Laboratório 4B: Fundamentos para Inferência Estatística - Intervalos de Confiança

Amostragem de Ames, Iowa

Se você tem acesso aos dados de um população inteira, por exemplo o tamanho de cada casa na cidade de Ames, Iowa, Estado Unidos, é fácil e direto responder a questões como "Qual é o tamanho de uma casa típica na cidade de Ames?" e "Quanta variação existe no tamanho das casas?". Se você tem acesso somente a uma amostra da população, como costuma ser o caso, responder a essas perguntas fica mais complicado. Qual é sua melhor estimativa para o tamanho típico de uma casa se você só sabe o tamanho de algumas dezenas de casas? Esse tipo de situação requer que você use sua amostra para fazer inferências a respeito da população em geral.

Os Dados

Na laboratório anterior nós exploramos os dados populacionais das casa da cidade de Ames, Iowa. Vamos começar carregando esse conjunto de dados.

```
download.file("http://www.openintro.org/stat/data/ames.RData", destfile = "ames.RData")

load("ames.RData")
```

Neste laboratório começaremos com uma amostra aleatória simples de 60 elementos da população. Perceba que o conjunto de dados contém informações sobre várias variáveis relativas às casas, mas para a primeira parte do laboratório focaremos no tamanho da casa, representada pela variável `Gr.Liv.Area`.

```
population <- ames$Gr.Liv.Area

samp <- sample(population, 60)
```

Exercício 1 Descreva a distribuição da sua amostra. Qual é o tamanho "típico" dentro da sua amostra? Procure esclarecer também como você interpretou o significado de "típico".

Exercício 2 Você acha que a distribuição de outro aluno seria idêntica a sua? Você acha que ela seria similar? Por quê, ou por quê não?

Intervalos de Confiança

Uma das maneiras mais comuns para se descrever o valor típico ou central de uma distribuição é por meio da média. Neste caso podemos calcular a média da amostra utilizando

```
sample_mean <- mean(samp)
```

Retome agora a pergunta que motivou este laboratório: baseado nesta amostra, o que podemos inferir sobre a população? Baseado apenas nesta única amostra, a melhor estimativa da área habitacional das casas vendidas em Ames seria a média amostral, geralmente representada como $\bar{x}$ (aqui denominaremos de `sample_mean` ("média amostral")). A média amostral serve como uma boa *estimativa pontual*, mas seria interessante também deixar claro quanta incerteza temos desta estimativa. Isso pode ser feito pelo uso de um *intervalo de confiança*.

Podemos calcular um intervalo de confiança de 95% para a média amostral adicionando e subtraindo 1.96 erros padrão da estimativa pontual.[†]

```
se <- sd(samp)/sqrt(60)

lower <- sample_mean - 1.96 * se

upper <- sample_mean + 1.96 * se

c(lower, upper)
```

Acabamos de fazer uma inferência importante: mesmo que não saibamos como a população inteira se distribui, temos 95% de confiança de que a média verdadeira do tamanho das casas em Ames se encontra entre os valores `lower` (limite inferior do intervalo de confiança) e `upper` (limite superior do intervalo de confiança). Contudo, existem algumas condições que precisam ser atendidas para esse intervalo ser válido.

> **Exercício 3** Para o intervalos de confiança ser válido, a média amostral precisa ter distribuição normal e ter um erro padrão igual a $s/\sqrt{n}$. Quais condições precisam ser atendidas para isso ser verdadeiro?

Níveis de Confiança

> **Exercício 4** O que significa "95% de confiança"? Se você não tem certeza, retome a Seção 4.2.2.

Neste caso nós temos a comodidade de saber a verdadeira média populacional, uma vez que temos os dados da população inteira. Este valor pode ser calculado utilizando o seguinte comando:

```
mean(population)
```

> **Exercício 5** O seu intervalo de confiança contém a verdadeira média do tamanho das casas em Ames? Se você está trabalhando neste laboratório em uma sala de aula, o intervalo de seus colegas também contém esse valor?

> **Exercício 6** Cada aluno de sua turma deve ter obtido um intervalo de confiança um pouco diferente. Que proporção desses intervalos você espera que contenha a verdadeira média populacional? Por quê? Se você está trabalhando neste laboratório em um sala de aula, reúna informações sobre os intervalos criados pelos outros alunos da turma e calcule a proporção de intervalos que contém a verdadeira média populacional.

Utilizando o R, vamos criar várias amostra para aprender um pouco mais a respeito de como as médias

[†]Confira a seção 4.2.3 se você não está familiarizado com essa fórmula.

amostrais e os intervalos de confiança variam de uma amostra para outra. *Loops* são úteis para isso.[§]

Eis o esboço do processo:

(1) Obter uma amostra aleatória.

(2) Calcular a média e o desvio padrão da amostra.

(3) Utilizar estas estatísticas para calcular um intervalos de confiança.

(4) Repetir as etapas (1)-(3) 50 vezes.

Mas antes de implementar esse processo, precisamos primeiro criar vetores vazios nos quais possamos salvar as médias e desvios padrão que serão calculados para cada amostra. Ao mesmo tempo, vamos também armazenar o tamanho da amostra como `n`.

```
samp_mean <- rep(NA, 50)

samp_sd <- rep(NA, 50)

n <- 60
```

Agora estamos prontos para o *loop*, com o qual calculamos as médias e desvios padrão de 50 amostras aleatórias.

```
for(i in 1:50){
  samp <- sample(population, n) # obtém uma amostra de n = 60 elementos da população
  samp_mean[i] <- mean(samp)    # salva a média amostral no i-ésimo elemento de samp_mean
  samp_sd[i] <- sd(samp)        # salva o dp da amostra como o i-ésimo elemento de samp_sd
}
```

Por fim, construímos os intervalos de confiança.

```
lower_vector <- samp_mean - 1.96 * samp_sd / sqrt(n)

upper_vector <- samp_mean + 1.96 * samp_sd / sqrt(n)
```

Os limites inferiores destes 50 intervalos de confiança são armazenados na vetor `lower_vector`, e o limites superiores são armazenados no vetor `upper_vector`. Vamos visualizar o primeiro intervalo.

```
c(lower_vector[1],upper_vector[1])
```

Sua Vez

1. Utilizando a seguinte função (que foi carregada junto com o conjunto de dados), crie gráficos de todos os intervalos. Que proporção dos intervalos de confiança contém a verdadeira média populacional?

[§]Se você não está familiarizado com *loops*, revise o Laboratório 4A.

Essa proporção é exatamente igual ao nível de confiança? Se não, explique por quê.[†]

```
plot_ci(lower_vector, upper_vector, mean(population))
```

2. Escolha um intervalo de confiança de sua preferência, desde que não seja de 95%. Qual é o valor crítico apropriado?
3. Calcule 50 intervalos de confiança utilizando o nível de confiança que você escolheu na questão anterior. Você não precisa obter novas amostras: simplesmente calcule os novos intervalos baseado nas médias amostrais e desvios padrão que você já coletou. Utilizando a função `plot_ci`, crie gráficos de todos os intervalos e calcule a proporção de intervalos que contém a verdadeira média populacional. Compare essa proporção com o nível de confiança escolhido para os intervalos.
4. Quais conceitos do livro são abordados neste laboratório? Quais conceitos, se houver algum, que não são abordados no livro? Você viu esses conceito em algum outro lugar, p.e., aulas, seções de discussão, laboratórios anteriores, ou tarefas de casa? Seja específico em sua resposta.

[†]Essa figura pode parecer familiar (Verifique a Seção 4.2.2.)