Outliers: O que são?

- As observações que apresentam um grande afastamento das restantes ou são inconsistentes com elas são habitualmente designadas por **outliers**.
- Estas observações são também designadas por observações *"anormais"*, *contaminantes*, *estranhas*, *extremas* ou *aberrantes*.

Outliers: O que fazer com este tipo de observações?

- A preocupação com observações outliers é antiga e data das primeiras tentativas de analisar um conjunto de dados. Inicialmente pensava-se que a melhor forma de lidar com este tipo de observações seria através da sua eliminação da análise.
- As opiniões não eram unânimes: uns defendiam a rejeição das observações "inconsistentes com as restantes", enquanto outros afirmavam que as observações nunca deveriam ser rejeitadas simplesmente por parecerem inconsistentes com os restantes dados e que todas as observações deviam contribuir com igual peso para o resultado final.

Outliers: Causas do seu aparecimento.

Antes de decidir o que deverá ser feito às observações outliers é conveniente ter conhecimento das causas que levam ao seu aparecimento. Em muitos casos as razões da sua existência determinam as formas como devem ser tratadas. Assim, as principais causas que levam ao aparecimento de outliers são:

- Erros de medição;
- Erros de execução;
- Variabilidade inerente dos elementos da população.

Outliers: Aplicação Práticas.

- Detecção de fraudes.
- Comportamento de gastos de consumidores.
- Em análises médicas (resultados não esperados de tratamentos).
- Pesquisa farmacêutica.
- Marketing.
- Etc.

Outliers: Estudo.

O estudo de outliers, independentemente da(s) sua(s) causa(s), pode ser realizado em várias fases:

- A fase inicial é a da identificação das observações que são potencialmente aberrantes. A identificação de outliers consiste na detecção, com métodos subjectivos, das observações surpreendentes. A identificação é feita, geralmente, por análise gráfica ou, no caso de um número de dados ser pequeno, por observação directa dos mesmos. São assim identificadas as observações que têm fortes possibilidades de virem a ser designadas por outliers.

- Na segunda fase, tem-se como objectivo a eliminação da subjectividade inerente à fase anterior. Pretende-se saber se as observações identificadas como outliers potenciais o são, efectivamente. São efectuados testes à ou às observações "preocupantes". Devem ser escolhidos os testes mais adequados para a situação em estudo. As observações suspeitas são testadas quanto à sua discordância. Se for aceite a hipótese de algumas observações serem outliers, elas podem ser designadas como discordantes. Uma observação diz-se discordante se puder considerar-se inconsistente com os restantes valores depois da aplicação de um critério estatístico objectivo. Muitas vezes o termo discordante é usado como sinónimo de outlier.

- Na última fase é necessário decidir o que fazer com as observações discordantes. A maneira mais simples de lidar com essas observações é eliminá-las. Como já foi dito, esta abordagem, apesar de ser muito utilizada, não é aconselhável. Ela só se justifica no caso de os outliers serem devidos a erros cuja correcção é inviável. Caso contrário, as observações consideradas como outliers devem ser tratadas cuidadosamente pois contêm informação relevante sobre características subjacentes aos dados e poderão ser decisivas no conhecimento da população à qual pertence a amostra em estudo.

Outliers: Métodos de identificação.

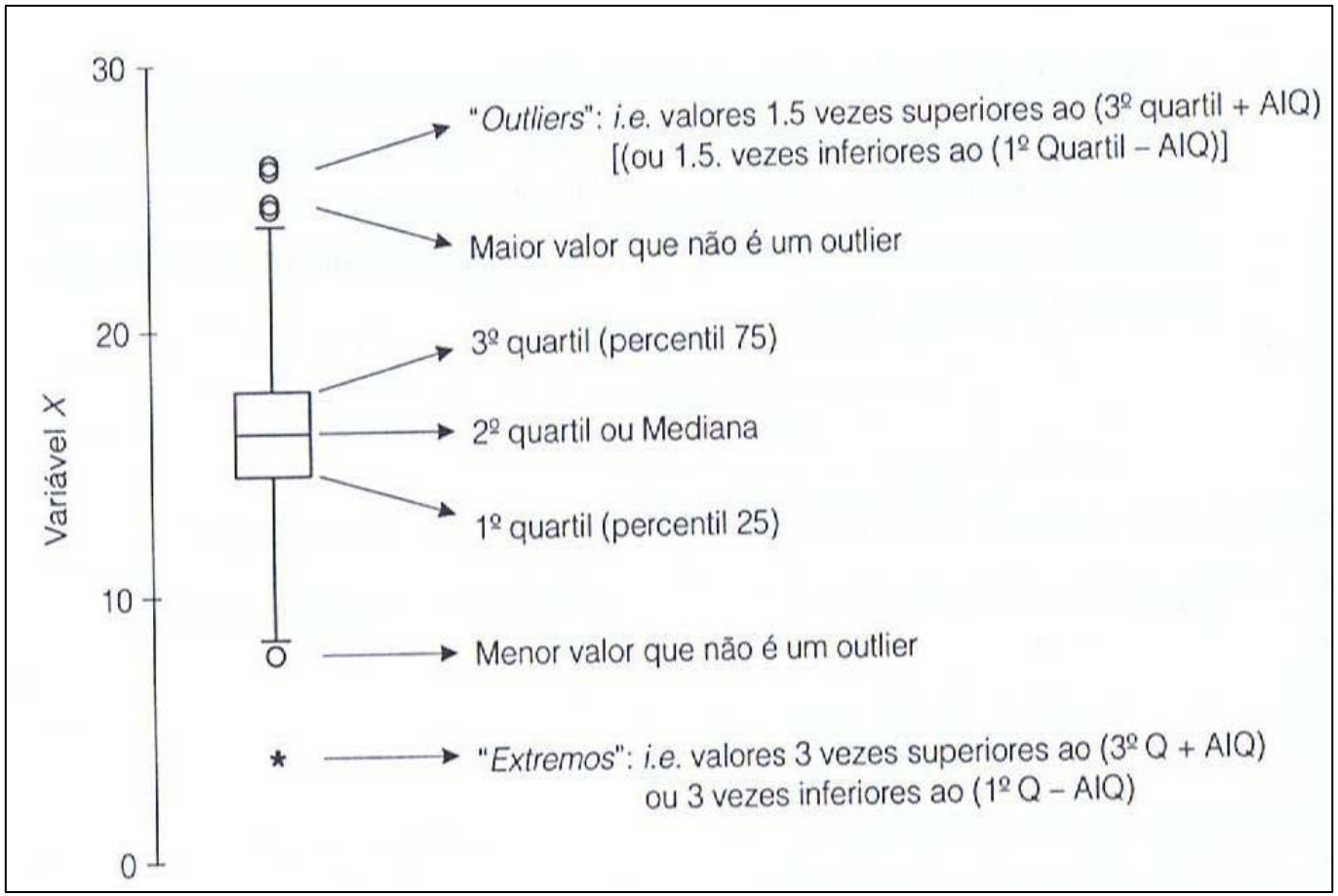
- Gráfico de Box
- Modelos de discordância
- Teste de Dixon
- Teste de Grubbs
- Z-scores
- etc

Gráfico de Box-Plot

O gráfico de Box é construído da seguinte forma:

- Calcula-se a mediana, o quartil inferior (Q1) e o quartil superior (Q3);
- Subtrai-se o quartil superior do quartil inferior = (L)
- Os valores que estiverem no intervalo de Q3+1,5L e Q3+3L e no intervalo Q1-1,5L e Q1-3L, serão considerados outliers podendo, portanto ser aceitos na população com alguma suspeita;
- Os valores que forem maiores que Q3+3L e menores que Q1-3L devem ser considerados suspeitos de pertencer à população, devendo ser investigada a origem da dispersão. Estes pontos são chamados de extremos.

OUTLIERS
Conceitos básicos



Moroco, J. (2003), Análise Estatística de dados – com utilização do SPSS, Edições Sílabo, Lisboa, pág. 36

Modelos de discordância:

Num modelo de discordância considera-se que num dado conjunto de dados, se existirem observações aberrantes elas têm distribuição diferente das restantes observações ou distribuições idênticas mas com parâmetros diferentes.

H_0: a amostra foi retirada de uma população com distribuição específica que pode ou não ser conhecida e ser especificada completamente ou não, e onde não existem observações "anormais"

H_1: todas as observações ou apenas as "anormais" têm distribuição diferente da da hipótese nula.

A hipótese nula será rejeitada a favor da hipótese alternativa se existirem observações aberrantes.

Para decidir pela aceitação ou rejeição da hipótese nula, da não existência de outliers é necessário utilizar testes de discordância que tenham distribuição desconhecida ou valores críticos tabelados. Na utilização de testes formais de outliers deve ter-se em conta que eles dividem-se em duas classes:

- aqueles em que as observações discordantes da amostra são identificadas como sendo outliers, e
- aqueles que testam a presença de outliers mas não identificam

observações particulares.

Teste de Dixon

- Distribuição normal; teste bilateral.
- Ordenar os valores de forma crescente de “1” a “H”.
- Supor a hipótese de que o menor valor, 1, ou o maior valor, H, são suspeitos como valores outliers.

OUTLIERS
Conceitos básicos

- Critérios :

	Extremo Inferior	Extremo superior
- n=3 a 7	$D = \frac{z(2) - z(1)}{z(H) - z(1)}$	$D = \frac{z(H) - z(H-1)}{z(H) - z(1)}$
- n=8 a 12	$Q = \frac{z(2) - z(1)}{z(H-1) - z(1)}$	$D = \frac{z(H) - z(H-1)}{z(H) - z(2)}$
- n > 13	$D = \frac{z(3) - z(1)}{z(H-2) - z(1)}$	$D = \frac{z(H) - z(H-2)}{z(H) - z(3)}$

- Se D > valor crítico, temos a presença de um outlier.

n	Valor crítico de D para P=0,05
3	0,970
4	0,829
5	0,710
6	0,628
7	0,569
8	0,608
9	0,504
10	0,530
11	0,502
12	0,479
13	0,611
14	0,589

Teste de Grubbs

- Distribuição normal;
- Calcular desvio di de cada ponto em relação à média

$$d_i = \left| x_i - \overline{x} \right|$$

- Calcular o desvio-padrão s
- Calcular G=di/s

$$G = \frac{\left| x_i - \overline{x} \right|}{s}$$

- Um valor é considerado como outlier quando G é maior do que o valor crítico correspondente na tabela.

OUTLIERS
Conceitos básicos

n	Gcrit 95 %
3	1,154
4	1,481
5	1,715
6	1,887
7	2,020
8	2,127
9	2,215
10	2,290
11	2,355
12	2,412
14	2,507
16	2,586
18	2,652
20	2,708
50	3,128

Z-Scores

- Calcular os z-scores, isto é, os valores z-standardizados dos dados.
- Se o conjunto dos dados é pequeno (inferior a 50), valores que tenham z-scores inferiores a -2.5 ou superiores a 2.5 devem ser considerados outliers.
- Se o conjunto dos dados é grande, valores que tenham z-socres inferiores a -3.3 ou superiores a 3.3 são tipicamente considerados outliers.
- Se o conjunto dos dados é muito grande (1000 ou mais), também valores mais extremos do que +-3.3 podem ser considerados dados normais e não outliers.

Exemplo:

Olhemos para este conjunto de 10 observações:

Observação	*X*	Y
1	111	68
2	92	46
3	90	50
4	107	59
5	98	50
6	150	66
7	118	54
8	110	51
9	117	59
10	94	97

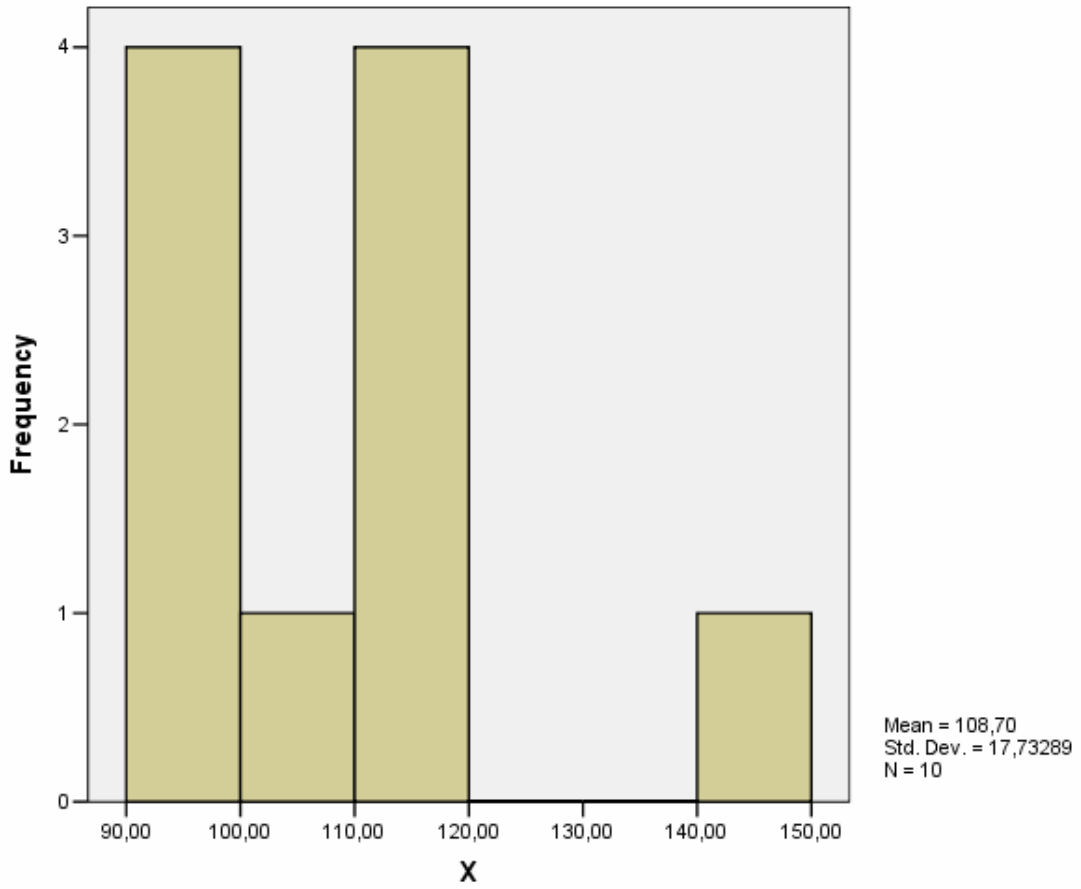
Utilizando a técnica dos Z – Scores a observação 5 da variável X é um outlier, o mesmo acontece para a observação 10 da variável Y.

Observação	*z-score de X*	z-score de Y
1	0.129702456075883	0.538815906080324 7
2	-0.941752615855318 8	-0.942927835640568 3
3	-1.05453736026913	-0.673519882600405 9
4	-0.0958670327517384 5	-0.067351988260040 59
5	-0.603398382613886 7	-0.673519882600405 9
6	**2.32900497214519 2**	0.404111929560243 5
7	0.524449061524220 4	-0.404111929560243 5
8	0.073310083868977 6	-0.606167894340365 3
9	0.468056689317315 1	-0.067351988260040 59
10	-0.828967871441508 1	**2.49202356562150 2**

OUTLIERS
Conceitos básicos

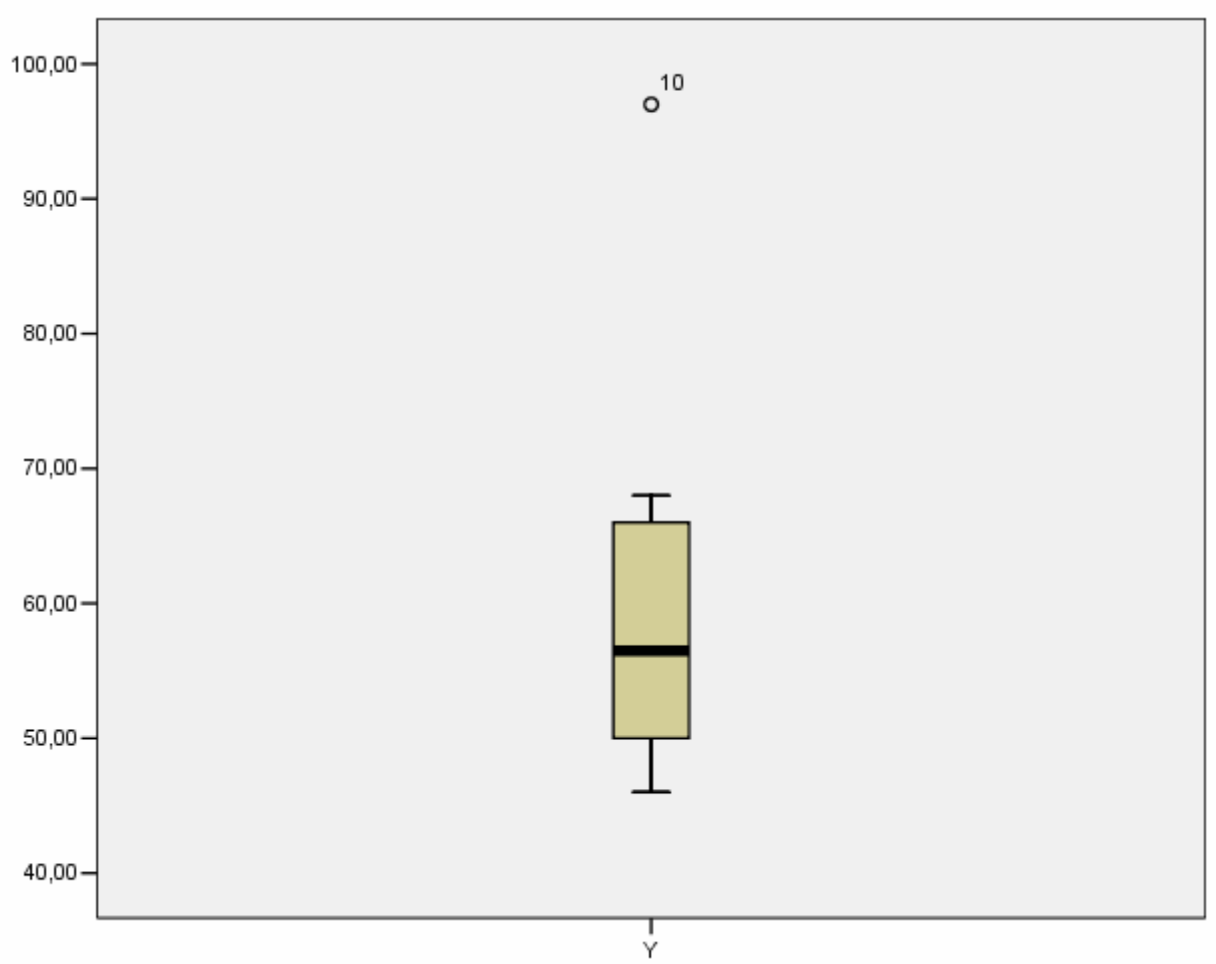
É possível detectar os outliers usando histogramas e box-plots:

Histograma de X



OUTLIERS
Conceitos básicos

Box – Plot de Y



Para aplicar o Teste de Dixon à variável X é necessário ordenar os valores por ordem crescente: 90, 92, 94, 98, 107, 110, 111, 117, 118, 150. O último valor é suspeito com outlier. Como temos n=10 observações, calculemos

$$D = (150-118) / (150 – 92) = 0.5517$$

Da tabela apresentada anteriormente para o teste de Dixon, para uma amostra de tamanho 10, o valor crítico de D é igual a 0.530 (para p=0.05). Como o valor de D excede esse valor, a observação suspeita é efectivamente um outlier.

OUTLIERS
Conceitos básicos

Da tabela apresentada anteriormente para o teste de Grubbs, para uma amostra de tamanho 10, o valor crítico de G é igual a 2.290 (para p=0.05). Como o valor de G, para a observação 97, excede esse valor, a observação é um outlier.

Yi	di	G
68	8	0,54
46	14	0,94
50	10	0,67
59	1	0,07
50	10	0,67
66	6	0,4
54	6	0,4
51	9	0,61
59	1	0,07
97	37	2,49

Média = 60

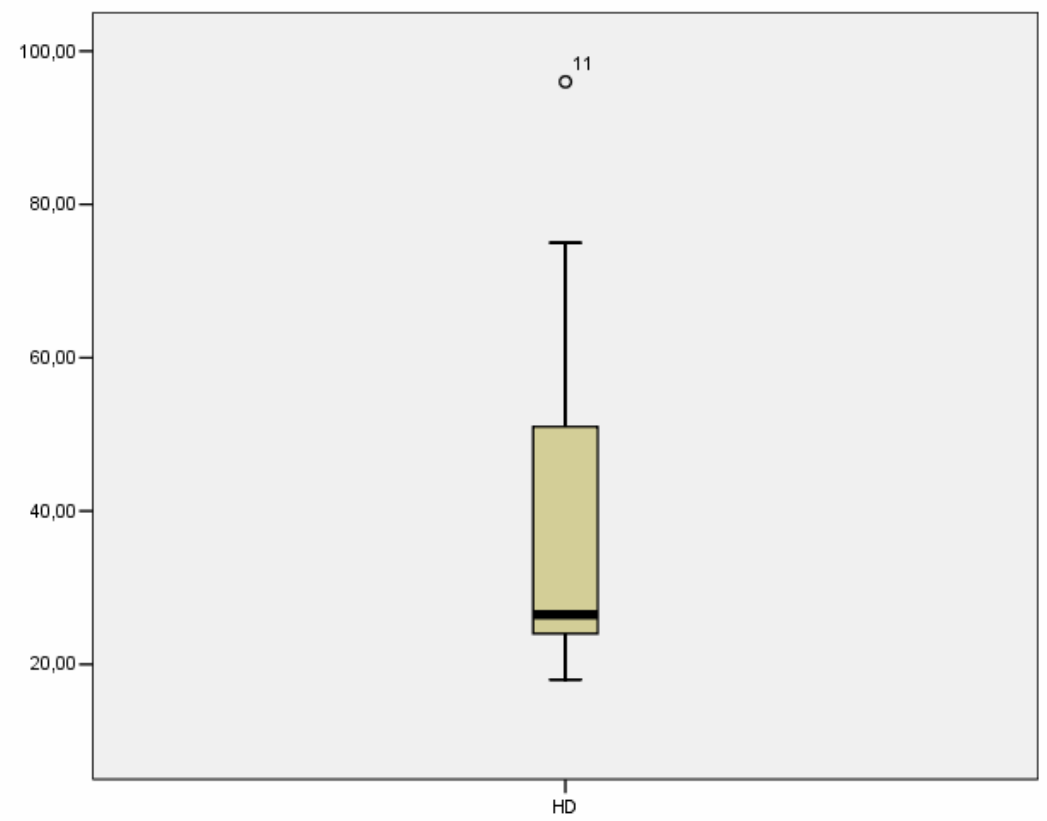
Desvio padrão = 14.8474

Exercícios:

1. Os valores seguintes referem-se às concentrações de nitrito numa amostra de água de um rio: 0.403, 0.410, 0.401 e 0.380. A última observação é suspeita: deverá ser considerada um outlier?
2. Os dados que se seguem referem-se à precipitação (em mm) caída num determinada cidade durante 5 meses: 53.5, 61.5, 62.3, 64.9, 40.6. Algum dos valores referidos anteriormente pode ser considerado um outlier?
3. Os valores seguintes referem-se à produção de trigo: 12.0, 12.4, 13.0, 11.8, 14.0, 12.8, 14.0, 13.5, 12.6, 13.0, 12.6, 12.7. Algum dos valores referidos anteriormente pode ser considerado um outlier?

OUTLIERS
Conceitos básicos

4. Considere os seguintes tempos de hemodiálise (em meses) em 14 doentes transplantados: 51, 24, 55, 75, 24, 27, 22, 23, 48, 18, 96, 24, 26 e 35. Verifique se alguma destas observações pode ser considerada um outlier.



Bibliografia

- Figueira, M.M.C, Identificação de Outliers, MILLENIUM nº12 – Outubro de 1998.
- Morel P., Validação e Incerteza na Medição Analítica, Ministério da Saúde, ANVISA / GGLAS

  http://www.anvisa.gov.br/reblas/cursos_gglas/validacao_incertezas_pierre_2.pdf
- Andrade, E.A. e Robin J., Seminário - Mineração de Exceções

  www.cin.ufpe.br/~compint/aulas-IAS/kdd-012/Outliers.ppt
- Miler, J.C. e Miler, J.N. (1988), Statistics for Analytical Chemistry – second edition, John Wiley & Sons, New York, Chichester, Brisbane, Toronto.