Introdução e Estatística Descritiva

Fernando B. Sabino da Silva

Rstudio

- Faça uma pasta no seu computador onde você deseja manter os arquivos para usar no **Rstudio**.
- Defina o diretório de trabalho nesta pasta: `Session -> Set Working Directory -> Choose Directory` (atalho: Ctrl+Shift+H).
- Torne a alteração permanente definindo o diretório padrão em: `Tools -> Global Options -> Choose Directory`.

R básico

- Cálculos simples:

```
4.6 * (2 + 3)^4
```

```
## [1] 2875
```

- Defina um objeto (escalar) e o imprima:

```
a <- 4
a
```

```
## [1] 4
```

- Defina um objeto (vetor) e o imprima:

```
b <- c(2, 5, 7)
b
```

```
## [1] 2 5 7
```

Extensões do **R**

- O **R** não precisa ser usado apenas como calculadora ou para atribuição de objetos simples. A sua funcionalidade pode ser estendida atráves de bibliotecas ou pacotes (muito similar a utilização de Plug-ins nos navegadores ou baixar aplicativos no google play). Alguns já vem instalados (automaticamente, by default) no **R** e você precisa apenas carregá-los (como fazemos depois que baixamos um aplicativo no celular e queremos usá-lo, por exemplo).
- Para instalar um novo pacote no **Rstudio** você pode usar o menu: `Tools -> Install Packages`
- Você precisa saber o nome do pacote que deseja instalar. Você também pode fazê-lo através do comando install.packages como abaixo:

```
install.packages("mosaic")
```

- Uma vez que o pacote esteja instalado, você pode carregá-lo através do comando `library` (ou `require`):

Ajuda do **R**

- Você pode receber ajuda (help) via `?<command>`:

```
?sum
```

- Procurando por ajuda:

```
help.search("plot")
```

- Você pode encontrar um cheat sheet com funções do **R** que usaremos neste curso aqui. Caso o arquivo não apareça, clique com o botão direito em cima do link e escolha `Open link in a new tab`.
- Você pode salvar os comandos que você porventura tenha digitado em um arquivo para uso posterior:
  - Selecione o guia History no painel superior direito no **Rstudio** .
  - Marque os comandos que você deseja salvar.
  - Pressione o botão `To Source`.
- Pratique as suas habilidades básicas em:
  http://tryr.codeschool.com

Dados: Exemplos

- Data: Legibilidade de Anúncios em Revistas
- Trinta revistas foram classificadas pelo nível educacional de seus leitores.
- Três revistas foram selecionadas **aleatoriamente** de cada um dos seguintes grupos:
  - Grupo 1: maior nível educacional
  - Grupo 2: nível educacional médio
  - Grupo 3: n;ivel educacional mais baixo.
- Seis anúncios foram selecionados **aleatoriamente** de cada uma das nove revistas selecionadas:
  - Grupo 1: [1] Scientific American, [2] Fortune, [3] The New Yorker
  - Grupo 2: [4] Sports Illustrated, [5] Newsweek, [6] People
  - Grupo 3: [7] National Enquirer, [8] Grit, [9] True Confessions
- Logo, os dados contém informações sobre um total de 54 anúncios.

Objetivos do Capítulo

- Identificar o tipo de variável (por exemplo, numérica ou categórica; discreta ou contínua; ordenada ou não)
- Usar visualizações apropriadas para diferentes tipos de dados (por exemplo, histograma, gráfico de barras (barplot), gráfico de dispersão (scatterplot), boxplot, etc.)
- Criar e interpretar tabelas de contingência e de distribuições de frequência (tabelas uni e bidirecionais - de uma e duas entradas)
- Usar diferentes medidas de tendência central e dispersão e ser capaz de descrever a robustez de diferentes estatística (por exemplo, quando devemos usar cada uma e até que ponto elas podem ser usadas)
- Descrever a forma das distribuições (usando também gráficos como o histograma e o boxplot)

Exemplo (continuação) - variáveis e formato

- Para cada anúncio (54 casos), os dados abaixo foram observados.
- **Nome das variáveis**:
  - WDS = número de palavras na propaganda
  - SEN = número de frases na propaganda
  - 3SYL = número de palavras com 3 ou mais sílabas no anúncio
  - MAG = revista (1 a 9 como na página anterior)
  - GROUP = nível educacional (1 a 3 como na página anterior)
- Dê uma olhada nos dados usando **Rstudio**:

```
magAds <- read.delim("C:/Users/fsabino/Desktop/Codes/papers
head(magAds)
```

```
##   WDS SEN X3SYL MAG GROUP
## 1 205   9    34   1     1
## 2 203  20    21   1     1
## 3 229  18    37   1     1
## 4 208  16    31   1     1
```

Tipos de Dados

Variáveis Quantitativas

- Medições contém valores numéricos.
- Os dados quantitativos geralmente surgem das seguintes maneiras:
  - **Variáveis contínuas**: `medições` de, por exemplo, tempo de espera em uma fila, receitas, preços de ações, etc.
  - **Variáveis discretas**: `contagens` de, por exemplo, palavras em um texto, acessos de um website, números de chegadas em uma fila em uma hora, etc.
- Medidas como esta têm um escala bem definida e no **R** elas são armazenadas como numéricas (**numeric**).

Variáveis Categóricas/Qualitativas

- A medida é um fator proveniente de um conjunto de determinadas categorias. Exemplos: sexo (masculino/feminino), classe social, escore de satisfação (baixo/médio/alto), etc.

População e Amostra

Objetivo da Estatística

- O objetivo da Estatística é "dizer algo" sobre a população.
- Tipicamente, isso é feito utilizando as informações de uma amostra aleatória retirada da população de interesse.
- Antes de retirar a amostra podemos ter alguma hipótese sobre a população. A amostra é então analisada como o objetivo de testar esta hipótese.
- O processo de fazer conclusões para uma população com base em uma amostra é chamado de **inferência estatística**.

Seleção **aleatória**

- Exemplo: Para os dados das revistas:
  - Primeiro nós selecionamos **aleatoriamente** 3 revistas de cada grupo.
  - Na sequência, nós selecionamos, **aleatoriamente**, 6 anúncios de cada revista.
  - Um detalhe importante é que a seleção é feita de maneira completamente **aleatória**, i.e.
    - cada revista dentro de um grupo tem a mesma chance de ser escolhida e
    - cada anúncio dentro de uma revista tem a mesma chance de ser escolhido.
- No que veremos neste curso é fundamental que os dados coletados respeitem o princípio da aleatoriedade. Sempre que utilizarmos a palavra **amostra** daqui em diante, estaremos nos referindo a uma a.a. (amostra aleatória).
- Mais geralmente:
  - Nós temos uma **população** de objetos.
  - Nós escolhemos aleatoriamente n destes objetos, e do j-ésimo objeto nós obtemos a medição y_j, $j = 1, 2, \ldots, n$.

Tabelas de agrupamento e frequência

Dividir toda a gama de valores em uma série de intervalos: "Binning"

- A função `cut` irá dividir o intervalo de uma variável numérica em vários intervalos de tamanho igual e registrar a qual intervalo pertence cada observação. Por exemplo, para a variável `X3SYL` (o número de palavras com mais de 3 sílabas):

```
# Antes de 'cortar':
magAds$X3SYL[1:5]
```

```
## [1] 34 21 37 31 10
```

```
# Após 'cortar' (dividir) em 4 intervalos:
syll <- cut(magAds$X3SYL, 4)
syll[1:5]
```

```
## [1] (32.2,43]     (10.8,21.5]   (32.2,43]     (21.5,32.2
## Levels: (-0.043,10.8] (10.8,21.5] (21.5,32.2] (32.2,43]
```

Tabelas

- Para resumir os resultados nós podemos utilizar a função `tally` (contagem) do pacote`mosaic` (relembre que o pacote **deve ser carregado** escrevendo `library(mosaic)` se você ainda não o fez):

```
tally( ~ syll, data = magAds)
```

```
## syll
##  poucas algumas  muitas  demais
##      26      14      10       4
```

- Em porcentagem:

```
tally( ~ syll, data = magAds, format = "percent")
```

```
## syll
##  poucas algumas  muitas  demais
##    48.1    25.9    18.5     7.4
```

2 fatores: Tabulação Cruzada

- Para fazer uma tabela da combinação de dois fatores nós utilizamos a função `tally` novamente:

```
tally( ~ syll + GROUP, data = magAds)
```

```
##           GROUP
## syll        1  2  3
##   poucas    8 11  7
##   algumas   4  2  8
##   muitas    3  5  2
##   demais    3  0  1
```

- Frequências relativas (em porcentagem) por coluna:

```
tally( ~ syll | GROUP, data = magAds, format = "percent")
```

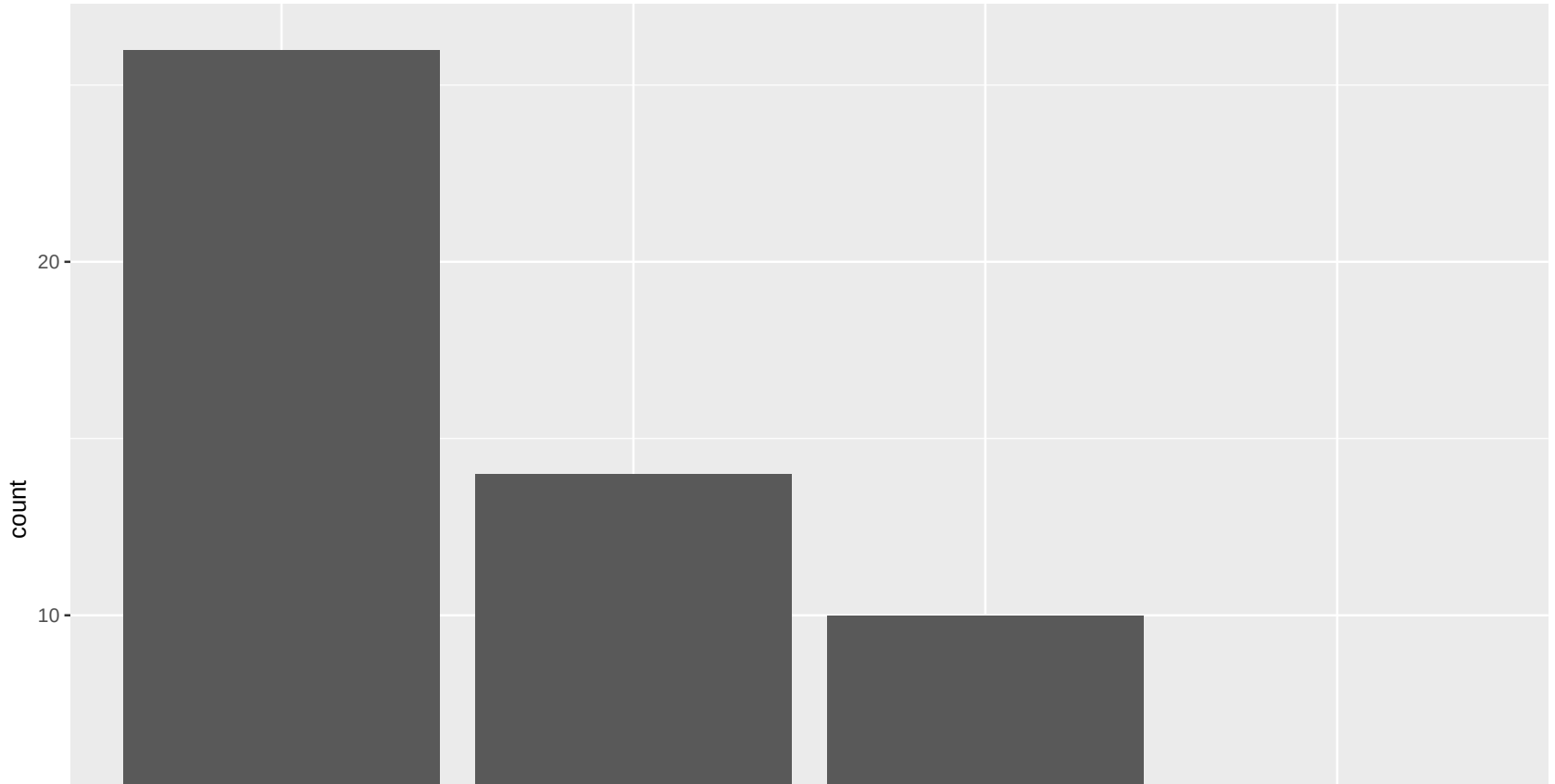
```
##           GROUP
## syll         1    2    3
```

Gráficos

Gráfico de barras

- Para criar um gráfico de barras com os dados da tabela nós usamos a função `gf_bar` do pacote `mosaic`. Para cada nível do fator uma caixa é desenhada com a altura proporcional a frequência (contagem) daquele nível.

```
gf_bar( ~ syll, data = magAds)
```



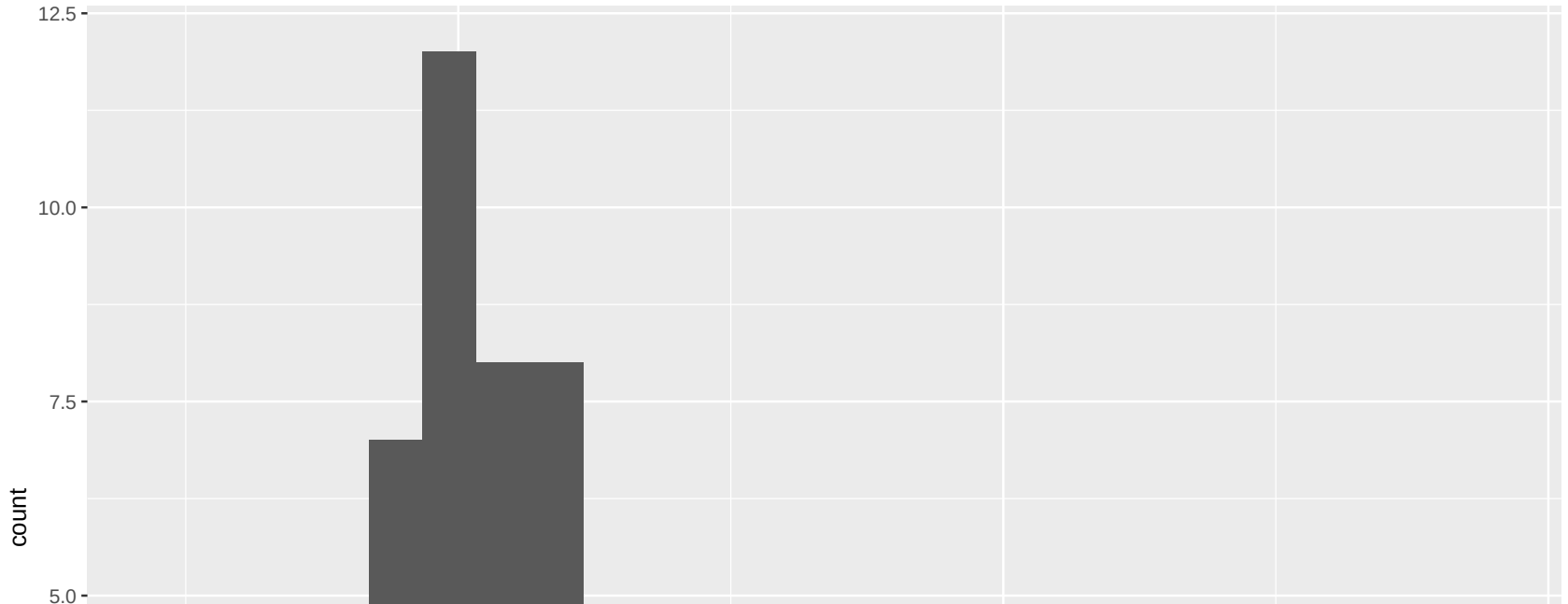
Os dados de Ericksen

- Descrição dos dados: Ericksen 1980 U.S. Census Undercount.
- Este conjunto de dados contém as seguintes variáveis:
  - `minority`: Percentual de negros ou hispânicos.
  - `crime`: Taxa de crimes graves por 1000 indivíduos na população.
  - `poverty`: Percentual de pobres.
  - `language`: Percentual com dificuldade em falar ou escrever Inglês.
  - `highschool`: Percentual com idade igual ou superior a 25 anos que não terminou o ensino médio.
  - `housing`: Percentual de habitação em pequenos edifícios de unidades múltiplas.
  - `city`: Um fator com níveis: `city` (cidade principal) ou `state` (estado or estado-resto).
  - `conventional`: Percentual de domicílios contados por enumeração pessoal convencional.
  - `undercount`: Estimativa preliminar de subentendimento percentual.
- Os dados de Ericksen têm 66 linhas/observações e 9 colunas/variáveis.

Histograma (usado para variáveis quantitativas)

- Como fazer um histograma para alguma variável `x`:
  - Divida o intervalo do valor mínimo de `x` para o valor máximo dex em um número apropriado de sub-intervalos de tamanho igual.
  - Desenhe uma caixa em cada sub-intervalo, sendo a altura proporcional ao número de observações no subintervalo.
- Histograma de taxas de criminalidade para os dados de Ericksen

```
gf_histogram( ~ crime, data = Ericksen)
```



Resumo de Variáveis Quantitativas

Medidas de centro dos dados (tendência central/posição): Média, Mediana e Moda

- Retornemos ao exemplo de anúncios da revista (`WDS` = número de palavras no anúncio). Uma série de resumos numéricos para `WDS` pode ser encontrada usando a função `favstats`:

```
favstats( ~ WDS, data = magAds)
```

```
##  min Q1 median  Q3 max mean sd  n missing
##   31 69     96 202 230  123 66 54       0
```

- Os valores observados da variável `WDS` são $y_1 = 205$, $y_2 = 203, \ldots, y_n = 208$, onde existe um total de $n = 54$ valores. Conforme definido anteriormente, isso constitui uma **amostra**.
- 123 é a **média** da amostra, que é calculada por

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Nós chamamos $\bar{y}$ de **média amostral**.

Medidas de variabilidade: amplitude, amplitude interquartílica, variância, desvio padrão, e coeficiente de variação

- Nós queremos saber "Quanto as observações estão desviadas do seu valor central?"
  - Ao olhar os dados e gráficos podemos ter uma sensação disto.
  - Porém, é comum estarmos interessados em um número para que possamos comparar as distribuições amostrais.
- **Amplitude** é a diferença entre o maior (máximo) e o menor (mínimo) valor.
  - Ela só usa dois valores para o seu cálculo, isto é, não leva todos em consideração.
  - Como trabalhamos com uma amostra, a amplitude que encontraremos será a amostral, isto é, em geral, temos uma subestimativa da verdadeira amplitude.
- A **amplitude interquartílica** é a diferença entre os valores do terceiro quartil e do primeiro quartil, isto é, $Q_3 - Q_1$.
  - Ela utiliza 50% dos valores para o seu cálculo.
- The **variância (empírica)** é a média dos desvios quadrados

Cálculo da média, mediana, amplitude interquartílica e desvio-padrão usando a função favstats do pacote mosaic

- Medidas Resumo de `WDS`:

```
favstats( ~ WDS, data = magAds)
```

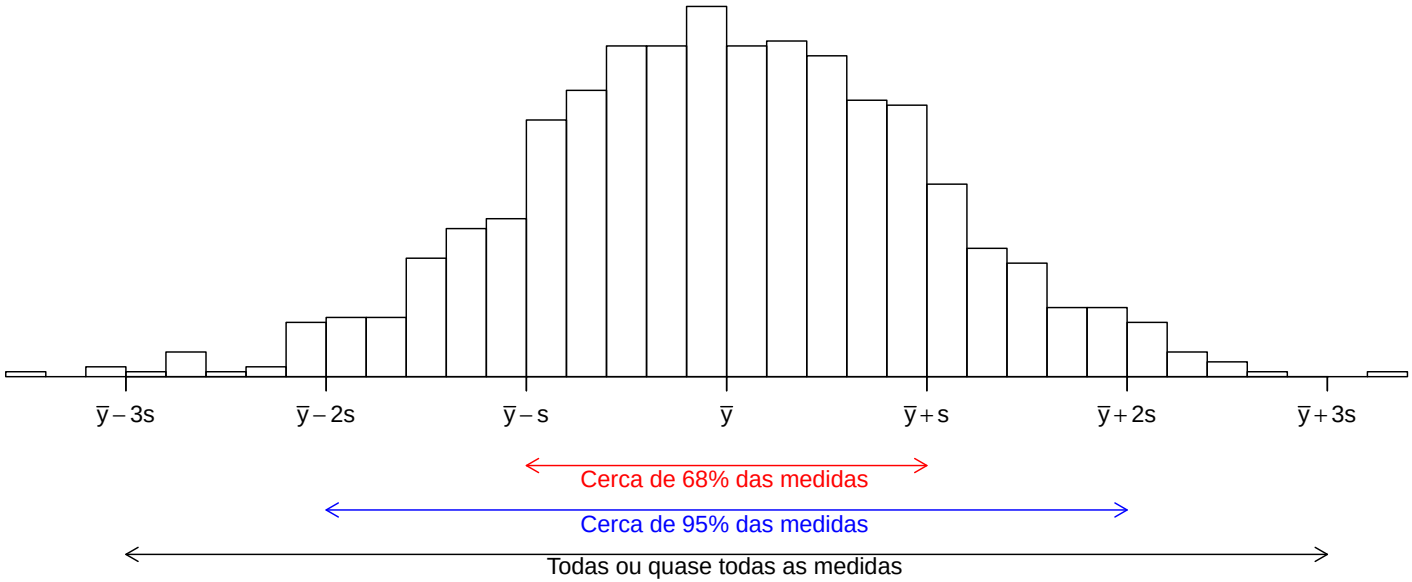
```
##  min Q1 median  Q3 max mean sd  n missing
##   31 69     96 202 230  123 66 54       0
```

Exercício: Interprete os resultados acima.

Uma palavra sobre terminologia

- **Desvio padrão**: uma medida de variabilidade de uma variável na amostra (ou população).
- **Erro padrão**: uma medida de variabilidade de uma estimativa (um particular valor de uma função da amostra). Por exemplo, uma medida de variabilidade da média amostral.

Uma regra empírica (veremos detalhes mais à frente)



Se o histograma com base na amostra parece uma função em forma de sino, então

- cerca de 68% das observações estão entre $\bar{y} - s$ e $\bar{y} + s$.
- acerca de 95% das observações estão entre $\bar{y} - 2s$ e $\bar{y} + 2s$.
- Todas ou quase todas as observações (99.7%) estão entre $\bar{y} - 3s$ e $\bar{y} + 3s$.

Percentis

- **O p-ésimo percentil** é um valor tal que pelo menos $p\%$ das observações são menores ou iguais a esse valor e pelo menos.
- Veja como calcular os percentis nas páginas 75-77 do livro texto.

Mediana, quartis e amplitude interquartílica

Recordando

```
favstats( ~ WDS, data = magAds)
```

```
##  min Q1 median  Q3 max mean sd  n missing
##   31 69     96 202 230  123 66 54       0
```

- 50-percentil = 96 é a **mediana** e é uma medida de tendência central/posição (centro dos dados).
- 0-percentil = 31 é o valor **mínimo**.
- 25-percentil = 69 é o **primeiro quartil** ou **quartil inferior** (Q1). Mediana dos 50% menores valores.
- 75-percentil = 201.5 é o **terceiro quartil** ou **quartil superior** (Q3). Mediana dos 50% maiores valores.
- 100-percentil = 230 é o valor **máximo**.
- **Amplitude Interquartílica (IQR)**: uma medida de variabilidade dada pela diferença entre o quartil superior e o quartil inferior: 201.5 - 69 = 132.5.

Mais gráficos

Box plots

Como desenhar um box plot:

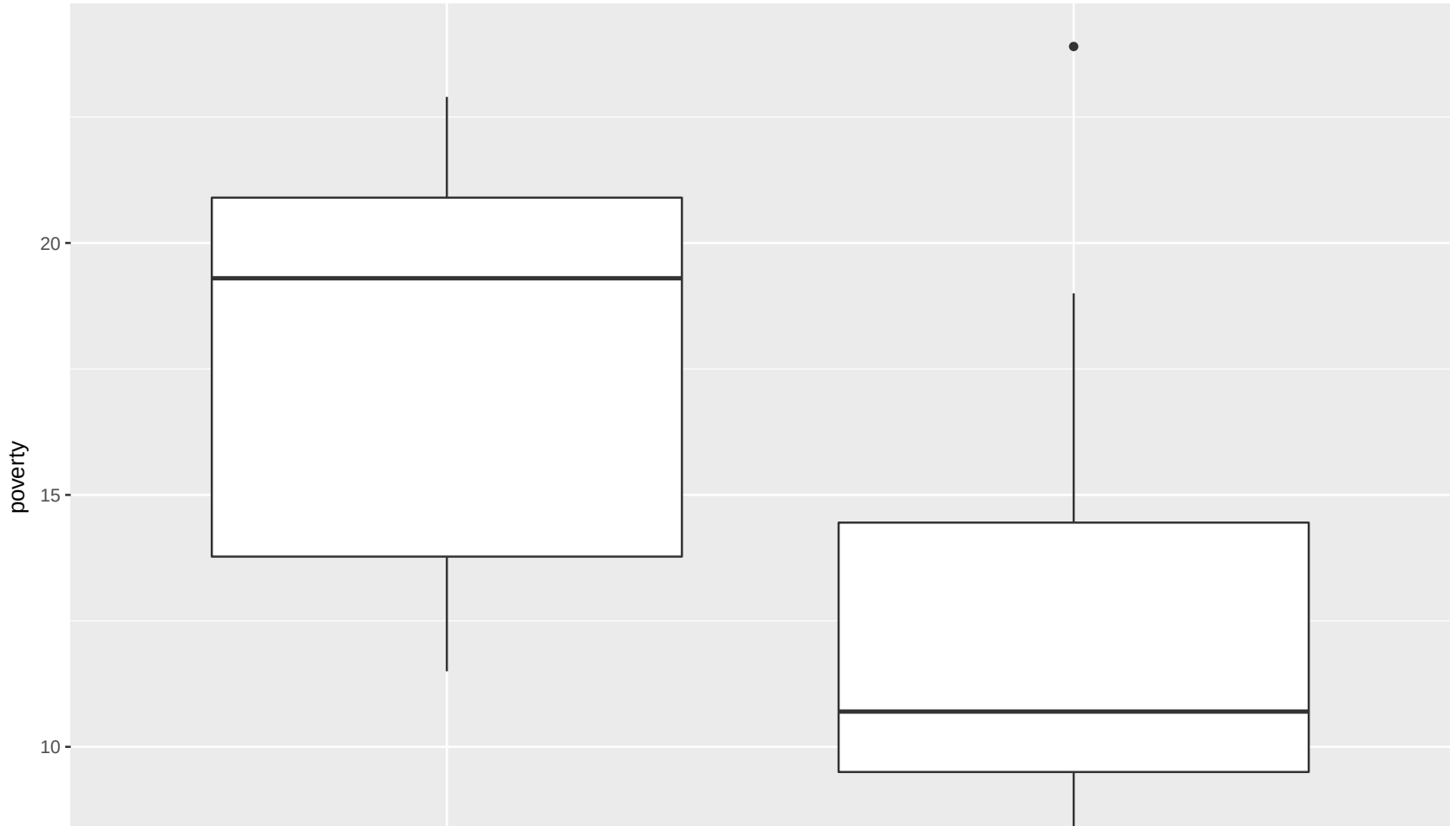
- Box:
  - Calcule a mediana, e os quartis inferior e superior.
  - Trace uma linha na mediana e desenhe uma caixa entre os quartis superior e inferior.
  - Calcule a amplitude interquartílica e a chame de IQR.
  - Calcule os seguintes valores:
    - L = quartil inferior - 1.5*IQR
    - U = quartil superior + 1.5*IQR
  - Desenhe uma linha ligando o quartil inferior até a menor medida que seja maior do que L.
  - Similarmente, desenhe uma linha ligando o quartil superior até a maior medida que seja inferior a U.
- Outliers: Observações com valor menor do que L ou maior do que U são desenhadas como círculos.

Nota: As caixas são fechadas (em inglês, as extremidades são chamadas de "Whiskers") no mínimo e no máximo das observações que não são consideradas outliers.

Boxplot para os dados de Ericksen

Boxplot das taxas de pobreza separadamente para cidados e estados (variável `city`):

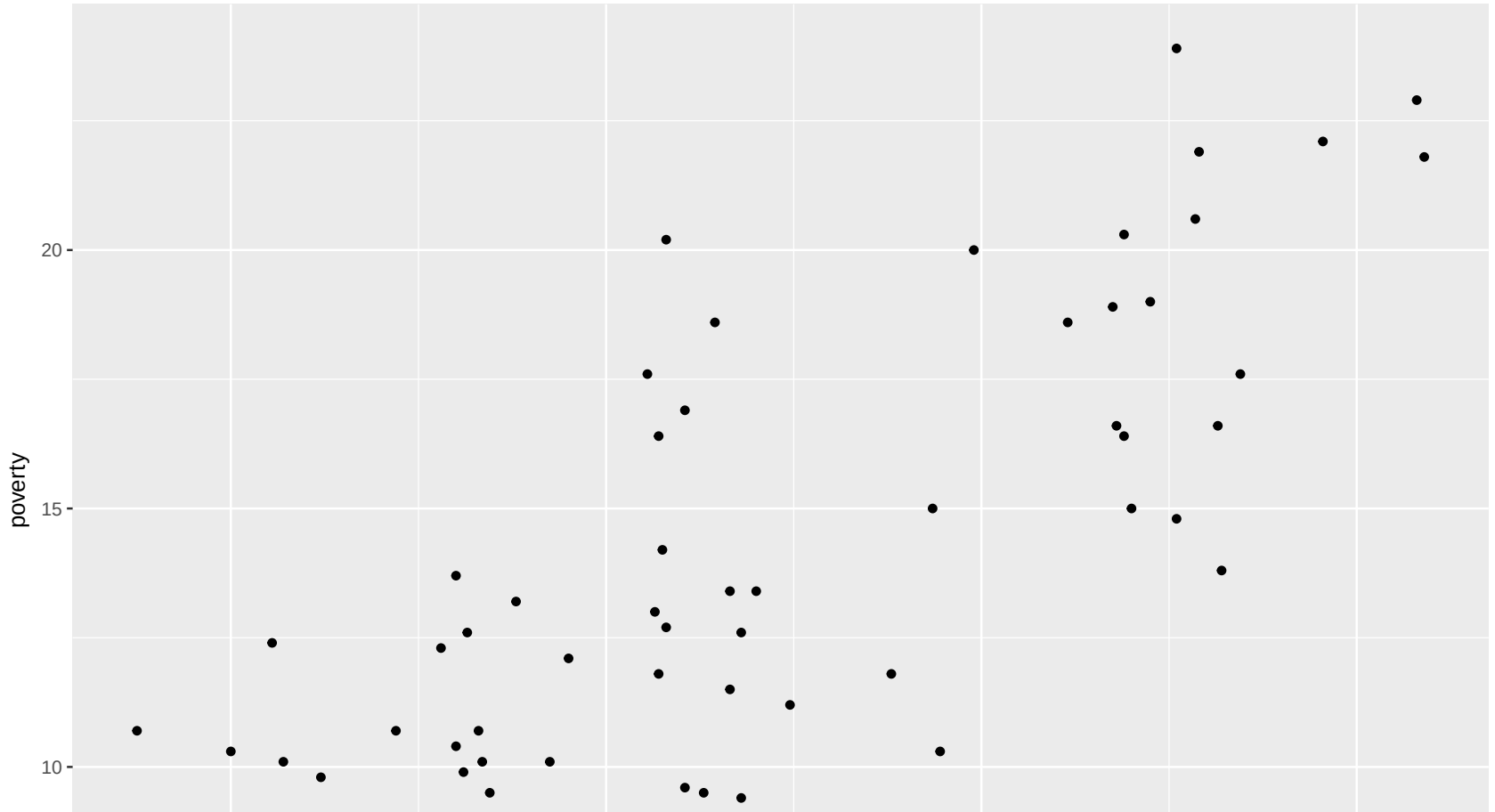
```
gf_boxplot(poverty ~ city, data = Ericksen)
```



2 variáveis quantitativas variables: Gráfico de dispersão ("Scatter plot")

Para duas variáveis quantitativas, um gráfico frequentemente utilizado é o de dispersão:

```
gf_point(poverty ~ highschool, data = Ericksen)
```



Apêndice

Recodificando variáveis

- A função `factor` converterá diretamente um vetor em uma variável qualitativa (escala nominal). Por exemplo:

```
head(magAds$GROUP)
```

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

```
class(magAds$GROUP)
```

```
## [1] "integer"
```

```
f <- factor(magAds$GROUP)
class(f)
```

```
## [1] "factor"
```

```
# magAds$GROUP <- f
# head(magAds$GROUP)
```

Apontar e clicar no gráfico

mplot

- Se os pacotes `mosaic` e `manipulate` forem instalados e estiverem carregados, nós podemos construir gráficos usando a função `mplot` simplesmente apontando e clicando.
- Usando `mplot` você pode fazer alterações pressionando o botão de configurações (uma roda dentada) no canto superior esquerdo da janela gráfica.

```
mplot(Ericksen)
```

- No final, você pode pressionar "Mostrar expressão" (Show expression) para obter o código.