Lista 4

Ítallo Silva - 118110718 | Thiago Nascimento - 118110804 | João Marcelo Junior - 117110448

Questão 1

a)

É uma ciência que trata da coleta, interpretação, análise e representação de dados númericos. A partir disso, a Estatística consegue transformar dados em conhecimento, sendo assim possível estimar ou prever fenônemos futuros. Por exemplo, a partir de uma pesquisa de uma parte da população eleitoral (amostra) é possível estimar, com um certo nível de confiança, qual candidato será vencedor em uma eleição.

b)

A estatística se divide basicamente em três grandes áreas, que são **Estatística Descritiva**, **Estatística Inferencial** e **Probabilístca**.

Estatística Descritiva: responsável por organizar e resumir os dados obtidos em uma pesquisa ou estudo, de modo a descreve-los de forma adequada. Por exemplo, apresentar em um gráfico de pizza a parcela de estudantes do curso de computação que fazem exercício físico semanalmente.

Estatística Inferencial: responsável por analisar os dados de uma amostra e fazer interpretações de forma que as informações geradas possam ser expandidas para a população. Por exemplo, a partir da análise dos dados obtidos de uma pesquisa eleitoral, qual a porcentagem de votos um candidato irá obter.

Probabilístca: a teoria de probabilidade permite a descrição de fenômenos aleatórios oriundos das incertezas.

Questão 2

a)

Se pretende então calcular o tamanho de uma amostra n e para isso temos que: O tamanho da populção é $N = 8311$ O erro amostral máximo é $\varepsilon = 0,02$ E com confiança $\gamma = 0,95 \rightarrow z_\gamma = 1,96$

Como não conhecemos o valor de p vamos utilizar $n \geq \frac{z^2}{4\varepsilon^2}$

$n \geq \frac{1,96^2}{4(0,02)^2} = 2401$

Portanto, o tamanho da amostra com as restrições requeridas é de 2401

b)

A amostragem estratificada seria mais adequada, pois visto dessa forma é possível manter a representatividade da população, visto que o conjunto populacional é divido nas classes de alunos, professores e servidores.

c)

Temos que $N = 150$, $\gamma = 0,99 \rightarrow z_\gamma = 2,57$ e $\overline{p} = 0,53$, então:

O erro amostral $\varepsilon = z\sqrt{\frac{\overline{p}(1-\overline{p})}{n}} = 2,57\sqrt{\frac{0,53\ .\ 0,47)}{150}} = 0,1047$

E o intervalo de confiança $IC(p; 99\%) = (0,4253; 0,6347)$

Portanto, isso significa que com 99% de confiança e erro amostral de 10,47% a verdadeira proporção de eleitores favoráveis ao candidato está entre 42,53% e 63,47%, ou seja, podemos dizer que 53% dos eleitores são favoráveis ao candidato, com uma margem de erro de 11% para mais ou para menos.

d)

A partir do enunciado temos que $N = 150$ e $\overline{p} = 0,63$

Considerando um erro amostral de 5% a probabilidade z_γ de que pelo menos 63% sejam favoráveis ao candidato é $z = \frac{0,05}{\sqrt{\frac{0,63\ .\ 0,37)}{150}}} = 95\%$

Portanto, considerando esse erro amostral é razoável o candidato fazer a afirmação.

Questão 3

a)

Sabendo que $X \sim N(348, 50^2)$ e os reservatórios tem 350000 m^3 por dia, em uma cidade com 1000 famílias temos que a cidade comporta 350 m^3 por familia/dia. Sendo assim, a probabilidade da capacidade ser ultrapassada é $P(X > 350) = 1 - P(X < 350) = 1 - P(Z < \frac{350-348}{50}) = 1 - P(Z < 0,04)$. Sendo $Z \sim N(0,1)$, então $1 - P(Z < 0,04) = 1 - 0,5160 = 0,484$. Logo, a chance de ultrapassar o limite é de 48,4%, sendo assim relativamente preocupante.

b)

Sabendo que $X \sim N(348, 50^2)$ e a capacidade c = 1000 familias * x m^3 por familia/dia. Com isso $P(X > x) = P(Z > \frac{x-348}{50}) = P(Z > z) = 1 - P(Z < z)$. Sendo $Z \sim N(0,1)$, então $P(Z < z) = 1 - 0,005 = 0,995$ e $z \simeq 2,58$, temos assim que $z = 2,58 \rightarrow \frac{x-348}{50} = 2,58 \rightarrow x = 477$. Sendo assim, a capacidade precisa ser c = 1000 * 477 = 477000 m^3 por dia.

Questão 4

a)

Sabendo que $\gamma = 0,95$ e $\overline{\mathrm{p}} = \frac{10}{30} = 0,333$, então temos que $\overline{\mathrm{p}} \pm \epsilon$, sendo $\epsilon = z_\gamma\sqrt{\frac{\overline{\mathrm{p}}(1-\overline{\mathrm{p}})}{n}} = 1,96\sqrt{\frac{0,333*0,666}{30}} \simeq 0,168$. Logo o IC = [0,333 - 0,168; 0,333 + 0,168] = [0,165; 0,501]. Sendo assim, a medida se mostrou válida, uma vez que, a proporção anteriormente de 60% caiu para entre 16,5% e 50,1%.

b)

Sendo,

```
ocorrencias_30 = c(7, 11, 8, 9, 10, 14, 6, 8, 8, 7, 8, 10, 10, 14, 12,
                   14, 12, 9, 11, 13, 13, 8, 6, 8, 13, 10, 14, 5, 14, 10)

num_dia = length(ocorrencias_30)

dias_violento = which(ocorrencias_30 >= 12)

num_dia_violento = length(dias_violento)

proporcao = num_dia_violento/num_dia
```

$\overline{\mathrm{p}} = 0{,}3333$ e $\gamma = 0,95$, então temos que $\overline{\mathrm{p}} \pm \epsilon$, sendo $\epsilon = z_\gamma \sqrt{\frac{\overline{\mathrm{p}}(1-\overline{\mathrm{p}})}{n}}$

```
erro = qnorm(0.975) * sqrt((proporcao * (1 - proporcao))/num_dia)

IC = c(proporcao - erro, proporcao + erro)
```

Logo temos que o intervalo de confiança, com 95% de confiança, está entre 16,4646% e 50,202%.

Questão 5

Nosso conjunto de dados, é o seguinte:

```
dados <- c(2.9, 3.4, 3.5, 4.1, 4.6, 4.7, 4.5, 3.8, 5.3, 4.9,
           4.8, 5.7, 5.8, 5.0, 3.4, 5.9, 6.3, 4.6, 5.5, 6.2)

tempo_medio <- mean(dados)
desvio_padrao_tempo <- sd(dados)
```

Temos uma média amostral de 4,745 e um desvio padrão amostral de 0,996.

a)

Queremos um intervalo de 95% de confiança. Calculemos então o erro $\epsilon = t_\alpha \frac{S}{\sqrt{n}}$. Temos que $S = 0{,}996$ e $n = 20$. Precisamos então descobrir o t_α.

Sabemos que $\alpha = 1 - \gamma \to \alpha = 1 - 0.95 = 0.05$ e que $gl = n - 1 \to gl = 20 - 1 = 19$. Assim, vamos olhar na tabela da distribuição t-Student. Assim $t_\alpha = 2,093$. Logo:

$\epsilon = 2,093 \frac{0{,}996}{\sqrt{20}} = 2,093 \cdot 0,2227 = 0,475$.

Sendo assim nosso intervalo é $4,745 \pm 0,475 = [4.27; 5.22]$. Sendo assim, podemos afirmar com 95% de certeza que a média da população está entre $[4.27; 5.22]$. Em outras palavras, em 95% das amostras retiradas o valor da média estará nesse intervalo.

b)

A seguir definimos uma função para o cálculo do intervalo de confiança para a média.

```
conf.int <- function(dados, desvio.p = NULL, level = 0.95) {

  n <- length(dados)
  q <- level + (1 - level)/2
  m <- mean(dados)

  if (is.null(desvio.p)) {

    dp <- sd(dados)
    fator.mult <- qt(q, df = n - 1)

  } else {

    dp <- desvio.p
    fator.mult <- qnorm(q)

  }

  erro <- fator.mult * dp / sqrt(n)

  list("limite.inferior" = m - erro, "limite.superior" = m + erro, "erro" = erro)

}
```

Temos então aplicando a função:

```
conf.int(dados)
```

```
## $limite.inferior
## [1] 4,279
##
## $limite.superior
## [1] 5,211
##
## $erro
## [1] 0,4662
```

Podemos ver que o valor calculado manualmente do calculado pela função foi bem próximo. Sendo a diferença entre os erros (calculado manualmente e pela função) igual à 0,0088.

Questão 6

```
library(tidyverse)
library(here)
```

O dataset

O conjunto de dados escolhido para este estudo foi encontrado no Kaggle e trata do preço de casas para alugar no Brasil. A seguir temos a leitura dos dados.

```
dataset <- read_csv(here("lista4/houses_to_rent_v2.csv"), col_types = cols_only(
  city = col_character(),
  area = col_double(),
  rooms = col_integer(),
  bathroom = col_integer(),
  floor = col_integer(),
  animal = col_character(),
  furniture = col_character(),
  `total (R$)` = col_double()
))
```

O conjunto tem um total de 10692 casas e possui 8 *features*, que estão descritas na tabela a seguir:

Nome	Tipo	Descrição
city	Caractere	Cidade de localização do imóvel
area	Real	Área do imóvel
rooms	Inteiro	Número de quartos do imóvel
bathroom	Inteiro	Número de banheiros do imóvel
floor	Inteiro	Número de pisos do imóvel
animal	Binária	Variável binária indicando se é permitido morar com animais no imóvel
furniture	Binária	Variável binária indicando se a casa é mobiliada ou não
total (R$)	Real	Valor do aluguel em reais

Neste estudo estaremos interessados em estimar dois parâmetros: **média do valor total de aluguel** e **proporção de casas disponíveis mobiliadas**. Iniciaremos realizando algumas manipulações nos dados para simplificar nossa análise no futuro.

Primeiramente, vamos mudar a *feature* forniture que atualmente é representada por caracteres para uma representação em 0 ou 1. Onde 0 indicará que a casa não é mobiliada e 1 que ela é.

```
dataset <- dataset %>% mutate(furniture = ifelse(furniture == "furnished", 1, 0))
```

Para facilitar a manipulação iremos mudar o nome da coluna que contém o valor do aluguel de **total (R$)** para simplesmente **aluguel**.

```
dataset <- dataset %>% rename(aluguel = "total (R$)")
```

Vamos então observar se existe algum dado faltante dentre as nossa variáveis de interesse, e se houver iremos remover as linhas em que elas ocorrem.

```
dataset <- dataset %>% filter(!is.na(furniture) & !is.na(aluguel))
```

Valor do aluguel

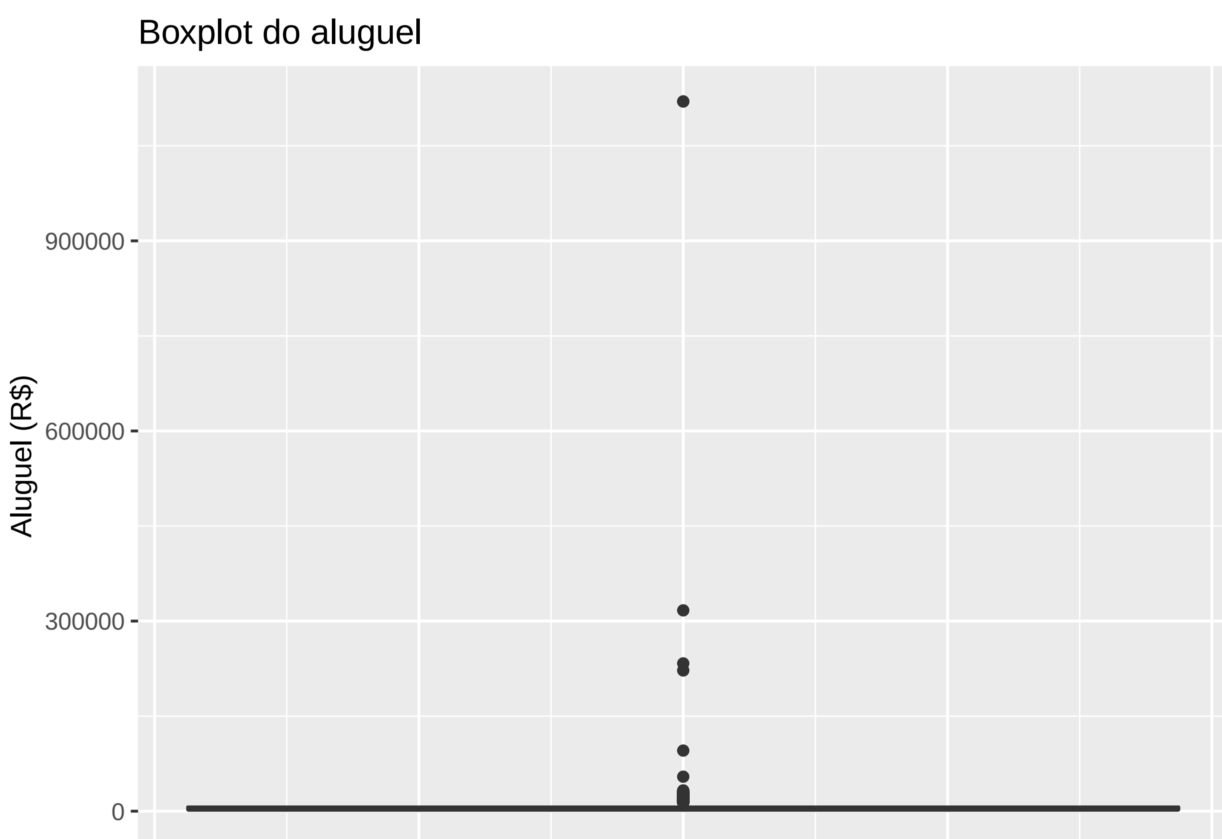
Inicialmente façamos uma análise descrita da variável. Comecemos observando sua distribuição:

```
dataset %>% pull(aluguel) %>% summary()
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     499    2062    3582    5490    6768 1120000
```

Podemos ver que os valores variam de 499 a 112000, com uma média de 5490 e mediana de 3582. Façamos um boxplot.

```
dataset %>% ggplot(aes(y = aluguel)) +
  geom_boxplot() +
  theme(axis.line.x = element_blank(),
        axis.ticks.x = element_blank(),
        axis.text.x = element_blank()) +
  labs(y = "Aluguel (R$)",
       title = "Boxplot do aluguel")
```



Pelos boxplot vemos que existem outliers em nossos dados, sendo assim iremos filtrar os dados com base no intervalo interquartil e remover esses dados. E em seguida repetimos o boxplot.

```
q1 <- quantile(dataset$aluguel, probs = 0.25)
q3 <- quantile(dataset$aluguel, probs = 0.75)
iiq <- q3 - q1

lim_inf <- q1 - 1.5 * iiq
lim_sup <- q3 + 1.5 * iiq
```

```
dataset <- dataset %>%
  filter(aluguel >= lim_inf & aluguel <= lim_sup)
```

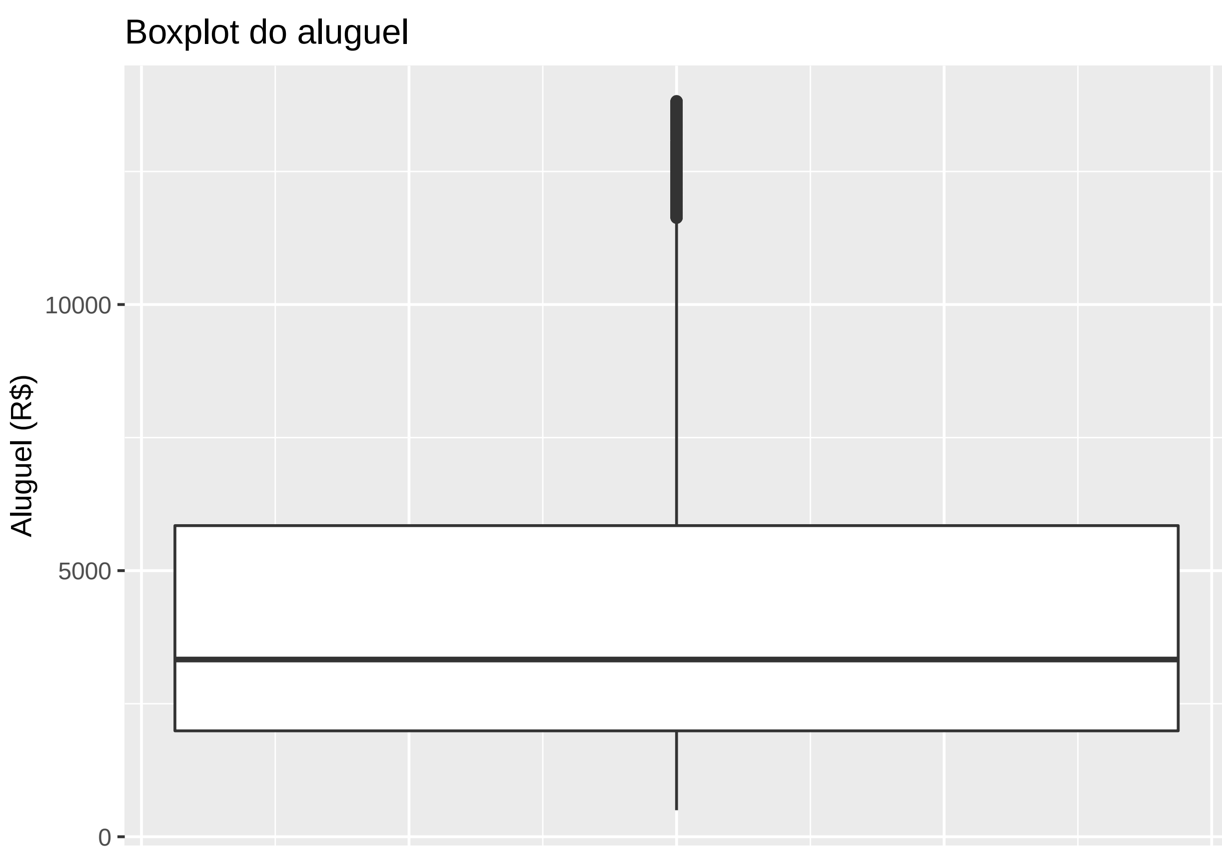
Vamos rever a distribuição dos dados.

```
dataset %>% pull(aluguel) %>% summary()
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     499    1990    3330    4340    5846   13820
```

Podemos ver que nosso valor máximo foi bastante reduzido. Vejamos mais uma vez o boxplot.

```
dataset %>% ggplot(aes(y = aluguel)) +
  geom_boxplot() +
  theme(axis.line.x = element_blank(),
        axis.ticks.x = element_blank(),
        axis.text.x = element_blank()) +
  labs(y = "Aluguel (R$)",
       title = "Boxplot do aluguel")
```

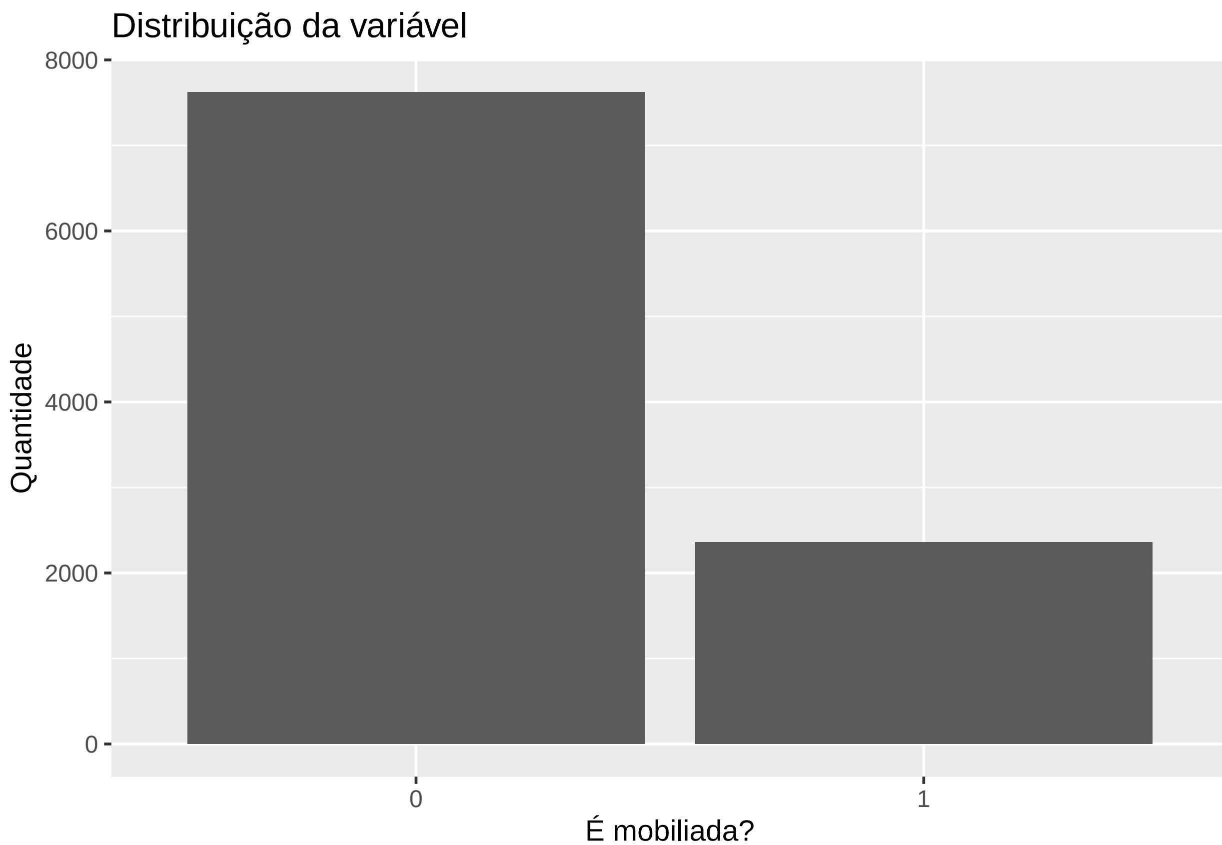


Vemos que ainda existem dados considerados outliers entretanto seus valores são significativas menores que os anteriores, e portanto eles serão considerados como válidos.

Variável binária - imóvel mobiliada

Como é uma variável binária, vejamos sua distribuição.

```
dataset %>% ggplot(aes(x = as.factor(furniture))) +
  geom_bar(stat = "count", position = "dodge") +
  labs(title = "Distribuição da variável",
       y = "Quantidade",
       x = "É mobiliada?")
```



Vemos uma predominância de imóveis não-mobiliados, calculemos então as proporções:

```
p_mobiliada <- nrow(dataset %>% filter(furniture == 1))/nrow(dataset)
p_nao_mobiliada <- 1 - p_mobiliada
```

Temos um total de 23,6531% de imóveis mobiliados e 76,3469% de imóveis não mobiliados.

Realizando uma amostragem no dataset

Para realizar nossa estimativa, iremos coletar uma amostra de 10% do dataset.

```
set.seed(4757)
tamanho_amostra <- round(nrow(dataset) * 0.1)
amostra <- dataset %>% slice_sample(n = tamanho_amostra)
```

Cálculo da estimativa intervalar da média do aluguel

Iniciemos com o cálculo da média da amostra.

```
media_amostra <- mean(amostra$aluguel)
```

Obtemos então o desvio padrão da população, considerando o dataset por completo.

```
desvio_padrao_pop <- sd(dataset$aluguel)
```

Calculemos então o erro, considerando o intervalo de 95% de confiança.

```
z_gamma <- qnorm(0.975) # Z_gamma para 95% de confiança
erro <- z_gamma * desvio_padrao_pop / sqrt(nrow(amostra))
```

Temos então um erro de R$ 190,8833, calculemos os limites do intervalo.

```
lim_inf <- media_amostra - erro
lim_sup <- media_amostra + erro
```

Assim nossa estimativa intervalar é [4251,8855;4633,652]. Calculemos a média da população para verificar se ela está contida no intervalo:

```
media_pop <- mean(dataset$aluguel)
```

Temos que a média da população é de R$ 4340,0648 e que ela está contida no intervalo. Isso era esperado uma vez que como o valor de confiança é de 95%, temos uma probabilidade 95% da média está contida nele. Isso torna a nossa estimativa uma boa estimativa para a média do aluguel.

Cálculo da estimativa da proporção da casas mobiliadas nos dados

Comecemos calculando a proporção na nossa amostra.

```
p_mob_amostra <- nrow(amostra %>% filter(furniture == 1))/nrow(amostra)
```

Sigamos então para o cálculo do erro.

```
z_gamma <- qnorm(0.975) # Z_gamma para 95% de confiança
erro <- z_gamma * sqrt((p_mob_amostra * (1 - p_mob_amostra))/nrow(amostra))
```

Definamos o intervalo de confiança:

```
lim_inf <- p_mob_amostra - erro
lim_sup <- p_mob_amostra + erro
```

Assim nosso intervalo é de [21,2784%;26,5694%]. Temos que a proporção na população é de 23,6531% e que ela está contida no intervalo. Pelo mesmo fator citado na análise da média.s