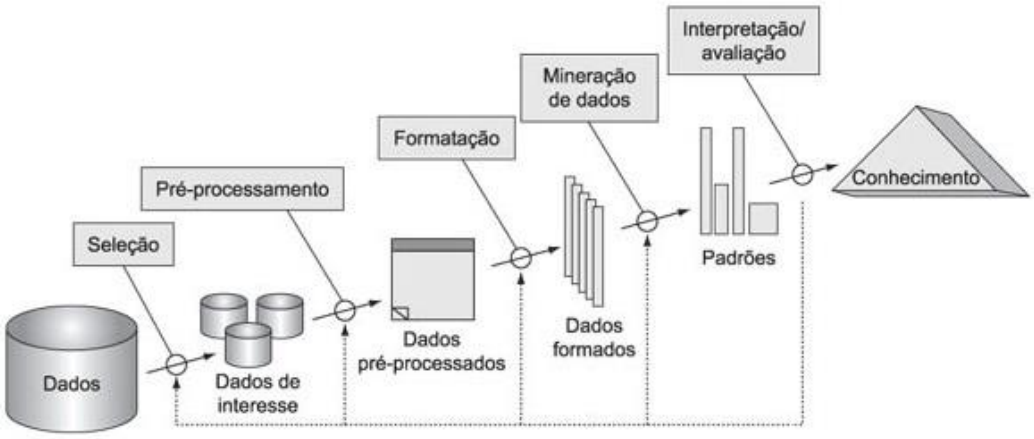
Outliers são os registros que em uma determinada série de números possuem, prevalecendo um determinado elemento do conjunto muito maior ou menor que os restos dos demais números. Com isso, pressupomos uma matriz de dados com determinado padrão e cujas dispersões serão os dados a serem pesquisados e analisados.

2. MINERAÇÃO DE DADOS – DATA MINING

Segundo Fayyad *et al.* (1996, p. 20), a mineração de dados é uma etapa do processo de KDD (Knowledge Discovery in Database), que consiste na aplicação da análise de dados e algoritmos de descoberta, os quais produzem uma enumeração particular de padrões ou modelos que visam o conhecimento sobre os dados.

Muitas são as maneiras de se analisar um grande volume de informações. Entender o objetivo que se espera alcançar com a mineração de dados é um dos primeiros passos. Como os dados são criados em diversos formatos e têm origens diferentes, é necessário identificar as informações relevantes e apresentar, de forma clara e objetiva, o instrumento de estudo. A preparação dos dados é o estágio seguinte de estudo; por eles terem comportamentos diferentes, algumas ações podem ser necessárias para que os dados estejam preparados adequadamente para a aplicação das técnicas de mineração.

Figura 1 - Etapas do processo KDD



Fonte: Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic.

É fundamental no processo de mineração de dados a identificação de pontos colocados fora de limites razoáveis. Esses pontos serão chamados de *outliers* e são utilizados principalmente na detecção de fraudes.

O *outlier* é definido, em uma série de números, como aquele que apresenta uma grande variação ou inconsistência em relação aos demais valores da série. Este método é utilizado em diferentes estudos, como na análise de irregularidade em votações, detecção de fraude em cartões de crédito ou previsão meteorológica. A seguir, teremos um exemplo preliminar de como identificar um *outlier*. Para encontrar *outliers*, basta identificar os números que sumarizam a série, por exemplo:

34	63	12	71	53	35	7	17	77

Passo 1: Ordenar os números em ordem crescente

7	12	17	34	35	53	63	71	77

Passo 2: Identificar a mediana

7	12	17	34	35	53	63	71	77

Passo 3: Identificar o menor e o maior número

REDECA – Revista Eletrônica do Departamento de Ciências Contábeis & Departamento de Atuária e Métodos Quantitativos da FEA-PUC/SP

7	12	17	34	35	53	63	71	77

Passo 4: Identificar a mediana entre o menor número e a mediana geral de toda a série de dados, a mediana entre a mediana geral e o maior número da série

7	12	17	34	35	53	63	71	77

Passo 5: Identificar os cinco números que sumarizam a série:

- 7 Menor número no conjunto;

- 17 Mediana entre o menor número e a mediana geral;

- 35 Mediana de toda a série;

- 63 Mediana entre o maior número e a mediana de toda a série;

- 77 Maior número no conjunto.

7	12	17	34	35	53	63	71	77

Q1	[7, 17]	14.5
Q2	[17, 63]	35
Q3	[63, 77]	67

O quartil inferior (Q1) é o segundo dos 5 números que sumarizam a série, representando 25% de todos os números da série que são aqueles menores que Q1.

7	12	17	34	35	53	63	71	77

O quartil superior (Q3) é o quarto dos 5 números que sumarizam os dados, representando 25% de todos os números da série que são maiores que Q3. A porcentagem restante de números que estariam entre Q1 e Q3 equivale a 50% de todos os números entre Q1 e Q3.

7	12	17	34	35	53	63	71	77

O Intervalo $[14.5 - 67]$ é chamado de Inter-Quartil (IQR).

REDECA – Revista Eletrônica do Departamento de Ciências Contábeis & Departamento de Atuária e Métodos Quantitativos da FEA-PUC/SP

O tamanho do IQR é a distância entre Q1 e Q3: $67 - 14.5 = 52.5$.

Para determinar se um número é um *outlier*, multiplique o IQR por 1.5 ($52.5 * 1.5 = 78.75$).

Os valores inferiores a Q1 - 1.5*IQR e superiores a Q3 + 1.5*IQR são caracterizados como outliers.

O crescimento das atividades empresariais fez com que os dados gerados e armazenados pelas empresas também aumentassem, e sua gestão tornou-se complexa devido ao grande volume de informações. Na implantação de um sistema de gestão, as companhias buscam, além da otimização dos processos, a redução de custos com o armazenamento de informações utilizadas na tomada de decisão.

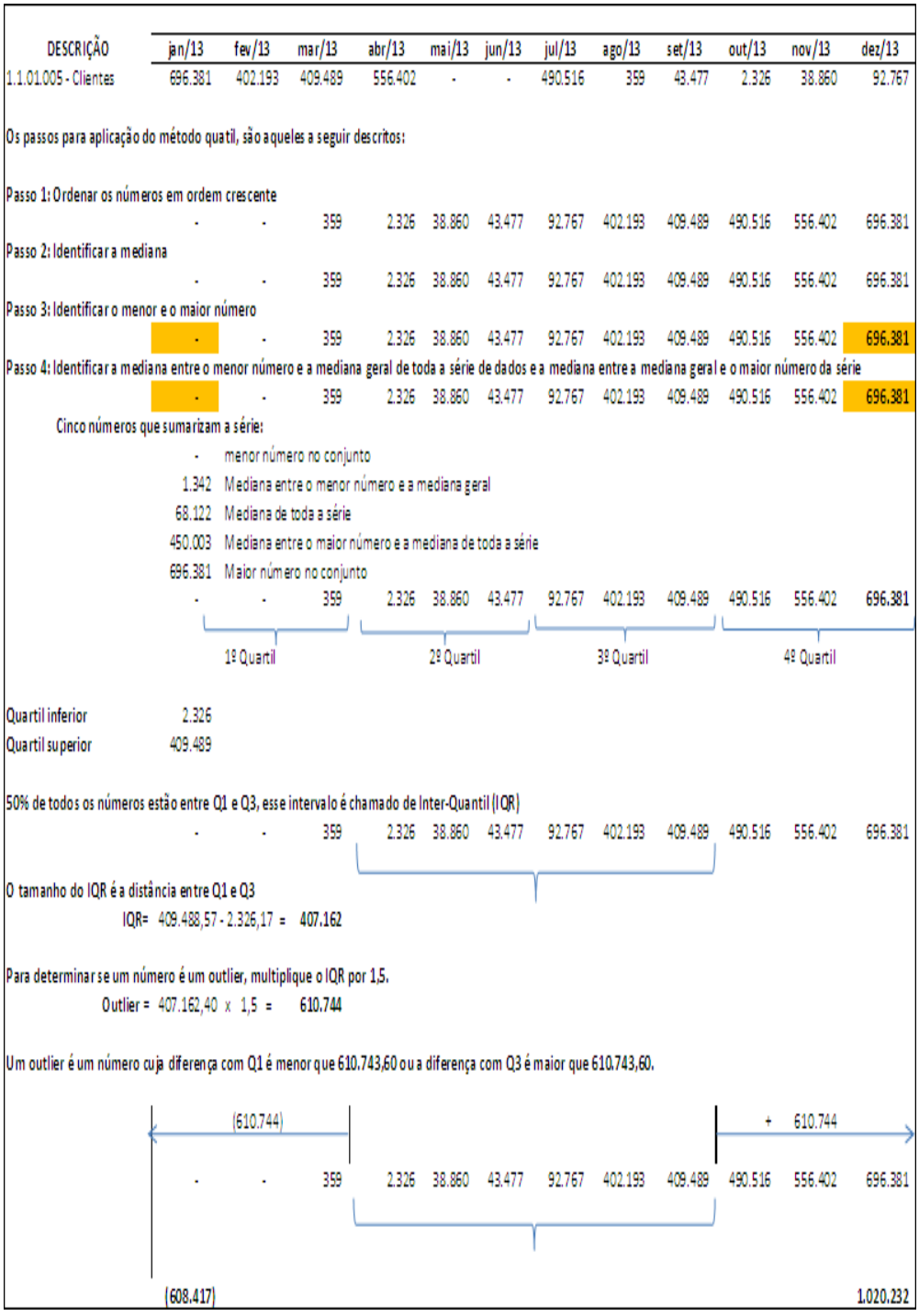
A mineração de dados surgiu como uma ferramenta fundamental de apoio às empresas, pois através dela serão analisadas e validadas grandes quantidades de dados. Os resultados obtidos com a mineração de dados são bons, porém esta área ainda precisa ser desenvolvida, a fim de acompanhar o nível de complexidade das empresas e dos negócios; logo, a popularização do uso de mineração poderá consolidar a precisão do mecanismo.

REDECA – Revista Eletrônica do Departamento de Ciências Contábeis & Departamento de Atuária e Métodos Quantitativos da FEA-PUC/SP

3. APLICAÇÃO PRÁTICA DO MÉTODO QUARTIL APLICADO A AUDITORIA

Seja uma conta de clientes extraída do razão auxiliar:

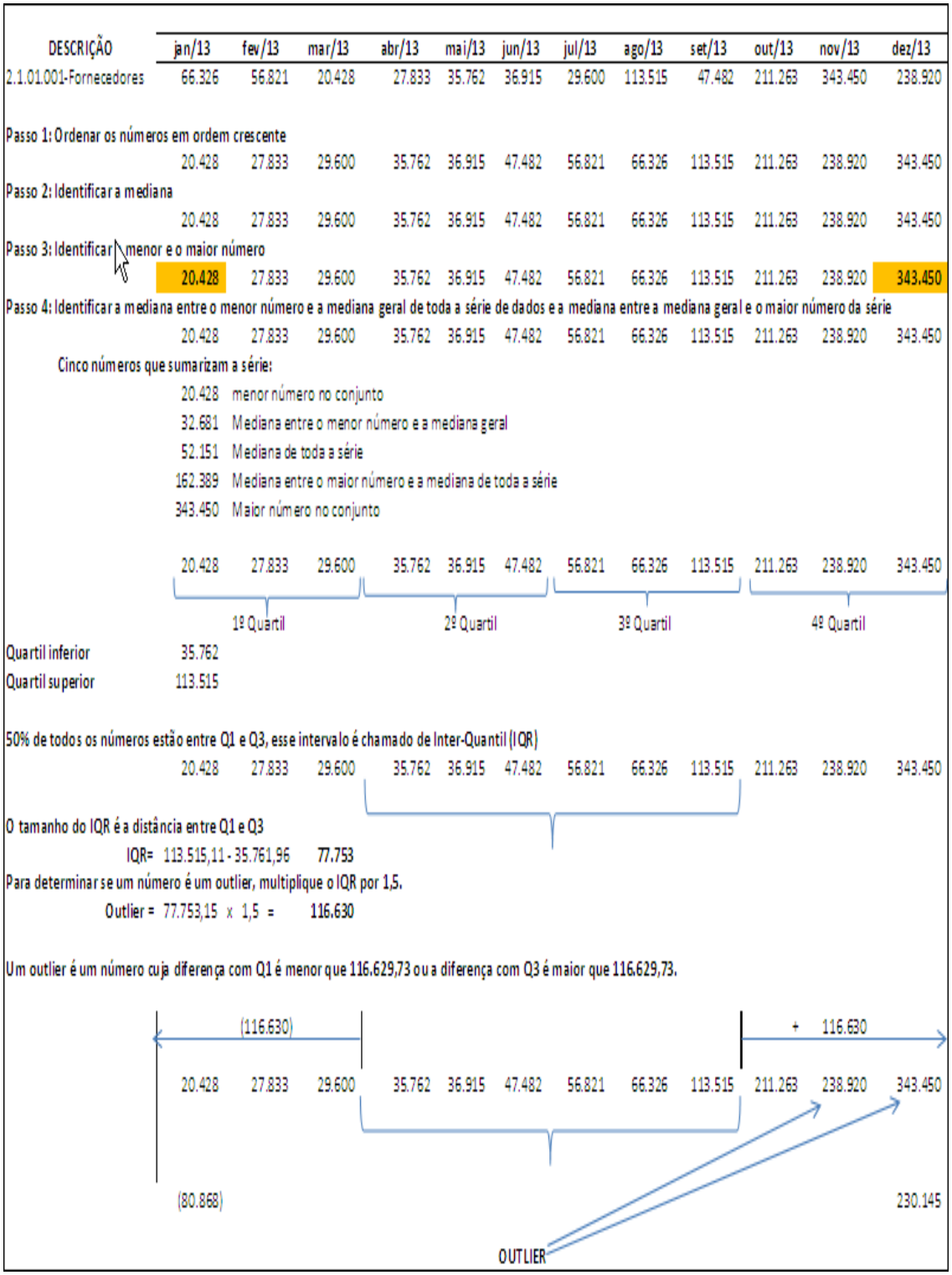
Figura 2 - Quadro demonstrativo de outliers, conta de clientes



Fonte: Elaborado pelos autores

Seja uma conta de fornecedores extraída do razão auxiliar:

Figura 3 - Quadro demonstrativo de outliers, conta de fornecedores



Fonte: Elaborado pelos autores

No caso da análise da conta de clientes, observa-se que não houve discrepância, pois tanto o *outlier* inferior quanto o superior estiveram dentro do limite analisado. Quanto à análise da conta de fornecedores, o *outlier* apresentou discrepância na análise do patamar

REDECA – Revista Eletrônica do Departamento de Ciências Contábeis & Departamento de Atuária e Métodos Quantitativos da FEA-PUC/SP

superior.

Infere-se dos resultados obtidos que o estudo de *outliers* exige uma análise compreensiva para que se tome decisão mais concreta sobre os valores obtidos, tornando-se um ferramental robusto para que seja aplicado aos processos de auditoria.

No caso presente, na ocorrência de *outliers*, trata-se da existência de notas fiscais emitidas pelos diretores da empresa analisada. Verificou-se que esta prestação de serviços estava contabilizada como serviços de digitação em desacordo com a atividade fim daquela sociedade.

Utilizaram-se dados obtidos de um caso real, registrados na contabilidade de uma determinada empresa.

4. OS MÉTODOS MAIS USADOS PARA A DETECÇÃO DE *OUTLIERS*

Barnett e Lewis (1978) definiram *outlier* como uma amostra de tamanho moderado feita a partir de certa população, apresentando que um ou dois valores são surpreendentemente longe do grupo principal. D. M. Hawkins (1980) define um *outlier* como uma observação que se desvia de outras observações, podendo levantar suspeitas do mecanismo usado diferentemente.

Basicamente, existem dois tipos de métodos de *outlier*: Método Formal e Método Informal. Normalmente são chamados “testes de Discordância e Métodos de rotulagem”, respectivamente. O procedimento do teste de detecção deve precisar um teste estatístico, denominado aqui como teste de discordância. Eles normalmente baseiam-se na assunção de alguma distribuição bem comportada, sendo o testar do alvo de ponto do valor extremo um *outlier* naquela distribuição. A seguir, descrevemos alguns métodos que usaremos neste trabalho:

4.1 Método Quartil

No método quartil há a necessidade de usar quadros estatísticos. Para encontrarmos o valor extremo usando o método de quartil, é necessário realizarmos os seguintes passos:

REDECA – Revista Eletrônica do Departamento de Ciências Contábeis & Departamento de Atuária e Métodos Quantitativos da FEA-PUC/SP

a) Calcular o quartil superior: Q3 - 75% dos dados do conjunto de dados são inferiores a este.

b) Calcular o quartil mais baixo: Q1 - 25% dos dados do conjunto de dados são mais elevados do que este.

c) Calcular a distância entre os quartis: H = Q3 - Q1. Um valor mais baixo do que Q1 - 1.5.H e maior do que Q3 1.5. H é considerado um *outlier* leve. Um valor mais baixo do que Q1-3.H e maior do que Q3 3. H é considerado um extremo *outlier*.

4.2 Método Hampel

No teste de Hampel, não são necessários quadros estatísticos. Teoricamente, esse método é resistente, o que significa que ele não é sensível a *outliers*; ainda assim, não tem nenhuma restrição quanto à abundância do conjunto de dados.

O teste de Hampel executa os passos seguintes para os conjuntos de dados:

1) Calcula-se a mediana (Me) para o conjunto de dados total. A mediana é descrita como o valor numérico que separa a maior metade de um conjunto de dados a partir da menor metade;

2) Calcula-se o valor do desvio do valor médio r_i; esse cálculo deve ser feito para todos os elementos do conjunto de dados: $r_i = (x_i - \mathrm{Me})$, em que x_i - dados simples do conjunto da série i - pertence ao conjunto $\{1,\ldots,n\}$, n - Número de todos os elementos do conjunto, e Me - mediana.;

3) Calcula-se a mediana do desvio $\mathrm{Me}_{|r_i|}$;

4) Verifica-se as condições de: $|r_i| \geq 4.5\mathrm{Me}_{|r_i|}$;

5) Se a condição ocorrer, então o valor dos dados conjunto pode ser aceito como um *outlier*.

Algumas definições importantes:

Hipóteses Nula e Alternativa: a hipótese nula (H_0), muitas vezes representa tanto uma

perspectiva cética ou uma alegação a ser testada. A hipótese alternativa (H_A) representa um pedido subsidiário sob consideração e é frequentemente representada por uma gama de possíveis valores de parâmetros. Estrutura de testes de hipóteses: o cético não rejeitaria a hipótese nula (H_0), a menos que a evidência a favor da hipótese alternativa (H_A) seja tão forte que ela rejeitaria H_0 em favor de H_A.

O p-valor como uma ferramenta em testes de hipóteses: o p-valor quantifica quão fortemente os dados favorecem HA sobre H0. Um p-valor pequeno (geralmente <0.05) corresponde a evidência suficiente para rejeitar H0 em favor de HA.

O teste estatístico é uma estatística resumo particularmente útil para a avaliação de um teste de hipóteses ou identificar o p-valor. Quando uma estimativa pontual é quase normal, utilizamos o Z escore da estimativa pontual como o teste estatístico.

O teste do Qui-quadrado: suponha que estamos a avaliar se existe evidência convincente de que um conjunto de contagens observadas $O_1,\ O_2,\ldots,\ O_k$ em k categorias seja diferente do que poderia ser esperado sob uma hipótese nula. Verificamos as contagens esperadas, que são com base na hipótese nula $E_1,\ E_2,\ldots,\ E_k$. Se cada contagem esperada é de pelo menos 5 e a hipótese nula é verdadeira, então a estatística de teste abaixo segue um qui-quadrado distribuição com $k-1$ graus de liberdade:

$$X^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \ \ldots + \frac{(O_k - E_k)^2}{E_k}$$

O p-valor para esse teste estatístico é determinado olhando para o limite superior da distribuição qui-quadrado. Consideramos o limite superior, porque para valores maiores de X^2 , proporcionaria uma maior evidencia contra da hipótese nula.

REDECA – Revista Eletrônica do Departamento de Ciências Contábeis & Departamento de Atuária e Métodos Quantitativos da FEA-PUC/SP

4.3 Método distribuição t de *Student*

A distribuição t é uma distribuição de probabilidade teórica. É simétrica e é semelhante à curva normal padrão, porém com caudas mais largas; uma simulação da t de *Student* pode gerar valores mais extremos que uma simulação da normal. O único parâmetro ν que a define e caracteriza a sua forma é o número de graus de liberdade. Quanto maior for esse parâmetro, mais próxima da normal ela será.

Suponha que Z tenha a distribuição normal, com média 0 e variância 1, que V tenha a distribuição Qui-quadrado com ν graus de liberdade, e que Z e V sejam independentes. Então:

$$t = \frac{Z}{\sqrt{V/\nu}}$$

tem a distribuição t de *Student* com ν graus de liberdade.

Grande parte dos livros estatísticos trazem uma tabela com valores para a distribuição t de *Student*. Essas tabelas apresentam valores arredondados, os quais podem ser grosseiros demais, dependendo do tipo de análise que está sendo feita. *Softwares* estatísticos e planilhas como Microsoft Excel e Open Office Calc possuem técnicas mais precisas para a estimação desses valores.

4.4 Método distribuição Qui-quadrado

A distribuição χ^2 ou chi-quadrado é uma das mais utilizadas em estatística inferencial, principalmente para realizar testes de χ^2, os quais avaliam quantitativamente a relação entre o resultado de um experimento e a distribuição esperada para o fenômeno. Isto é, ele nos informa com quanta certeza os valores observados podem ser aceitos como regidos pela teoria em questão. Muitos outros testes de hipótese usam, também, a distribuição χ^2.

Seja X uma variável aleatória com distribuição normal padronizada. Então X^2 tem distribuição χ^2, com um grau de liberdade.

Sejam $X_1, X_2, \ldots, X_n$ variáveis aleatórias independentes normalmente distribuídas com média 0 e variância 1. Então $Z = \sum_{i=1}^{n} X_i$ segue uma distribuição qui-quadrado com n graus de liberdade.

Seja $X_1, X_2, \ldots, X_n$ uma amostra aleatória de uma distribuição normal, com média μ e variância σ^2, então

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

4.5 Método *Boxplot*

O *boxplot* resume um conjunto de dados, usando o resumo dos cinco números ao mesmo tempo, e traça observações incomuns. O primeiro passo na construção de um *boxplot* é desenhando uma linha escura que denota a mediana, dividindo os dados em duas metades. Se os dados forem ordenados do menor ao maior, a mediana é uma observação bem no meio. Se houver um número par de observações, haverá dois valores no meio, e a mediana será tomada como a média.

O segundo passo na construção de um *boxplot* é desenhar um retângulo para representar o meio de 50% dos dados. O comprimento total da caixa, mostrada na Figura 1.25 verticalmente, é chamada o intervalo interquartil (IQR). Como o desvio padrão, é uma medida de variabilidade nos dados: a variável mais os dados, quanto maior for o desvio padrão e IQR. Os dois limites da caixa são chamados de primeiro quartil (percentil 25, ou seja, 25% da queda dados abaixo desse valor) e terceiro quartil (percentil 75), os quais são muitas vezes rotulados Q1 e Q3, respectivamente.

Intervalo interquartil (IQR): O IQR é o comprimento da caixa num *boxplot*. É calculado como IQR = Q3 - Q1, em que Q1 e Q3 são os percentis 25 e 75.

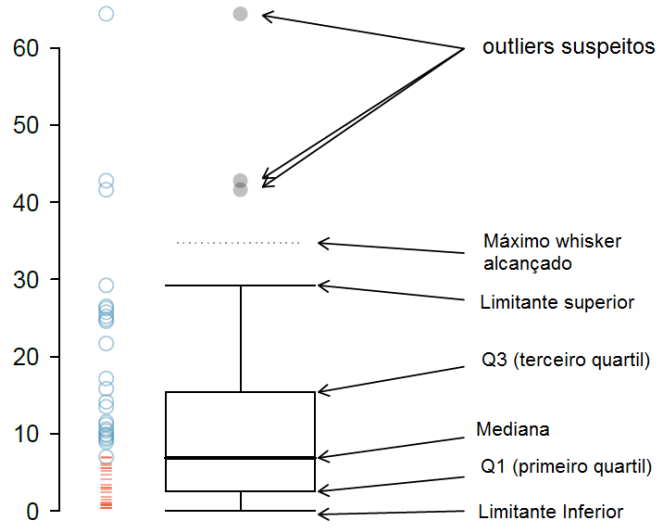
Estendendo para fora da caixa, os limitantes tentam capturar os dados fora da caixa, no entanto seu alcance nunca permite ser mais do que 1.5*IQR. Eles capturam tudo dentro desse intervalo. Na Figura 1, o limitante superior não se prolonga até os últimos três pontos, que estão além do Q3 + 1.5*IQR, e por isso se estende apenas a este último ponto, abaixo do limite. O limitante inferior para no valor mais baixo, de 33 anos, uma vez que não há dados

adicionais a alcançar; limite do limitante inferior não é mostrado na figura, porque o enredo não faz estenderem-se até Q1 - 1.5*IQR. Em certo sentido, a caixa é como o corpo do gráfico de caixa e os limitantes são como seus braços tentando chegar ao resto dos dados.

Os *outliers* são extremos: Um *outlier* é uma observação que aparece extrema em relação ao resto dos dados. A análise dos dados para possíveis valores atípicos serve a muitos propósitos úteis, incluindo

1. Identificação de forte assimetria na distribuição;
2. Identificar os erros de coleta de dados ou entrada;
3. Fornecer informações sobre propriedades interessantes dos dados.

Figura 4 – Gráfico *boxplot* mostrando seus elementos



Fonte: Elaborado pelos autores

4.6 Método gráfico Quantil-Quantil

Suponha dados os valores $\{x_1, x_2, \ldots, x_n\}$ da variável X, e valores $\{y_1, y_2, \ldots, y_m\}$ da variável Y, todos medidos pela mesma unidade.

REDECA – Revista Eletrônica do Departamento de Ciências Contábeis & Departamento de Atuária e Métodos Quantitativos da FEA-PUC/SP

Por exemplo, notas parciais de uma disciplina, ou temperaturas de duas cidades, ou ainda porcentagens da renda familiar gastas com saúde e educação. O gráfico $q-q$ é um gráfico dos quantis da variável X contra os quantis da variável Y. Se $m = n$, o gráfico $q-q$ será um gráfico dos dados ordenados de X contra os dados ordenados Y.

Se as distribuições dos dois conjuntos de dados fossem idênticas, os pontos estariam sobre a reta $y = x$. Enquanto que um gráfico de dispersão fornece uma possível relação global entre as variáveis, o gráfico $q-q$ mostra se valores pequenos de X estão relacionados com valores pequenos de Y, se valores intermediários de X estão relacionados com valores intermediários de Y, se valores grandes de X estão relacionados com valores grandes de Y num gráfico de dispersão, podemos ter $x_1 < x_2$ e $y_1 > y_2$. Num gráfico $q-q$, não é possível ter $x_1 < x_2$ e $y_1 > y_2$, pois os valores em ambos os eixos estão ordenados do menor para o maior.

Os Gráficos quantil-quantil também são uma forma de estudar o comportamento de variáveis, mas utilizando as propriedades que emergem de uma variável quando trabalhamos com os seus quantis. O gráfico quantil-quantil mais tradicional é aquele usado para verificar se uma variável possui distribuição Normal.

No R, isso é realizado com a função 'qqnorm', associada à função 'qqline' que adiciona uma linha ao gráfico:

```
qqnorm( cax$dap )

qqline( cax$dap )
```

A ideia central do gráfico quantil-quantil é a seguinte: quando uma variável segue uma dada distribuição (como a distribuição Normal), os quantis empíricos, isto é, calculados a partir de uma amostra, formam uma linha reta contra os quantis teóricos, calculados a partir das estimativas dos parâmetros da distribuição (no caso da Normal: média e desvio padrão). É isso que a função 'qqnorm' faz para distribuição Normal.

5. TESTES USADOS PARA A DETECÇÃO DE *OUTLIERS*

5.1 Teste de Grubbs

REDECA – Revista Eletrônica do Departamento de Ciências Contábeis & Departamento de Atuária e Métodos Quantitativos da FEA-PUC/SP

Grubbs (1969) é utilizado para se detectar um único *outlier* num conjunto de dados uni variada. O conjunto de dados segue uma distribuição aproximadamente normal e o teste de Grubbs é definido como as duas seguintes hipóteses:

H0: Não há *outlier* no conjunto de dados;

H1: Existe, pelo menos, único *outlier* no conjunto de dados.

A fórmula geral para teste estatístico Grubbs é definida como:

$$G = \frac{\max|Y_i - \overline{Y}|}{s}$$

em que Y_i é o elemento do conjunto de dados, $\overline{Y}$ e *s*, denotando a média da amostra e o desvio padrão; o teste estatístico é o maior desvio absoluto da média amostral em unidades do desvio padrão da amostra. O valor calculado do parâmetro G é comparado com o valor crítico para o teste de Grubbs. Quando o valor calculado é mais elevado ou mais baixo do que o valor crítico de escolha estatística significante, então o valor calculado pode ser aceito como um *outlier*. O significado estatístico (α) descreve o nível máximo de erro que uma pessoa procura num *outlier* aceito.

5.2 Teste de Dixon

O teste desenvolvido por Dixon (1950) é utilizado em testes de pequeno tamanho de espaço amostral. O teste tem algumas limitações para $n \leq 30$, sendo mais tarde estendido para $n \leq 40$ (UNI 9225: 1988).

A primeira etapa do teste serve para organizar os dados em ordem crescente; em seguida, o próximo passo é contar o parâmetro R.

O teste tem vários resultados no parâmetro R.

Supõe-se, por testes, grande conjunto de elemento ser um *outlier*; a amostra é disposta em ordem crescente $X_1 \leq X_2 \leq \cdots \leq X_n$, implicando que o maior elemento da amostra é dado por X_n. Dixon propôs as seguintes estatísticas de teste definido como:

$$R_{10} = \frac{x_n - x_{n-1}}{x_n - x_1}, \qquad \text{para } 3 \leq n \leq 7$$

$$R_{11} = \frac{x_n - x_{n-1}}{x_n - x_2}, \qquad \text{para } 8 \leq n \leq 10$$

$$R_{21} = \frac{x_n - x_{n-2}}{x_n - x_2}, \qquad \text{para } 11 \leq n \leq 13$$

$$R_{22} = \frac{x_n - x_{n-2}}{x_n - x_3}, \qquad \text{para } 14 \leq n \leq 30$$

Para verificar se o elemento menor da amostra é um *outlier* ou não, ordena-se a amostra em ordem decrescente, o que implica na rotulação menor elemento da amostra como x_n. Toda a seleção das estatísticas de teste depende dos critérios de Dixon.

A variável X_n é marcada como um *outlier*, quando o correspondente estatística $R^{(n)}$ exceder um valor crítico, o qual depende do nível α de significância escolhido.

O valor calculado para o parâmetro R é comparado com valor crítico do teste de Dixon para a escolha de estatística significante. Quando o valor calculado para o parâmetro R é maior do que o valor crítico, é possível aceitar os dados do conjunto obtido como um *outlier*.

5.3 Teste Generalizado ESD para *Outlier*s

Rosner (1983) utilizou a generalização (teste do desvio extremo da distribuição t de *Student*) ESD para detectar um ou mais valores discrepantes em um conjunto de dados uni variado que segue uma aproximadamente uma distribuição normal.

O teste generalizado ESD (Rosner, 1983) exige apenas que exista um limite superior para o número suspeito de *outlier*s a ser especificado.

Definição: dado o limitante superior r, o teste generalizado ESD essencialmente realiza r testes por separado, um teste para um *outlier*, outro teste para dois *outlier*s, e assim por diante, até a r *outlier*s. O teste generalizado ESD é definido por meio de hipóteses:

H_0: Não há *outlier* encontrado no conjunto de dados

H_a: Há até r *outlier*s no conjunto de dados

O teste estatístico:

REDECA – Revista Eletrônica do Departamento de Ciências Contábeis & Departamento de Atuária e Métodos Quantitativos da FEA-PUC/SP

$$R_i = \frac{\max |x_i - \bar{x}|}{s}$$

em que $\bar{x}$ e s são a media amostral e o desvio padrão amostral respectivamente.

Retira-se a observação que maximiza $|x_i - \bar{x}|$ e depois calcula-se a estatística anterior com $n - 1$ observações. Fazendo-se isso, de forma continua, o processo repete-se até que r observações sejam removidas. Em seguida, observam-se os resultados em r testes estatísticos $R_1, R_2, \ldots, R_r$.

Nível de significância: α em uma distribuição Normal

Região crítica: Correspondente aos r testes estatísticos, calcula-se os seguintes r valores críticos

$$\lambda_i = \frac{(n-i)t_{p,n-i-1}}{\sqrt{(n-i-1+t^2_{p,n-i-1})(n-i+1)}}$$

em que $i = 1, 2, \ldots, r$, $t_{p,v}$ é o ponto percentual a partir de 100_p da distribuição t de *Student* com ν graus de liberdade e

$$p = 1 - \frac{\alpha}{2(n-i+1)}$$

O número de outliers é determinado por encontrar o maior I, tal que $I > \lambda_i$. Estudos de simulação por Rosner (1983) indicam que esse valor crítico de aproximação é muito preciso para $n \geq 25$. Ele é usado para testar com maior número de outliers do que o esperado quando o teste para *outlier*s entre os dados provenientes ocorre a partir de uma distribuição normal.

6. EXPERIMENTO COMPUTACIONAL DE UMA AMOSTRA ALEATÓRIA REVISADO PELO TESTE GENERALIZADO ESD

Seja o conjunto de dados:

Tabela 1. Conjunto de dados gerados Aleatoriamente

-0.25	0.68	0.94	1.15	1.20	1.26	1.26	1.34	1.38
1.56	1.58	1.65	1.69	1.70	1.76	1.77	1.81	1.91
2.10	2.14	2.15	2.23	2.24	2.26	2.35	2.37	2.40
2.92	2.92	2.93	3.21	3.26	3.30	3.59	3.68	4.30
1.43	1.49	1.49	1.55	2.09	4.64	5.34	5.42	6.01
1.94	1.96	1.99	2.06	2.90	2.47	2.54	2.62	2.64

Fonte: Elaborado pelos autores

Na seguinte tabela, apresentam-se os *outlier*s:

Tabela 2. *Outlier*s encontrados do conjunto dos dados da Tabela 1

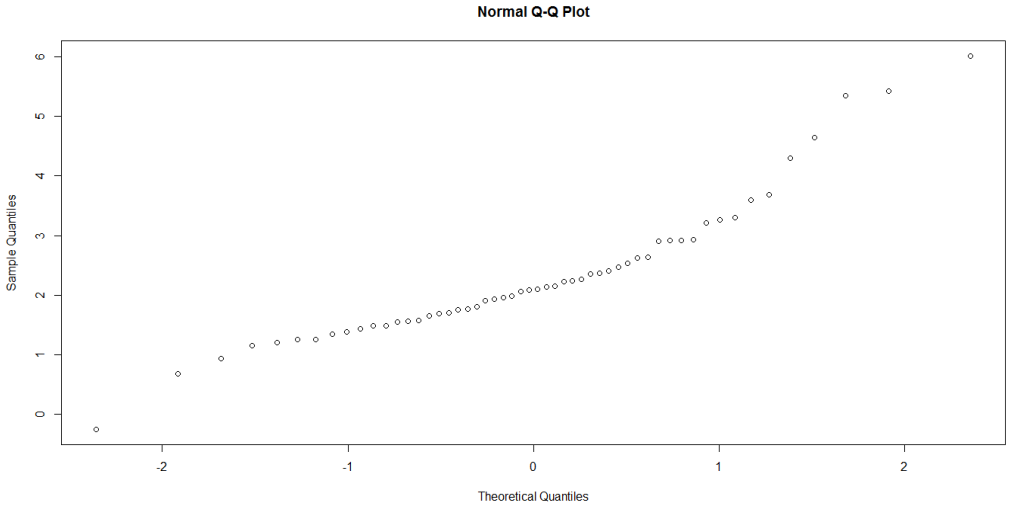
ID Dado	Ponto X	Ponto Y
Ponto 1	3.118906	3.158794
Ponto 2	2.942973	3.151430
Ponto 3	3.179424	3.143890
Ponto 4	2.810181	3.136165
Ponto 5	2.815580	3.128247
Ponto 6	2.848172	3.120128
Ponto 7	2.279327	3.111796
Ponto 8	2.310366	3.103243
Ponto 9	2.101581	3.094456
Ponto 10	2.067178	3.085425

Fonte: Elaborado pelos autores

REDECA – Revista Eletrônica do Departamento de Ciências Contábeis & Departamento de Atuária e Métodos Quantitativos da FEA-PUC/SP

A seguir, observa-se o gráfico da distribuição normal usando o teste generalizado ESD para detecção de *outlier*s, incluídos os 10 *outlier*s apresentados na Tabela 2.

Figura 5 – Gráfico da distribuição normal Quantil - Quantil



Fonte: Elaborado pelos autores

7. EXPERIMENTO COMPUTACIONAL DE UMA AMOSTRA ALEATÓRIA REVISADO PELO TESTE GENERALIZADO ESD USANDO O MÉTODO GRÁFICO QUANTIL - QUANTIL

Seja o conjunto de dados:

Tabela 3. Conjunto de dados gerados aleatoriamente

0.0	0.0	358.59	2326.17	38859.86	43477.41
402193.19	409488.57	490516.46	556401.55	696381.11	92767.38

Fonte: Elaborado pelos autores

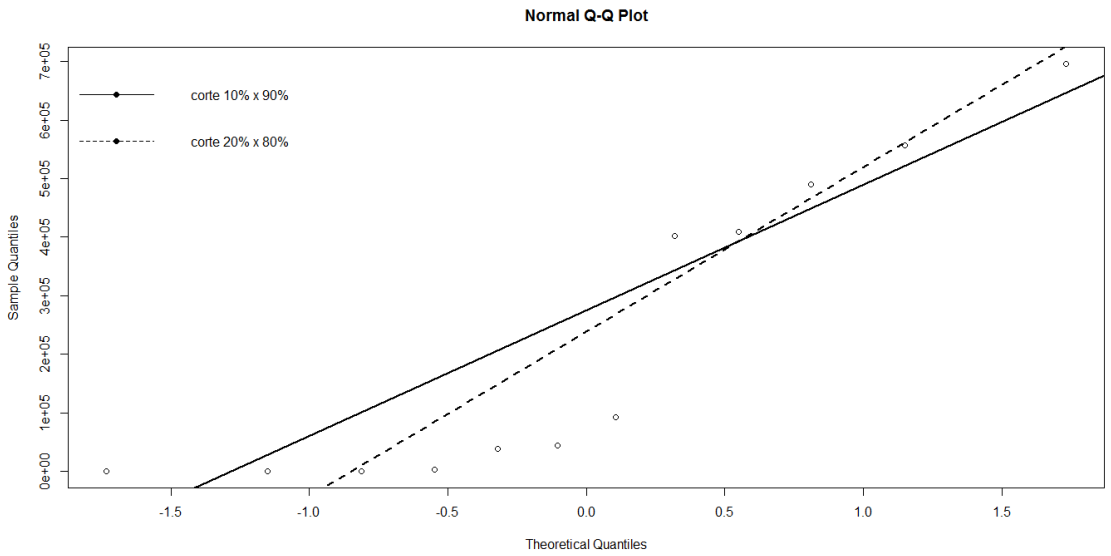
Tabela 4. *Outlier*s encontrados do conjunto de dados da Tabela 3

Quartil	Limite
1°	0.00
2°	1342.38
3°	68122.40
4°	450002.52
5°	696381.11

Fonte: Elaborado pelos autores

REDECA – Revista Eletrônica do Departamento de Ciências Contábeis & Departamento de Atuária e Métodos Quantitativos da FEA-PUC/SP

Figura 6 – Gráfico Quantil-Quantil



Fonte: Elaborado pelo autor

Observam-se as retas de corte para $10-90$ e $20-80$ tem um cruzamento e aproximação das retas, isso quer dizer não existência de *outliers* para o conjunto de dados.

Sejam os dados:

Tabela 5. Conjunto de dados gerados aleatoriamente

20427.52	27833.26	29599.65	35761.96	36915.11	47481.90
56820.99	66325.67	113515.11	211262.68	238920.02	343449.77

Fonte: Elaborado pelos autores

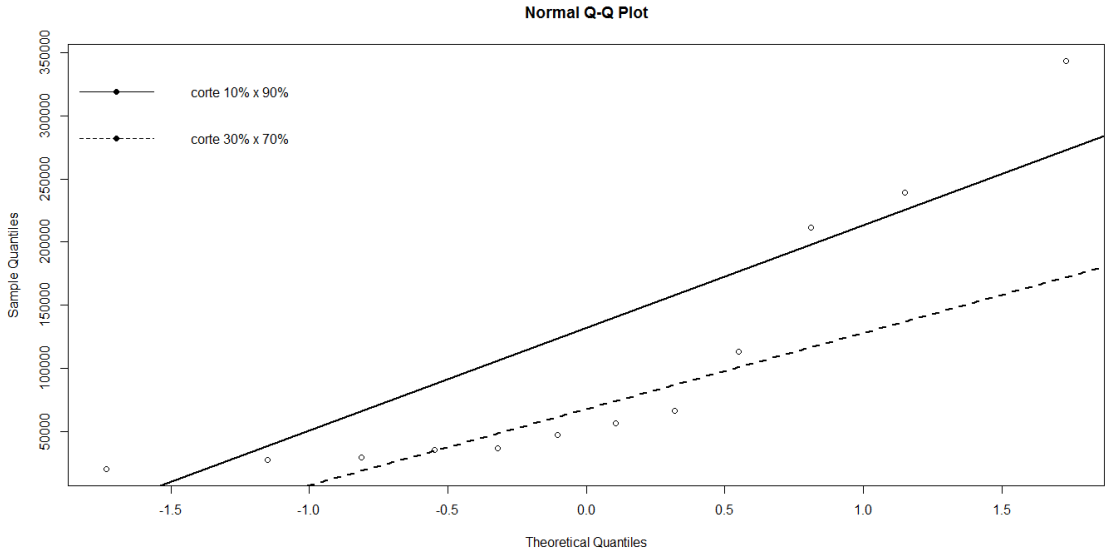
Tabela 6. *Outlier*s encontrados do conjunto de dados da tabela 5

Quartil	Limite
1	20427.52
2	32680.81
3	52151.44
4	162388.89
5	343449.77

Fonte: Elaborado pelos autores

REDECA – Revista Eletrônica do Departamento de Ciências Contábeis & Departamento de Atuária e Métodos Quantitativos da FEA-PUC/SP

Figura 7 – Gráfico Quantil - Quantil



Fonte: Elaborado pelos autores

Observa-se que, nesse caso, temos 4 *outlier*s determinados pela reta corte $10-90$; usando o teste generalizado ESD para *Outlier*s, tem-se:

Tabela 7. O teste estatístico e valor critico para os *Outlier*s da Tabela 6

N° *Outlier*s	Test Stat.	Critical Val.
Ponto 1	2.294721	2.411560
Ponto 2	2.080701	2.354730
Ponto 3	2.524213	2.289954
Ponto 4	2.290881	2.215004

Fonte: Elaborado pelos autores

8. EXPERIMENTO COMPUTACIONAL DE UMA AMOSTRA ALEATÓRIA COM TESTES QUI-QUADRADO, DIXON E GRUBBS

Seja o conjunto 120 dados gerados aleatoriamente:

Tabela 8. Conjunto de dados gerados aleatoriamente

3,57382627	0,99570096	-3,93323431	1,4027118	-0,94558282	-2,13564741
-0,43594983	-2,0520089	-1,45778246	-1,25007854	-3,37338662	1,67557409
0,30674624	-2,27627387	2,50762984	0,85292844	-0,59014297	1,79025132
1,75626698	1,64316216	1,37728051	1,10783531	-0,12382342	-0,61192533
-0,760942	-1,38941396	-0,41583456	-2,5307927	4,33791193	2,415924
-2,24621717	-0,80576967	-0,93331071	1,55993024	-0,16673813	0,50663703
-0,05709351	-0,08574091	2,73720457	-0,45154197	3,03294121	-3,09750561
1,1692275	0,24770849	0,43188314	0,75927897	-1,00464691	-0,66641477
-2,03715077	-2,14358245	0,60705728	0,89641956	0,10600845	1,84453494
4,10016937	-0,98206233	-4,61833775	2,01147705	-1,41840153	-1,37601723
2,05114274	-0,56954601	-2,44143542	0,36260696	-0,55556545	0,02305674
1,5411216	-1,48264013	2,57750619	-0,88194625	1,32712786	4,38735605
1,74072596	-1,30372634	4,59523047	3,97401542	2,19358784	0,95492694
-2,5116243	5,44260979	-2,40103835	8,74933197	6,1304425	-0,94280144
-4,1056836	-2,84162625	1,02753484	-0,98676751	-1,3901704	-3,80647427
-0,1801109	-3,13961788	-6,67176775	-1,52090608	3,67598644	-2,30138785
3,64778593	-9,70729625	-0,33337179	3,11644322	1,80692017	0,63405716
-3,84423605	-5,09822608	-6,14477274	0,70587958	-5,68484769	-2,94334466
-1,53655315	11,06317203	-3,91169941	1,41231943	0,6236868	-7,69485307
-0,57046469	11,55640687	3,61203242	0,32986338	-3,37997466	-16,42597777
11,31337213	-14,60640071	7,39947511	19,09103569		

Fonte: Elaborado pelos autores

Para cada teste para detecção de *outliers*, apresentamos seus parâmetros, p-valor e *Outlier*s resumo para os dados gerados aleatoriamente apresentados na tabela

REDECA – Revista Eletrônica do Departamento de Ciências Contábeis & Departamento de Atuária e Métodos Quantitativos da FEA-PUC/SP

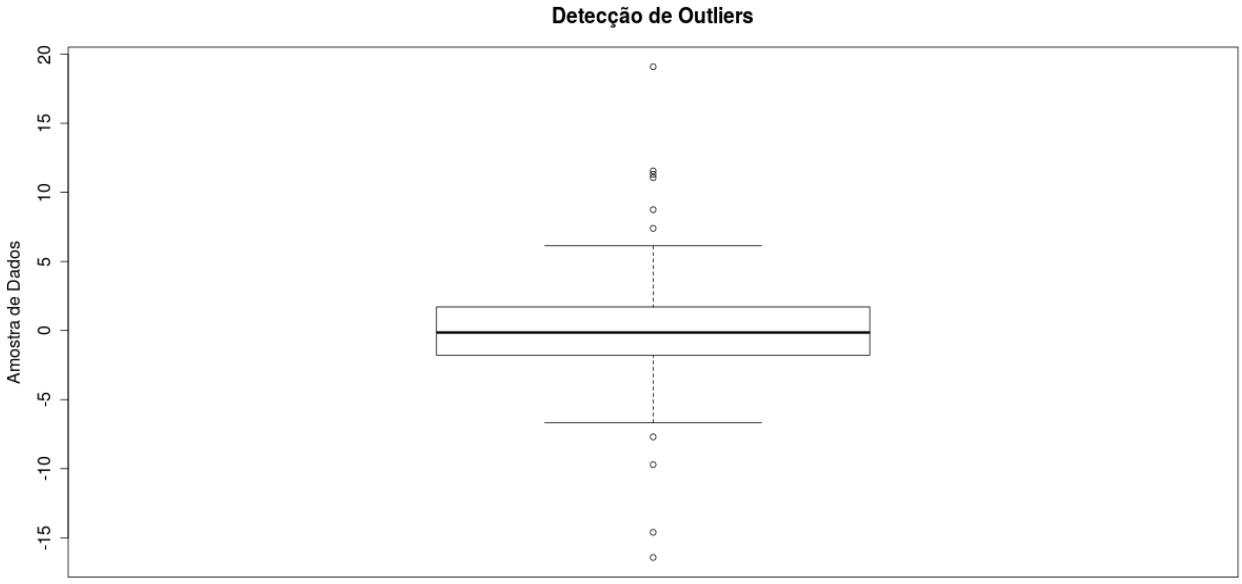
Tabela 9. Comparação dos testes para detecção de *Outliers*

Método Teste	Parâmetro	p-valor	Outiliers
Qui-quadrado	20,1534	7,15E-03	19,091036
Qui-quadrado	15,0434	0,0001051	-16,425978
Dixon	8,9892	0,003383	-16, 425978
Dixon	2,8286	0,003317	19, 091036
Grubbs	3,8786	0,003951	-16,425978
Grubbs	8,3678	2,20E-16	19,091036

Fonte: Elaborado pelos autores

O gráfico a seguir é um caixa preta, da aglomeração dos dados; como pode se observar, os dados mais afastados do conjunto de dados são os *outlier*s detectados pelos métodos identificados anteriormente.

Figura 8 – Gráfico *Boxplot* do conjunto de dados da Tabela 8



Fonte: Elaborado pelos autores

9. CONSIDERAÇÕES FINAIS

Cada vez mais, os registros contábeis passam a fazer parte de grandes bancos de dados, com automação dos sistemas, minimizando-se a utilização de documentos. Desta forma, o contador ou o auditor tem acesso a esses registros ou aos documentos, tendo que

REDECA – Revista Eletrônica do Departamento de Ciências Contábeis & Departamento de Atuária e Métodos Quantitativos da FEA-PUC/SP

confiar nas informações presentes; surge então a necessidade de utilização de ferramentas e mecanismos para auxílio a fim de se constatar a veracidade de tais registros.

As técnicas de *outlier*s são ferramentas robustas no auxílio ao trabalho do auditor ou também do contador, que visa basicamente a identificação de anomalias, erros, fraudes e apuração de detecção de intrusões.

Nos resultados obtidos através dos experimentos apresentados neste artigo, constatou-se uma precisão na identificação de anomalias de registros, embora tais recursos estejam muito incipiente para o auditor e para o contador.

Apresentou-se um método operacional que poderá subsidiar esses profissionais, com técnicas avançadas de identificação de valores contabilizados de forma distorcidas e com resultados positivos de detecção.

Foi possível identificar valores de *outlier*s, que comprovadamente representaram desvios não justificados na análise efetuada, o que justifica a robustez do método.

Como possíveis estudos futuros, indica-se o aprofundamento de tais técnicas e procedimentos que utilizem mineração de dados para obtenção de análises de valores enviesados, muito presente nos registros contábeis, o que possibilitará, além da rapidez, a identificação desses registros.

10. REFERÊNCIAS

ALMEIDA, M. C. *Auditoria: um curso moderno e completo*. São Paulo: Atlas, 1996.

BARNETT, V.; LEWIS, T. *Outliers in statistical data*. Wiley series in probability and mathematical statistics. 3. ed., 1963, pp. 143-264.

BONINI, E. E.; BONINI, S. E. *Estatística: teoria e exercícios*. São Paulo: Nobel 1972. 439 páginas

CAULCUTT, R. Statistic in research and development. Chapman & Hall/CRC Texts in Statistical Science, 2. ed. , pp 100-116.

DIEZ, D. M.; BARR, C. D.; RUNDEL, M. Ç.. *OpenIntro statistics*. 2. ed. 2014, pp. 221-235, 273-287. Disponível em <https://www.openintro.org/stat/textbook.php> Acesso em 06.04.2015.

REDECA – Revista Eletrônica do Departamento de Ciências Contábeis & Departamento de Atuária e Métodos Quantitativos da FEA-PUC/SP

DIXON, W. J. Analysis of extreme values. Ann. Math. Statist. Vol. 21, n. 4, 488-506, 1950. Disponível em <projecteuclid.org/download/pdf_1/euclid.aoms/1177729747> Acesso em 06.04.2015.

ELLISON, S. L. R.; FARRANT, T. J.; BARWICK, V. *Practical statistics for the analytical scientist: A bench guide valid analytical measurement series*. Royal Society of Chemistry, 2009.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. *From data mining to knowledge discovery in databases*. American Association for Artificial Intelligence, 1996.

GRUBBS F. E. *Procedures for detecting outlying observations in samples. Technometrics*, Vol. 11. Issue 1, 1969, pp. 1-21. Disponível em <http://amstat.tandfonline.com/toc/utch20/11/1#.VSLU1GaSCn8> Acesso em 06.04.2015.

HAWKINS, D. M. Identification of *Outlier*s. Chapman and Hall, London, 1980. Disponível em <http://professor.ufabc.edu.br/~ronaldo.prati/DataMining/*Outlier*s.pdf> Acesso em; 28.03.2014.

ROSNER, B. *Percentage points for a generalized ESD many-outlier procedure*. *Technometrics*, Vol. 25, n. 2, 1983, pp. 165-172. Disponível em http://www.jstor.org/discover/10.2307/1268549?sid=21105902910091&uid=2129&uid=2&uid=70&uid=4&uid=3737664 Acesso em 06.04.2015.

TAN, P-N. ; STEINBACH, M.; KUMAR, V. *Introdução ao DATAMINING mineração de dados*. Rio de Janeiro: Editora Ciência Moderna Ltda, 2009.

UNI 9225. *Precision of test methods: determination of repeatability and reproductivity*, 1988.

WANG, J. *Encyclopedia of data warehousing and mining*. Montclair State University. 2. ed., 2008, pp. 214-289,

11. ANEXO I

O resultado obtido foi elaborado através de um programa, em linguagem R.

Apresentamos, a seguir, um *script* que pode ser utilizado no R para utilização de estudo de *outliers*. Inicialmente, deve-se instalar o pacote da livraria "kmeans", disponível na linguagem R.

#R commands and output:
Input data.
y =

-0.25	0.68	0.94	1.15	1.20	1.26	1.26	1.34	1.38
1.56	1.58	1.65	1.69	1.70	1.76	1.77	1.81	1.91
2.10	2.14	2.15	2.23	2.24	2.26	2.35	2.37	2.40
2.92	2.92	2.93	3.21	3.26	3.30	3.59	3.68	4.30
1.43	1.49	1.49	1.55	2.09	4.64	5.34	5.42	6.01
1.94	1.96	1.99	2.06	2.90	2.47	2.54	2.62	2.64

```
## Generate normal probability plot.
qqnorm(y)
## Create function to compute the test statistic.
rval = function(y){
  ares = abs(y - mean(y))/sd(y)
df = data.frame(y, ares)
r = max(df$ares)
  list(r, df)}
## Define values and vectors.
n = length(y)
alpha = 0.05
lam = c(1:10)
R = c(1:10)
## Compute test statistic until r=10 values have been
## removed from the sample.
for (i in 1:10){
  if(i==1){
    rt = rval(y)
    R[i] = unlist(rt[1])
    df = data.frame(rt[2])
    newdf = df[df$ares!=max(df$ares),]}
  else if(i!=1){
    rt = rval(newdf$y)
    R[i] = unlist(rt[1])
    df = data.frame(rt[2])
    newdf = df[df$ares!=max(df$ares),]}
## Compute critical value.
  p = 1 - alpha/(2*(n-i+1))
  t = qt(p,(n-i-1))
  lam[i] = t*(n-i) / sqrt((n-i-1+t**2)*(n-i+1))}
## Print results.
newdf = data.frame(c(1:10),R,lam)
names(newdf)=c("No. Outliers","Test Stat.", "Critical Val.")
newdf
```

N° *Outliers*	Test Stat.	Critical Val.
Ponto 1	3.118906	3.158794
Ponto 2	2.942973	3.151430
Ponto 3	3.179424	3.143890
Ponto 4	2.810181	3.136165
Ponto 5	2.815580	3.128247
Ponto 6	2.848172	3.120128

REDECA – Revista Eletrônica do Departamento de Ciências Contábeis & Departamento de Atuária e Métodos Quantitativos da FEA-PUC/SP