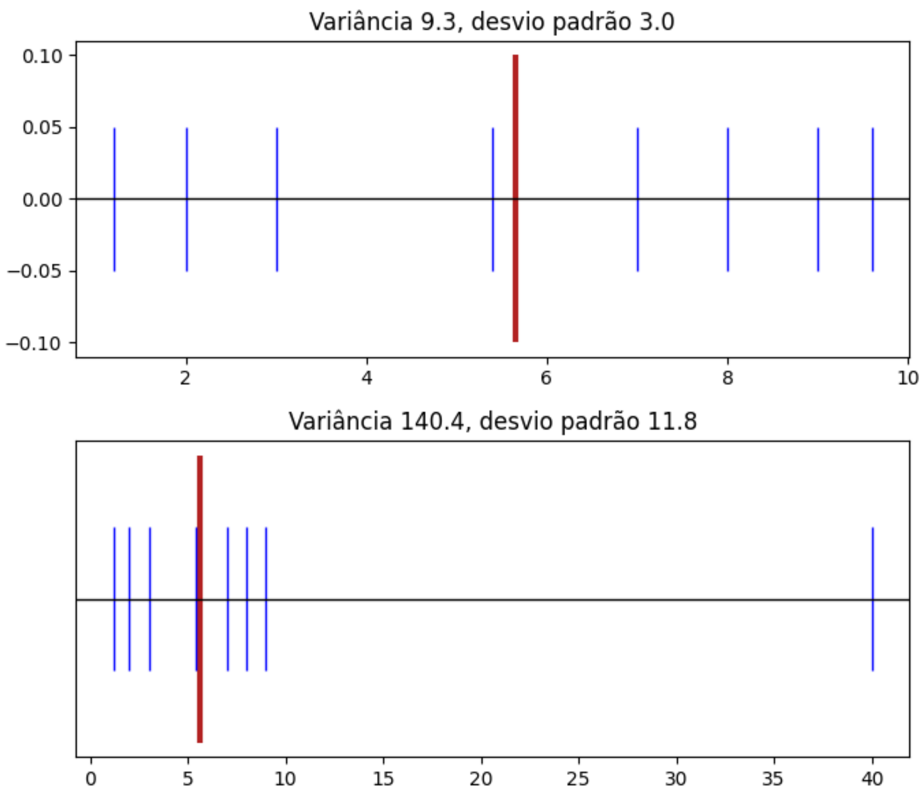
Erro amostral

Erro Padrão da Média
Estimativa de Desvio Padrão da População
Implicações para o planejamento da amostragem

Prof. Renato Paes de Almeida
Instituto de Geociências
Universidade de São Paulo

I - Medidas de variabilidade



Variância:

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$

Desvio padrão:
(mesma unidade da medida)

$$\sigma = \left[\frac{\sum (x_i - \bar{x})^2}{n}\right]^{\frac{1}{2}}$$

Coeficiente de variação:
(adimensional)

$$C_v = \frac{1}{\mu}\left[\frac{\sum (x_i - \bar{x})^2}{n}\right]^{\frac{1}{2}}$$

Sabemos que: $\sigma = \left[\frac{1}{n}\sum e_i^2\right]^{\frac{1}{2}}$

Porém não conhecemos os erros e_i em relação à média da população

Conhecemos apenas os *resíduos* entre cada medida e a média da amostra (d_i)

Podemos calcular o Desvio Padrão da Amostra:

$$S = \left[\frac{1}{n}\sum d_i^2\right]^{\frac{1}{2}}$$

numpy.std(Amostra)

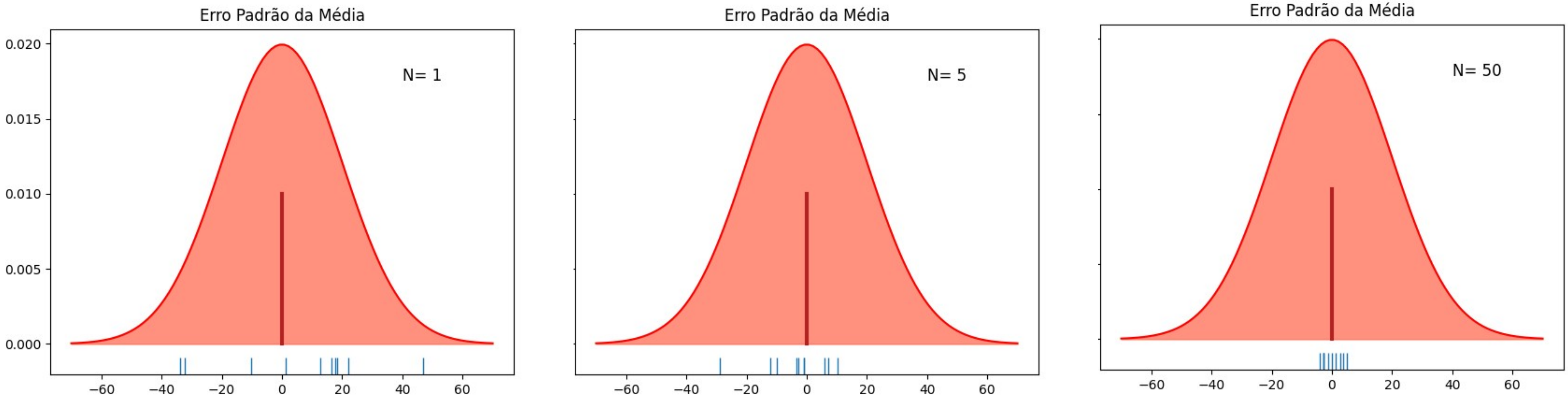
Mas S não é o erro da amostra!

II - O desvio padrão da amostra não é o erro da amostra

Quanto maior a amostra, menor o errro da média da amostra com relação à média da população.

Se retirarmos um número grande de amostras aleatórias de mesmo tamanho de uma dada população, podemos calcular o Erro Padrão da Média, que é uma estimativa da diferença entre a média da amostra e a média da população.

O EPM (σ_m) pode ser calculado com base em S e no tamanho da amostra



\# III - Entendendo a variância de subamostras

$$\sigma_m \approx \left(\frac{1}{n-1}\right)^{\frac{1}{2}} S$$

scipy.stats.sem(Amostra)

Essa é a estimativa do *erro* de uma amostra, e não o Desvio Padrão da Amostra.

Sabemos também que a relação entre o Desvio Padrão da População e o Erro Padrão da Amostra é:

$$\sigma_m = \frac{\sigma}{\sqrt{n}}$$

Portanto podemos estimar o Desvio Padrão da População por:

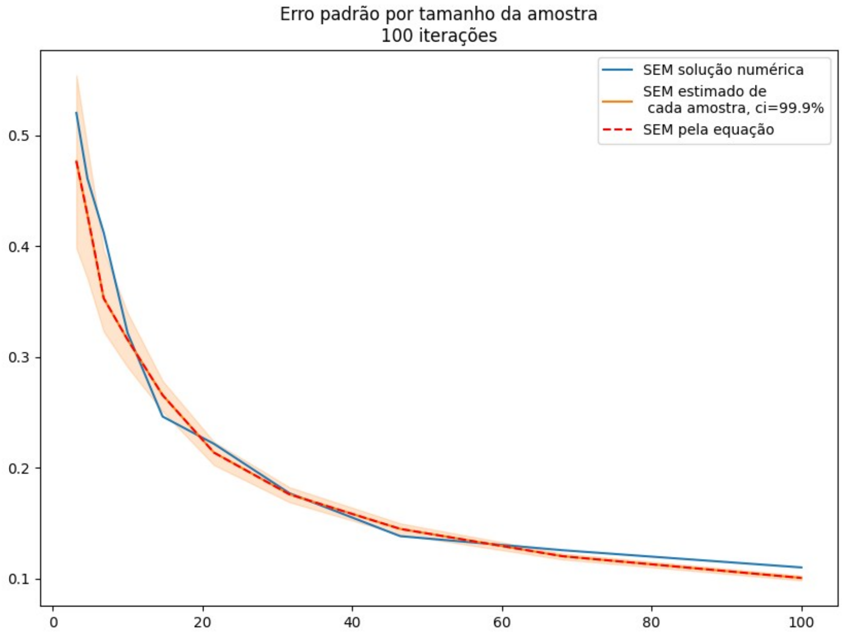
$$\sigma \approx \left(\frac{n}{n-1}\right)^{\frac{1}{2}} S$$

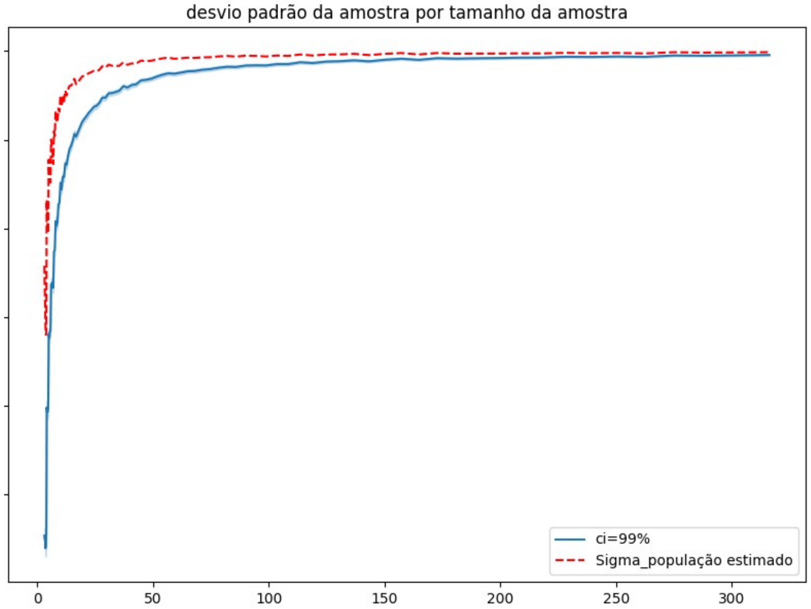
numpy.std(Amostra,ddof=1)

IV - Simulação numérica e estimativa pela equação

$$\sigma_m \approx \left(\frac{1}{n-1}\right)^{\frac{1}{2}} S$$

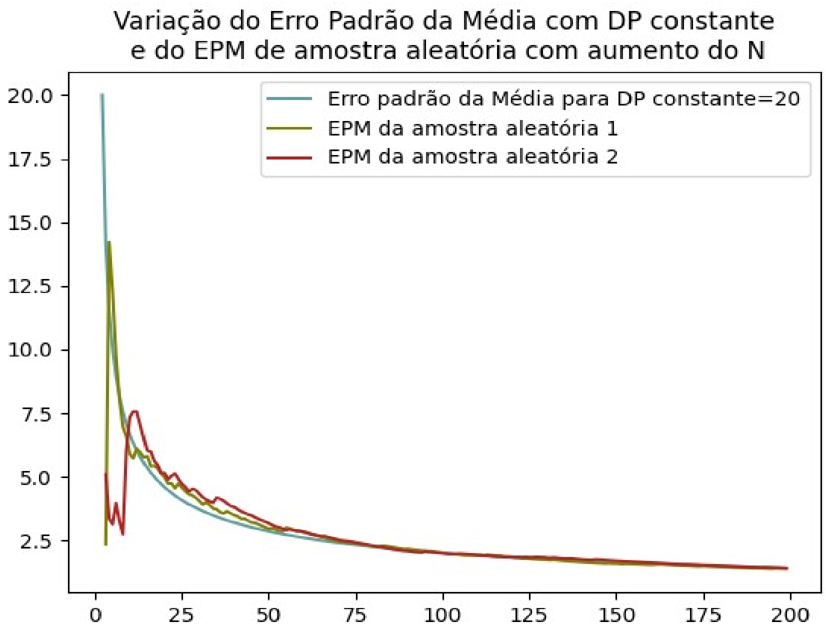
$$\sigma \approx \left(\frac{n}{n-1}\right)^{\frac{1}{2}} S$$





V- Aumentando o o número de análises

Tanto σ_m quanto σ tendem a cair com *n*, porém coletar mais dados (ampliar o tamanho da amostra) não necessariamente reduz essas estimativas.



Considere a amostra [3.3 , 3.5 , 3.2 , 3.4, 5.7 , 0.1]

As primeiras 4 medidas têm S pequeno (0.11) e portanto estimativas de:

$$\sigma_m \approx \left(\frac{1}{4-1}\right)^{\frac{1}{2}} * 0.11 \approx 0.065$$

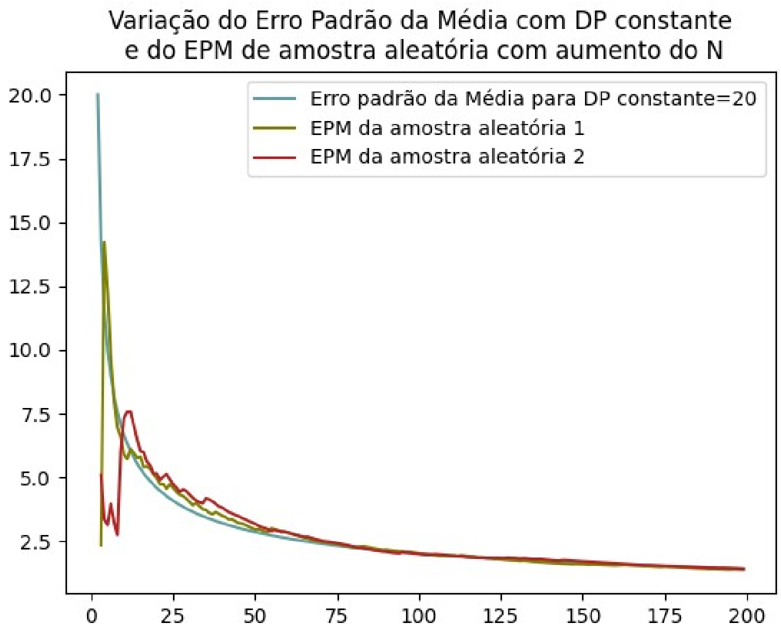
$$\sigma \approx \sigma_m * \sqrt{n} \approx 0.13$$

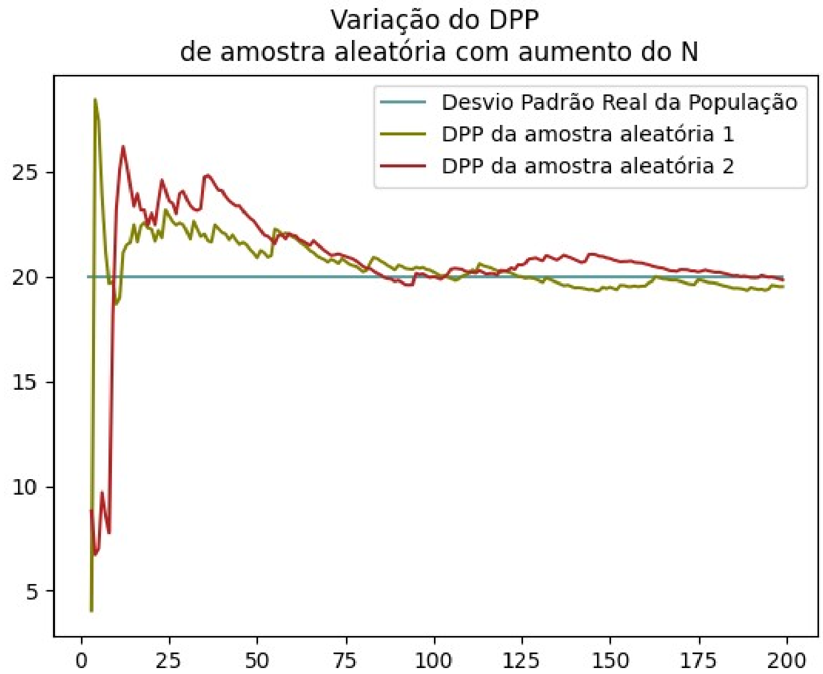
Ao adquirir a quinta e a sexta medidas os valores mudam para:

$$n = 5 \;\Rightarrow S = 0.95,\; \sigma_m \approx 0.47,\; \sigma \approx 1.06$$

$$n = 6 \;\Rightarrow S = 01.63,\; \sigma_m \approx 0.73,\; \sigma \approx 1.79$$

Nesse caso, vale descartar medidas?





Para a curva vermelha, o σ_m com 8 medidas é melhor estimativa do σ_m com 200, mas o σ é muito subestimado. Pior: a média pode não ser precisamente o valor real.

Se o estudo envolve conhecer a dispersão da população, descartar amostras piora a estimativa.

Excluir ponto isolado a 3 σ da média dos demais pode ser aceitável *se há explicação física* para considerar a medida expúria

VI - Exercício calcular erro amostral