

**Data Science
Academy**

www.datascienceacademy.com.br

Big Data Real-Time Analytics com Python e Spark

Caso de Uso

Desejamos estimar a média de dinheiro que uma pessoa de uma determinada cidade gasta comprando produtos anunciados em um canal de televisão. Para começar, precisamos coletar uma amostra aleatória. Vamos considerar que a média da amostra seja R\$ 129,20, com uma margem de erro de R\$11,80 e limites de confiança máximo e mínimo de R\$117.40 e R\$141.00 respectivamente.



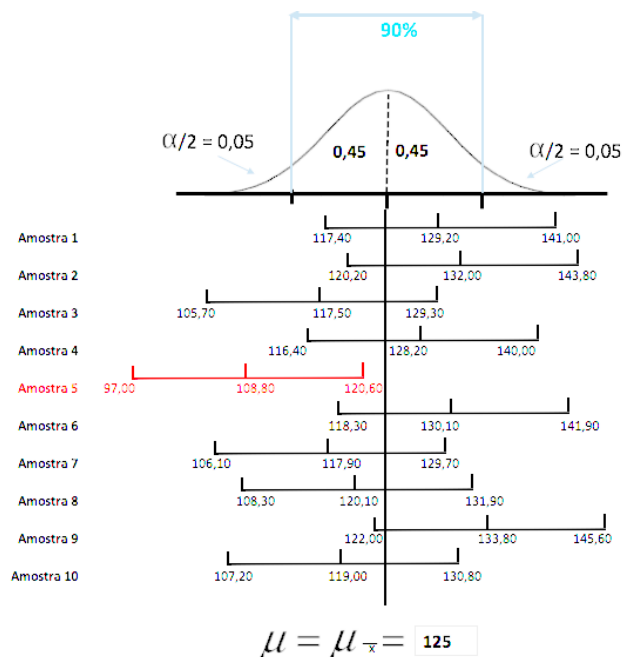
Interpretar o intervalo de confiança, não é simples como parece. Um erro muito comum, é fazer a seguinte afirmação:

“Há 90% de probabilidade de que a média de gastos com produtos anunciados em propagandas de TV esteja entre R\$117.40 e R\$141.00”.

Embora a afirmação acima pareça bastante razoável, ela não pode ser suportada com os cálculos do intervalo de confiança. Vamos coletar mais amostras e calcular os intervalos de confiança:

Amostra	Média da Amostra	Margem de Erro	90% de nível de confiança	
			Limite Mínimo	Limite Máximo
1	R\$ 129,20	R\$11,80	R\$117,40	R\$141,00
2	R\$ 132,00	R\$11,80	R\$120,20	R\$143,80
3	R\$ 117,50	R\$11,80	R\$105,70	R\$129,30
4	R\$ 128,20	R\$11,80	R\$116,40	R\$140,00
5	R\$ 108,80	R\$11,80	R\$97,00	R\$120,60
6	R\$ 130,10	R\$11,80	R\$118,30	R\$141,90
7	R\$ 117,90	R\$11,80	R\$106,10	R\$129,70
8	R\$ 120,10	R\$11,80	R\$108,30	R\$131,90
9	R\$ 133,80	R\$11,80	R\$122,00	R\$145,60
10	R\$ 119,00	R\$11,80	R\$107,20	R\$130,80

Um ponto importante que precisa ser esclarecido aqui é que cada amostra extraída da população tem seu próprio intervalo de confiança, conforme podemos ver na tabela anterior. Note que a margem de erro é a mesma, pois o tamanho da amostra e o desvio padrão da população não foram alterados e todos os intervalos representam 90% de nível de confiança.



Veja que a amostra 5 não contém a média da população. Sendo assim, fazer a afirmação abaixo seria um erro:

“Há 90% de probabilidade de que a média de gastos com produtos anunciados em propagandas de TV esteja entre R\$117.40 e R\$141.00”.

A lição aqui é: não há garantia que cada intervalo de confiança irá incluir a média da população. Esta é a correta definição do nível de confiança.

“Nós esperamos que 90% das médias das amostras de uma população, irão produzir um intervalo de confiança que inclua a média da população.”

Entretanto, não há garantia que 9 de cada 10 intervalos de confiança irão incluir a média da população. Isso é uma estimativa. O mesmo raciocínio pode ser aplicado para 95% e 99% de nível de confiança.

No caso anterior, lidamos com tamanho de amostra superior a 30 elementos. Sob estas condições, as médias das amostras tendem a seguir uma distribuição de probabilidade normal, independente do formato da distribuição de probabilidade da população. Mas o que acontece quando o tamanho da amostra é menor que 30, assumindo que o desvio padrão da população é conhecido? Bem, neste caso não podemos mais nos apoiar no Teorema do Limite Central.



Referências:

The Logic of Science: Principles and Elementary Applications Vol 1

E. T. Jaynes