Laboratório semana 1

```
---
title: "Foundations for inference - Sampling distributions"
---

### Load packages

Neste laboratório vamos explorar os dados usando o pacote 'dplyr' e
visualizá-los usando o pacote 'ggplot2' para visualização de dados. Os
dados podem ser encontrados no pacote complementar para este curso,
'statsr'.

```{r load-packages, message=FALSE}
library(statsr)
library(dplyr)
library(shiny)
library(ggplot2)
```
The data

Consideramos dados imobiliários da cidade de Ames, Iowa. Os detalhes de
cada transação imobiliária em Ames são registrados pela Assessoria da
Prefeitura. Nosso foco particular para este laboratório será todas as
vendas de casas residenciais
em Ames entre 2006 e 2010. Esta coleção representa a nossa população de
interesse. Neste laboratório, gostaríamos de aprender sobre essas vendas
de casas coletando amostras menores de toda a população. Vamos carregar os
dados.

```{r load-data}
data(ames)
```

Vemos que existem algumas variáveis no conjunto de dados, o suficiente
para fazer uma análise muito aprofundada. Para este laboratório,
restringiremos nossa atenção a apenas duas das variáveis: a área de estar
acima do solo da casa em metros quadrados ('área') e o preço de venda
('price').

Podemos explorar a distribuição de áreas de residências na população de
vendas de casas visualmente e com estatísticas resumidas. Vamos primeiro
criar uma visualização, um histograma:

```{r area-hist}
ggplot(data = ames, aes(x = area)) +
  geom_histogram(binwidth = 250)
```
```

Vamos também obter algumas estatísticas resumidas. Note que podemos fazer isso usando a função 'resumir'. Podemos calcular quantas estatísticas quisermos usando essa função, e apenas string ao longo dos resultados. Algumas das funções abaixo devem ser autoexplicativas (como 'média', 'mediana', 'sd', 'IQR', 'min' e 'max').

Uma nova função aqui é a função 'quantil' que podemos usar para calcular valores correspondentes a pontos de corte de percentis específicos na distribuição. Por exemplo, 'quantil(x, 0,25)' produzirá o valor de corte para o percentil 25 (Q1) na distribuição de x. Encontrar esses valores é útil para descrever a distribuição, pois podemos usá-los para descrições como *"os 50% médios das casas têm áreas entre tais e tais pés quadrados"*.

```{r area-stats}
ames %>%
  summarise(mu = mean(area), pop_med = median(area),
            sigma = sd(area), pop_iqr = IQR(area),
            pop_min = min(area), pop_max = max(area),
            pop_q1 = quantile(area, 0.25),  # first quartile, 25th percentile
            pop_q3 = quantile(area, 0.75))  # third quartile, 75th percentile
```

mu	pop_med	sigma	pop_iqr	pop_min	pop_max
<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>
1 1500.	1442	506.	617.	334	5642

1. Qual dos seguintes itens é falso?

A distribuição das áreas das casas em Ames é unimodal e inclinada para a direita

50% das casas em Ames são menores do que 1.499,69 pés quadrados.

O meio 50% das casas variam entre aproximadamente 1.126 pés quadrados e 1.742,7 pés quadrados.

O IQR é de aproximadamente 616,7 pés quadrados.

A menor casa tem 334 metros quadrados e a maior tem 5.642 metros quadrados.

A distribuição amostral desconhecida

Neste laboratório temos acesso a toda a população, mas este raramente é o caso na vida real. Reunir informações sobre toda uma população é muitas vezes extremamente caro ou impossível. Por causa disso, muitas vezes pegamos uma amostra da população e usamos isso para entender as

propriedades da população.

Se estivéssemos interessados em estimar a área de vida média em Ames com base em uma amostra, podemos usar o seguinte comando para fazer um levantamento da população.

```{r samp1}
samp1 <- ames %>%
  sample_n(size = 50)
```

Este comando coleta uma amostra aleatória simples de 'tamanho' 50 do conjunto de dados 'ames', que é atribuído a 'samp1'. Isso é como entrar no banco de dados da Prefeitura e pegar os arquivos sobre 50 vendas aleatórias de casas. Trabalhar com esses 50 arquivos seria consideravelmente mais simples do que trabalhar com todas as 2930 vendas de casas.

Exercício: Descrever a distribuição desta amostra? Como se compara à distribuição da população? **Dica:** A função 'sample_n' pega uma amostra aleatória de observações (ou seja, linhas) do conjunto de dados, você ainda pode se referir às variáveis no conjunto de dados com os mesmos nomes. O código que você usou no exercício anterior também será útil para visualizar e resumir a amostra, no entanto, tenha cuidado para não rotular mais os valores 'mu' e 'sigma', pois essas são estatísticas de amostra, não parâmetros populacionais. Você pode personalizar os rótulos de qualquer uma das estatísticas para indicar que elas vêm do exemplo.

```{r samp1-dist}
ggplot(data=samp1, aes(x=area))+
  geom_histogram()
```

A distribuição é mais próxima da normalidade

Se estivermos interessados em estimar a área média de vida em casas em Ames usando a amostra, nosso melhor palpite é a média da amostra.

```{r mean-samp1}
samp1 %>%
  summarise(x_bar = mean(area))
```

x_bar
1501

Dependendo de quais 50 casas você selecionou, sua estimativa pode estar um pouco acima ou um pouco abaixo da média real da população de 1.499,69 metros quadrados. Em geral, porém, a média amostral acaba sendo uma boa estimativa da área de vida média, e conseguimos obtê-la por amostragem de

menos de 3\% da população.

2. Suponha que tenhamos tomado mais duas amostras, uma de tamanho 100 e outra de tamanho 1000. Qual você acha que forneceria uma estimativa mais precisa da média populacional?

Tamanho da amostra de 50.

Tamanho da amostra de 100.

Tamanho da amostra de 1000.

Vamos pegar mais uma amostra de tamanho 50 e visualizar a área média nesta amostra:

```{r mean-samp2}
ames %>%
  sample_n(size = 50) %>%
  summarise(x_bar = mean(area))
```

Não surpreendentemente, cada vez que tomamos outra amostra aleatória, obtemos uma média de amostra diferente. É útil ter uma noção de quanta variabilidade devemos esperar ao estimar a média populacional dessa maneira.

A distribuição das médias amostrais, chamada de *distribuição amostral*, pode nos ajudar a entender essa variabilidade.

Neste laboratório, por termos acesso à população, podemos construir a distribuição amostral para a média amostral repetindo as etapas acima muitas vezes. Aqui vamos gerar 15.000 amostras e calcular a média amostral de cada uma. Note que estamos amostrando com substituição, 'replace = TRUE', uma vez que as distribuições amostrais são construídas com amostragem com reposição.

```{r loop}
sample_means50 <- ames %>%
                    rep_sample_n(size = 50, reps = 15000, replace = TRUE)
%>%
                    summarise(x_bar = mean(area))

ggplot(data = sample_means50, aes(x = x_bar)) +
  geom_histogram(binwidth = 20)
```

Aqui usamos R para tirar 15.000 amostras de tamanho 50 da população, calcular a média de cada amostra e armazenar cada resultado em um vetor chamado 'sample_means50'. Em seguida, analisamos como esse conjunto de código funciona.

Exercício: Quantos elementos existem em 'sample_means50'? Descreva a distribuição da amostragem e certifique-se de observar especificamente seu centro. Certifique-se de incluir um gráfico da distribuição em sua resposta.

```{r sampling-dist}
# type your code for the Exercise here, and Run Document

```

Interlúdio: Distribuições de amostragem

A ideia por trás da função 'rep_sample_n' é *repetição*. Anteriormente, tomamos uma única amostra de tamanho 'n' (50) da população de todas as casas em Ames. Com esta nova função, somos capazes de repetir este procedimento de amostragem 'rep' tempos, a fim de construir uma distribuição de uma série de estatísticas amostrais, que é chamado de **sampling distribution**.

Note-se que, na prática, raramente se consegue construir distribuições amostrais, porque raramente temos acesso a dados de toda a população.

Sem a função "rep_sample_n", isso seria doloroso. Teríamos que executar manualmente o seguinte código 15.000 vezes

```{r sample-code, eval=FALSE}
ames %>%
  sample_n(size = 50) %>%
  summarise(x_bar = mean(area))
```
bem como armazenar as médias de amostra resultantes cada vez em um vetor separado.

Note que para cada uma das 15.000 vezes que calculamos uma média, o fizemos a partir de uma amostra **diferente**!

Exercise: Para garantir que você entenda como as distribuições de amostragem são criadas e exatamente o que as funções 'sample_n' e 'do' fazem, tente modificar o código para criar uma distribuição de amostragem de **25 médias de amostra** de **amostras de tamanho 10**, e coloque-as em um quadro de dados chamado 'sample_means_small'. Imprima a saída. Quantas observações existem nesse objeto chamado 'sample_means_small'? O que representa cada observação?

```{r practice-sampling-dist}
# type your code for the Exercise here, and Run Document
```

```
sample_means_small <- ames %>%
  rep_sample_n(size = 10, reps = 25, replace = TRUE) %>%
  summarise(x_bar = mean(area))


ggplot(data = sample_means_small, aes(x = x_bar)) +
  geom_histogram(binwidth = 20)
```

3. How many elements are there in this object called `sample_means_small`?

0
3
25
100
5,000

4. Qual dos itens a seguir é **verdadeiro sobre os elementos nas distribuições de amostragem que você criou?**

Cada elemento representa uma metragem quadrada média de uma amostra aleatória simples de 10 casas.

Cada elemento representa a metragem quadrada de uma casa.

Cada elemento representa a verdadeira média populacional da metragem quadrada das casas.

Tamanho da amostra e distribuição amostral

Mecânica à parte, vamos voltar ao motivo pelo qual usamos a função 'rep_sample_n': para calcular uma distribuição amostral, especificamente, esta.

```{r hist}
ggplot(data = sample_means50, aes(x = x_bar)) +
  geom_histogram(binwidth = 20)
```

A distribuição amostral que calculamos nos diz muito sobre a estimativa da área média de vida nos domicílios em Ames. Como a média amostral é um estimador imparcial, a distribuição amostral está centrada na verdadeira área de vida média da população, e a dispersão da distribuição indica quanta variabilidade é induzida pela amostragem de apenas 50 casas vendidas.

No restante desta seção, trabalharemos para ter uma noção do efeito que o tamanho da amostra tem em nossa distribuição amostral.

Exercício: Use o aplicativo abaixo para criar distribuições de amostragem de médias de 'área' de amostras de tamanho 10, 50 e 100. Use 5.000 simulações. O que representa cada observação na distribuição amostral? Como a média, o erro padrão e a forma da distribuição amostral mudam à medida que o tamanho da amostra aumenta? Como (se é que esses valores mudam se você aumenta o número de simulações?

Cada observação representa a média amostral da área das casas. A mediada que o tamanho da amostra aumenta a distribuição fica mais próxima da normalidade com um intervalo mais curto (o gráfico fica mais fino), indicando que o erro padrão diminuiu, a medida que esse parâmetro aumenta.

```{r shiny, echo=FALSE}
shinyApp(
  ui <- fluidPage(

    # Sidebar with a slider input for number of bins
    sidebarLayout(
      sidebarPanel(

        selectInput("selected_var",
                    "Variable:",
                    choices = list("area", "price"),
                    selected = "area"),

        numericInput("n_samp",
                     "Sample size:",
                     min = 1,
                     max = nrow(ames),
                     value = 30),

        numericInput("n_sim",
                     "Number of samples:",
                     min = 1,
                     max = 30000,
                     value = 15000)

      ),

      # Show a plot of the generated distribution
      mainPanel(
        plotOutput("sampling_plot"),
        verbatimTextOutput("sampling_mean"),
        verbatimTextOutput("sampling_se")
      )
    )
```

```
  ),

  # Define server logic required to draw a histogram
  server <- function(input, output) {

    # create sampling distribution
    sampling_dist <- reactive({
      ames[[input$selected_var]] %>%
        sample(size = input$n_samp * input$n_sim, replace = TRUE) %>%
        matrix(ncol = input$n_samp) %>%
        rowMeans() %>%
        data.frame(x_bar = .)
      #ames %>%
      #  rep_sample_n(size = input$n_samp, reps = input$n_sim, replace =
TRUE) %>%
      #  summarise_(x_bar = mean(input$selected_var))
    })

    # plot sampling distribution
    output$sampling_plot <- renderPlot({
      x_min <- quantile(ames[[input$selected_var]], 0.1)
      x_max <- quantile(ames[[input$selected_var]], 0.9)

      ggplot(sampling_dist(), aes(x = x_bar)) +
        geom_histogram() +
        xlim(x_min, x_max) +
        ylim(0, input$n_sim * 0.35) +
        ggtitle(paste0("Sampling distribution of mean ",
                       input$selected_var, " (n = ", input$n_samp, ")")) +
        xlab(paste("mean", input$selected_var)) +
        theme(plot.title = element_text(face = "bold", size = 16))
    })

    # mean of sampling distribution
    output$sampling_mean <- renderText({
      paste0("mean of sampling distribution = ",
round(mean(sampling_dist()$x_bar), 2))
    })

    # mean of sampling distribution
    output$sampling_se <- renderText({
      paste0("SE of sampling distribution = ",
round(sd(sampling_dist()$x_bar), 2))
    })
  },

  options = list(height = 500)
)
```

5. Faz sentido intuitivo que, à medida que o tamanho da amostra aumenta, o centro da distribuição amostral se torna uma estimativa mais confiável para a verdadeira média da população. Também à medida que o tamanho da amostra aumenta, a variabilidade da distribuição amostral _____.

diminui
aumenta
permanece o mesmo

Exercício: Pegue uma amostra aleatória de tamanho 50 de 'preço'. Usando essa amostra, qual é a sua melhor estimativa pontual da média populacional?

```{r price-sample}
# type your code for this Exercise here, and Run Document

sampprice <- ames %>%

 sample_n(size = 50)


ggplot(data=sampprice, aes(x=area))+

 geom_histogram()


sampprice %>%

 summarise(media = mean(price))

```

* * *

Até agora, nos concentramos apenas em estimar a área média de vida nas casas em Ames. Agora você vai tentar estimar o preço médio da casa.

Observe que, embora você possa responder a algumas dessas perguntas usando o aplicativo, espera-se que você escreva o código necessário e produza os gráficos e estatísticas de resumo necessários. Você está convidado a usar o aplicativo para exploração.

6. Qual das seguintes afirmações é falsa?

A variabilidade da distribuição amostral com o menor tamanho de amostra ('sample_means50') é menor do que a variabilidade da distribuição amostral com o maior tamanho de amostra ('sample_means150').

As médias para as duas distribtuações de amostragem são aproximadamente semelhantes.

Ambas as distribuições amostrais são simétricas.

```{r price-sampling-compare}
# type your code for Question 6 here, and Run Document
```