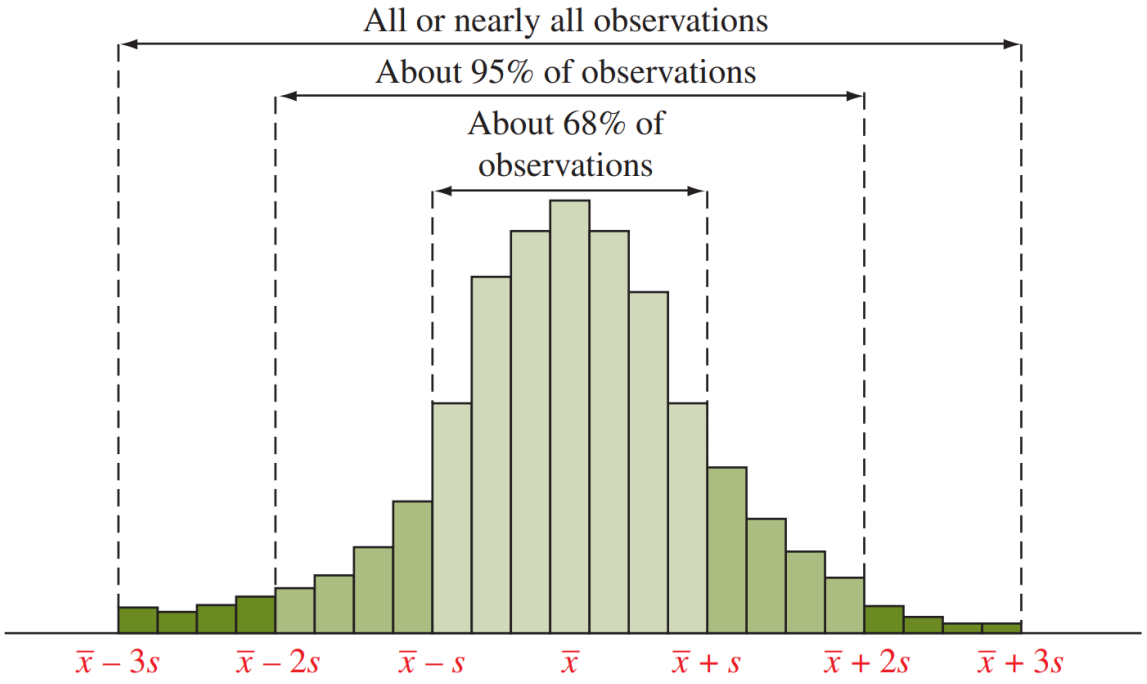
Outliers a partir da média e do desvio padrão - o Z-score

# Aula	32
☑ Preparada	☑
☑ Revisada	☑
☑ Lecionada	☑

▼ Já estudamos a Regra Empírica...



▼ E a partir dela vamos adotar um critério para outliers:

Observações a mais de 3 desvios padrão serão consideradas outliers.

▼ Mas para facilitar, vamos padronizar com o Z-score:

Considerando a variável x, calculamos a variável Z-score como sendo

$$Zscore = \frac{[x - média(x)]}{desviopadrão(x)}$$

▼ Que se "desmembra" em:

$$Zscore_{Populacional} = \frac{(x - \mu)}{\sigma}$$

ou

$$Zscore_{Amostral} = \frac{(x - \overline{x})}{s}$$

▼ Observe que o Z-score tem média igual a 0 e desvio padrão igual a 1.

Uma vez que temos o Z-score, é imediata a constatação de quantos desvios padrão uma determinada observação está da média.

▼ Observe que o Z-score é adimensional.

Isso quer dizer que com o Z-score, podemos comparar variáveis diferentes, já que esta padronização torna as observações comparáveis na distância em desvios padrão da média.

▼ Voltando ao exemplo do professor novato e as idades dos alunos…

▼ Dados originais:

- Repare que nos dados originais a média é de 45,58 anos de idade e o desvio padrão é de 17,62 anos de idade.
- Repare na observação de 70 anos de idade, a 1,39 desvios padrão da média dos dados originais…
- Repare que o Z-score possui (sempre) média igual a 0 e desvio padrão igual a 1…

Original	Z-Score			
57	0,65		Dados Originais	
68	1,27		45,5806	média amostral
66	1,16		17,6177	desvio padrão amostral
70	1,39			
31	-0,83			
30	-0,88		Z-Score	
23	-1,28		0,0000	média amostral
29	-0,94		1,0000	desvio padrão amostral
67	1,22			
61	0,88			
28	-1,00			
68	1,27			
67	1,22			
21	-1,40			
55	0,53			
24	-1,22			
26	-1,11			
55	0,53			
31	-0,83			
69	1,33			
41	-0,26			
31	-0,83			
29	-0,94			
69	1,33			
40	-0,32			
53	0,42			
48	0,14			
41	-0,26			
35	-0,60			
59	0,76			
21	-1,40			

▼ Ao invés de 70 anos, imagine agora que um erro de digitação listasse uma idade de 700 anos…

- Repare que com este outlier a média agora é de 65,90 anos de idade e o desvio padrão agora é de 118,91 anos de idade.
- Repare na observação de 700 anos de idade, a 5,33 desvios padrão da média dos dados originais…
- Repare que o Z-score possui (sempre) média igual a 0 e desvio padrão igual a 1…

Original	Z-Score			
57	-0,07		Dados Originais	
68	0,02		65,9032	média amostral
66	0,00		118,9085	desvio padrão amostral
700	5,33			
31	-0,29			
30	-0,30		Z-Score	
23	-0,36		0,0000	média amostral
29	-0,31		1,0000	desvio padrão amostral
67	0,01			
61	-0,04			
28	-0,32			
68	0,02			
67	0,01			
21	-0,38			
55	-0,09			
24	-0,35			
26	-0,34			
55	-0,09			
31	-0,29			
69	0,03			
41	-0,21			
31	-0,29			
29	-0,31			
69	0,03			
40	-0,22			
53	-0,11			
48	-0,15			

41	-0,21			
35	-0,26			
59	-0,06			
21	-0,38			