Reamostragem

A reamostragem é um processo baseado no conceito de “plug-in”, onde estimamos parâmetros da população a partir de uma amostra, considerando a mesma como uma distribuição empírica aproximada, logo é uma extrapolação.

Dessa amostra podemos gerar várias outras através de fatias ou sorteio com reposição. Não obtemos informação nova, porém podemos analisar várias características da população, independente da distribuição da população.

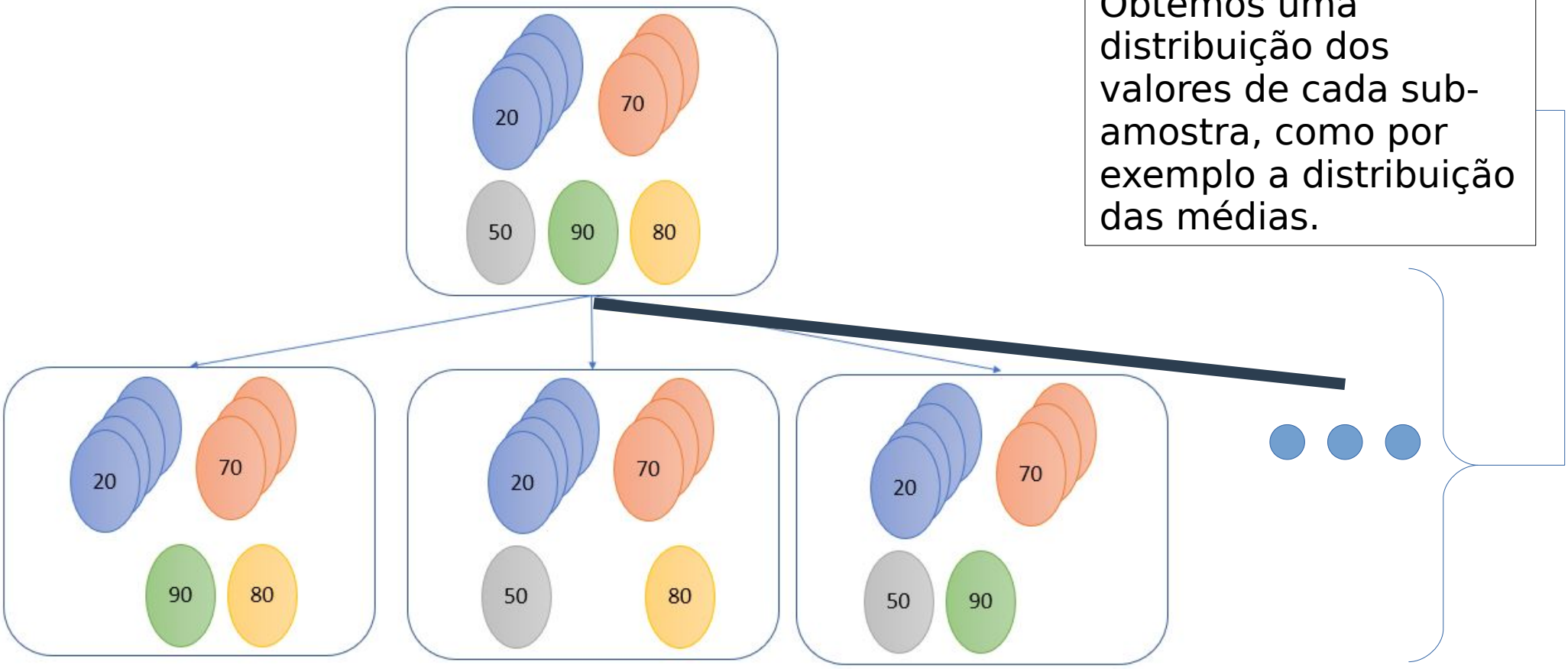
-Utilidade: Erro padrão, mediana (apenas o Bootstrap), intervalos de confiança (assim é possível saber a certeza com relação a extrapolação), proporções e estimativa de viés. ML.

-Benefícios: Não depende da distribuição, estimar parâmetros da população quando não é viável expandir a amostra, é pouco sensível a outliers.

-Cuidados: Não funciona para distribuições com variância indefinida/“cauda extensa”, por exemplo Leis de Potência. Se a amostra é enviesada os resultados de reamostragem também serão.

Jackknife (analisando fatias)

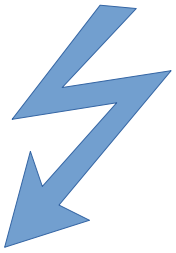
Fatiamos os resultados, **obtendo todas as possíveis amostras com um elemento retirado**.
Os resultados são sempre os mesmos (ideal para reprodução).



Obtemos uma distribuição dos valores de cada sub-amostra, como por exemplo a distribuição das médias.

Reamostragem com reposição

Passo a passo



Também podemos gerar novas amostras sorteando itens de uma amostra. Se repormos os itens sorteados, basicamente estaremos **utilizando a amostra original como distribuição de probabilidade** (pense em termos do número de elementos sobre espaço amostral). Podemos assim **gerar "novas" amostras com o mesmo tamanho que a original**, mas serão valores diferentes, pois o mesmo elemento pode ser sorteado mais de uma vez ou não ser sorteado.

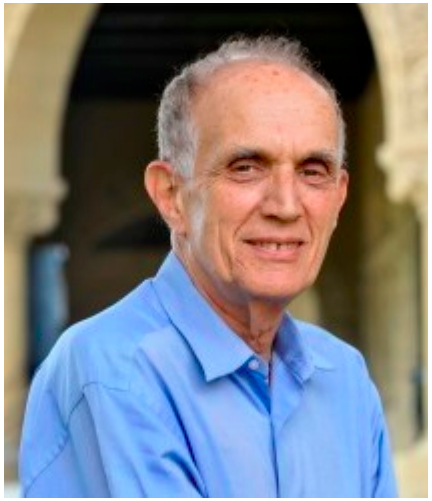
Não ganhamos nova informação com isso, porém podemos estimar o erro padrão através deste método.

Isso pode ser provado pela "Lei dos grandes números" da estatística. É um conhecimento estabelecido que **amostras com 30 elementos ou mais já fornecem uma boa aproximação, amostras com mais de 100 elementos são ideais**. O número de reamostragens ideal é de 1000 ou mais (embora 100 já forneça uma boa aprox.).

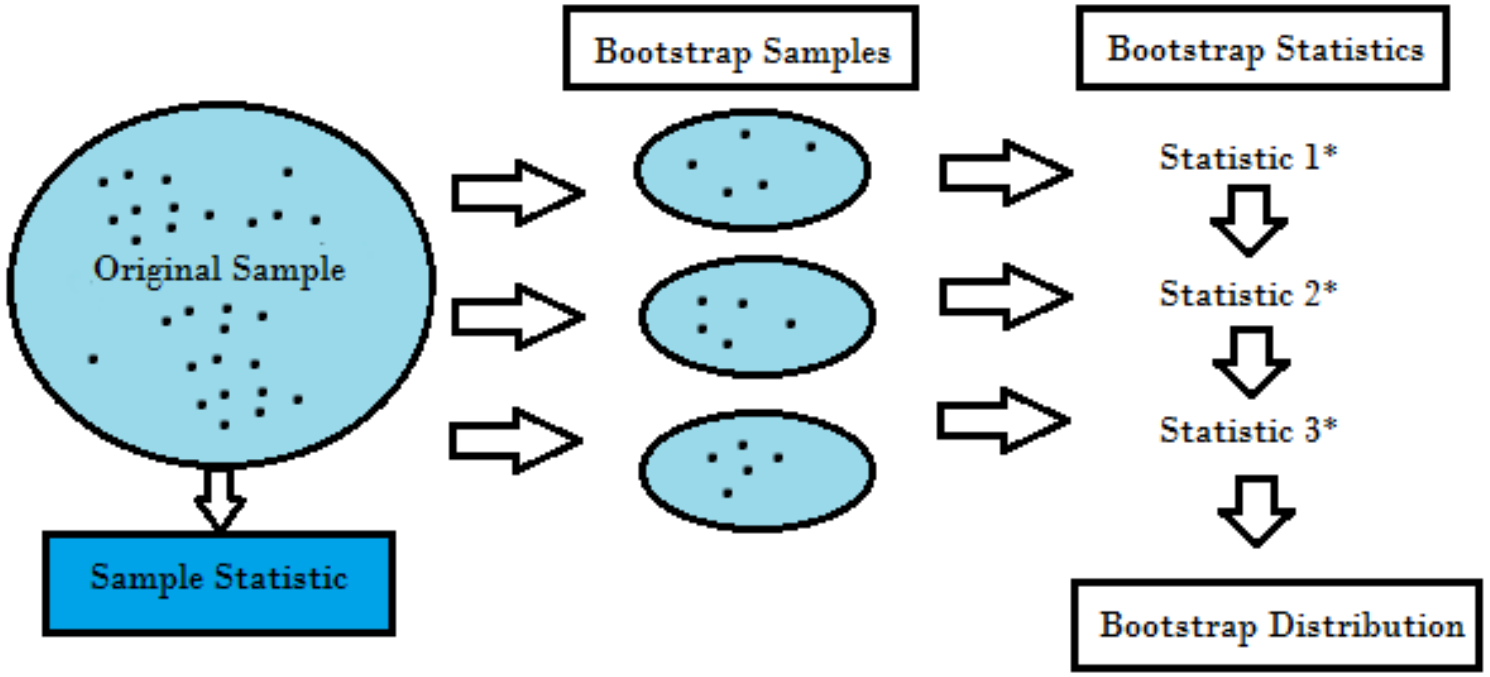
Bootstrap

No slide anterior descrevemos na verdade um método de reamostragem chamado de Bootstrap (alça da botina). Esse método foi um novo passo na era da estatística computacionalmente custosa (veremos em aulas futuras outros métodos custosos utilizadas atualmente).

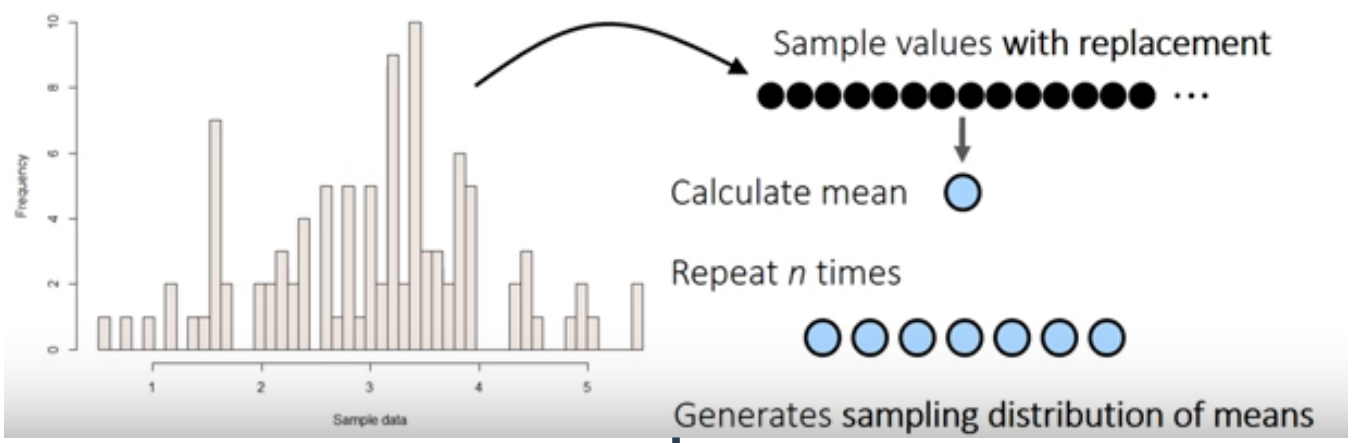
Os resultados nunca serão idênticos pois são realizações aleatórias.

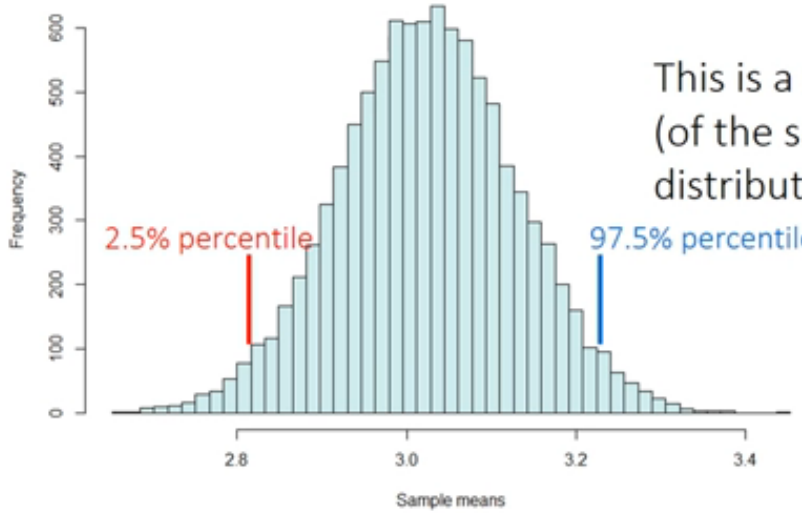


Criado por Bradley Efron em 1979



Bootstrap





This is a *sampling distribution* (of the sample means), not a distribution of data

Parametric calculation	Resampling with replacement
Mean = 3.025323	Mean = 3.024759
95% CI min = 2.820611	2.5th %ile = 2.817557

Bootstrap (parâmetros enviesados)

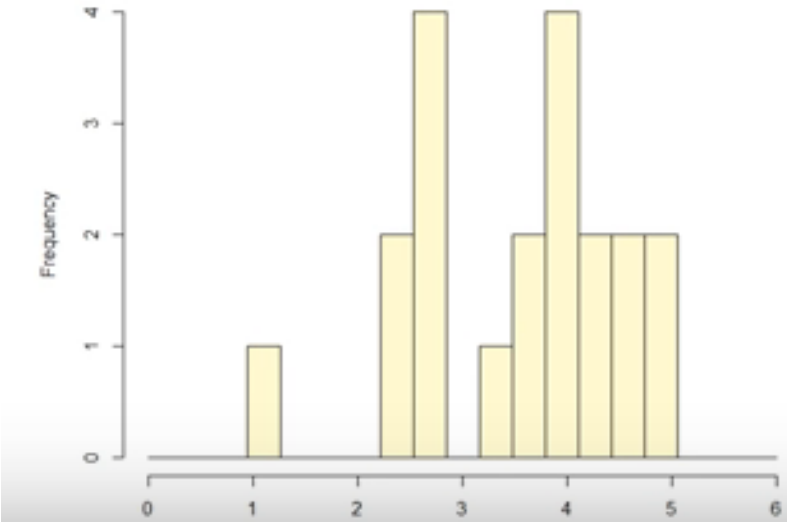
Inherently Biased Parameters

Some parameters are inherently biased during bootstrapping

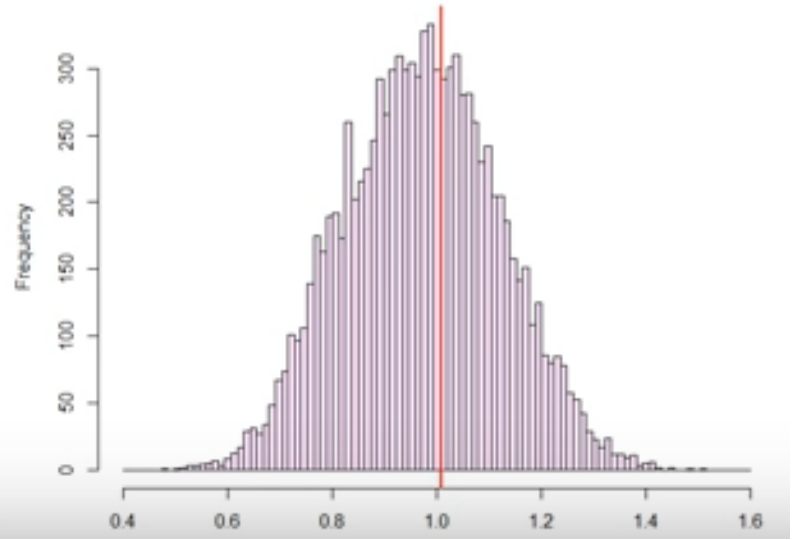
For example: standard deviation/variance

Quanto maior o número de amostras geradas menor o viés.

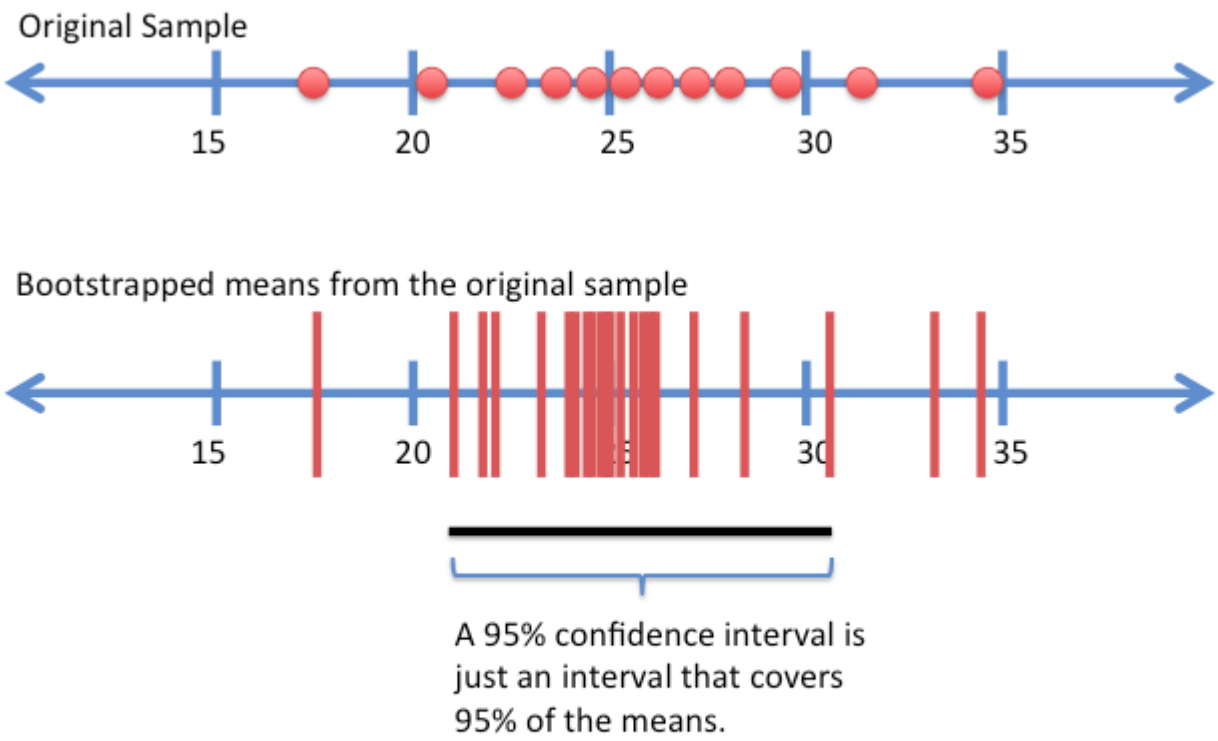
Data (sd = 1.008)



Bootstrapped (mean sd = 0.9275)



Aplicações Bootstrap



- Metrologia (industria)
- Eleições
- Estimativas do governo sobre a população (intervalo de confiança da altura média de homens e mulheres dada uma amostra)
- Intervalo de confiança de um resultado cientifico.