Lista 04 - MQ

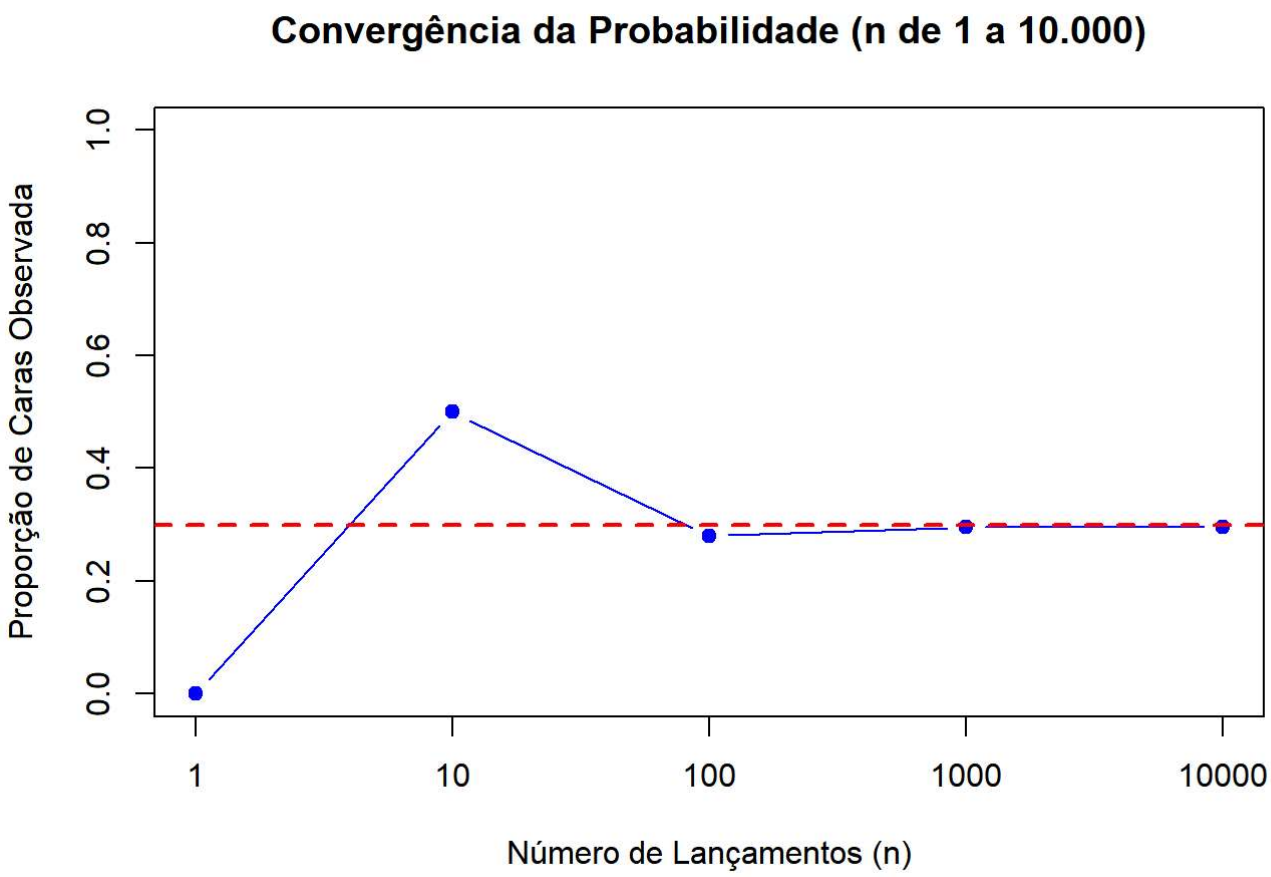
Martha Gaudencio

2026-02-01

```
#MQ101 - Lista 03
#Nome: Martha Gaudencio da Silva
#Data: 30/11/2025
#Descrição: Probabilidade e Inferência Estatística

# 1 Probabilidade como frequência de longo prazo (moeda viesada)

set.seed(123)

p_cara <- 0.3
n_vec  <- c(1, 10, 100, 1000, 10000)

result <- data.frame(
  n      = n_vec,
  p_cara = NA_real_
)

for (i in seq_along(n_vec)) {
  x <- rbinom(n = n_vec[i], size = 1, prob = p_cara)
  result$p_cara[i] <- mean(x)
}

print(result)
```

```
##       n p_cara
## 1     1 0.0000
## 2    10 0.5000
## 3   100 0.2800
## 4  1000 0.2950
## 5 10000 0.2956
```

```
plot(result$n, result$p_cara, type = "b", pch = 19, log = "x",
     ylim = c(0, 1), col = "blue",
     xlab = "Número de Lançamentos (n)",
     ylab = "Proporção de Caras Observada",
     main = "Convergência da Probabilidade (n de 1 a 10.000)")

abline(h = p_cara, lty = 2, col = "red", lwd = 2)
```




```
#O gráfico confirma a trajetória das probabilidades apresentadas de sair 'cara'
#a cada número de lançamentos, sendo que a partir de 100 lançamentos ela se
#aproxima de 29%, sendo a frequência relativa quando um evento acontece
#repetidamente um grande número de vezes.

# 2 - Bernoulli, Binomial e probabilidades exatas (satisfação em saúde)

set.seed(123)

p <- 0.65
n <- 20

y <- rbinom(n, size = 1, prob = p)

num_satisfeitos <- sum(y)

prob_12 <- dbinom(12, size = n, prob = p)

prob_12_ou_mais <- 1 - pbinom(11, size = n, prob = p)

print(paste("Número de satisfeitos na simulação:", num_satisfeitos))
```

```
## [1] "Número de satisfeitos na simulação: 12"
```

```
print(paste("Probabilidade exata de 12:", round(prob_12, 4)))
```

```
## [1] "Probabilidade exata de 12: 0.1614"
```

```
print(paste("Probabilidade de 12 ou mais:", round(prob_12_ou_mais, 4)))
```

```
## [1] "Probabilidade de 12 ou mais: 0.7624"
```

```
#Em uma amostra de 20 pessoas com a taxa de satisfação sendo de 65%, os dados
#mostram que há 76% de chance de 12 pessoas ou mais estarem satisfeitas,
#enquanto a porcentagem de serem exatamente 12 pessoas é de apenas 16%.

#3 - Probabilidade condicional e independência (base saúde)

library(dplyr)
```

```
##
## Anexando pacote: 'dplyr'
```

```
## Os seguintes objetos são mascarados por 'package:stats':
##
##     filter, lag
```

```
## Os seguintes objetos são mascarados por 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)

set.seed(123)
N <- 5000

dados_saude <- data.frame(
  sexo       = sample(c("F", "M"), size = N, replace = TRUE, prob = c(0.55, 0.45)),
  fumante    = rbinom(N, 1, 0.22),
  hipertenso = rbinom(N, 1, 0.30)
)

P_fumante   <- mean(dados_saude$fumante == 1)
P_fumante_F <- mean(dados_saude$fumante[dados_saude$sexo == "F"] == 1)
P_fumante_M <- mean(dados_saude$fumante[dados_saude$sexo == "M"] == 1)

cat("Probabilidade Geral de ser fumante:", round(P_fumante, 4), "\n")
```

```
## Probabilidade Geral de ser fumante: 0.2112
```

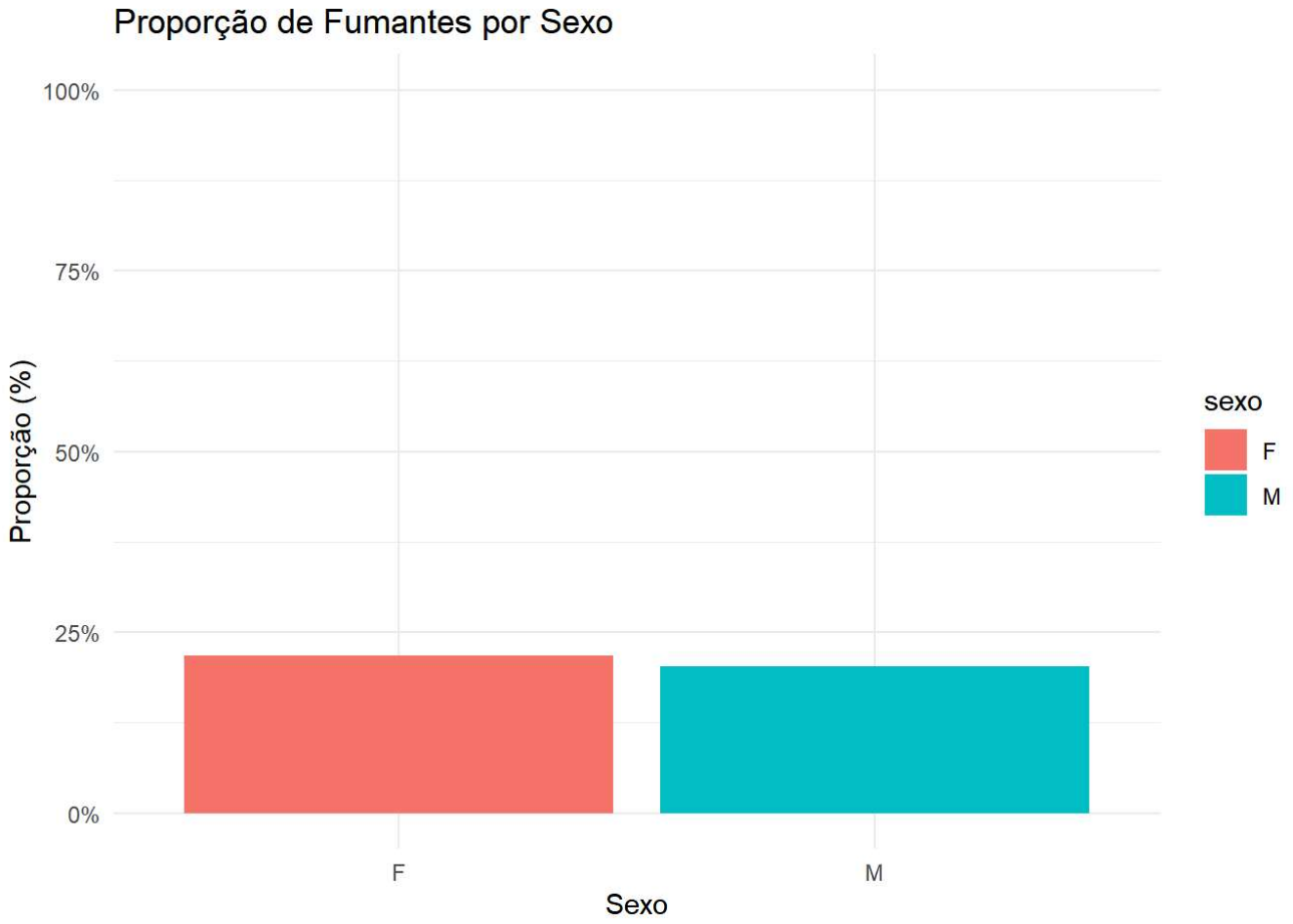
```
cat("Probabilidade de ser fumante dado que é Mulher:", round(P_fumante_F, 4), "\n")
```

```
## Probabilidade de ser fumante dado que é Mulher: 0.2179
```

```
cat("Probabilidade de ser fumante dado que é Homem:", round(P_fumante_M, 4), "\n")
```

```
## Probabilidade de ser fumante dado que é Homem: 0.2031
```

```
tab_fumo <- dados_saude |>
  group_by(sexo) |>
  summarise(prop_fumante = mean(fumante))

ggplot(tab_fumo, aes(x = sexo, y = prop_fumante, fill = sexo)) +
  geom_col() +
  scale_y_continuous(labels = scales::percent, limits = c(0, 1)) +
  labs(title = "Proporção de Fumantes por Sexo",
       x = "Sexo", y = "Proporção (%)") +
  theme_minimal()
```



```
#Os dados mostram que a probabilidade geral de ser fumante é de 21%, sendo que
#a de mulheres fumarem é ligeiramente maior que homens (0.2179 e 0.2031,
#respectivamente). É possível visualizar a proporção de fumantes por sexo no
#gráfico, onde a de mulheres é ligeiramente maior.

#4 – Probabilidade conjunta e regra do produto (saúde)

P_hipertenso <- mean(dados_saude$hipertenso == 1)
P_fumante    <- mean(dados_saude$fumante == 1)
P_hip_e_fum  <- mean(dados_saude$hipertenso == 1 & dados_saude$fumante == 1)

P_hip_dado_fum <- mean(dados_saude$hipertenso[dados_saude$fumante == 1] == 1)

P_produto <- P_hip_dado_fum * P_fumante

cat("P(Hipertenso):", round(P_hipertenso, 4), "\n")
```

```
## P(Hipertenso): 0.3014
```

```
cat("P(Fumante):", round(P_fumante, 4), "\n")
```

```
## P(Fumante): 0.2112
```

```
cat("P(Hipertenso E Fumante) [Empírico]:", round(P_hip_e_fum, 4), "\n")
```

```
## P(Hipertenso E Fumante) [Empírico]: 0.0658
```

```
cat("P(Hipertenso | Fumante):", round(P_hip_dado_fum, 4), "\n")
```

```
## P(Hipertenso | Fumante): 0.3116
```

```
cat("P(Hipertenso E Fumante) [Pela Regra do Produto]:", round(P_produto, 4), "\n")
```

```
## P(Hipertenso E Fumante) [Pela Regra do Produto]: 0.0658
```

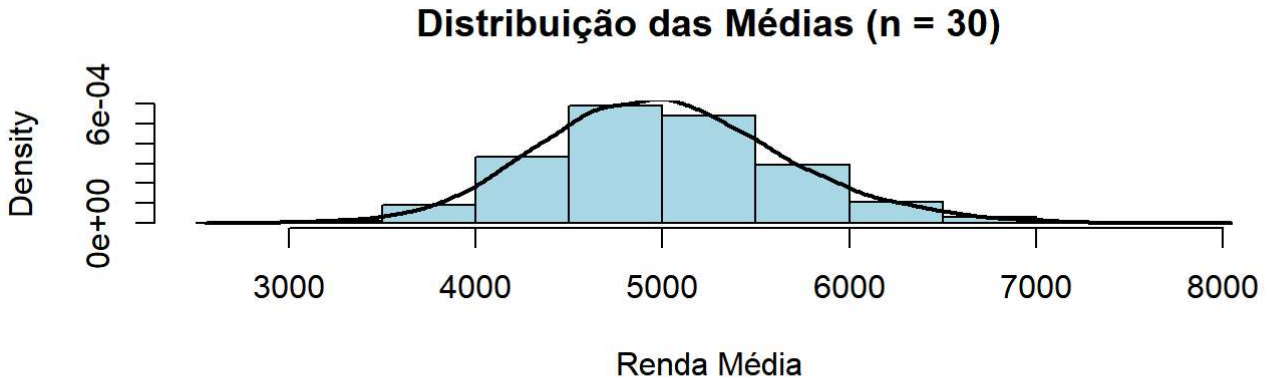
```
#Na base apresentada, a probabilidade de a pessoa ser hipertensa é de 30%, de
#ser fumante de 21%. Já a de ser fumante e hipertensa empiricamente é de 6,58%,
#exatamente o mesmo valor da regra do produto. Assim, considerando apenas estes
#dados, #não há uma evidência muito forte de que pessoas fumantes tenham mais
#chances de serem também hipertensas.

#5 - Bayes “de bolso” em triagem de benefícios

set.seed(123)

P_F         <- 0.02
P_T_dado_F  <- 0.9
P_T_dado_Fc <- 0.05
P_Fc        <- 1 - P_F

P_F_dado_T <- (P_T_dado_F * P_F) / (P_T_dado_F * P_F + P_T_dado_Fc * P_Fc)

N <- 100000
fraude <- rbinom(N, 1, P_F)
alerta <- ifelse(
  fraude == 1,
  rbinom(N, 1, P_T_dado_F),
  rbinom(N, 1, P_T_dado_Fc)
)

P_empirico <- mean(fraude[alerta == 1] == 1)

cat("Probabilidade Analítica (Bayes):", round(P_F_dado_T, 4), "\n")
```

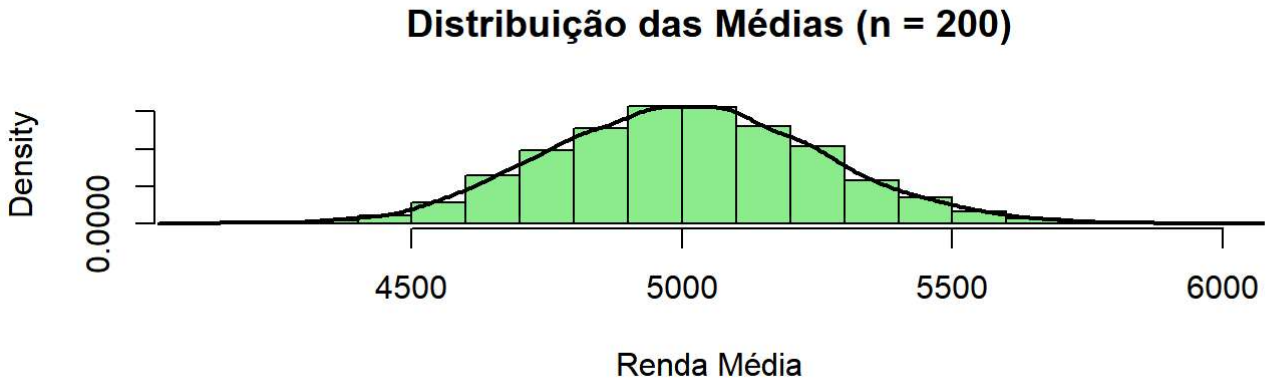
```
## Probabilidade Analítica (Bayes): 0.2687
```

```
cat("Probabilidade Empírica (Simulação):", round(P_empirico, 4), "\n")
```

```
## Probabilidade Empírica (Simulação): 0.2638
```

```r
#Pela fórmula de Bayes, a probabilidade de fraudes no programa considerada
#em uma amostragem com 100.000 beneficiários é de cerca de 26%, sendo então
#que a maioria da amostragem seria de 'alarmes falsos'.

#6 - Teorema Central do Limite com renda

set.seed(123)
N <- 100000
renda_pop <- rgamma(N, shape = 2, rate = 1/2500)

simular_medias <- function(n, n_rep = 5000) {
  medias <- numeric(n_rep)
  for (i in seq_len(n_rep)) {
        amostra <- sample(renda_pop, n, replace = TRUE)
    medias[i] <- mean(amostra)
  }
  medias
}

medias_n30  <- simular_medias(30)
medias_n200 <- simular_medias(200)

par(mfrow = c(2, 1))

hist(medias_n30,  main = "Distribuição das Médias (n = 30)",
     col = "lightblue", xlab = "Renda Média", prob = TRUE)
lines(density(medias_n30), lwd = 2)

hist(medias_n200, main = "Distribuição das Médias (n = 200)",
     col = "lightgreen", xlab = "Renda Média", prob = TRUE)
lines(density(medias_n200), lwd = 2)
```


Distribuição das Médias (n = 30)
Density
0e+00
6e-04
3000
4000
5000
6000
7000
8000
Renda Média


Distribuição das Médias (n = 200)
Density
0.0000
4500
5000
5500
6000
Renda Média

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

#O gráfico com n=30 tem a distribuição das médias em formato de sino (Normal),
#mas apresenta a base mais larga, o que indica uma maior variabilidade nas
#estimativas. Já o com n=200, a curva torna-se #é mais estreita e mais alta,
#com as médias amostrais mais concentradas em torno da média real da população.
#Isso significa que o erro amostral diminuiu e confirma o teorema central do
#limite, pois quanto maior o tamanho da amostra, menor é o desvio padrão da
#média amostral.


#7 - Intervalo de confiança para proporção (saúde)

set.seed(123)

n <- 400
amostra_saude <- dados_saude[sample(1:nrow(dados_saude), n), ]

p_hat <- mean(amostra_saude$hipertenso)
SE_p  <- sqrt(p_hat * (1 - p_hat) / n)

IC_95 <- c(
  inferior = p_hat - 1.96 * SE_p,
  superior = p_hat + 1.96 * SE_p
)

p_verdadeiro <- mean(dados_saude$hipertenso)

cat("Proporção na Amostra (p_hat):", round(p_hat, 4), "\n")
```

```
## Proporção na Amostra (p_hat): 0.285
```

```
cat("Erro Padrão (SE):", round(SE_p, 4), "\n")
```

```
## Erro Padrão (SE): 0.0226
```

```
cat("Intervalo de Confiança (95%): [", round(IC_95[1], 4), ",", round(IC_95[2], 4), "]\n")
```

```
## Intervalo de Confiança (95%): [ 0.2408 , 0.3292 ]
```

```
cat("Proporção Verdadeira na População:", round(p_verdadeiro, 4), "\n")
```

```
## Proporção Verdadeira na População: 0.3014
```

```
#A proporção de hipertensos na amostra é de 28%, valor próximo ao da população
#verdadeira da população, o que reflete os números trazidos pelo intervalo de
#de confiança que varia entre 24% e 32%. Para o campo das políticas públicas,
#conhecer o IC de uma amostra é fundamental para saber se os dados estão próximos
#da realidade antes de considerá-lo para a elaboração, a avaliação e o
#monitoramento de políticas.

#8 – Correlação, regressão simples e inferência (educação)

set.seed(123)
N <- 2000

dados_educacao <- data.frame(
  ideb        = rnorm(N, mean = 5.5, sd = 0.7),
  gasto_aluno = rnorm(N, mean = 6000, sd = 1500)
)

dados_educacao$ideb <- dados_educacao$ideb + 0.0002 * (dados_educacao$gasto_aluno - 6000)

cor_ideb_gasto <- cor(dados_educacao$ideb, dados_educacao$gasto_aluno)

modelo <- lm(ideb ~ gasto_aluno, data = dados_educacao)
resumo <- summary(modelo)
ic_modelo <- confint(modelo)

cat("Correlação de Pearson:", round(cor_ideb_gasto, 4), "\n")
```

```
## Correlação de Pearson: 0.3786
```

```
print(resumo$coefficients)
```

```
##                 Estimate   Std. Error  t value     Pr(>|t|)
## (Intercept) 4.3573583460 6.528995e-02 66.73858 0.000000e+00
## gasto_aluno 0.0001938357 1.060265e-05 18.28181 3.668237e-69
```

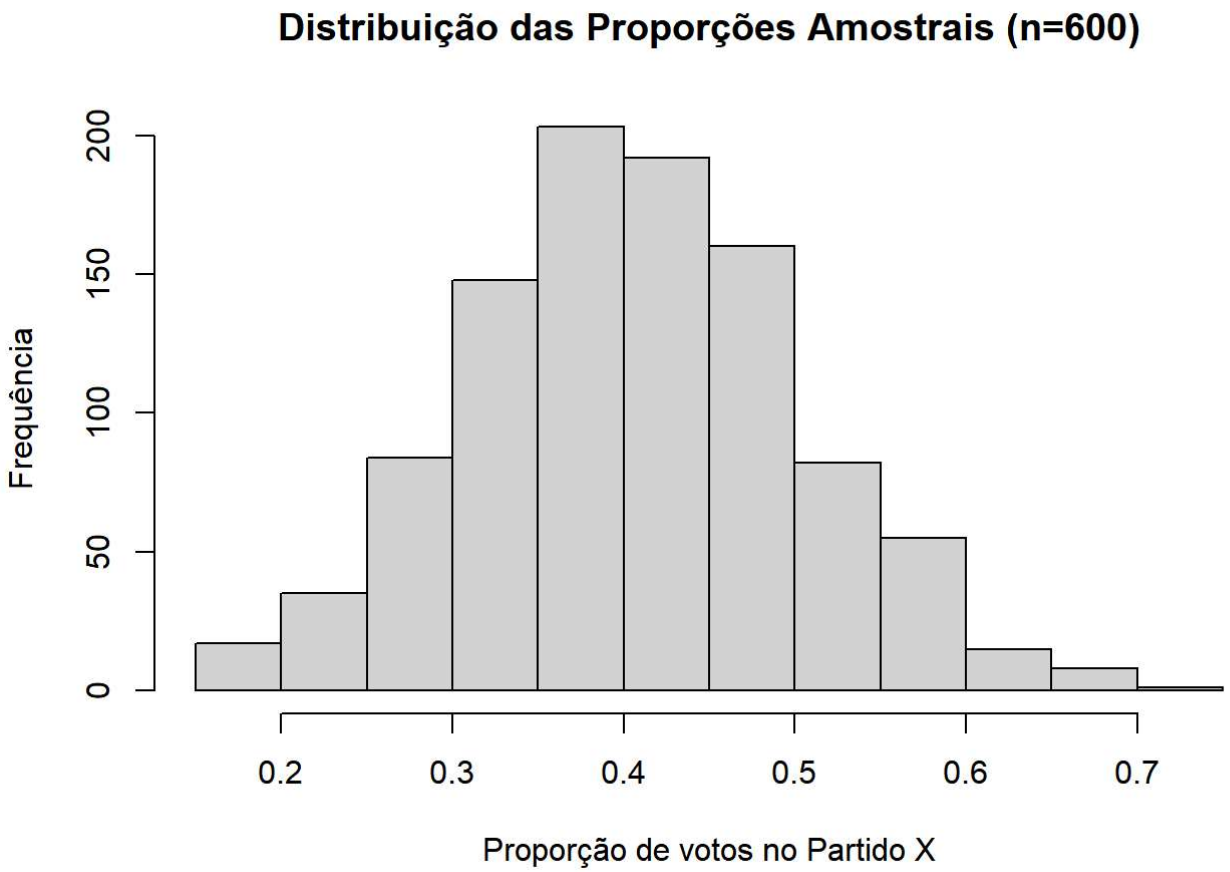
```
cat("\nIntervalo de Confiança para o Coeficiente:\n")
```

```
##
## Intervalo de Confiança para o Coeficiente:
```

```
print(ic_modelo)
```

```
##                    2.5 %       97.5 %
## (Intercept) 4.2293148316 4.4854018603
## gasto_aluno 0.0001730422 0.0002146291
```

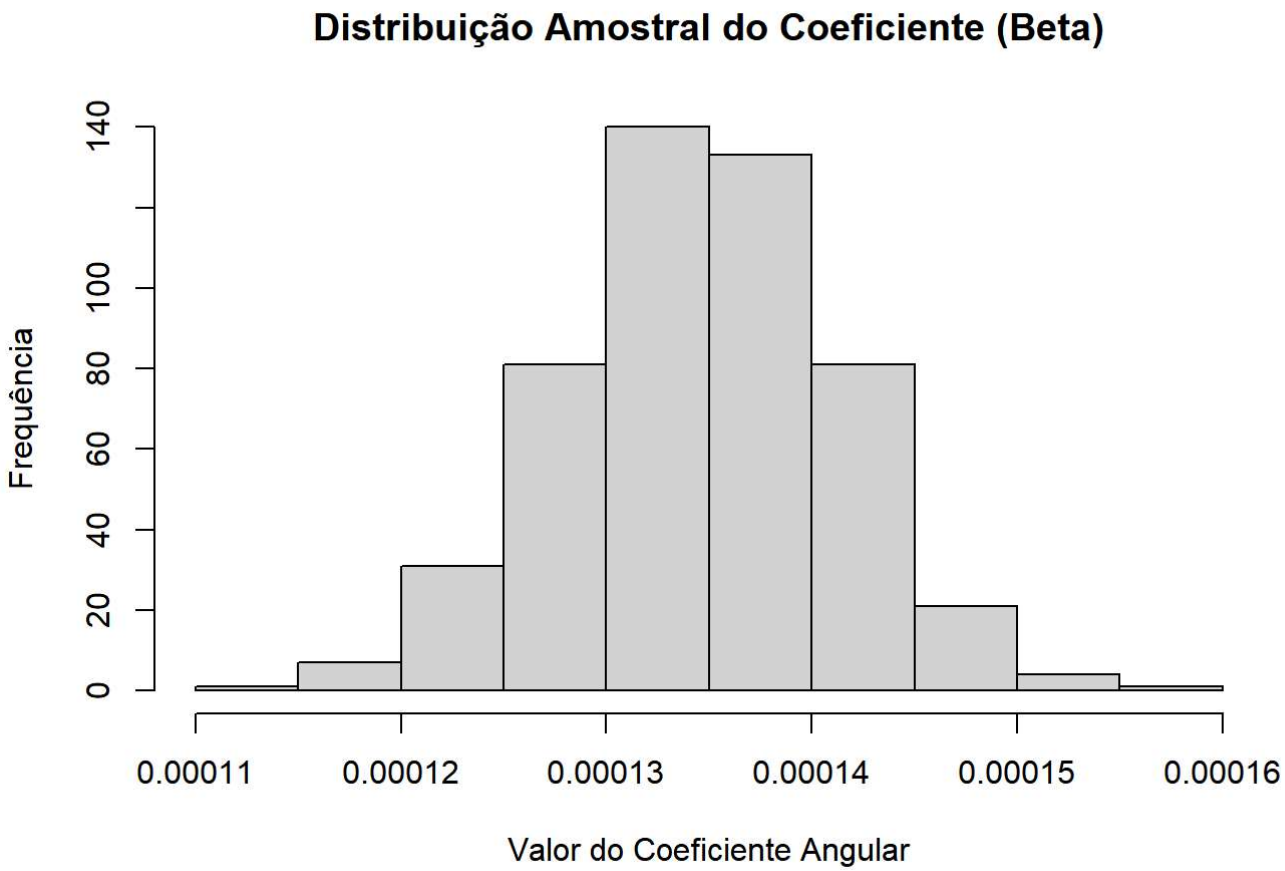
```
#O coeficiente angular é positivo, sendo + 0.0002, enquanto o valor p é baixo,
#mostrando uma relação significante. #Ademais, o intervalo de confiança não
#inclui o número zero. Esses três valores indicam assim que há uma correlação
#forte entre gasto por aluno e indicador de qualidade do ensino.

# 9 - Margem de erro e tamanho amostral (TSE)

set.seed(123)
N <- 5000
dados_tse <- data.frame(
  id_mun       = 1:N,
  prop_partido = rbeta(N, shape1 = 10, shape2 = 15)
)

n_amostra <- 600
n_rep     <- 1000
p_hat_vec <- numeric(n_rep)

for (r in 1:n_rep) {
  id_escolhido <- sample(dados_tse$id_mun, 1)
  p_mun <- dados_tse$prop_partido[dados_tse$id_mun == id_escolhido]

  votos <- rbinom(n_amostra, size = 1, prob = p_mun)
  p_hat_vec[r] <- mean(votos)
}

ME <- 1.96 * sqrt(0.25 / n_amostra)

hist(p_hat_vec,
     main = "Distribuição das Proporções Amostrais (n=600)",
     xlab = "Proporção de votos no Partido X",
     ylab = "Frequência")
```




```
cat("Margem de Erro Calculada (ME):", round(ME, 4), "\n")
```

```
## Margem de Erro Calculada (ME): 0.04
```

```
#A margem de erro apresentada está próxima a 0,04, em uma amostragem com 600
#eleitores repetida 1000 vezes. O histograma mostra que há uma frequência maior
#da proporção de votos no partido em torno de 40%, mostrando uma margem de erro
#segura, o que deve ser levado em consideração em análises eleitorais, como ao
#analisar a possibilidade de empates técnicos.

#10 - Regressão e distribuição amostral do coeficiente (TSE)

set.seed(123)
N <- 5000
dados_tse <- data.frame(
  id_mun      = 1:N,
  renda_media = rnorm(N, mean = 2500, sd = 600)
)
dados_tse$prop_partido <- plogis(
  -1 + 0.0006 * dados_tse$renda_media + rnorm(N, 0, 0.3)
)

n_mun_amostra <- 300
n_eleitores   <- 400
n_rep         <- 500
coef_angular  <- numeric(n_rep)

for (r in 1:n_rep) {
  mun_sorteados <- sample(dados_tse$id_mun, n_mun_amostra)
  base_pesq <- dados_tse[dados_tse$id_mun %in% mun_sorteados, ]

  p_hat <- numeric(n_mun_amostra)
  for (i in seq_len(n_mun_amostra)) {
    p_true   <- base_pesq$prop_partido[i]
    votos    <- rbinom(n_eleitores, size = 1, prob = p_true)
    p_hat[i] <- mean(votos)
  }

  base_pesq$p_hat <- p_hat

  modelo <- lm(p_hat ~ renda_media, data = base_pesq)
  coef_angular[r] <- coef(modelo)[2]
}

hist(coef_angular,
     main = "Distribuição Amostral do Coeficiente (Beta)",
     xlab = "Valor do Coeficiente Angular",
     ylab = "Frequência")
```




```
#A distribuição representa a 'distribuição amostral do estimador', mostrando
#como o coeficiente varia entre diferentes amostras. A média da distribuição
#indica o efeito real, no qual a variação mostrada no histograma) corresponde
#ao erro-padrão da estimativa. Essa é a base do intervalo de confiança, que
#delimita a faixa onde o coeficiente estaria em 95% das repetições possíveis.
```