Módulo VII: Estatística elementar

Revisão

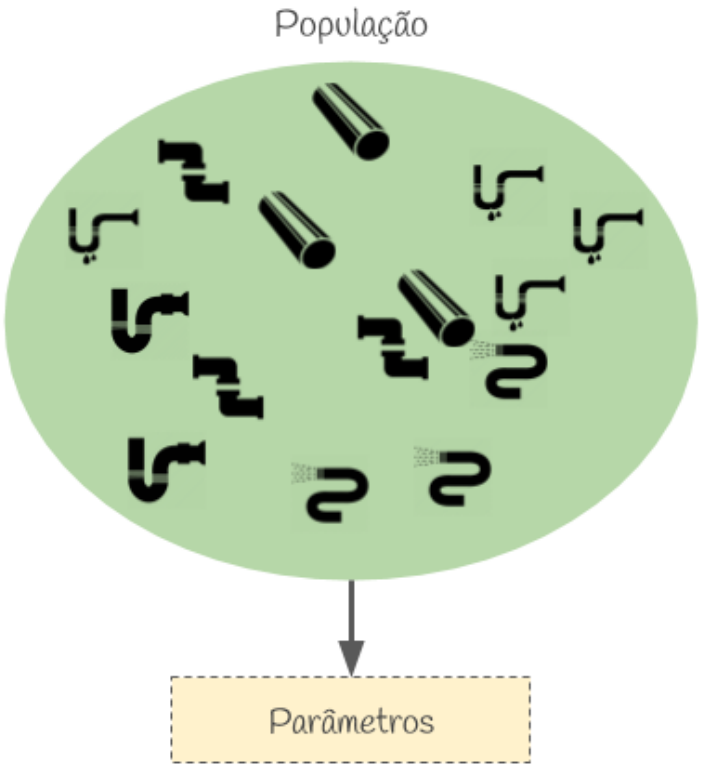
Palavras-chave em estatística

A estatística é composta por um conjunto de métodos destinados à coleta, à apresentação, à análise e à interpretação de dados, tendo como objetivo compreender fenômenos de natureza aleatória e ajudar o usuário na tomada de decisão.

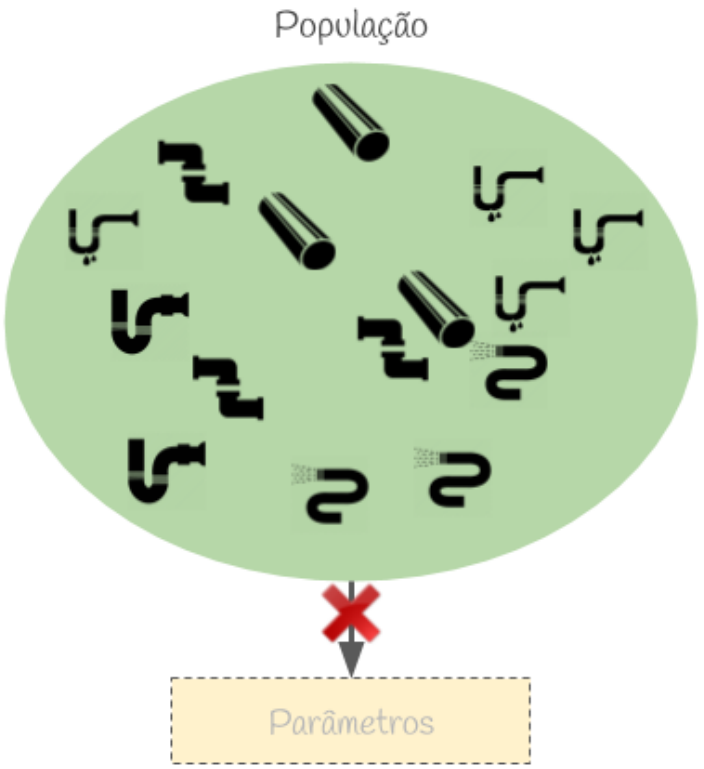
Palavras-chave

- `Variável:` O objeto/fenômeno/característica de interesse em estudo
- `População:` Conjunto de todas as observações da variável em estudo
- `Amostra:` Subconjunto de observações retiradas da população
- `Parâmetro:` Uma característica da população
- `Inferência:` Com base numa amostra inferimos os parâmetros associados à população

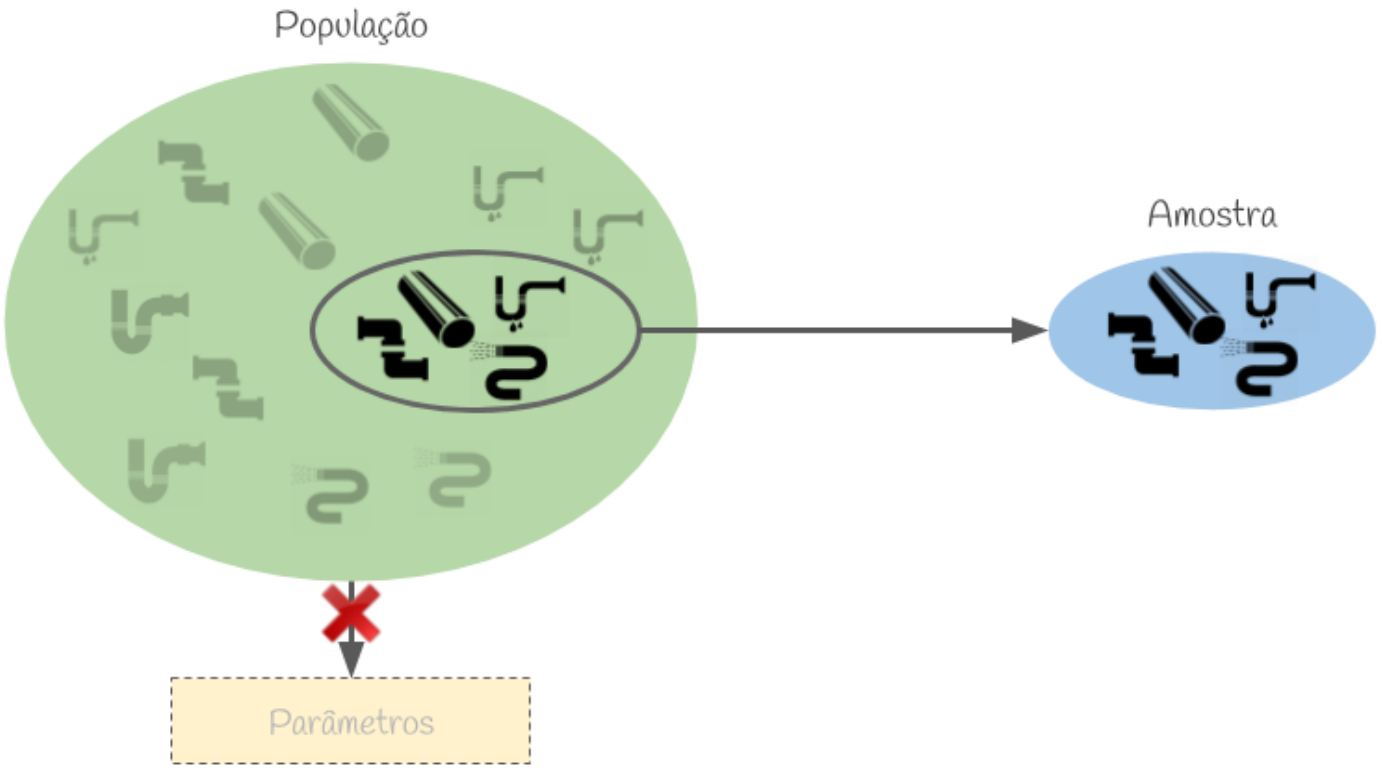
Estatística em um *slide*



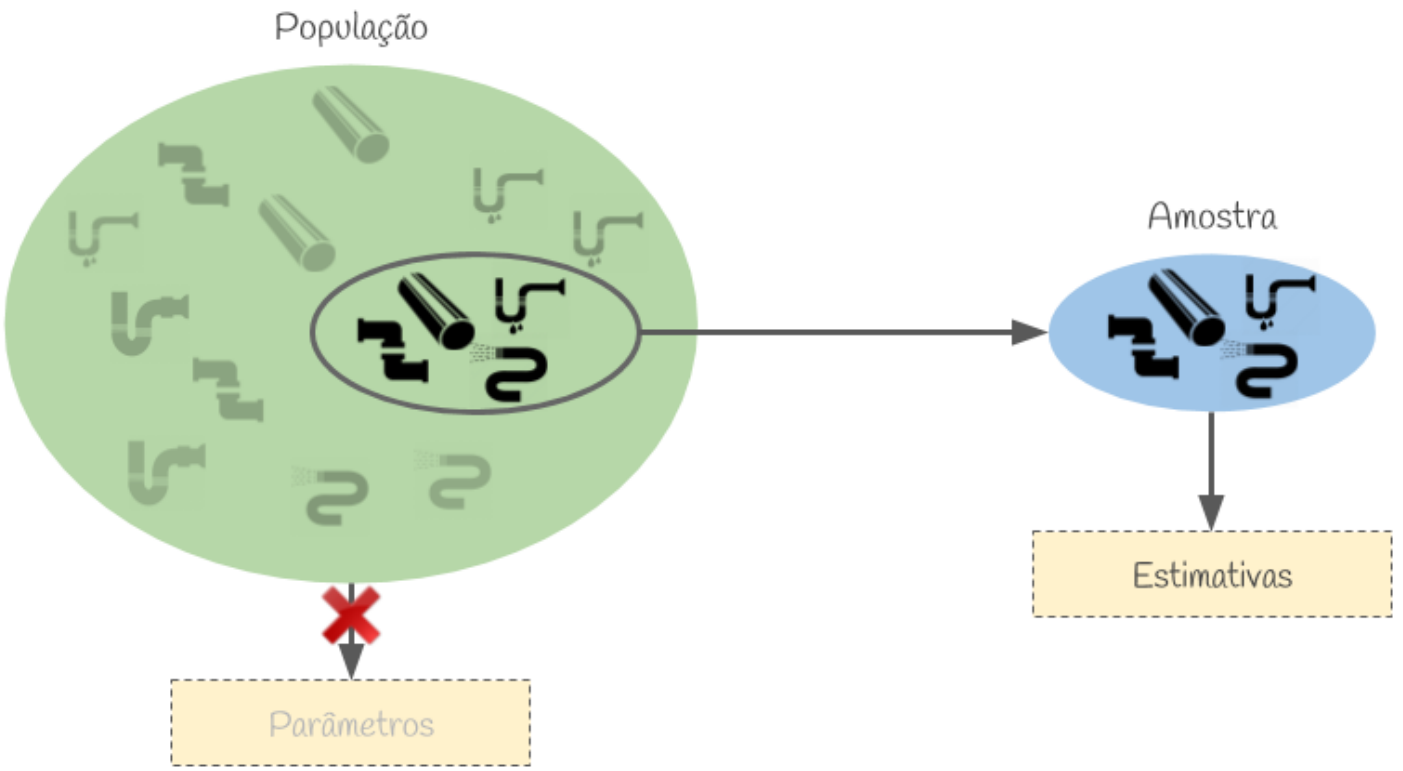
Estatística em um *slide*



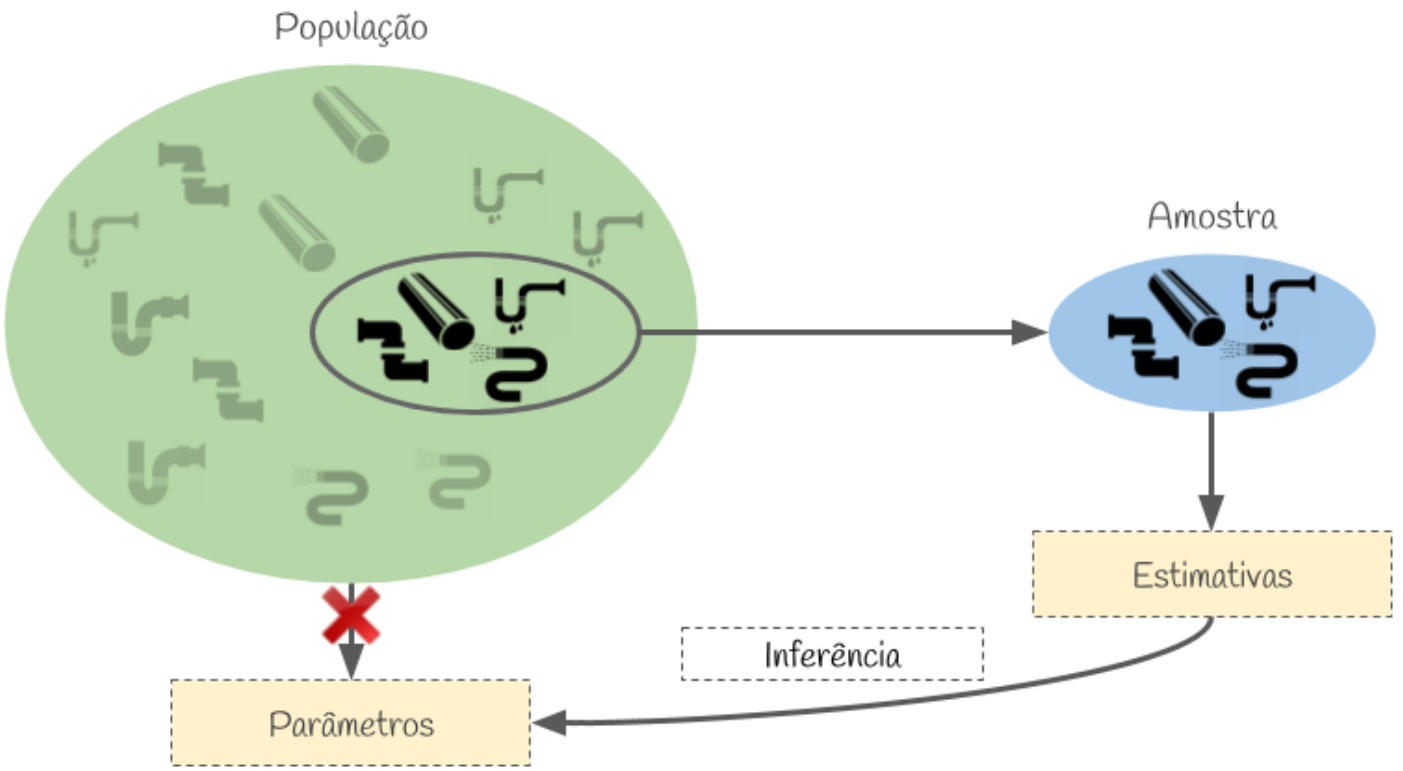
Estatística em um *slide*



Estatística em um *slide*



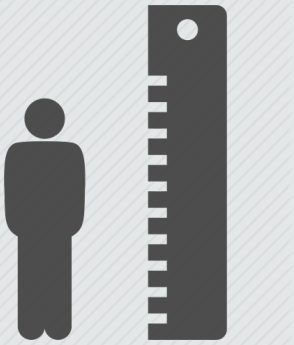
Estatística em um *slide*



Tipos de variáveis

`Variáveis quantitativas` `Variáveis qualitativas`

Contínuas



Discretas

Ordinais

Nominais

Para cada tipo de variável, diferentes tipos de análises (e gráficos) são indicados.

Estatística descritiva

Estatística descritiva

A estatística descritiva é uma área da estatística responsável por descrever e sumarizar um conjunto de dados.

Estatística descritiva

- `Medidas de tendência central:` Média, moda, mediana, ...
- `Medidas de variabilidade:` Variância, desvio padrão, amplitude, ...
- `Gráficos:` Boxplot, histograma, gráfico de barras, ...

Base para exemplificação

Vamos trabalhar com a base `dadosForno` e principalmente com as variáveis `PRS_GAS` e `PRS_FORNO`.

```
> load("/home/douglas/Documents/Dropbox/Cursos/Curso_R/data/Dados_VSB/dadosForno.RData")
> PRS_GAS <- dadosForno$PRS_GAS
> head(PRS_GAS)
[1] 100.1  99.5 100.1  99.9  99.5  99.9
```

```
> PRS_FORNO <- dadosForno$PRS_FORNO
> head(PRS_FORNO)
[1] 0.011 0.012 0.014 0.007 0.006 0.007
```

```
> miniBase <- cbind(PRS_FORNO,
+                   PRS_GAS)
```

Medidas de tendência central

As medidas de tendência central nos dão uma ideia de onde os dados estão concentrados.

Medidas de tendência central

- `mean`: Média de um vetor
- `median`: Mediana de um vetor
- `table`: Tabela de frequências/contingência
- `mlv`: Moda de um vetor
- `colSums`: Soma por colunas de base de dados
- `colMeans`: Média por coluna de uma base de dados

Medidas de variabilidade

As medidas de variabilidade nos dão uma ideia da dispersão dos dados.

Medidas de variabilidade

- `min`: Menor valor de um vetor
- `max`: Maior valor de um vetor
- `which.min`: Posição em que o mínimo ocorre
- `which.max`: Posição em que o máximo ocorre
- `quantile`: Quantis de um vetor
- `var`: Variância de uma matriz/vetor
- `sd`: Desvio padrão de um vetor
- `cov`: Covariância entre duas variáveis
- `cor`: Correlação entre duas ou mais variáveis

Resumo dos dados

Uma maneira simples de resumir os dados é utilizando a função `summary`

```
> summary(miniBase)
   PRS_FORNO            PRS_GAS
 Min.   :-0.037000   Min.   :  0.00
 1st Qu.: 0.007000   1st Qu.: 91.40
 Median : 0.010000   Median : 95.30
 Mean   : 0.009762   Mean   : 95.63
 3rd Qu.: 0.012000   3rd Qu.:100.00
 Max.   : 0.042000   Max.   :125.10
```

Agora é a sua vez

Fazendo uma análise descritiva

Análise descritiva

Utilizando a base de dados `dadosForno` encontre (quando possível):

- A média, mediana e variância
- O maior e o menor valor juntamente com a data em que entraram no forno
- Existem valores atípicos? O que ocorre com a média retirando 1
- Divida o desvio padrão pela média e multiplique por 100. O que podemos concluir?

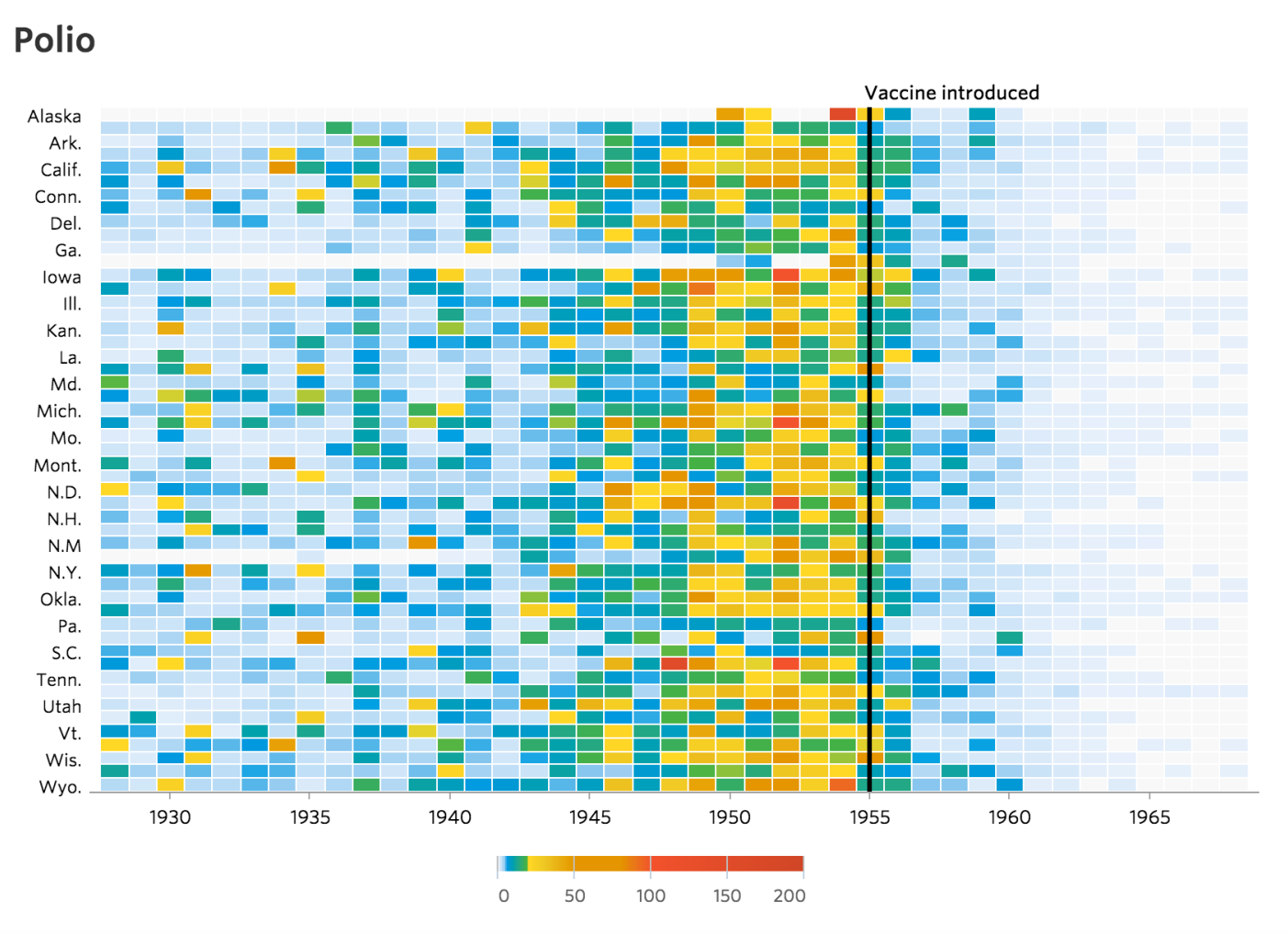
Utilize as variáveis:

- `DIS_ULT_BLOCO_ENFORNADO`,
- `CMP_BLOCO_ENFORNAMENTO` e
- `PER_O2_CANAL_FUMOS`

Gráficos estatísticos

Gráficos estatísticos

Visualizar os dados ajuda no entendimento do problema e também na apresentação dos resultados. Sendo assim é parte `essencial` de uma análise.



Gráficos estatísticos

Gráficos mostram de uma forma clara e objetiva o comportamento dos dados.

Gráficos estatísticos

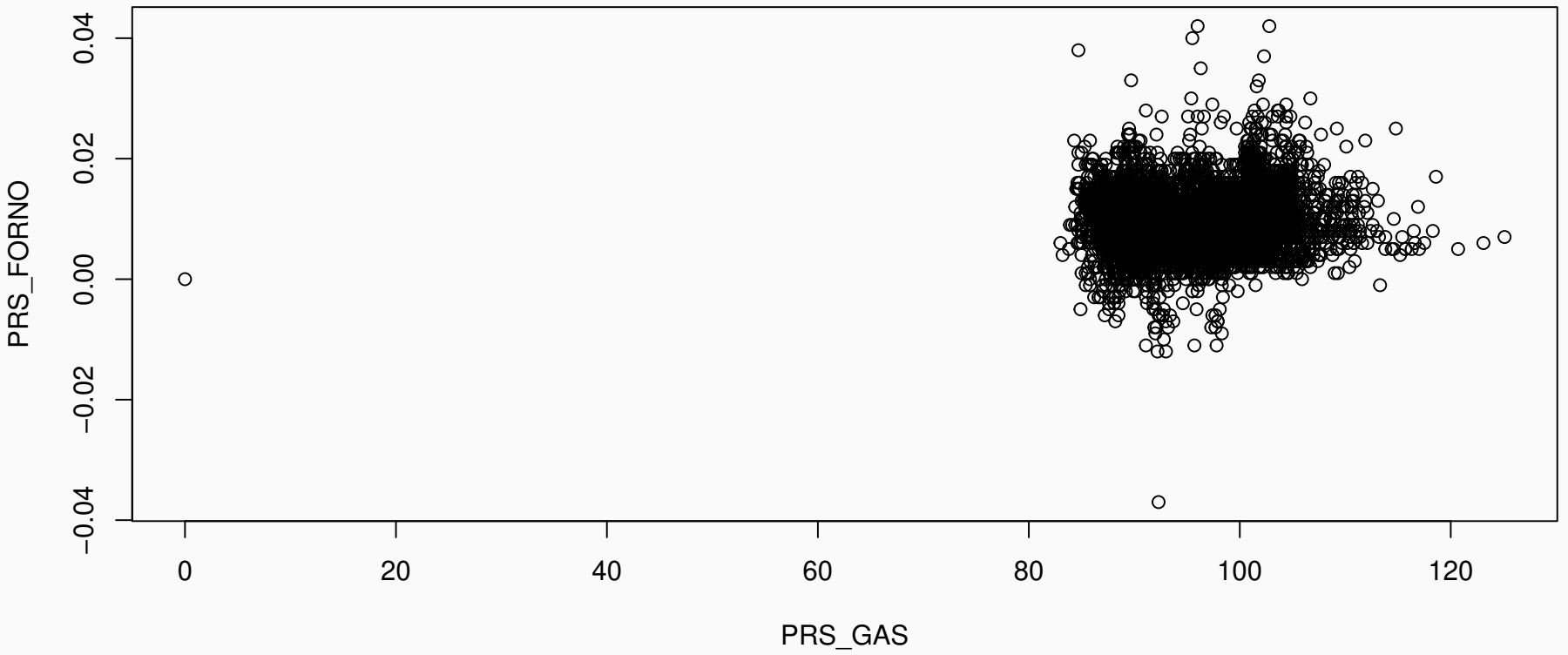
- `Dispersão`
- `Linhas`
- `Barras`
- `Setores/Pizza`
- `Histograma`
- `Boxplot`

Veremos no `próximo módulo` como `customizar` estes e outros gráficos.

Gráficos estatísticos - Dispersão

`Gráfico de dispersão`: Gráfico contrapondo duas varíaveis que pode evidenciar correlação entre as variáveis.

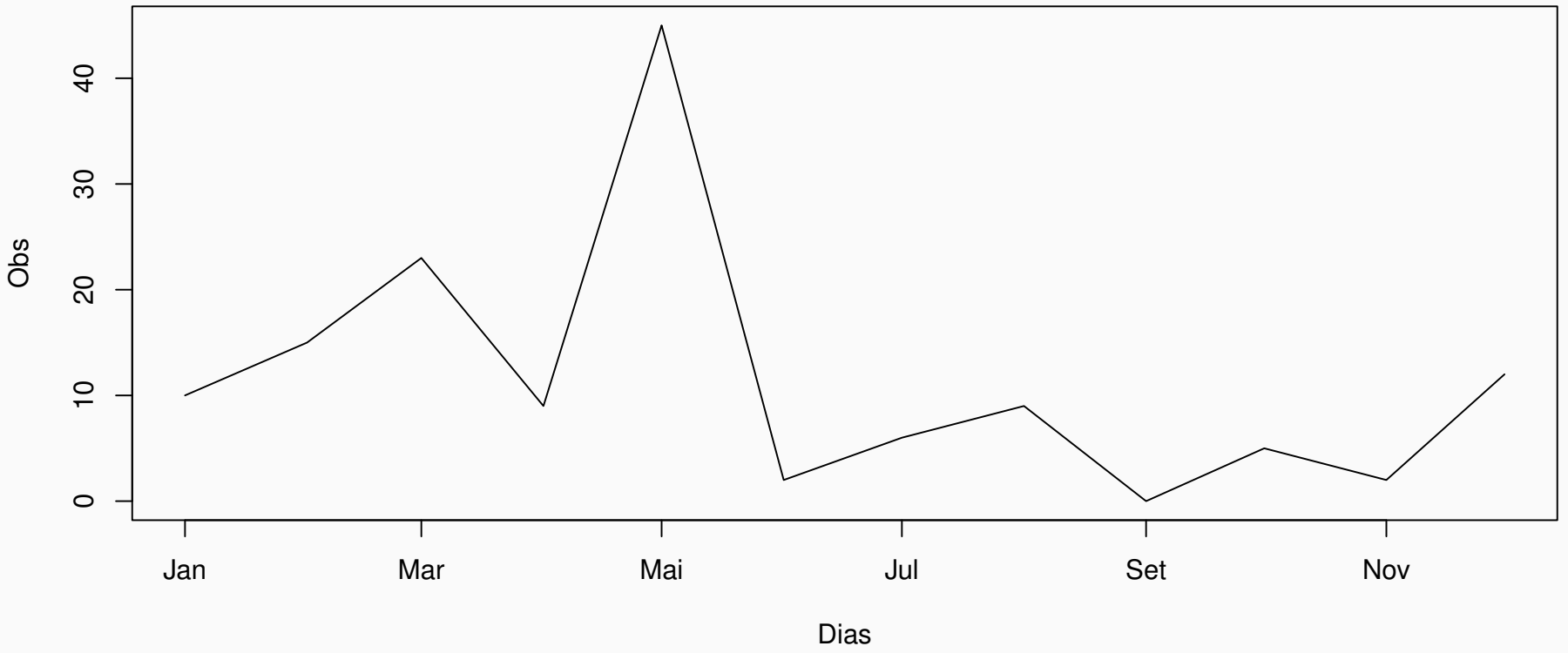
```
> plot(x = PRS_GAS, y = PRS_FORNO)
```



Gráficos estatísticos - Linhas

`Gráfico de linhas`: Gráfico contrapondo duas varíaveis em que normalmente o eixo x apresenta uma variável temporal.

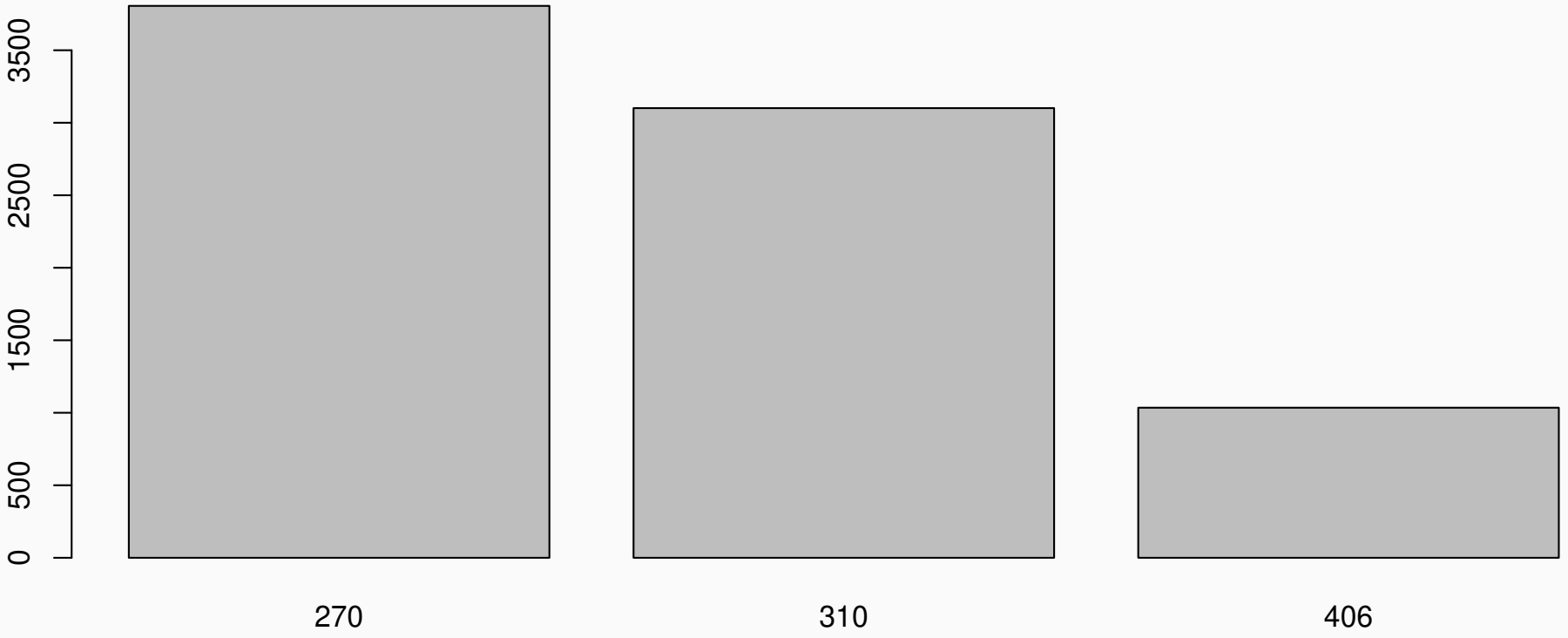
```
> Dias <- seq(as.Date("2016/01/01"), by = "month", length.out = 12)
> Obs <- c(10, 15, 23, 9, 45, 2, 6, 9, 0, 5, 2, 12)
> plot(x = Dias, y = Obs, type = "l")
```



Gráficos estatísticos - Barras

Gráfico de barras: É uma forma de resumir um conjunto de dados categóricos. Cada barra representa uma categoria particular e a altura de cada barra é proporcional à frequência de cada categoria.

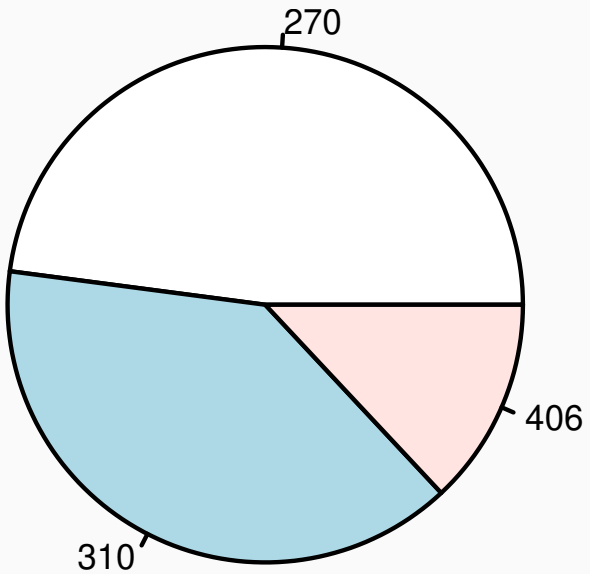
```
> DIM_BLOCO_ENFORNAMENTO <- dadosForno$DIM_BLOCO_ENFORNAMENTO
> cont <- table(DIM_BLOCO_ENFORNAMENTO)
> barplot(height = cont)
```



Gráficos estatísticos - Setores/Pizza

`Gráfico de setores`: Útil para visualizar a participação de cada categoria no total geral.

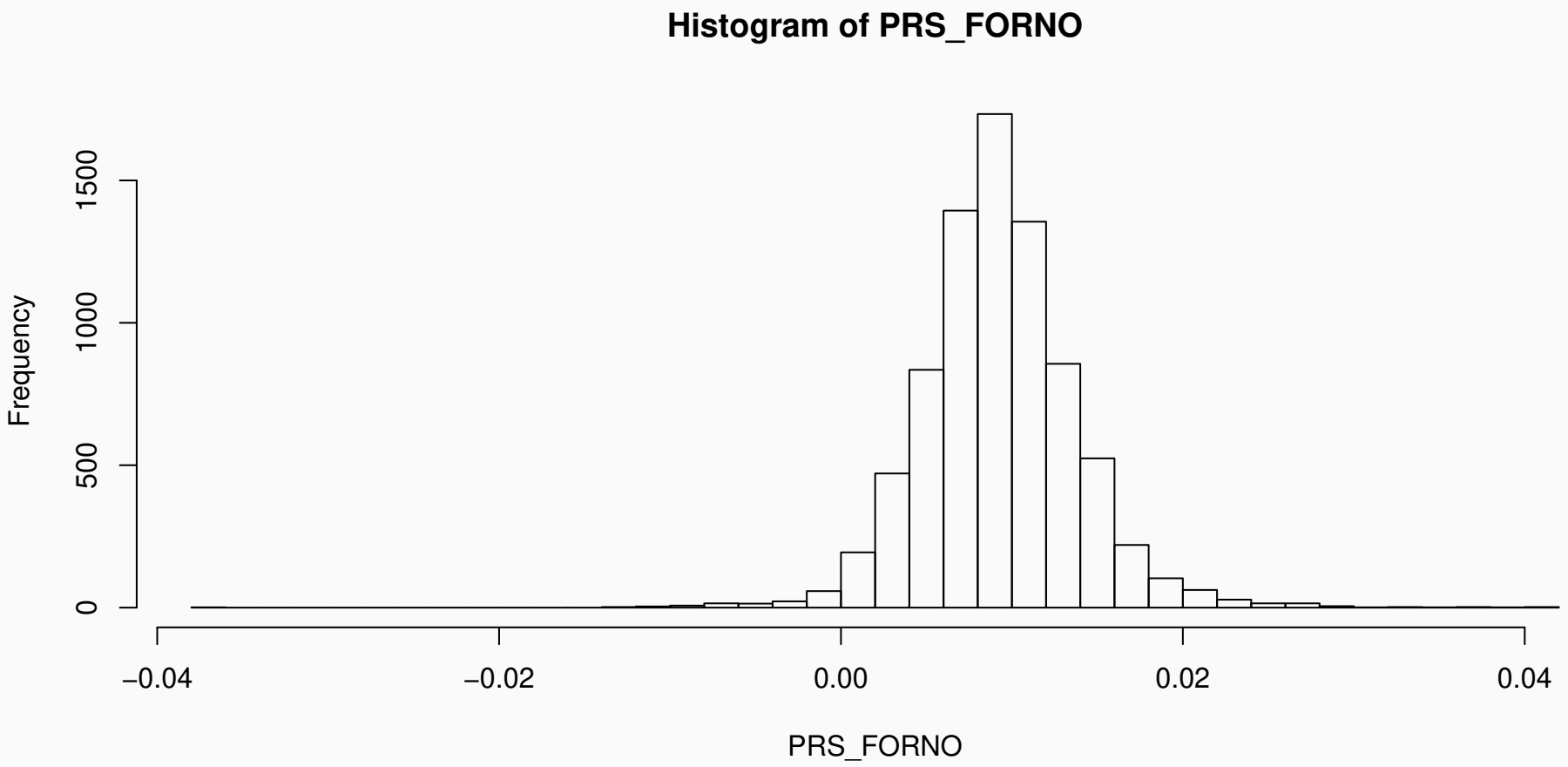
```
> pie(x = cont, cex = 0.5)
```



Gráficos estatísticos - Histograma

`Histograma`: Útil para caracterizar a distribuição dos dados.

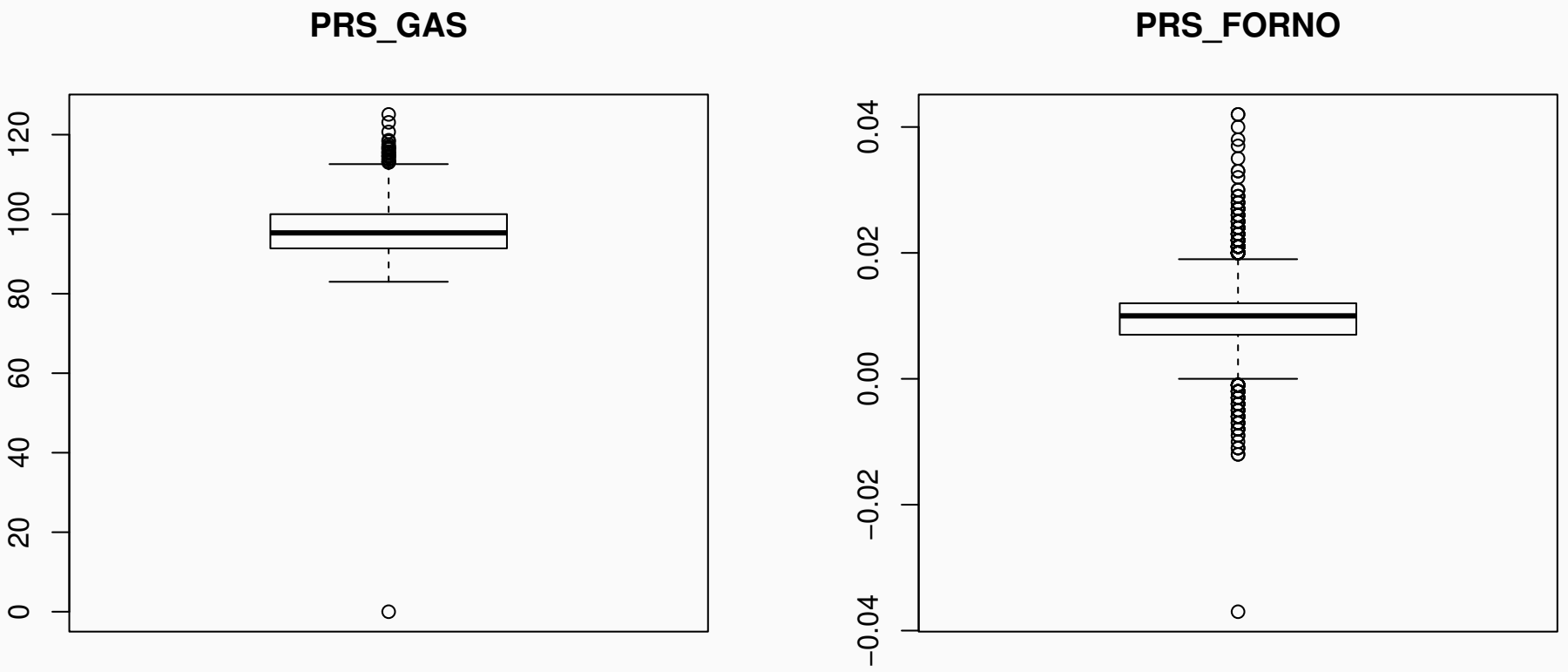
```
> hist(PRS_FORNO, breaks = 50)
```



Gráficos estatísticos - Boxplot

Boxplot: Útil para caracterizar a distribuição dos dados. Cada elemento do gráfico representa uma estatística e os pontos os *outliers*.

```
> par(mfrow=c(1,2))
> boxplot(PRS_GAS, main = "PRS_GAS")
> boxplot(PRS_FORNO, main = "PRS_FORNO")
```



Agora é a sua vez

Gráficos descritivos

Utilizando a base de dados `dadosForno` e as variáveis `DIS_ULT_BLOCO_ENFORNADO`, `CMP_BLOCO_ENFORNAMENTO` e `PER_O2_CANAL_FUMOS`, faça:

- O gráfico de dispersão entre DIS e CMP
- O histograma e o boxplot das variáveis DIS e CMP
- O gráfico de barras e de setores da variável PER

Distribuições de probabilidade

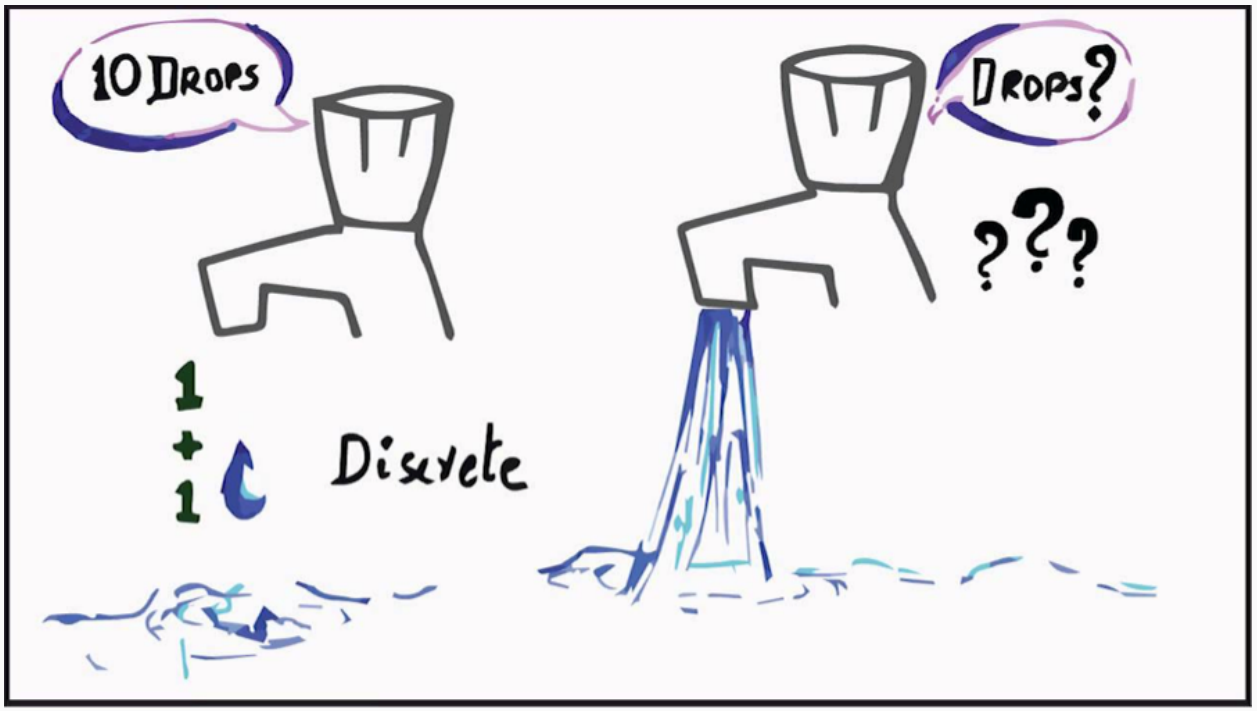
Distribuições de probabilidade

Distribuições contínuas

Assumem valores em uma escala contínua. Exemplo: Altura

Distribuições discretas

Assumem um número finito de valores. Exemplo: Número de filhos



Distribuições de probabilidade

Distribuições discretas

- `sample`: Uniforme
- `binom`: Binomial
- `multinom`: Multinomial
- `pois`: Poisson
- `geom`: Geométrica
- `nbinom`: Binomial negativa
- `...`

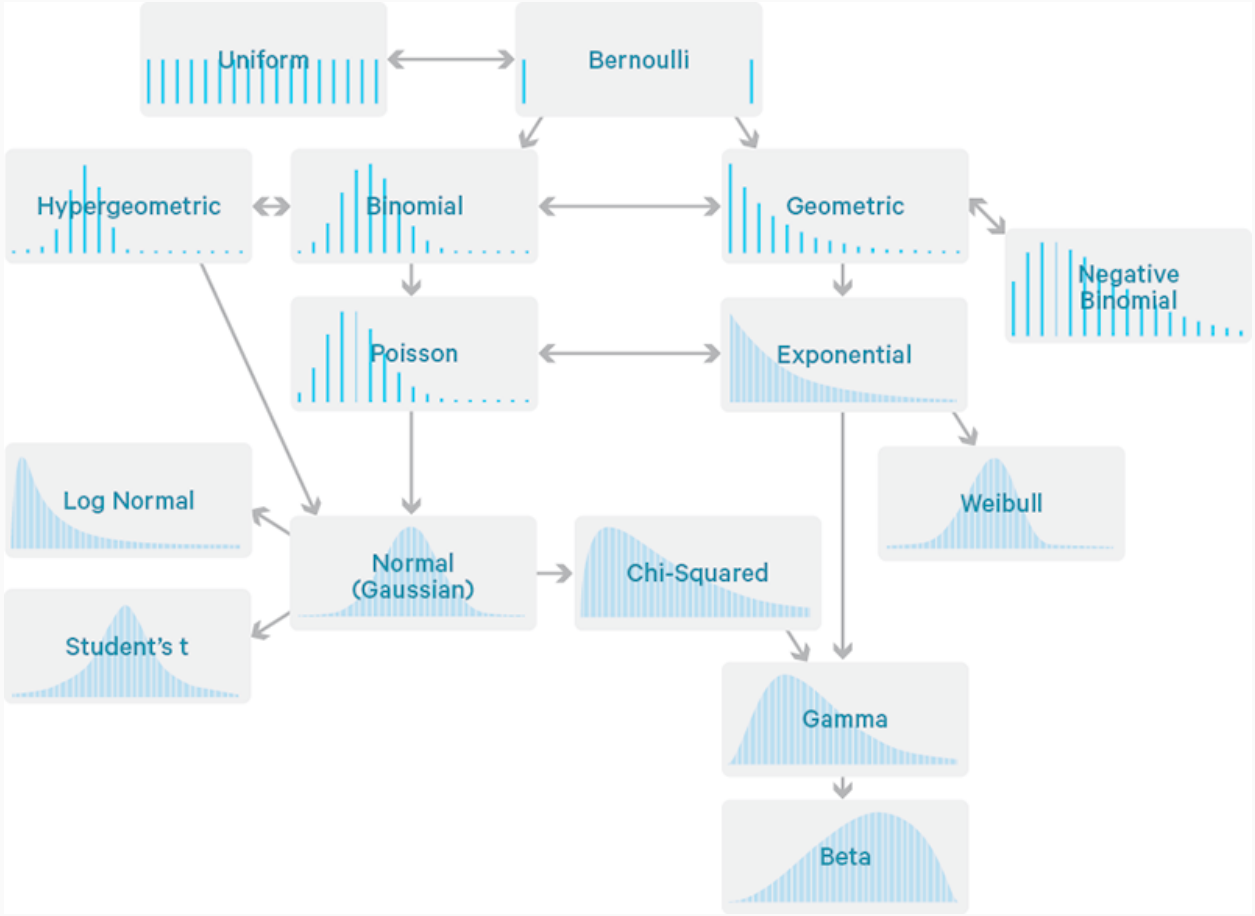
Distribuições contínuas

- `unif`: Uniforme
- `norm`: Normal
- `exp`: Exponencial
- `chisq`: Qui-quadrado
- `t`: t-student
- `f`: F de snedecor
- `...`

O uso de uma distribuição de probabilidade depende da natureza da variável em análise. Para isso é interessante conhecer o suporte das distribuições de probabilidade.

Distribuições de probabilidade

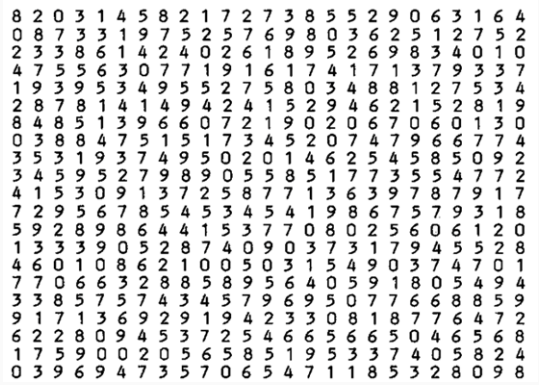
Existem muitas (`Muitas mesmo`) distribuições de probabilidade. Na prática algumas delas são mais utilizadas (como as citadas no *slide* anterior).

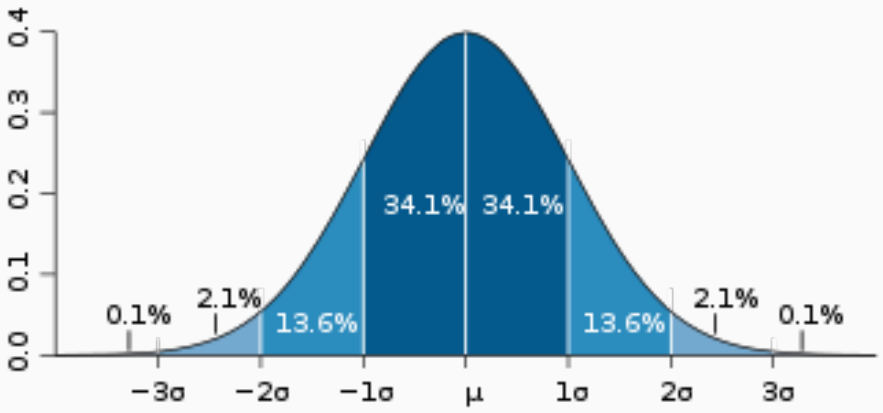


No R as distribuições são utilizadas com um dos prefixos abaixo.

Distribuições de probabilidade no R

- `Prefixo d`: Densidade
- `Prefixo q`: Quantil
- `Prefixo p`: Probabilidade
- `Prefixo r`: Números aleatórios

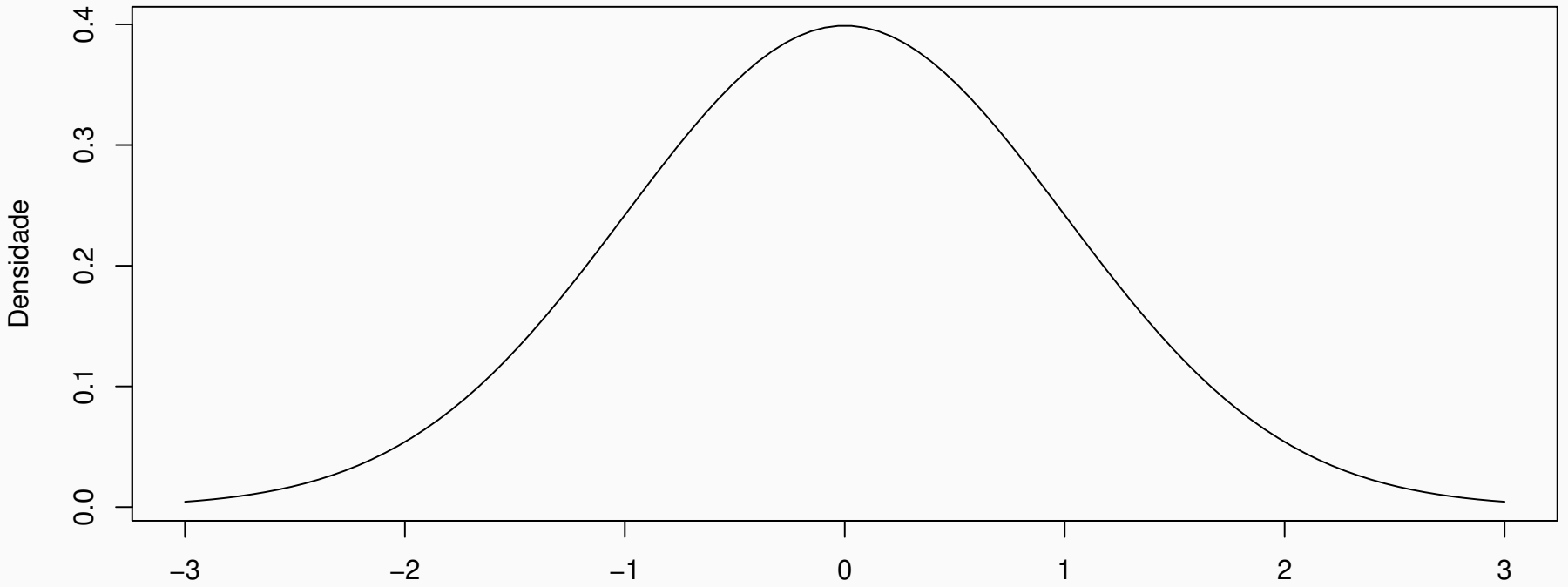




Distribuições de probabilidade - O prefixo d

O prefixo 'd' seguido do nome de uma distribuição corresponde à densidade da variável.

```
> x <- seq(-3, 3, length.out = 100)
> densidade <- dnorm(x = x, mean = 0, sd = 1)
> plot(x = x, y = densidade, type = "l",
+      xlab = "", ylab = "Densidade")
```



Distribuições de probabilidade - O prefixo q

O prefixo 'q' seguido do nome de uma distribuição corresponde ao `quantil` da variável.

```
> quantil1 <- qnorm(p = 0.025, mean = 0, sd = 1)
> quantil1
[1] -1.959964
```

```
> quantil2 <- qnorm(p = 0.975, mean = 0, sd = 1)
> quantil2
[1] 1.959964
```

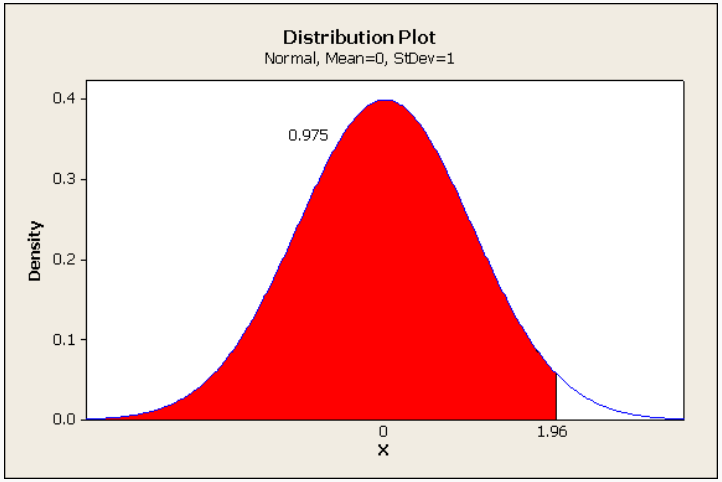
Ou seja, em uma distribuição Normal padrão, a probabilidade acumulada em $(-\infty$, -1.96$] \cup [$ 1.96, $\infty)$ é de 0.05.

Distribuições de probabilidade - O prefixo p

O prefixo 'p' seguido do nome de uma distribuição corresponde à probabilidade acumulada até determinado quantil.

```
> prob1 <- pnorm(q = quantil1, mean = 0, sd = 1)
> prob1
[1] 0.025
```

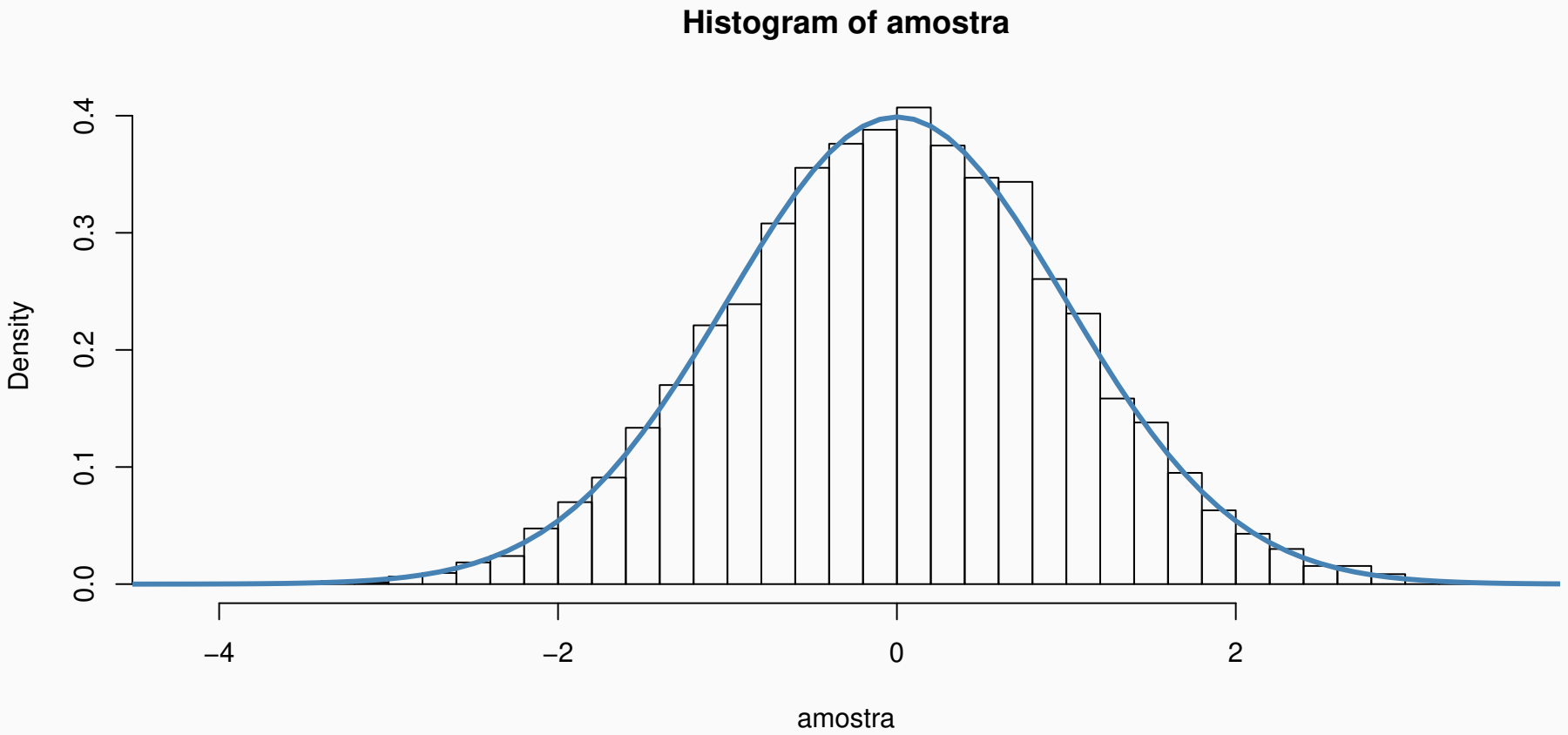
```
> prob2 <- pnorm(q = quantil2, mean = 0, sd = 1)
> prob2
[1] 0.975
```



Distribuições de probabilidade - O prefixo r

O prefixo 'r' seguido do nome de uma distribuição permite a geração de números aleátorios.

```
> amostra <- rnorm(n = 10000, mean = 0, sd = 1)
> hist(amostra, probability = T, breaks = 30)
> curve(dnorm, from = -5, to = 5, add = T,
+       col = "steelblue", lwd = 3)
```



Agora é a sua vez

Entendendo os prefixos ‘d’, ‘q’, ‘p’ e ‘r’

Gráficos de uma distribuição

Utilizando os prefixos ’d’, ’q’, ’p’ e ’r’ e se valendo dos nomes das distribuições vistas no slide Distribuições de probabilidade:

- Gere 1.000 observações de duas distribuições à sua escolha
- Faça o histograma e a curva da distribuição (ao mesmo tempo)
- Encontre o quantil 0.025 e o quantil 0.975 destas distribuições
- Qual a proporção de observações maiores e menores que estes quantis (2.5
- Qual a probabilidade de observarmos valores entre -1 e 1?
- Qual a proporção encontrada na nossa amostra?