Questão 76 - Lista 7 - Inferência Estatística

Alisson Moreira Ferreira - 11/0106946
Augusto Cesar Ribeiro Nunes - 13/0103004
Davi Souza Botelho - 12/0029057
Thalyta Brito dos Santos - 12/0023075

Novembro de 2015

Introdução

Este trabalho tem como intuito explicar ao leitor o conceito de nível de confiança $1 - \alpha$ de um intervalo de confiança. A Estimação Intervalar ocupa um lugar pivotal na Inferência Estatística, complementando, em certo sentido, a Estimação Pontual, e intimamente associada ao que é comumemente visto como o terceiro fundamento da Inferência Estatística: os Testes de Hipóteses (TH).

Uma parte considerável do cânone da disciplina foi estabelecida durante a primeira metade do século XX conjuntamente por Jerzy Neyman [em inglês] e Egon Pearson [idem]. Existe uma extensa bibliografia sobre a *saga* do desenvolvimento da disciplina, que inclui as contribuições de Neyman, Pearson e outros protagonistas, porém, em razão da brevidade, cito apenas o livro mais facilmente acessível aos falantes do português: *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*, de David Salsburg, publicado como uma tradução "aceitável" no Brasil como *Uma Senhora Toma Chá: Como a Estatística Revolucionou a Ciência no Século XX*: no Capítulo 12, e brevemente, no Capítulo 27 cita as contribuições seminais destes senhores.

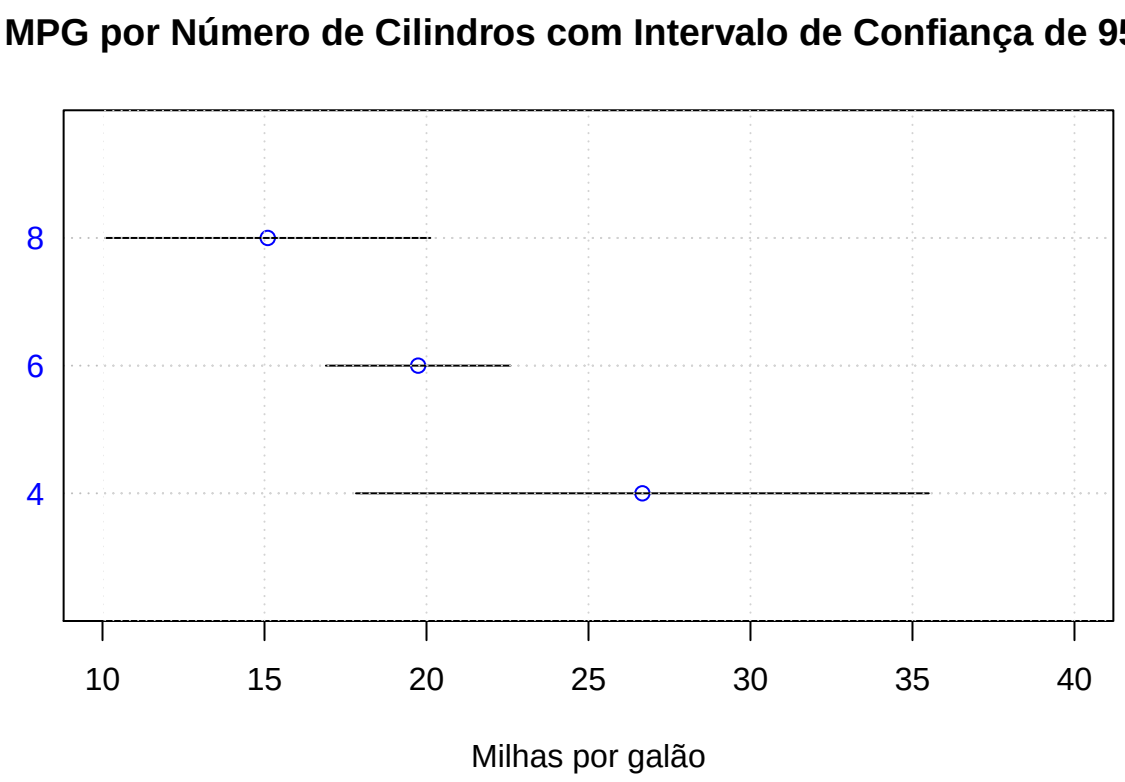
Exemplo 1: Estudo de caso usando conjunto de dados do R

Este exemplo inicial ilustra a utilização do conjunto de dados mtcars disponível do pacote base do R em uma situação hipotética. Um empresário americano do ramo automobilístico dispõe de um conjunto de dados coletados em sua linha de produção, que possui diversas variáveis de interesse no processo produtivo. Interessado em saber se existe relação entre as variáveis MPG (milhas por galão) e CYL (quantidade de cilindros), o empresário contratou uma empresa de consultoria em estatística para realizar um estudo. Esse estudo consiste em determinar um intervalo de confiança para a variável número de cilindros e analisar se os valores da variável MPG (milhas por galão) estão contidos nesse intervalo, considerando um intervalo de confiança de 95%, suficiente para analisar se a quantidade de milhas por galão de cada carro está em conformidade com o número de cilindros. Diante disso, os estatísticos que trabalham nessa empresa obtiveram o conjunto de dados **mtcars** e realizaram as seguintes programações utilizando o R.

```
mean <- tapply(mtcars$mpg, list(mtcars$cyl), mean)
sd <- tapply(mtcars$mpg, list(mtcars$cyl), sd)
x <- data.frame(mean, sd)

# Criando um conjunto de dados com média e desvio padrão.  Uso do comando
# tapply é para retornar uma lista de vetores.  Ao final criou-se um
# data-frame com as variáveis mean(média) e sd(desvio padrão)
x$LL <- x$mean - qnorm(0.975) * x$sd
x$UL <- x$mean + qnorm(0.975) * x$sd
# Estabelecendo os limites superior e inferior dos intervalos de confiança.
title <- "MPG por Número de Cilindros com Intervalo de Confiança de 95%"
dotchart(x$mean, col = "blue", xlim = c(floor(min(x$LL)/10) * 10, ceiling(max(x$UL)/10) *
    10), main = title, xlab = "Milhas por galão", ylab = "Número de cilindros")  # Plota o gráfico dotc
```

```r
for (i in 1:nrow(x)) {
    lines(x = c(x$LL[i], x$UL[i]), y = c(i, i))
}
grid()  # adiciona as linhas horizontais no gráfico
```



Conclusão: Diante do gráfico acima podemos observar que as variáveis são inversamente proporcionais, ou seja, quanto maior o número de cilindros, menor a quantidade de combustível o veículo necessita para deslocar-se por uma milha. E mais, percebe-se que intervalos com mesmo nível de confiança (0,95) possuem comprimentos diferentes, pois quando o número de cilindros é igual a 4, o intervalo de confiança é o de maior amplitude e quando o número de cilindros é 6, o intervalo de confiança é o de menor amplitude.

Exemplo 2: IC para a Normal fixada com dados simulados

Os exemplos mais utilizados para Intervalos de Confiança envolvem a Distribuição Normal, por vários motivos, e citamos dois:

- ela é uma distribuição frequente em muitos fenômenos naturais, sociais, econômicos etc.;
- outras distribuições podem ser *aproximadas* a partir dela, graças a resultados ditos *assintóticos* (para *grandes* amostras) graças ao essencial Teorema Central do Limite, quando temos por exemplo a distribuição Binomial (com p próximo de 0.5), Poisson, Soma de Postos de Wilcoxon.

Ou seja, podemos trabalhar com a Normal em muitos (**não todos!**) os casos com certo grau de acurácia que pode ser mais ou menos aceitável, dependendo do nosso escrutínio com o problema.

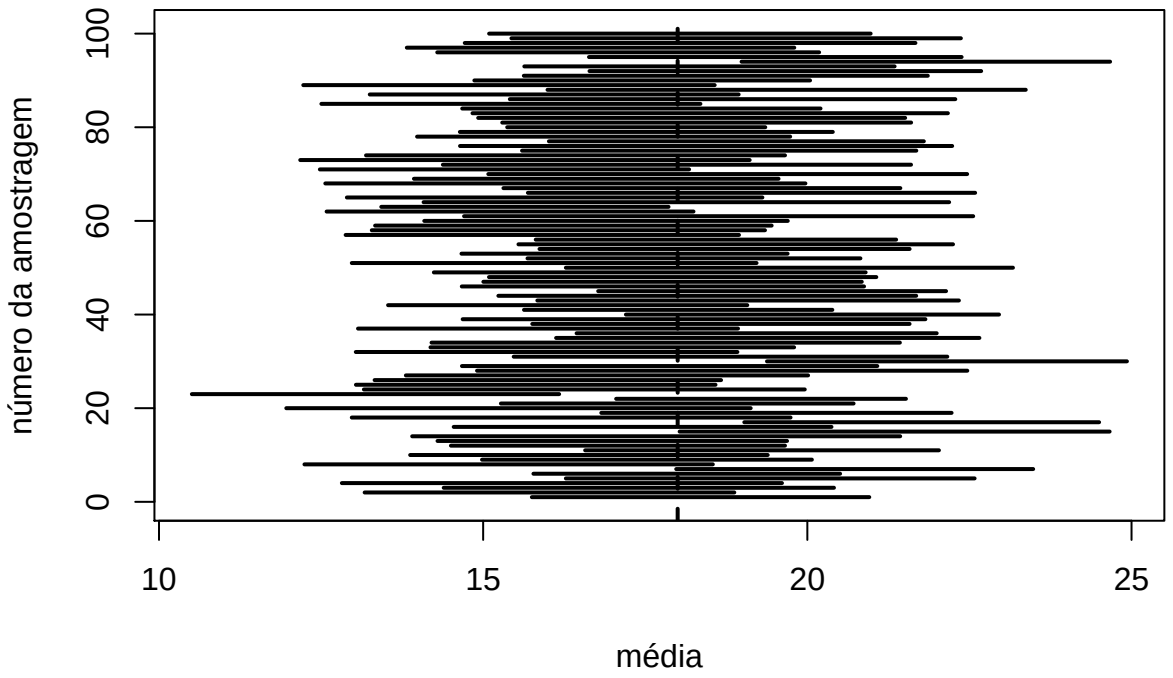
```r
x=rnorm(1000, 18, 9) # Vetor Aleatório de 1000 observações da N com média 18 e dp 9
norm.interval = function(data, variance = var(data), conf.level = 0.95) {

  z = qnorm((1 - conf.level)/2, lower.tail = FALSE)
  xbar = mean(data)
  sdx = sqrt(variance/length(data))
  inter=c(xbar - z * sdx, xbar + z * sdx)
  return(inter)}
norm.interval(x)
```

```
## [1] 17.63593 18.73144
```

```r
# Exemplo de simulação
n.draw = 100  # número de realizações da amostragem
mu = 18 # media da amostra aleatória gerada
n = 30  # tamanho de cada amostra
SD = sd(x) # dp da população (conhecido)
draws = matrix(rnorm(n.draw * n, mu, SD), n) #criação de conjunto de dados com todas as amostras, cada

intervalo = apply(draws, 2, norm.interval) #criando os intervalos
plot(range(intervalo), c(0, 1 + n.draw), type = "n", xlab = "média", ylab = "número da amostragem")
for (i in 1:n.draw) lines(intervalo[, i], rep(i, 2), lwd = 2)
abline(v = 18, lwd = 2, lty = 2)
```



Exemplo 3: Aplicativo Interativo

Como a didaticidade do trabalho é essencial, escolhemos apresentar um aplicativo interativo escrito utilizando o pacote *Shiny*. Livremente disponível e graças ao serviço de hospedagem shinyapps, a escolha foi fácil. O aplicativo é **fortemente** inspirado no trabalho disponibilizado livremente no repositório ConfidenceIntervals, mantido pelo usuário JackStat no serviço github. O github é um serviço de hospedagem de projetos de programação que comporta um rico ecossistema escrito em uma miríade de linguagens de programação, inclusive R.

O programa abaixo gera 100 intervalos de confiança aleatórios para a média da distribuição Normal a partir dos parâmetros informados: o tamanho amostral (usado para gerar cada IC), a média e a variância usadas para simular os dados, e o nível de confiança.

Shiny applications not supported in static R Markdown documents

Conclusão

Talvez seja adequada uma *definição negativa* sobre o que é o nível de confiança de um IC : ele **não** é a probabilidade de que os dados amostrais coincidam com o intervalo.

Após a realização do experimento (amostragem) e a obtenção do IC, este contém, ou não, o parâmetro de interesse. O nível de confiança do IC pode ser entendido como uma *medida de confiabilidade* do procedimento de estimação: como em muitos resultados da estatística *clássica* (frequentista) a confiabilidade $1 - \alpha$ refere-se à probabilidade de que um IC obtido em um experimento futuro (como uma simulação subsequente dos exemplos 2 e 3 acima), sob os mesmos procedimentos, forneça um IC que contém o verdadeiro parâmetro de interesse da distribuição.

Em um contexto de Teste de Hipóteses, poderíamos fazer uma equivalência entre o nível de significância do IC e o nível de significância do TH: o intervalo de confiança representa valores para os quais a diferença entre o valor do parâmetro e a estimativa obtida não é estatisticamente significante a um nível $1 - \alpha$.