Laboratório 4A: Fundamentos para Inferência Estatística - Distribuições Amostrais

Neste laboratório, investigaremos os meios pelos quais as estatísticas de uma amostra aleatória de dados podem servir como estimativas pontuais de parâmetros populacionais. Estamos interessados em formular uma *distribuição amostral* de nossa estimativa para aprender sobre as propriedades da estimativa, como sua distribuição.

Os Dados

Vamos analisar dados do setor imobiliário da cidade de Ames, no estado de Iowa, Estados Unidos. Os detalhes de cada transação imobiliária na cidade de Ames é registrada pelo escritório da Secretaria Municipal da Receita da cidade. Nosso foco particular para este laboratório será todas as vendas de casa em Ames entre 2006 e 2010. Essa coleção representa nossa população de interesse. Neste laboratório queremos aprender sobre essas vendas de casa retirando pequenas amostra da população completa. Vamos importar os dados.

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

load("ames.RData")
```

Vemos que há muitas variáveis em nosso conjunto de dados, o suficiente para realizar uma análise aprofundada. Para este laboratório, restringiremos nossa atenção para somente duas variáveis: a área habitável da casa acima do nível do solo em pés quadrados (`Gr.Liv.Area`) e o preço da venda (`SalePrice`). Para economizar esforços ao longo do laboratório, crie duas variáveis com nomes curtos para representar essas duas variáveis do conjunto de dados.

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

price <- ames$SalePrice
```

Vamos dar uma olhada na distribuição da área em nossa população de vendas de casas calculando algumas estatísticas sumárias e criando um histograma.

```
summary(area)

hist(area)
```

Exercício 1 Descreva a distribuição da população.

A Distribuição Amostral Desconhecida

Neste laboratório nós temos acesso à população inteira, mas isso raramente acontece na vida real. Reunir informação sobre uma população inteira costuma ser muito custoso ou impossível. Por essa razão,

frequentemente retiramos uma amostra da população e a utilizamos para compreender propriedades da população.

Se estivermos interessados em estimar a área habitável média na cidade de Ames com base numa amostra, podemos utilizar o seguite comando para sondar a população.

```
samp1 <- sample(area, 50)
```

Esse comando retira uma amostra aleatória simples de tamanho 50 do vetor `area`, que é atribuída à variável `samp1`. É como se fôssemos ao banco de dados da Secretaria Municipal da Fazendo e retirássemos os arquivos de 50 vendas de casas aleatoriamente. Trabalhar com esses 50 arquivos seria consideravelmente mais simples do que lidar com todas as 2930 vendas de casas.

> **Exercício 2** Descreva a distribuição desta amostra. Como ela se compara à distribuição da população?

Se estamos interessados em estimar a área habitável média nas casas da cidade de Ames utilizando esta amostra, nossa melhor suposição é a média da amostra.

```
mean(samp1)
```

Dependendo de quais foram as 50 casas que foram sorteadas, sua estimativa como estar um pouco acima ou abaixo da média populacional verdadeira de 1499,69 pés quadrados. De maneira geral, mesmo assim, a média da amostra costuma ser uma estimativa muito boa da média da área habitável, e nós a obtemos por meio de uma amostra de menos de 3% da população.

> **Exercício 3** Retire uma segunda amostra, também de 50 casos, e a atribua a uma variável de nome `samp2`. Como a média de `samp2` se compara à média de `samp1`? Vamos supor que retiremos mais duas amostras, uma de de 100 casos e outra de 1000 casos. Qual você acha que daria uma estimativa mais precisa da média populacional?

Não é surpreendente que, a cada vez que retiramos uma nova amostra aleatória, obtemos uma média amostral diferente. É útil ter uma ideia de quanta variabilidade podemos esperar quando estimamos a média populacional desta maneira. A distribuição das médias amostrais, denominada de *distribuição amostral*, pode nos ajudar a compreeder essa variabilidade. Neste laboratório, uma vez que temos acesso à população, podemos elaborar a distribuição amostral para a média amostral repetindo os passos acima várias vezes. Agora geraremos 5000 amostras e calcularemos a média amostra de cada uma.

```
sample_means50 <- rep(0, 5000)

for(i in 1:5000){
   samp <- sample(area, 50)
   sample_means50[i] <- mean(samp)
   }

hist(sample_means50)
```

Se você quiser ajustar a largura dos segmento do seu histograma para exibir um pouco mais de detalhes, você pode fazê-lo mudando o argumento `breaks`.

```
hist(sample_means50, breaks = 25)
```

Nós utilizamos o R para retirar 5000 amostras de 50 casos da população geral, calcular a média de cada amostra, e registrar cada resultado num vetor denominado `sample_means50`. Na próxima página, compreenderemos como esse conjunto de códigos funciona.

> **Exercício 4** A variável `sample_means50` contém quantos elementos? Descreva a distribuição amostral, e certifique-se de prestar atenção especificamente em seu centro. Você acha que a distribuilção mudaria se coletássemos 50.000 médias amostrais?

Interlúdio: O Comando `for` para Repetições

Vamos nos afastar da estatística por um momento para comprender melhor o último bloco de código. Você acabou de rodar seu primeiro *loop*, uma repetição de uma mesma sequência de instruções que é fundamental para a programação de computadores. A ideia por trás do *loop* é a noção de *iteração*: ele permite que você execute um código quantas vezes quiser sem ter que digitar cada iteração. Na caso acima, nós queríamos iterar as duas linhas de código que estão dentro das chaves, que retiram uma amostra aleatória de 50 casos da variável `area` e então salva a média da amostra no vetor `sample_means50`. Sem o *loop*, programar isso seria tedioso:

```
sample_means50 <- rep(0, 5000)

samp <- sample(area, 50)
sample_means50[1] <- mean(samp)

samp <- sample(area, 50)
sample_means50[2] <- mean(samp)

samp <- sample(area, 50)
sample_means50[3] <- mean(samp)

samp <- sample(area, 50)
sample_means50[4] <- mean(samp)
```

e assim por diante...

Usando o comando `for` para implementar um *loop*, essas milhares de linhas de código são comprimidas em um punhado de linhas. Adicionamos uma linha extra ao código abaixo, que imprime a variável `i` em cada iteração do *loop*. Rode este código.

```
sample_means50 <- rep(0, 5000)

for(i in 1:5000){
   samp <- sample(area, 50)
   sample_means50[i] <- mean(samp)
   print(i)
   }
```

Vamos examinar este código linha a linha para compreender o que ele faz. Na primeira linha nós *inicializamos um vetor*. Nesse caso, criamos um vetor com 5000 zeros denominado `sample_means50`. Esse vetor armazenará os valores gerados dentro do *loop*.

A segunda linha executa o *loop*. A sintaxe pode ser lida mais ou menos como, "para cada elemento `i` de 1 a 5000, execute as seguintes linhas de código". Você pode interpretar o `i` como um contador que mantém o registro de qual *loop* você está. Portanto, mais precisamente, o *loop* será executado uma vez quando `i=1`, e então mais uma vez quando `i=2`, e assim por diante até `i=5000`.

A parte principal do *loop* se encontra dentro das chaves, e esse conjunto de linhas de código é executado para cada valor de `i`. Aqui, em cada iteração, selecionamos uma amostra aleatória de 50 elementos a partir da variável `area`, calculamos sua média, e registramos seu valor como o $i^{\text{ésimo}}$ elemento do vetor `sample_means50`.

Para demonstrar que isso está de fato acontecento, pedimos ao R para imprimir o valor de `i` em cada iteração. Esta linha de código é opcional e é usada somente para mostrar o que está acontecendo enquando o *loop* do comando `for` está em execução.

O *loop* nos permite não somente rodar o código 5000 vezes, mas também armazenar os resultados ordenadamente, elemento por elemento, num vetor vazio que inicializamos nas primeiras linhas.

> **Exercício 5** Para certificar que você compreendeu o que você fez neste *loop*, experimente rodar uma versão menor. Inicialize um vetor com 100 zeros com o nome `sample_means_small`. Execute um *loop* que retira uma amostra de 50 elementos da variável `area` e armazena o média amostral no vetor `sample_means_small`, mas que repete a iteração de 1 a 100. Imprima o resultado em sua tela (basta digitar `sample_means_small` no console e pressionar enter). Há quantos elementos no objeto `sample_means_small`? O que cada elemento representa?

Tamanho da Amostra e Distribuição Amostral

À parte dos aspectos mecânicos de programação, vamos retomar a razão pela qual utilizamos o *loop* do comando `for`: para calcular uma distribuição amostral, especificamente, esta aqui:

```
hist(sample_means50)
```

A distribuição amostral que calculamos nos informa bastante sobre as estimativas da área habitável nas casas na cidade de Ames. Uma vez que a média amostral é um estimador não-enviesado, a distribuição amostral está centrada na verdadeira média da área habitável da população, e a dispersão da distribuição indica quanta variabilidade é possível ao se amostra somente 50 vendas de casas.

Para ter uma ideia melhor do efeito do tamanho da amostra na distribuição amostral, vamos construir mais duas distribuições amostrais: uma baseada numa amostra de 10 elementos e outra baseada numa amostra de 100.

```
sample_means10 <- rep(0, 5000)
sample_means100 <- rep(0, 5000)

for(i in 1:5000){
  samp <- sample(area, 10)
  sample_means10[i] <- mean(samp)
  samp <- sample(area, 100)
  sample_means100[i] <- mean(samp)
}
```

Aqui podemos utilizar um único *loop* para construir duas distribuições adicionando mais algumas linhas dentro das chaves. Não se preocupe com o fato de que `samp` é utilizado como o nome de dois objetos diferentes. No segundo comando do *loop*, a média de `samp` é salva em seu devido lugar no vetor `sample_means10`. Com a média já salva, podemos sobrescrever o objeto `samp` com uma nova amostra,

desta vez de com 100 elementos. De maneira geral, quando você cria um objeto utilizando um nome que já está em uso, o objeto antigo será substituído pelo novo.

Para verificar o efeito que diferentes tamanhos de amostra tem na distribuição amostral, crie gráficos das três distribuições, um em cima do outro.

```
par(mfrow = c(3, 1))

xlimits = range(sample_means10)

hist(sample_means10, breaks = 20, xlim = xlimits)
hist(sample_means50, breaks = 20, xlim = xlimits)
hist(sample_means100, breaks = 20, xlim = xlimits)
```

O primeiro comando especifica que você quer dividir a área do gráfico em três linhas e uma coluna para cada um dos gráficos[†]. O argumento `breaks` ("quebras") especifica o número de segmentos utilizados para construir o histograma. O argumento `xlim` especifica o intervalo no eixo x no hisograma, e ao definilo como igual a `xlimits` para cada histograma, certificamo-nos de que todos os três histogramas serão criados com os mesmos limites no eixo x.

> **Exercício 6** Quando o tamanho da amostra é maior, o que acontece com o centro da distribuição? E com a dispersão?

Sua Vez

Até agora, nós nos ocupamos em estimar a média da área habitável nas casas do município de Ames. Agora você tentará estimar a média dos preços das casas.

1. Retire uma amostra aleatória de 50 elementos da variável `price` (`preço`). Com essa amostra, qual é sua melhor estimativa pontual para a média populacional?
2. Já que você tem acesso à população, simule a distribuição amostral de $\bar{x}_{price}$ retirando 5000 amostras de 50 elementos da população e calculando 5000 médias amostrais. Armazene essas médias em um vetor com o nome `sample_means50`. Crie um gráfico com os resultados, e então descreva a forma dessa distribuição amostral. Baseado nessa distribuição amostral, qual seria seu palpite para a média dos preços das casas na população? Por fim, calcule e informe a média populacional.
3. Mude o tamanho da sua amostra de 50 para 150, e então calcule a distribuição amostral utilizando o mesmo método descrito acima, e guarde as médias em um novo vetor com o nome `sample_means150`. Descreva a forma dessa distribuição amostral e compare-a com a distribuição amostral para a amostra de 50 elementos. Com base nessa distribuição amostral, qual seria seu palpite sobre a média dos preços de vendas de casas no município de Ames?
4. Das distribuições amostrais calculadas nos exercícios 2 e 3, qual tem menor dispersão? Se estamos interessados em estimativas que estão mais próximas do valor verdadeiro, preferiríamos uma distribuição com uma dispersão pequena ou grande?

[†]Talvez você precise esticar um pouco sua janela com os gráficos para acomodar os gráficos extras. Para retornar para a configuração padrão de criar um gráfico por vez, rode o seguinte comando:

```
par(mfrow = c(1, 1))
```

5. 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.