Escore Z na detecção de Outliers

O termo Escore Z, valor Z ou simplesmente Z é uma medida estatística de um número em relação à média do grupo de números. Refere-se a pontos ao longo da base da curva normal padronizada. O ponto central da curva tem um valor Z de 0. Valores Z à direita de 0 são positivos e os valores de Z à esquerda são valores negativos. Um escore Z está acima da média se estiver à direita de 0 e abaixo da média se estiver à esquerda do 0 como ponto central. A distância da média é medida por desvios padrão. Se o escore Z for 0, é 0 o desvio padrão em relação a média, significando que o valor é igual à média.

O escore Z é calculado considerando a diferença entre o número e a média e depois dividindo a diferença obtida pelo desvio padrão.

$Z = (X – \mu)/\sigma$

X representa o número bruto. A média da população é representada por μ e o desvio padrão é o símbolo σ.

Um escore Z é padronizado. É irrelevante se você está comparando dólares canadenses, dólares americanos, euros ou libras. De fato, a unidade pode estar medindo altura, peso, níveis de escolaridade ou notas nos testes. O escore Z é sempre relativo à média que é o centro ou designada como zero. O escore Z informa muito sobre a distribuição dos números no seu conjunto de dados e pode destacar extremos.

Com o escore Z, a área sob a curva normal pode ser determinada utilizando-se de algum software ou olhando para as tabelas. Você pode pesquisar essas tabelas na Internet. Um exemplo está localizado em https://cursosextensao.usp.br/mod/resource/view.php?id=24361

Um escore Z de 1,50 indica que 43,32% da área sob a curva normal está localizada entre a média e o escore Z de 1,50. Quanta área sob a curva seria a pontuação Z entre –1,96 e +1,96? Como observamos os lados negativo e positivo da média, multiplicamos a área entre a média e Z por 2 (0,4750 * 2), resultando em 95%. Quanto maior o valor absoluto do escore Z, mais longe o número está da média ou do padrão. O auditor pode desejar examinar as transações que são extremos.

A teoria estatística afirma que, em 99,7% das vezes, o Escore Z estará entre –3,00 e +3,00. Estará entre –2,00 e +2,00 por 95% das vezes e 68% das vezes, entre –1,00 e +1,00.

Vamos analisar 2 datasets para ver o Escore Z em ação:

In [1]:

```
# Importando as bibliotecas necessárias para a análise.
import pandas as pd
import numpy as np
from scipy import stats
import statistics
pd.set_option('display.max_columns', 10) #Limitar o numero de colunas mostradas
```

Primeiro vamos analisar um conjunto de dados contendo pagamentos. Analisaremos a coluna 'AMOUNT'(montante). Primeiro, veremos qual a média dessa coluna, depois veremos qual o desvio-padrão, e por último criaremos uma coluna 'Z_Score', contendo o escore Z de cada pagamento, o que vai nos permitir ter uma noção de quão longe cada valor está em relação a média, para que você possamos avaliar a magnitude da anomalia caso verifiquemos alguma.

In [2]:

```
# Importanto o conjunto de dados 'Payments'
dates = ['INVOICE_DATE_DATE','PAY_DATE_DATE']
pagamentos = pd.read_csv('Payments.csv', parse_dates=dates)
pagamentos.head() # Mostrando as 5 primeiros linhas do Dataset.
```

Out[2]:

	POSTED_BY	SUPPNO	SUPPNAME	INVOICE_DATE_DATE	INVOICE_DATE_TIME	...	PAY_DATE_TIME	PURCH_ORDE	AUTH	AMOUNT	MEMO
0	KSA	10500	AASMAN DESIGN INC	2011-12-29	00:00:00	...	00:00:00	100092100	CW	545.09	Nal
1	MIA	10900	ACKLANDS - GRAINGER INC	2011-12-29	00:00:00	...	00:00:00	100095400	BC	485.19	Sponsor Wome Labo Associatio
2	SOUB	11400	AG JOYERIA	2011-12-29	00:00:00	...	00:00:00	100095800	BC	227.62	Adjustment by th refiner
											Adjustment

	POSTED_BY	SUPPNO	SUPPNAME	INVOICE_DATE_DATE	INVOICE_DATE_TIME	...	PAY_DATE_TIME	PURCH_ORDE	AUTH	AMOUNT	MEMO
3	KSA	11810	DKNY INC.	2011-12-29	00:00:00	...	00:00:00	100091700	HMV	4839.92	confirme by th refiner
4	ERIC	12203	ANKE KRUSE ORGANICS INC	2011-12-29	00:00:00	...	00:00:00	100097400	HMV	820.63	Nal

5 rows × 13 columns

In [3]:

```
# Calculando a média da coluna 'AMOUNT'
pagamentos.AMOUNT.mean()
```

Out[3]:

20270.35329729729

In [4]:

```
# Calculando o desvio padrão da coluna 'AMOUNT'
np.std(pagamentos.AMOUNT)
```

Out[4]:

26006.28833536601

In [5]:

```
# Criando a colua 'Z_Score' para a coluna 'AMOUNT' e colocando a coluna 'Z_Score' em ordem decrescente.
pagamentos['Z_Score'] = stats.zscore(pagamentos.AMOUNT, axis = 0)
pagamentos.sort_values('Z_Score', ascending= False).head(10) # Mostrando os 10 primeiros resultados
```

Out[5]:

	POSTED_BY	SUPPNO	SUPPNAME	INVOICE_DATE_DATE	INVOICE_DATE_TIME	...	PURCH_ORDE	AUTH	AMOUNT	MEMO	Z_Score
109	MIA	99999	PENNY CILLIN	2011-01-04	00:00:00	...	100088900	HMV	97376.40	NaN	2.964900
19	KSA	20535	JOHN PETERSON	2011-01-05	00:00:00	...	100084100	V.S.T	96166.49	NaN	2.918376
35	MIA	20849	TIMEX INC.	2011-01-10	00:00:00	...	100084900	HMV	94329.52	NaN	2.847741
91	ERIC	92411	SAAN STORES LTD.	2011-01-08	00:00:00	...	100086700	HMV	86441.66	NaN	2.544435
126	MIA	99999	POLLY GUNN	2011-01-07	00:00:00	...	100082400	BC	86117.39	NaN	2.531966
70	CW	60600	MULTI-LINGUAL TEK INC.	2011-01-09	00:00:00	...	100081400	H.M.V.	85728.78	NaN	2.517023
113	CW	99999	WANDA FARR	2011-01-06	00:00:00	...	100085800	BC	83880.73	NaN	2.445961
57	MIA	40205	LOCKSMITH SERVICES LTD	2011-01-16	00:00:00	...	100080200	BC	79571.88	NaN	2.280276
137	MIA	99999	N RICH	2011-01-01	00:00:00	...	100086900	H.M.V.	79500.00	NaN	2.277512
38	CW	21174	SWISS ARMY INC.	2011-01-15	00:00:00	...	100086000	BC	79237.49	NaN	2.267418

10 rows × 14 columns

In [6]:

```
pagamentos.sort_values('Z_Score', ascending= False).tail(10) # Mostrando os 10 últimos resultados
```

Out[6]:

	POSTED_BY	SUPPNO	SUPPNAME	INVOICE_DATE_DATE	INVOICE_DATE_TIME	...	PURCH_ORDE	AUTH	AMOUNT	MEMO	Z_Sc
73	KSA	61900	NORTHERN HOSPITAL SUPPLIES LTD	2011-12-29	00:00:00	...	100097600	V.S.T	59.90	NaN	0.777
119	KSA	99999	P GREEN	2011-12-29	00:00:00	...	100092600	BC	54.32	NaN	0.777
68	MIA	60005	MSS LTD.	2011-01-15	00:00:00	...	100090000	HMV	50.63	NaN	

	POSTED_BY	SUPPNO	SUPPNAME	INVOICE_DATE_DATE	INVOICE_DATE_TIME	...	PURCH_ORDE	AUTH	AMOUNT	MEMO	Z_Sc
78	KSA	92211	RICARDO BAL	2011-12-29	00:00:00	...	100095400	BC	45.34	NaN	0.777
63	DES	40713	MACKENZIE PETROLEUM LTD.	2011-12-29	00:00:00	...	100096100	BC	11.98	NaN	0.778
181	MIA	99999	A MEADOW	2011-12-29	00:00:00	...	100095200	HMV	4.34	Petty cash	0.779
123	DES	99999	MICROCOMPUTERS	2011-12-29	00:00:00	...	100096900	BC	0.00	NaN	0.779
135	WJT	99999	A RAID	2011-12-29	00:00:00	...	100091800	CB	-695.50	Adjustments from refinery	0.806
173	MIA	99999	P GREEN	2011-12-29	00:00:00	...	100095000	WJT	-1198.00	Adjustments refinery	0.825
65	SOUB	40713	MACKENZIE PETROLEUM LTD.	2011-12-29	00:00:00	...	100090600	HMV	-1994.67	Adjustments after refinery	0.856

10 rows × 14 columns

Como podemos verificar, não houve nenhum caso que o Escore Z superou + 3.00 e nem - 3.00.

No próximo dataset, temos dados sobre faturas, nele faremos os mesmos procedimentos do dataset anterior, porém, veremos que temos valores de Escore Z bem elevados.

In [7]:

```
# Importando o dataset
faturas = pd.read_csv('Invoices.csv', sep =';')
faturas['Amount'] = faturas['Amount'].apply(lambda x: x.replace('$', '').replace(',', '.')).astype('float') # Transformando a
# coluna 'Amount' em uma coluna numérica.
faturas.head()
```

Out[7]:

	Date	InvoiceNo	CustomerNo	SalesPerson	ProductNo	UnitPrice	Quantity	Amount
0	09/07/2003	20000	10220	8	8	9,20	41	377.2
1	21/08/2003	20001	10491	4	48	14,00	30	420.0
2	27/08/2003	20002	10704	3	43	15,00	25	375.0
3	28/05/2003	20003	10430	5	54	24,00	22	528.0
4	06/12/2003	20004	10841	17	11	15,00	21	315.0

In [8]:

```
# Calculando a média da coluna 'Amount'
faturas.Amount.mean()
```

Out[8]:

785.5941268253656

In [9]:

```
# Calculando o desvio padrão da coluna 'Amount'
np.std(faturas.Amount)
```

Out[9]:

1293.9044106281592

In [10]:

```
# Criando a colua 'Z_Score' para a coluna 'Amount' e colocando a coluna 'Z_Score' em ordem decrescente
faturas['Z_Score'] = stats.zscore(faturas.Amount, axis = 0)
faturas.sort_values('Z_Score', ascending= False).head(20) # Mostrando os 20 primeiros resultados
```

Out[10]:

	Date	InvoiceNo	CustomerNo	SalesPerson	ProductNo	UnitPrice	Quantity	Amount	Z_Score
4046	28/02/2003	24047	10073	22	60	263,50	51	13438.5	9.778857
3183	24/04/2003	23184	10913	9	14	263,50	51	13438.5	9.778857
251	11/07/2003	20252	10636	9	60	263,50	49	12911.5	9.371562
1548	12/06/2003	21549	10181	12	14	263,50	49	12911.5	9.371562
271	06/07/2003	20272	10934	15	60	263,50	48	12648.0	9.167915
158	30/07/2003	20159	10894	2	60	263,50	47	12384.5	8.964268
4283	05/05/2003	24284	10655	23	60	263,50	47	12384.5	8.964268
1460	19/04/2003	21461	10473	23	14	263,50	46	12121.0	8.760621
350	27/06/2003	20351	10208	10	60	263,50	46	12121.0	8.760621
3210	06/05/2003	23211	10365	25	14	263,50	46	12121.0	8.760621
3140	29/05/2003	23141	10885	6	14	263,50	46	12121.0	8.760621
2483	05/11/2003	22484	10360	7	14	263,50	46	12121.0	8.760621
1280	04/12/2003	21281	10022	25	60	263,50	44	11594.0	8.353326
3252	12/12/2003	23253	10097	2	60	263,50	44	11594.0	8.353326
1497	19/03/2003	21498	10590	22	60	263,50	43	11330.5	8.149679
356	06/11/2003	20357	10098	5	14	263,50	43	11330.5	8.149679
1597	09/07/2003	21598	10029	5	14	263,50	43	11330.5	8.149679
193	25/12/2003	20194	10151	25	14	263,50	42	11067.0	7.946032
3571	30/01/2003	23572	10200	20	60	263,50	41	10803.5	7.742385
4026	22/06/2003	24027	10980	14	14	263,50	41	10803.5	7.742385

In [11]:

```
faturas.sort_values('Z_Score', ascending= False).tail(10) # Mostrando os 10 últimos resultados
```

Out[11]:

	Date	InvoiceNo	CustomerNo	SalesPerson	ProductNo	UnitPrice	Quantity	Amount	Z_Score
4144	20/09/2003	24145	10938	23	55	9,00	1	9.0	-0.600194
4361	01/06/2003	24362	10987	12	34	2,50	3	7.5	-0.601354
2371	22/08/2003	22372	10617	22	34	2,50	3	7.5	-0.601354
3720	13/10/2003	23721	10325	15	42	7,00	1	7.0	-0.601740
493	27/12/2003	20494	10354	17	47	7,00	1	7.0	-0.601740
628	06/01/2003	20629	10304	25	53	6,00	1	6.0	-0.602513
3025	14/03/2003	23026	10487	12	34	2,50	2	5.0	-0.603286
2565	08/01/2003	22566	10344	6	34	2,50	2	5.0	-0.603286
2692	25/03/2003	22693	10666	17	34	2,50	2	5.0	-0.603286
477	21/06/2003	20478	10597	16	34	2,50	2	5.0	-0.603286

Nesse dataset, vemos que há vários pagamentos com Escore Z maior que 8, ou seja, demonstra uma grande variação com relação a média. O auditor precisa julgar quantos *outiliers* devem ser examinados detalhadamente ou determinar um *cut-off* com relação ao Escore Z.

In [ ]:

```
```