Estimação, Viés e Variância. Distribuição Amostral de uma Estatística.

Fernando B. Sabino da Silva

March 7, 2018

Parâmetros de uma Distribuição

Todas as distribuições que foram discutidas em Estatística I contém um conjunto de parâmetros que descreve completamente a função densidade de probabilidade (ou função massa de probabilidade). Por exemplo, uma variável aleatória gaussiana, $X \sim N(\mu, \sigma^2)$, tem como parâmetros a média μ e a variância σ^2 . Uma variável aleatória com distribuição exponencial, $X \sim Exp(\lambda)$, tem a taxa λ como parâmetro. Uma variável aleatória Bernoulli, $X \sim Ber(p)$, tem como único parâmetro a probabilidade de sucesso p . É importante relembrar que parâmetros são constantes desconhecidas. Em Estatística I, os parâmetros eram informados (“God gave us”).

Notação: Quando queremos fazer uma declaração genérica sobre parâmetros, é costume usar a letra grega θ .

Estimação dos Parâmetros de uma Distribuição. Algumas vezes nós fazemos suposições de que os dados provém de uma determinada distribuição. Por exemplo, se nós acreditamos que os dados são provenientes da soma de efeitos aleatórios, nós argumentamos que a distribuição que rege o fenômeno é a gaussiana como resultado do Teorema Central do Limite (CLT). Outras vezes, porém, é necessário tomar uma decisão de modelagem, isto é, nós devemos decidir qual modelo (qual particular tipo de distribuição) se ajusta melhor aos nossos dados. Às vezes, podemos decidir até olhando um histograma.

Uma vez que decidimos qual distribuição utilizar, a próxima questao é: quais devem ser os parâmetros para esta distribuição? Por exemplo, se nós escolhemos modelar nossos dados utilizando a distribuição normal, nós temos que escolher entre um número infinito de distribuições gaussianas, pois existe um número infinito de parâmetros μ e σ^2 . Nós poderíamos, talvez, fazer algumas suposições adicionais e escolher valores particulares para μ e σ^2 . No entanto, se assim o fizermos, as decisões começaram a ser extremamente restritivas e muito difíceis de justificá-las. Felizmente, podemos usar a amostra para *estimar* estes parâmetros. No caso de uma distribuição normal, a média amostral $\bar{X}_n$ e a variância s_n^2 parecem candidatos naturais para estimar os parâmetros μ e σ^2 .

Definição 1: Seja $X_1, X_2, \ldots, X_n$ variáveis aleatórias iid provenientes de uma distribuição com parâmetro θ . Um estimador de θ é a estatística $\hat{\theta} = T(X_1, X_2, \ldots, X_n)$.

Nota: A notação de “chapéu” serve para indicar que estamos estimando um parâmetro particular. Por exemplo, se estivermos tentando estimar o parâmetro μ de uma distribuição normal, nós costumamos chamar o estimador de $\hat{\mu}$.

Definição 2: O estimador $\hat{\theta}$ para o parâmetro θ é dito **não-viesado** se $E[\hat{\theta}] = \theta$.

O **viés** de $\hat{\theta}$ é definido por $b(\hat{\theta}) = E[\hat{\theta}] - \theta$.

Exercício 1: Estime a média μ de uma distribuição normal. Se nós escolhermos a média amostral como nosso estimador, isto é, $\hat{\mu} = \bar{X}_n$, mostre que este estimador é não-viesado, isto é, $E[\bar{X}_n] = \mu$.

Exercício 2: Estime a variância σ^2 de uma distribuição normal. Se nós escolhermos a variância amostral como nosso estimador, isto é, $\hat{\sigma}^2 = s_n^2$, iremos encontrar um motivo para usar $(n-1)$ no denominador. O objetivo é tornar o estimador não-viesado. Primeiro, relembre que $Var(X) = E[X^2] - [E(X)]^2$. Usando isto, é fácil deduzir que

$$E[X_i^2] = Var(X_i) + [E(X_i)]^2 = \sigma^2 + \mu^2$$

, e

$$E[X_n^2] = Var(X_n) + [E(X_n)]^2 = \frac{\sigma^2}{n} + \mu^2.$$

Usando os resultados acima **mostre que**

$$E[s_n^2] = E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - X_n)^2\right]$$

é um estimador não-viesado para σ^2 .

- Obs: Se tivéssemos colocado n no denominador, nós teríamos encontrado

$$E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - X_n)^2\right] = \frac{n-1}{n}\sigma^2.$$

Exercício 3: Se X_i são variáveis aleatórias iid com distribuição $Ber(p)$, mostre que $E[X_i] = p$ e que $E[X_n] = p$, isto é, X_n é um estimador não-viesado para p .

É possível que dois estimadores, $\hat{\theta}_1$, $\hat{\theta}_2$, para um parâmetro θ sejam *ambos* não-viesados (vocês conseguem pensar em um exemplo?). Se não-viés for algo que consideramos necessário, como decidimos qual dos dois estimadores iremos usar? Bem, neste caso é usual preferir aquele que tenha menor variabilidade, isto é, que tenha uma probabilidade maior de estar mais perto do verdadeiro valor do parâmetro (o nosso trabalho é feito geralmente com uma amostra).

Definição 3: Sejam dois estimadores $\hat{\theta}_1$ e $\hat{\theta}_2$ não-viesados para θ . O estimador $\hat{\theta}_1$ é dito ser **mais eficiente** que o estimador $\hat{\theta}_2$ se

$$Var(\hat{\theta}_1) < Var(\hat{\theta}_2).$$

Consistência: Outro conceito importante para ser definido é o de consistência de um estimador. Por exemplo, conforme o tamanho amostral n aumenta, a distribuição da média amostral $\bar{X}_n$ torna-se cada vez mais concentrada em torno da média populacional μ . Quando um estimador converge (em probabilidade) para um parâmetro, nós dizemos que este estimador é consistente para o parâmetro (ou converge em probabilidade para ele).

Definição 4: Um estimador $\hat{\theta}_n$ é dito consistente se:

$$\lim_{n \to \infty} P\left(\left| \hat{\theta}_n - \theta \right| < \epsilon \right) = 1$$

, isto é,

$$plim\hat{\theta}_n = \theta$$

, ou ainda,

$$\hat{\theta}_n \xrightarrow{p} \theta$$

, ou seja, se $\hat{\theta}_n$ converge em probabilidade para a constante θ , que é o valor verdadeiro do parâmetro.

Proposição 1: Um estimador $\hat{\theta}_n$ é dito consistente se:

$$\lim_{n \to \infty} E(\hat{\theta}_n) = \theta$$

e

$$\lim_{n \to \infty} Var(\hat{\theta}_n) = 0$$

, ou

$$\lim_{n \to \infty} EQM(\hat{\theta}_n) = 0$$

, onde

$EQM(\hat{\theta}_n) = Var(\hat{\theta}_n) + b^2(\hat{\theta}_n)$. EQM é chamado de *erro quadrático médio*.

- Nota: Consistência do EQM implica consistência de $\hat{\theta}_n$, mas o inverso não necessariamente é verdadeiro, isto é, $\hat{\theta}_n$ pode ser consistente para θ sem que a proposição acima seja válida. Mas se valer o estimador $\hat{\theta}_n$ será consistente.

Distribuição Amostral de uma Estatística

Considere uma amostra aleatória $X_1, X_2, \ldots, X_n$. Relembre que então as realizações X_i são iid (independentes e identicamente distribuídas). Considere a estatística (qualquer função apenas da amostra é dita uma estatística) média amostral,

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

, no exemplo a seguir.

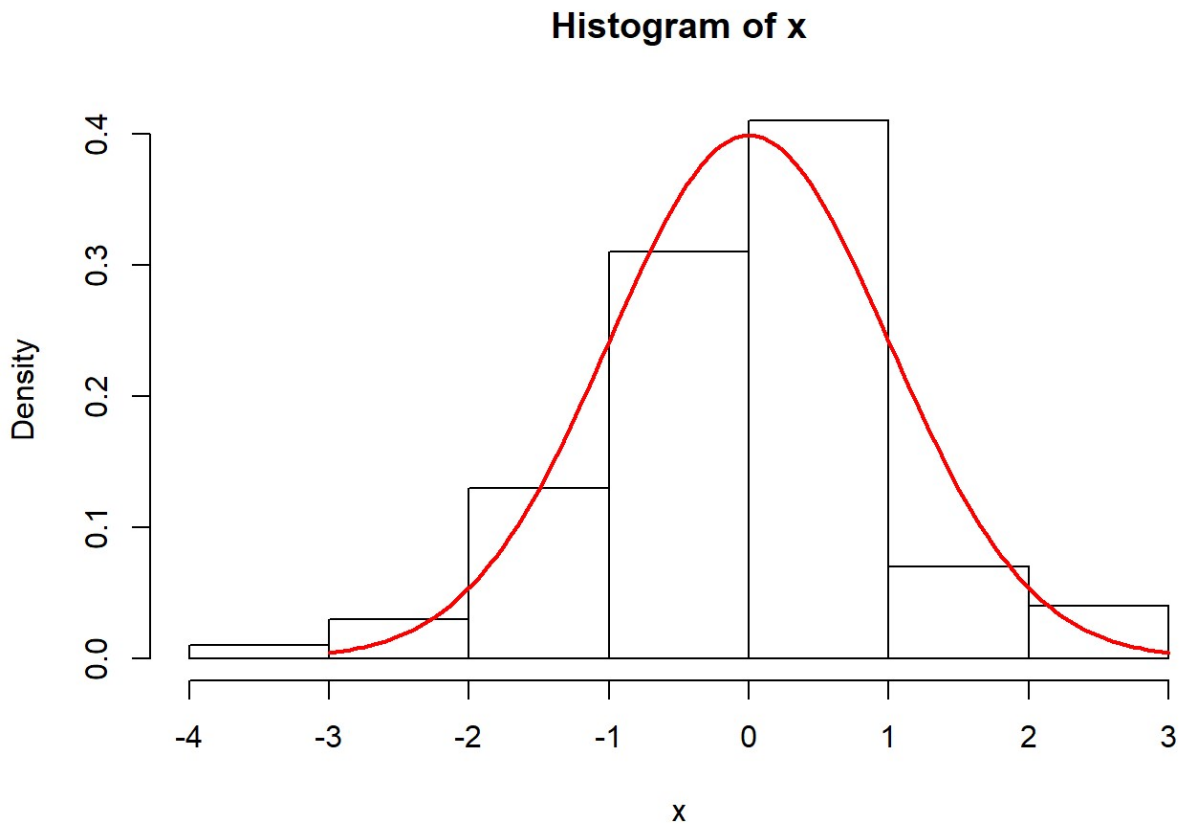
Lembre-se de que uma estatística é uma variável aleatória (se é função da amostra ela depende daquela particular amostra). Isto significa $\bar{X}_n$ tem uma distribuição de probabilidade. Em outras palavras, se repetirmos o experimentos várias vezes, nós iremos obter resultados diferentes para os X_i . Isso por sua vez resultaria em diferentes valores para a estatística média amostral, $\bar{X}_n$. A distribuição de uma estatística é chamada de **distribuição amostral**.

Vamos fazer simulações no R para ilustrar isto. Assuma, inicialmente, que $X_i \sim N(0, 1)$ e simulamos 100 realizações da variável aleatória:

```
n = 100
x = rnorm(n)
```

Um histograma representando a distribuição amostral destas 100 realizações (aleatórias) de X_i está a seguir. Nós também traçamos a densidade da distribuição normal (a verdadeira, pois os dados foram simulados de uma distribuição normal padrão) sobre o histograma para checar quão perto ela está.

```
hist(x, freq = FALSE)

s = seq(-3,3,0.05)
lines(s, dnorm(s), col = 'red', lwd = 2)
```



Vejamos qual a média amostral destas 100 realizações

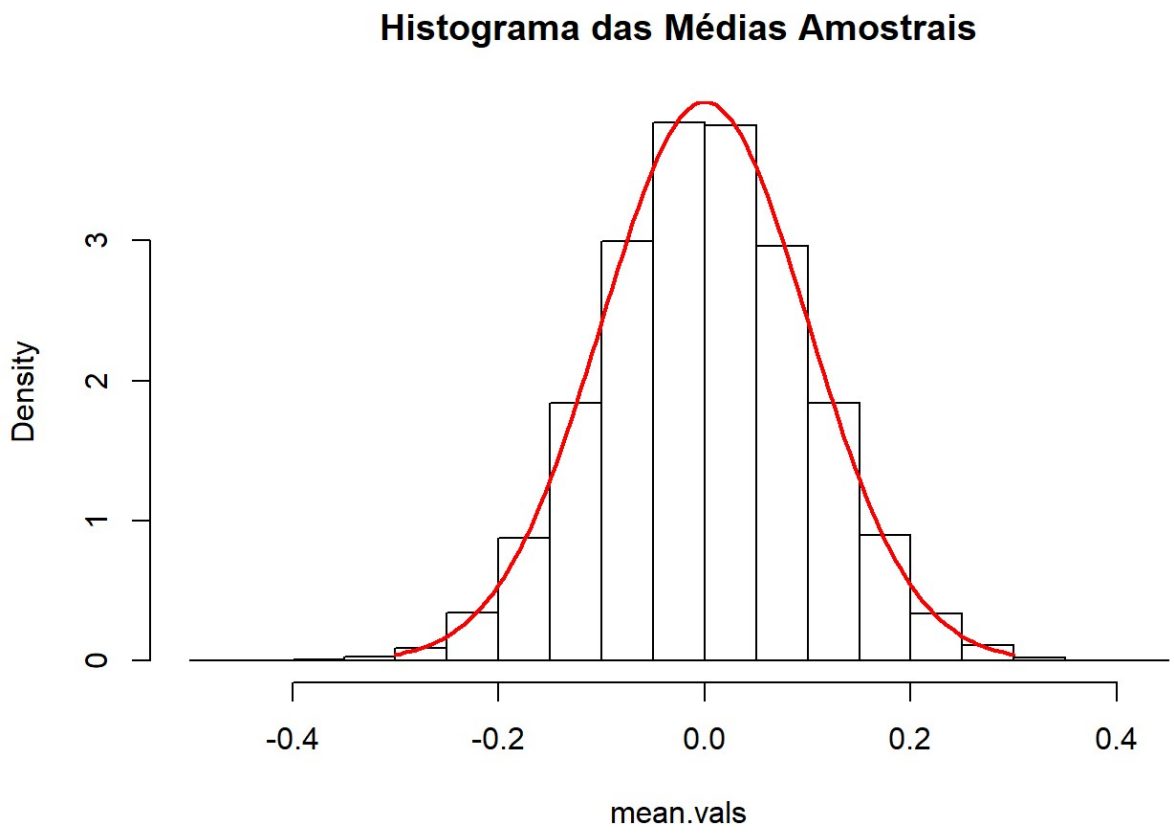
```
mean(x)
```

```
## [1] -0.04288582
```

E se nós repetíssemos este “experimento” muitas vezes? Qual seria a distribuição amostral da estatística média? Anteriormente, nós mostramos que a média amostral também tem uma distribuição normal com a mesma média μ e variância dividida por n . Em outras palavras, $\bar{X}_n \sim N(\mu, \sigma^2/n)$.

Abaixo fazemos 100000 simulações (repetições do experimento anterior)

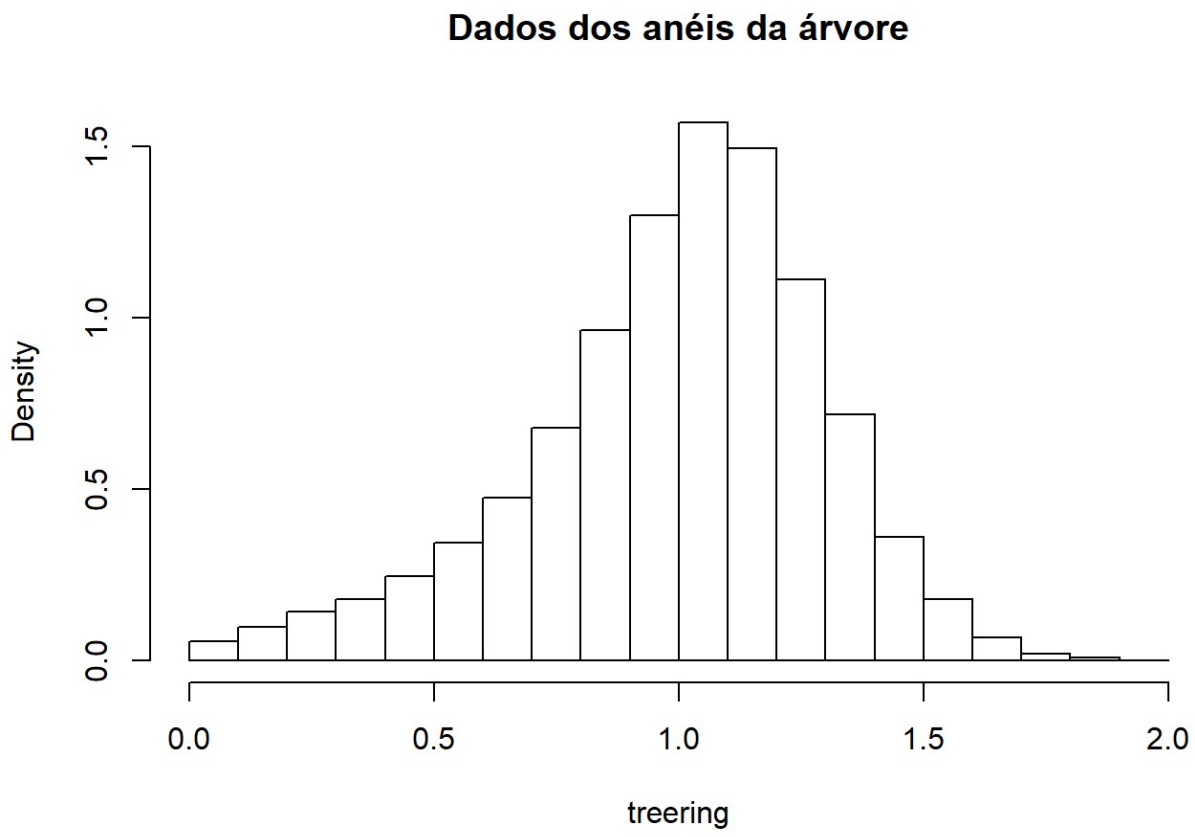
```
numSims = 100000

## Cada coluna nesta matriz representa uma simulação
sims = matrix(rnorm(n * numSims), n, numSims)
mean.vals = colMeans(sims)

hist(mean.vals, freq = FALSE, main = "Histograma das Médias Amostrais")
s = seq(-0.3, 0.3, 0.005)
lines(s, dnorm(s, 0, 0.1), col = 'red', lwd = 2)
```



Exemplo 1: Estimando os Parâmetros de uma Distribuição Normal

Abaixo temos o histograma de um conjunto de dados contido no R. Ele consiste de larguras de anéis de árvores de um determinado tipo de pinheiro (bristlecone) na Califórnia. Veja uma breve explicação para medirmos o tamanho dos anéis aqui (http://www.ebc.com.br/infantil/voce-sabia/2012/09/voce-sabia-que-e-possivel-descobrir-a-idade-de-uma-arvore-olhando-o).

```
hist(treering, main = "Dados dos anéis da árvore", freq=FALSE)
```



Os dados têm um unico “pico” e são aproximadamente simétricos, então talvez possamos nos sentir confortáveis com o uso de variável aleatória com distribuição normal para modelá-los. Mas e quanto aos parâmetros μ e σ^2 ? Como discutimos anteriormente, a média e a variância amostral são estimadores não-viesados para esses parâmetros.

```
(mu.hat = mean(treering))
```

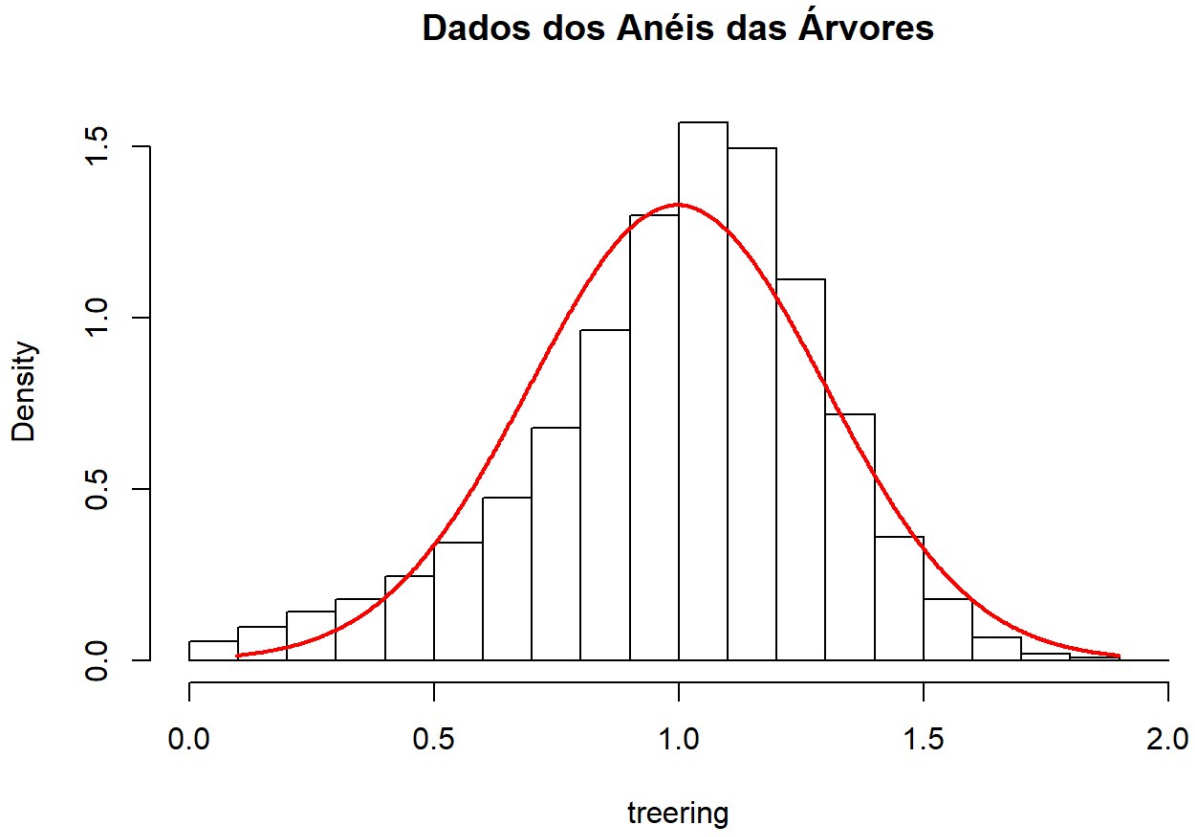
```
## [1] 0.9968362
```

```
(sigma2.hat = var(treering))
```

```
## [1] 0.09021466
```

Vamos traçar a função densidade de probabilidade da distribuição normal com base nos valores amostrais sobre o histograma:

```
hist(treering, main = "Dados dos Anéis das Árvores", freq = FALSE)
s = seq(mu.hat - 3*sd(treering), mu.hat + 3*sd(treering), 0.01*sd(treering))
lines(s, dnorm(s, mean(treering), sd(treering)), col = 'red', lwd = 2)
```



Observe que este não é um ajuste perfeito. Os dados são ligeiramente assimétricos à direita.

Exemplo 2: Estimando o Parâmetro de uma Distribuição Exponencial

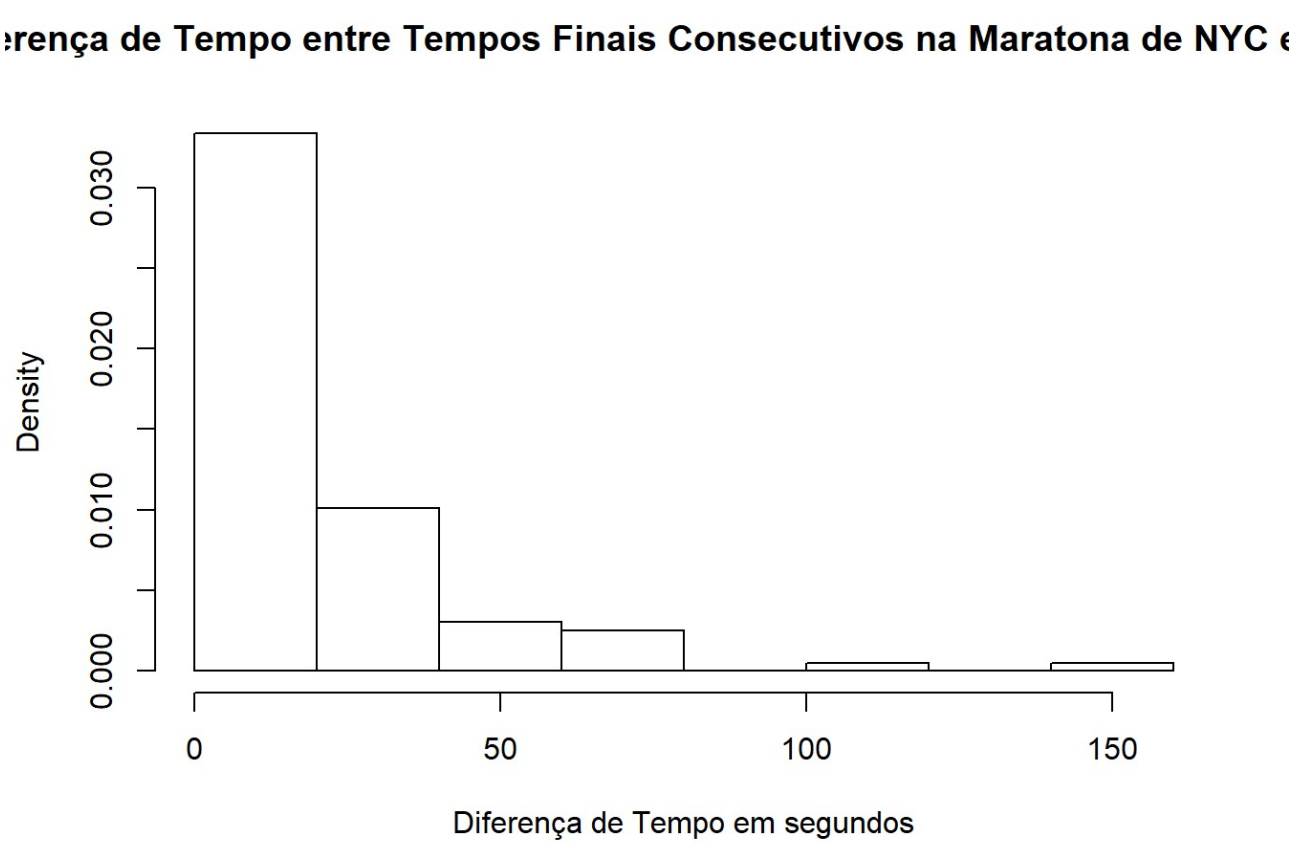
Para este exemplo, vamos usar dados dos tempos dos primeiros 100 colocados na maratona de New York em 2016 e calcular a diferença entre os tempos (consecutivos) entre eles em segundos. Os dados podem ser encontrados na página do curso no moodle. Similarmente, veja os dados para a última maratona em 2017 no link abaixo:

http://www.tcsnycmarathon.org/about-the-race/results/overall-men (http://www.tcsnycmarathon.org/about-the-race/results/overall-men)

```
marathon = read.csv("marathon.csv", header = FALSE, sep = "\t")

times = as.difftime(as.character(marathon$V4))
diffs = as.numeric(times[2:100] - times[1:99]) * 60 * 60
```

Vejamos o histograma deste conjunto de dados:

```
hist(diffs, freq = FALSE, main = "Diferença de Tempo entre Tempos Finais Consecutivos
na Maratona de NYC em 2016", xlab = "Diferença de Tempo em segundos")
```



Aparentemente, uma distribuição exponencial se ajustaria bem a estes dados, isto é, $T_i \sim \text{Exp}(\lambda)$. Mas e quanto ao parâmetro (taxa) λ ? Nós sabemos que o valor esperado para uma variável aleatória com distrição exponencial é $E[T_i] = \frac{1}{\lambda}$. Portanto, um estimador não-viesado para $\frac{1}{\lambda}$ é a média amostral e, portanto,

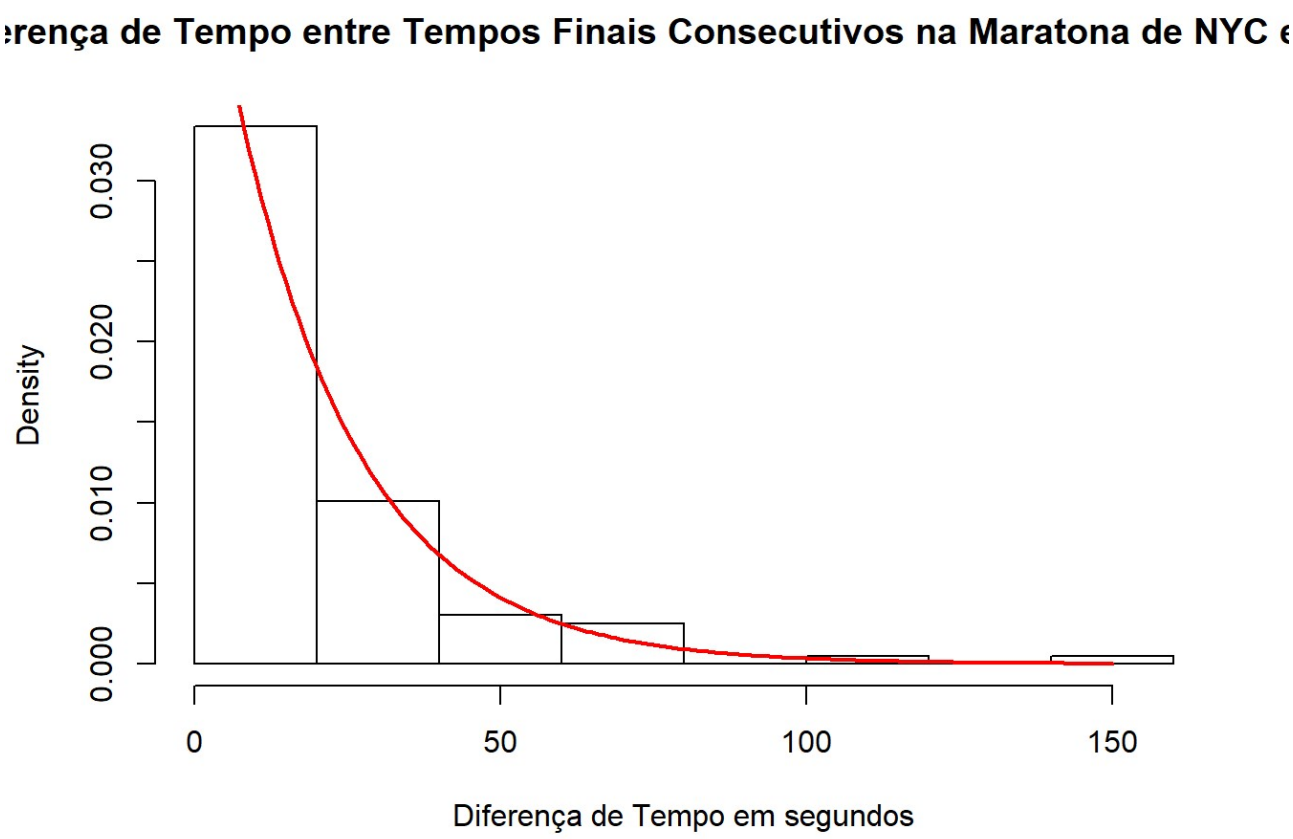
$$\hat{\lambda} = \frac{1}{\bar{X}_n}.$$

```
(lambda.hat = 1 / mean(diffs))
```

```
## [1] 0.04982386
```

Vejamos quão bem esta particular distribuição se ajusta aos nossos dados:

```
hist(diffs, freq = FALSE, main = "Diferença de Tempo entre Tempos Finais Consecutivos
na Maratona de NYC em 2016", xlab = "Diferença de Tempo em segundos")
s = 0:max(diffs)
lines(s, dexp(s, rate = lambda.hat), col = 'red', lwd = 2)
```



Veja a página a seguir do Wikipedia para ver mais exemplos de modelagem com a distribuição exponencial:

https://en.wikipedia.org/wiki/Exponential_distribution (https://en.wikipedia.org/wiki/Exponential_distribution)

Lista de Exercícios

Questão 1

a. O que é um parâmetro e o que é um estimador?
b. Defina viés e o que significa um estimador ser não-viesado?
c. Calcule o viés e a variância da média amostral assumindo que a amostra foi retirada de uma população com (i) distribuição normal com média μ e variância σ^2 e (ii) bernoulli com probabilidade de sucesso p .

Questão 2

Sua amiga arremessou uma moeda justa n vezes e lhe disse o número de caras que apareceram no experimento. Porém, ela não lhe contou quantas vezes ela arremessou a moeda. Ela repetiu este experimento 10 vezes, isto é, em cada experimentos ela arremessou a moeda n vezes, e lhe informou o número de caras em cada experimento: $x_1, x_2, \ldots, x_{10}$.

a. Indique uma estatística não-viesada, $\hat{n}$ para estimar n. Explique por que o viés será zero.
b. Simule este experimento 10000 vezes com $n = 25$. Cada simulação deverá produzir uma lista com 10 números. Use a estatística qud você sugeriu em (a) para estimar n em cada simulação, isto é, no final você deverá ter 10000 valores para $\hat{n}$. Isto é semelhante a simulação feita acima para a média de uma distribuição normal.

Dica: Use a função rbinom do **R** da seguinte forma: rbinom(10, 25, 0.5), onde 10 representa o número de experimentos, $n = 25$ e $p = 0.5$ é a probabilidade de sucesso de uma moeda justa.

c. Faça um boxplot para os 10000 valores de $\hat{n}$. Desenhe uma linha horizontal vermelha representando o verdadeiro valor de n. Os valores que você encontrou estão centrados em torno do verdadeiro valor de n?

Dica: Para fazer um boxplot você pode usar, por exemplo, as funções boxplot ou gf_boxplot (esta contida na library ggplot2) do **R** (dentre outras, não esqueça que o **R** tem em torno de 10000 pacotes e alguns devem conter a possibilidade de fazer este gráfico). Veja um exemplo do uso da função gf_plot aqui (https://rpubs.com/fsabino_da_silva/367934).

d. Faça um histograma para os seus 10000 valores de $\hat{n}$. Usando a função lines, trace a função de distribuição de probabilidade (pdf) de uma distribuição normal em cima do seu histograma. Use como parâmetros μ e σ^2 a média e a variância amostrais que você encontrou para os seus $\hat{n}$ valores.