



UNIVERSIDADE FEDERAL
DO RIO DE JANEIRO



Astroestatística

Aula 006 -- Prof. Walter Martins-Filho

Relembrando

$$\frac{d}{dx} \ln L = 0 \rightarrow x = x_0$$

$$\sigma^2 = \frac{d^2}{dx^2} \ln L$$

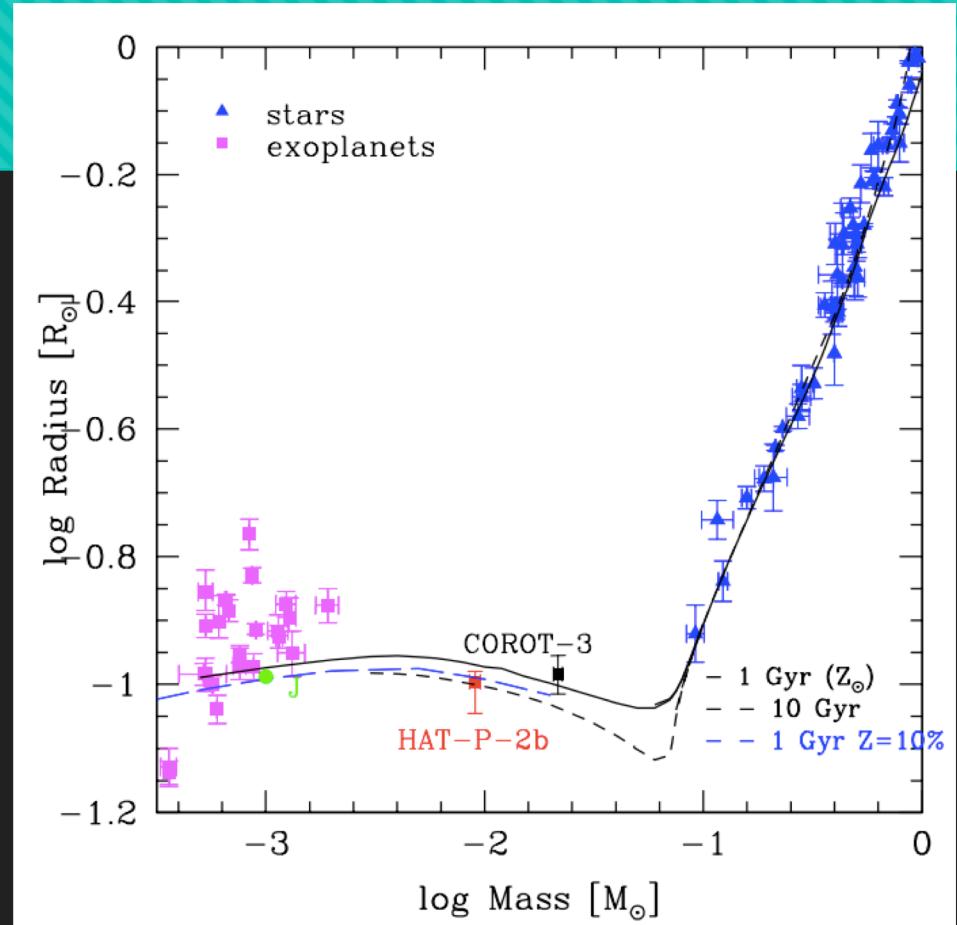
$$Cov(x, y) = \sigma_{x,y}^2 = \frac{d^2}{dxdy} \ln L$$

Ajuste de Modelo

Exemplo: relação massa-raio

- As relações de estrutura interna permitem derivar uma relação entre raio e Massa
- Para estrelas, a eq. de estrutura interna permite estabelecer que $\log R \propto \log M$, uma relação linear
- Podemos assumer então modelo linear para região estelar:

$$\log R = a \log M + b$$



<https://arxiv.org/abs/0810.5085>

$$\log R = a \log M + b$$

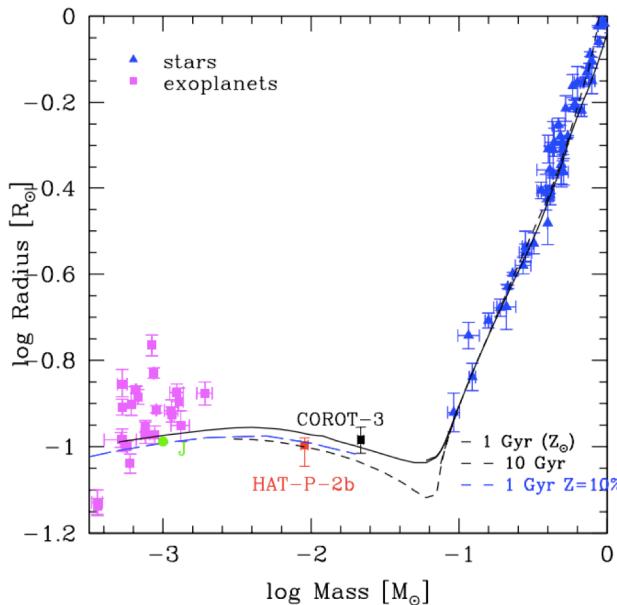
Então: $\{x_i\} = \{\log M_i\}$ aonde $i = 1, \dots, N$

$\{y_i\} = \{\log R_i\}$ aonde $i = 1, \dots, N$

$\{\mu_y\} = \{a \log M_i + b\}$ aonde $i = 1, \dots, N$ modelo

E assumimos uma distribuição Normal para cada valor x_i . Por simplificação, vamos considerar que todas tem o mesmo $\sigma_i = \sigma$.

Assim, para cada y_i teremos que:



$$p(y|X = x, a, b, \sigma^2) = N(\mu_i, \sigma | \mu_i = ax_i + b)$$

Assim, $\ln L$ é definido como:

$$L = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left(-\frac{(y_i - (ax_i + b))^2}{2\sigma^2} \right)$$

$$L = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - (ax_i + b))^2}{2\sigma^2}\right)$$

$$\ln L = \ln \left(\prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - (ax_i + b))^2}{2\sigma^2}\right) \right)$$

$$\ln L = \sum_{i=1}^N \ln \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - (ax_i + b))^2}{2\sigma^2}\right) \right)$$

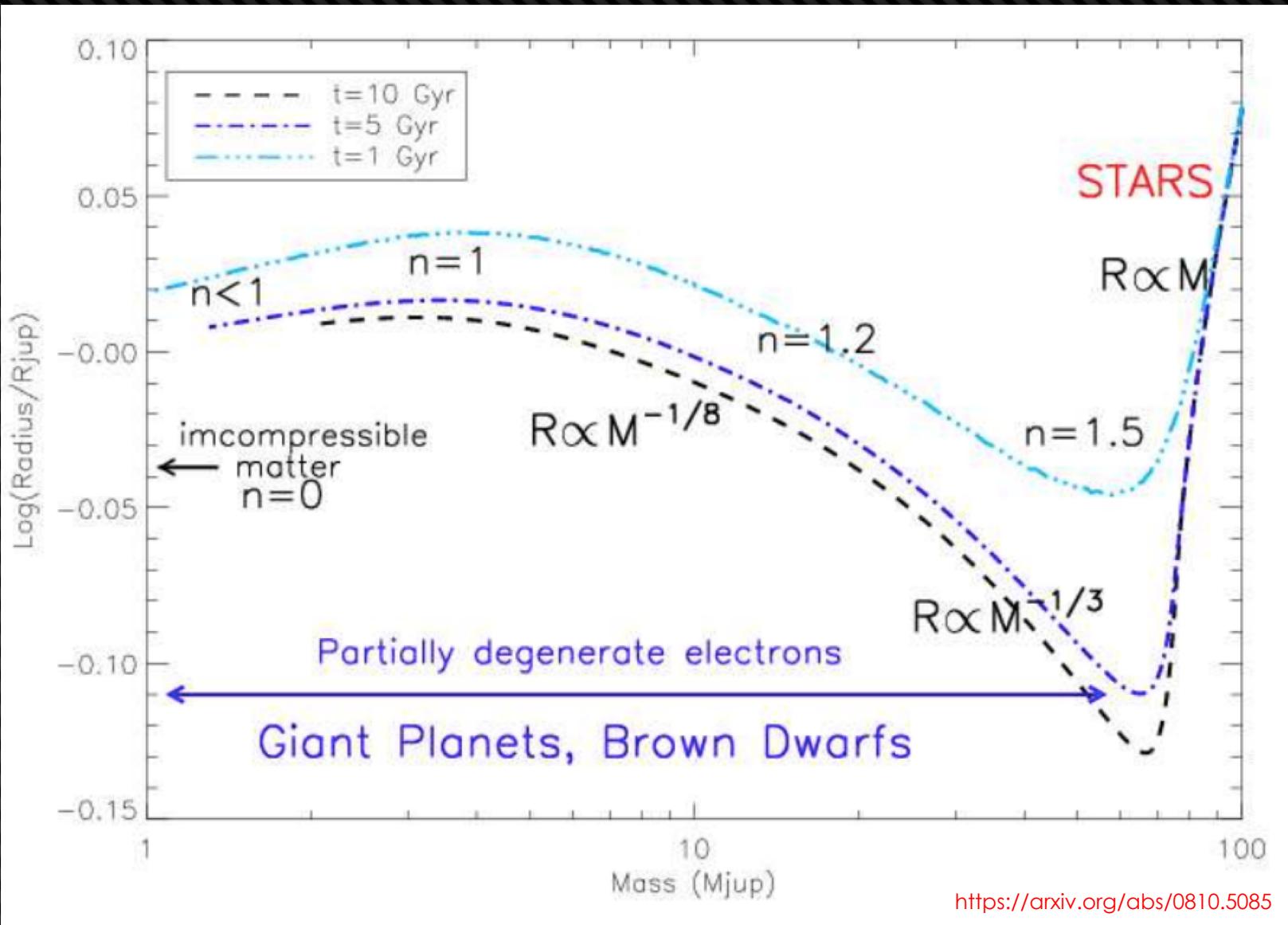
$$\ln L = -\frac{N}{2} \ln 2\pi - N \ln \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^N (y_i - (ax_i + b))^2$$



$$\left. \frac{\partial}{\partial a} \ln L \right|_{\theta=\theta^0} \equiv 0 \quad \left. \frac{\partial}{\partial b} \ln L \right|_{\theta=\theta^0} \equiv 0$$

Exercício

- Assuma uma lei de potência ($y = A x^b$) sugerida pelo paper o qual veio a figura no próximo slide (ou a figura do exemplo dos slides anteriores) e escreva as hipóteses, suposições e a função de máxima verossimilhança. Obtenha a equação que solucione os parâmetros $\{A, b\}$, assim como as suas respectivas variâncias σ_A e σ_b , e a covariância $\sigma_{A,b}$.
- As equações que surgem da derivação de L devem ser solucionadas simultaneamente. Dado isto, sempre cairemos em conjunto de equações de mesmo número de variáveis? Teste com a lei de potência e um ajuste linear para comparar as equações dadas pela derivação de L .



Superestimação ou subestimação

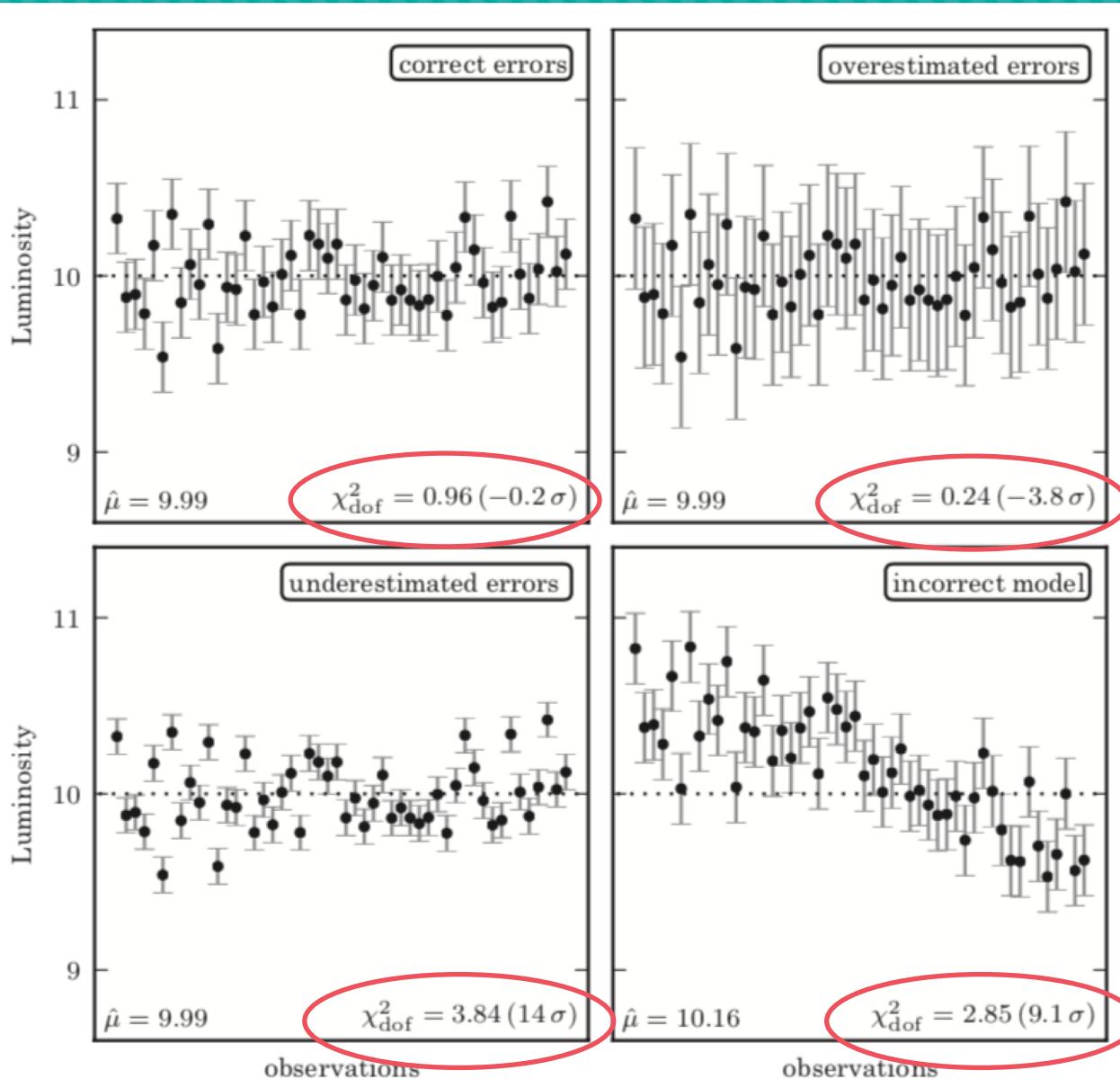
Caso de estudo: Luminosidade

- Obtermos o valor da luminosidade da estrela várias vezes
- Assumimos que essa estrela tem modelo de luminosidade em que não há variabilidade de seu brilho
- Estamos sobreestimando ou superestimando os erros?

Ajuste de modelo: uso do χ^2_{dof}

- Avaliar subestimação ou superestimação é equivalente a avaliar o quão provável é a nossa função de verossimilhança dado modelo e resultados de θ_0
- Para tanto, devemos levar em consideração o ajuste de modelo, pesado pelos parâmetros livres
- A distribuição que combina a χ^2 por graus de liberdade é a χ^2_{dof} :
 - Se $\chi^2_{dof} \sim 1$: o modelo está bem ajustado
 - Se $\chi^2_{dof} \gg 1$: o modelo está subestimando
 - Se $\chi^2_{dof} \ll 1$: o modelo está superestimado

Caso de estudo: Luminosidade



Critério de Informação de Aikake (AIC)

- Quanto maior número de parâmetros a serem ajustados, maior torna-se a quantidade de soluções para θ_0
- Nesta situação, devemos procurar aquele que melhor adequa-se considerando informação contida.
- AIC define que o modelo que tiver menor valor é aquele que condiz com maior informação possível. Desta forma, o vetor de soluções θ_0 maximiza a função de verossimilhança sem tender a uma dimensão específica em θ
- k : quant. de parâmetros a serem ajustados(livres)

$$\text{AIC} \equiv -2 \ln (L^0(M)) + 2k + \frac{2k(k+1)}{N-k-1}$$

Exercícios

- Diversos trabalhos astronômicos baseiam-se na ideia de uma função de custo, o qual define-se um risco.
 - Dado sua área de interesse, ou de iniciação científica, você consegue reconhecer alguma função de custo?

Intervalos de Confiança

Bootstrap

Jackknife

Intervalos de Confiança

- Historicamente, a suposição de que a variável $\{x_i\}$ segue uma distribuição específica, comumente Normal, permite que defina-se um intervalo assumindo um fator de confiabilidade α (visto na disciplina de Intr. A Estatística), e voltaremos a falar em Teste de Hipóteses.
- Com avanço do poder de processamento, pode-se usar metodologias baseadas na reamostragem dos dados $\{x_i\}$
- Duas metodologias baseadas na reamostragem são o **bootstrap** e o **jackknife**.

Bootstrap

- Foi proposto por Efron(1979)
- Metodologia:
 - Obtemos de amostra N elementos até $N!$ re-amostras seguindo uma determinada função que maximize a probabilidade de obter os valores observados.
 - A probabilidade associada de que a amostra obtida seja idêntica a anterior é tão pequena (se $N = 10 \rightarrow p = 0.00036$) que dificilmente teremos amostras completamente idênticas
 - Assim, tendo conjunto $\{x_i\}$, re-amostramos b -vezes, resultando em b -amostras. Avaliamos os valores de média, mediana, desvio-padrão, etc. Usando a ideia do Teorema do Limite Central, a média desses valores é o valor que se aproxima mais dos valores corretos da distribuição da população

Bootstrap

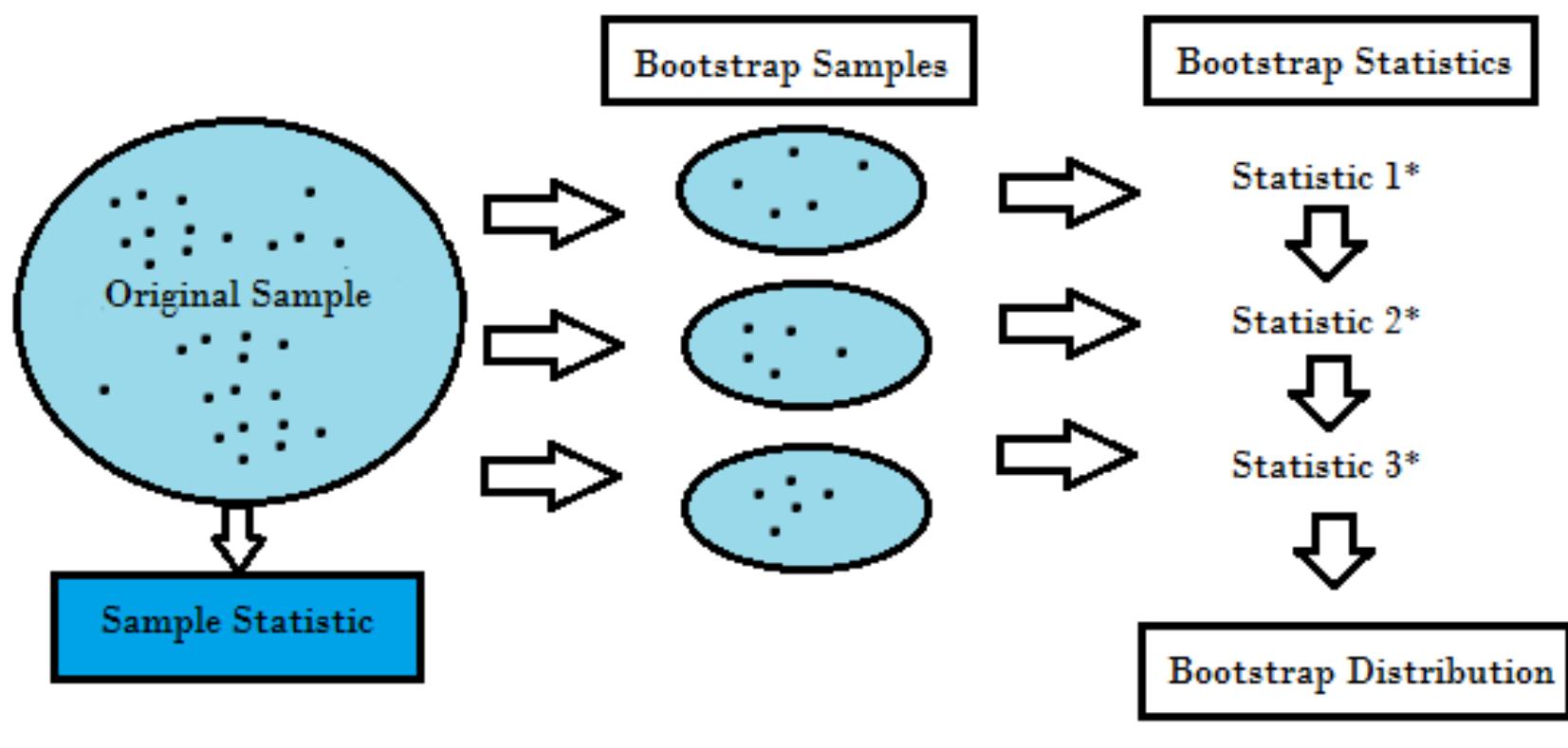
- Desconhecemos a verdadeira distribuição $h(x)$, dado isto, cada amostra obtida $f(x)$ é uma aproximação
- O Bootstrap assume que se aumentarmos a quantidade de sub-amostras, encontraremos ao final a verdadeira $h(x)$
- Assim, se considerarmos uma amostra inicial com N -valores, teremos que a distribuição $f(x) \rightarrow h(x)$ quando amostramos $f(x)$ por:

$$f(x) = \frac{1}{N} \sum_{i=1}^N \delta(x - x_i)$$

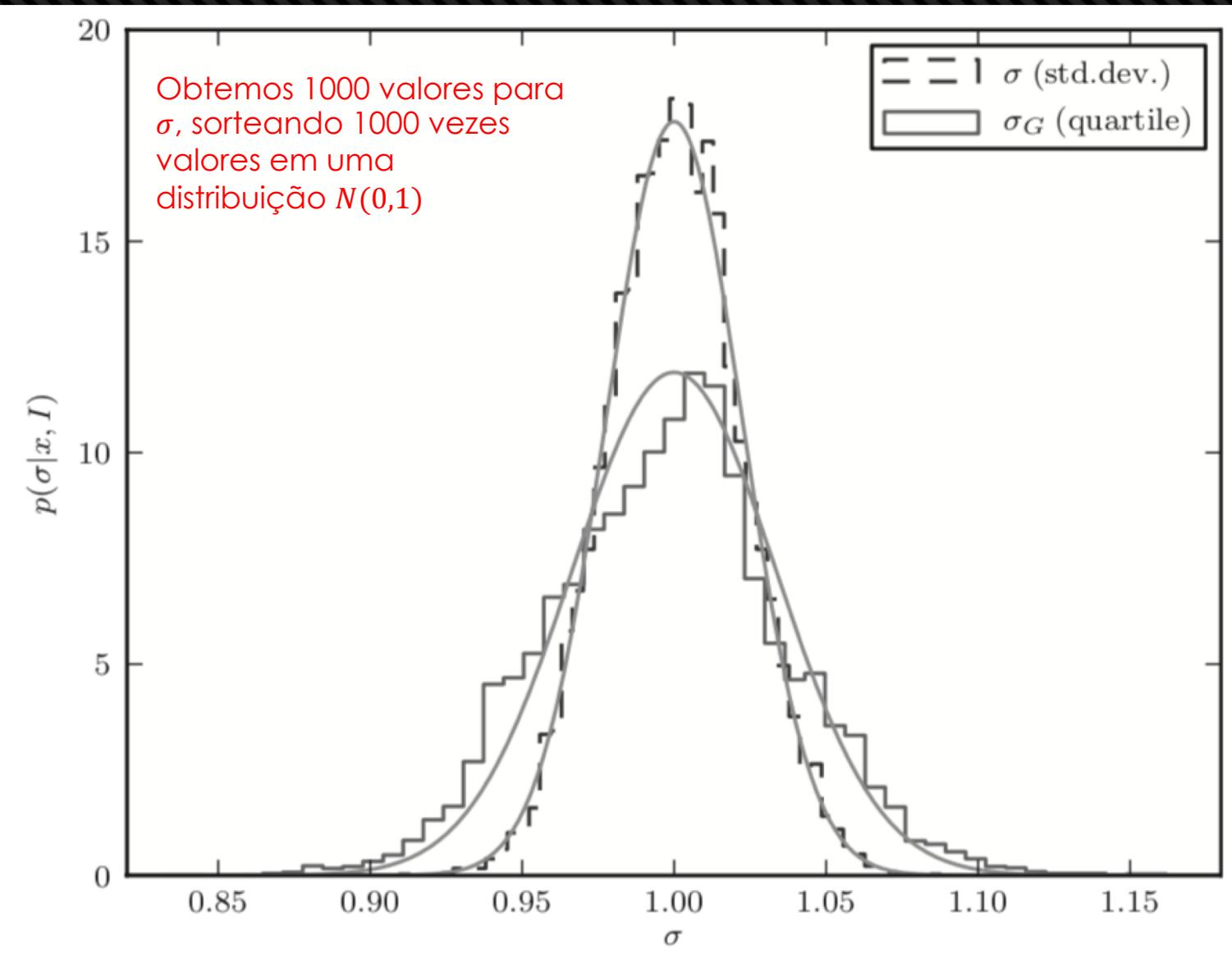
Exercício prático

- Temos o seguinte conjunto amostral: $\{1,2,4,4,10\}$, com $N = 5$
- Vamos sortear 5 elementos dentro desse conjunto amostral, 20 vezes, i.e., $b = 20$
- Obtemos a média e o desvio-padrão de cada um dessas reamostras, e obtemos assim distribuição para média e o desvio-padrão da população.

- 2, 1, 10, 4, 2
- 4, 10, 10, 2, 4
- 1, 4, 1, 4, 4
- 4, 1, 1, 4, 10
- 4, 4, 1, 4, 2
- 4, 10, 10, 10, 4
- 2, 4, 4, 2, 1
- 2, 4, 1, 10, 4
- 1, 10, 2, 10, 10
- 4, 1, 10, 1, 10
- 4, 4, 4, 4, 1
- 1, 2, 4, 4, 2
- 4, 4, 10, 10, 2
- 4, 2, 1, 4, 4
- 4, 4, 4, 4, 4
- 4, 2, 4, 1, 1
- 4, 4, 4, 2, 4
- 10, 4, 1, 4, 4
- 4, 2, 1, 1, 2
- 10, 2, 2, 1, 1



<http://www.statisticshowto.com/bootstrap-sample/>



Pequenas Considerações

- NO fundo, fazemos isso quando obtemos valores numa rotina probabilística qualquer dentro de algum programa/linguagem.
- **Qual maior problema para uso do bootstrap durante uma pesquisa?**

Determinar o tamanho da nossa re-amostragem

- Bootstrap usa o conceito de quanto mais, melhor. Ou a ideia da Lei dos Números Grandes.
- Como falado inicialmente, podemos ter até $N!$ re-amostras.
- Desta forma, o intervalo de confiança é determinado tanto quanto reaproxima-se de $b \rightarrow N!$ quanto por estabelecer qual a confiabilidade que se queira
- **Se estamos procurando 90% de confiança, então determinamos q_5 e q_{95} são o que restringe nosso intervalo amostral final dos parâmetros a serem avaliados.**

Exercício

- Usando diversos passos, desenhe o processo de bootstrap para uma distribuição Normal de média zero e variância 1. Demostre com esses desenhos de distribuições como obtemos os valores finais para média e variância da população
- Encontre a rotina em Python e em R que faça bootstrap para determinado conjunto amostral dado.

Jacknife

- Inventado por Tukey(1958).
- Assemelha-se ao bootstrap, porém ele considera que algum porcentagem do valor observado na distribuição amostra $\{x_i\}$ não fará parte do sorteio para a re-amostra.
- Contudo, diferente do bootstrap, a técnica de Jacknife tem seus intervalos de confiança determinados por uma distribuição t -Student, i.e., esperaríamos que fosse normal, mas a reamostra pode assumir valores muito discrepantes, logo, a t -Student levaria em consideração esses valores.

Vamos assumir que avaliamos o parâmetro α

$$\{\alpha_i\}, i = 1, \dots, N$$

Correção devido o Jackknife é dado por:

$$\alpha^J = \alpha_N + \Delta\alpha$$

$$\Delta\alpha = (N - 1) \left(\alpha_N - \frac{1}{N} \sum_{i=1}^N \alpha_i^* \right)$$

Sendo α_i^* parâmetro ajustado a cada re-amostra, e α_N o parâmetro considerado da amostra original.

$$\sigma_\alpha = \sqrt{\frac{1}{N(N-1)} \sum_{i=1}^N [N\alpha_N - \alpha^J - (N-1)\alpha_i^*]^2}$$

E o intervalo de confiança seguindo t -Student com $t = (\alpha - \alpha^J)/\sigma_\alpha$ e $N - 1$ graus de liberdade.

Bootstrap ou Jackknife?

- Para amostras muito grandes, a probabilidade de reamostragem entre ambos os métodos torna-se irrelevante a discrepância, retornando valores similares.
- Jackknife é útil pois leva em consideração, tanto a existência de bias quanto esquemas complexos de amostragem
- Por sua vez, bootstrap é simples de computar intervalos de confiança.
- Ambos podem ser usados como forma de ***cross-validation*** entre si.

Referências

- Livros-texto
- <http://www.statisticshowto.com/bootstrap-sample/>
- <https://www.thoughtco.com/example-of-bootstrapping-3126155>
- https://www.encyclopediaofmath.org/index.php/Likelihood_equation
- https://www.encyclopediaofmath.org/index.php/Sequential_approximation,_method_of