Oficina de Estatística Tutorial 4

Virgilio Mendes

19/07/2019

Tutorial 4

Neste tutorial 4, vamos iniciar nossos trabalhos com a ideia de inferência estatística. Vamos então estimar parâmetros (proporções ou médias), calcular suas margens de erro e intervalos de confiança a partir de diferentes níveis de confiança.

1. Estimando proporções

Como vimos, podemos estimar a proporção de algum atributo de um elemento (variável) da população.

Em nosso tutorial, vamos estimar a aprovação do prefeito de Belo Horizonte a partir de uma simulação de dados. Para tanto, vamos simular 100 amostras de 1.200 cidadãos de Belo Horizonte acerca da aprovação do trabalho do prefeito Alexandre Kalil. Suponha uma pergunta como "Você aprova ou desaprova o desempenho de Alexandre Kalil como prefeito de Belo Horizonte?", em que todos os entrevistados responderam uma ou outra opção (isto é, não houve respostas "não sei" ou recusa de resposta). Além disso, nossa suposição é a de que 70% dos entrevistados aprovarão o trabalho do prefeito. Dessas 200 amostras, vamos selecionar uma com a qual vamos trabalhar.

Como discutimos, distribuição de uma proporção é conhecida como distribuição binomial. Portanto, para simularmos a resposta a essa pergunta, vamos gerar as 100 amostras de 1.200 cidadãos a partir de uma distribuição randomizada de uma distribuição binomial que assume dois valores (0 para a resposta "desaprova" e 1 para a resposta "aprova"), sendo que a aprovação assume uma probabilidade de 0.7 com a função `rbinom`. Todas essas 100 amostras serão guardadas numa lista (chamada aqui de `amostras_pesquisas`), das quais sortearemos 1 de forma aleatória (com a já conhecida função `sample`).

```
set.seed(1234)

amostras_pesquisas <- list()

n_amostra <- 1200

for (i in 1:100) {
  banco_pesquisa <- rbinom(n_amostra, 1, prob = 0.7)
  amostras_pesquisas[[i]] <- banco_pesquisa
}

amostra_sorteada <- sample(amostras_pesquisas, 1)[[1]]
```

Qual a proporção de indivíduos na amostra sorteada que aprovam o trabalho do prefeito Alexandre Kalil?

```
proporcao <- prop.table(table(amostra_sorteada))[[2]]
```

E qual é o erro-padrão desta proporção?

```
erro_padrao <- sqrt((proporcao*(1-proporcao))/n_amostra)
```

Vamos calcular agora o tamanho da margem de erro com um nível de confiança de 95%:

```
margem_erro <- 1.96*erro_padrao
```

Com a nossa estimativa da proporção, podemos calcular o intervalo de confiança dessa estimativa:

```
limite_superior_ic <- proporcao + margem_erro
limite_inferior_ic <- proporcao - margem_erro
```

Sendo assim, qual é o intervalo de confiança da estimativa da proporção de cidadãos belo-horizontinos que aprovam o trabalho do prefeito Alexandre Kalil na nossa amostra sorteada? Considerando que eu utilizo uma versão 3.6 do R (que afeta o set.seed em relação a versões anteriores), esta estimativa está entre 0.683 e 0.734, com um nível de confiança de 95%.

- Isto é, com um nível de confiança de 95%, entre 68,3% e 73,4% dos eleitores aprovam o trabalho do prefeito.

Para versões anteriores do R, com um nível de confiança de 95%, entre 67,7% e 72,8% dos eleitores aprovam o trabalho do prefeito.

Vamos nos lembrar, porém, que podemos alterar o nível de confiança. Sendo assim, vamos estimar os intervalos de confiança para dois níveis bastante comuns nas pesquisas científicas, 90% e 99%:

```
margem_erro_90 <- 1.645*erro_padrao
margem_erro_99 <- 2.576*erro_padrao

limite_inferior_ic_90 <- proporcao - margem_erro_90
limite_superior_ic_90 <- proporcao + margem_erro_90

limite_inferior_ic_99 <- proporcao - margem_erro_99
limite_superior_ic_99 <- proporcao + margem_erro_99
```

Quais os intervalos de confiança estimados para estes dois outros níveis de confiança? (v. 3.6 do R)

- 90%: entre 68,7% e 73,0% dos eleitores aprovam o trabalho do prefeito;
- 99%: entre 67,5% e 74,2% dos eleitores aprovam o trabalho do prefeito.

No caso de versões anteriores do R, os intervalos de confiança estimados são:

- 90%: entre 68,1% e 72,4% dos eleitores aprovam o trabalho do prefeito;
- 99%: entre 66,9% e 73,6% dos eleitores aprovam o trabalho do prefeito.

Portanto, como havíamos discutido, há um *trade-off* entre precisão e "confiança" na escolha do intervalo de confiança.

Uma maneira de também representar este tipo de *trade-off* está no tamanho da amostra. O que aconteceria se, ao invés de lidarmos com amostras de n = 1200, só conseguíssemos entrevistar 800 indivíduos? Para tanto, vamos sortear 100 novas amostras, agora com n = 800, e sortearmos uma para trabalharmos.

```
set.seed(1234)

amostras_pesquisas <- list()

n_amostra <- 800

for (i in 1:100) {
  banco_pesquisa <- rbinom(n_amostra, 1, prob = 0.7)
  amostras_pesquisas[[i]] <- banco_pesquisa
}

amostra_sorteada <- sample(amostras_pesquisas, 1)[[1]]
```

Em seguida, vamos calcular a proporção, seu erro-padrão, margens de erro e intervalos de confiança para os três níveis de confiança usados anteriormente (90%, 95% e 99%):

```
proporcao <- prop.table(table(amostra_sorteada))[[2]]
erro_padrao <- sqrt((proporcao*(1-proporcao))/n_amostra)

margem_erro_90 <- 1.645*erro_padrao
margem_erro_95 <- 1.96*erro_padrao
margem_erro_99 <- 2.576*erro_padrao

limite_inferior_ic_90 <- proporcao - margem_erro_90
limite_superior_ic_90 <- proporcao + margem_erro_90

limite_inferior_ic_95 <- proporcao - margem_erro_95
limite_superior_ic_95 <- proporcao + margem_erro_95

limite_inferior_ic_99 <- proporcao - margem_erro_99
limite_superior_ic_99 <- proporcao + margem_erro_99
```

Agora, as novas estimativas para os intervalos de confiança são (v. 3.6 do R):

- 90%: entre 67,2% e 72,5% dos eleitores aprovam o trabalho do prefeito;
- 95%: entre 66,7% e 73,0% dos eleitores aprovam o trabalho do prefeito;
- 99%: entre 65,7% e 74,0% dos eleitores aprovam o trabalho do prefeito.

Em versões anteriores (utilizando o mesmo set.seed), essas estimativas são:

- 90%: entre 66,4% e 71,8% dos eleitores aprovam o trabalho do prefeito;
- 95%: entre 65,9% e 72,3% dos eleitores aprovam o trabalho do prefeito;
- 99%: entre 64,9% e 73,3% dos eleitores aprovam o trabalho do prefeito.

Com a redução no tamanho da amostra, os novos intervalos de confiança são maiores! Isto ocorre porque o erro-padrão agora é maior (lembre-se que o tamanho da amostra influencia no cálculo do erro-padrão). Ou seja, o tamanho de uma amostra também importa para a precisão das nossas estimativas.

2. Estimando médias

Para estimarmos uma média, vamos simular um banco de dados com a votação percentual de 1300 candidatos a vereador na cidade de Belo Horizonte, da qual retiraremos uma amostra aleatória de 150 candidatos. Para tanto, vamos supor que estes votos se distribuem de forma normal (sim, sabemos que não são assim, mas isso não importa para nossos fins, importa?), com média de 600 votos e desvio-padrão de 40 dos votos, com a função `rnorm`.

```
set.seed(1234)

banco_candidatos <- rnorm(1300, mean = 600, sd = 40)

n_amostra <- 150

amostra_sorteada <- sample(banco_candidatos, n_amostra, replace = F)
```

Qual a média estimada de votos na amostra de 150 candidatos que nós sorteamos? E o erro-padrão desta média?

```
media_amostra <- mean(amostra_sorteada)
erro_padrao <- sd(amostra_sorteada)/sqrt(n_amostra)
```

Calcule agora as margens de erro dessa média da amostra para os níveis de confiança de 90%, 95% e 99%.

```
margem_erro_90 <- 1.645*erro_padrao
margem_erro_95 <- 1.96*erro_padrao
margem_erro_99 <- 2.576*erro_padrao

limite_inferior_ic_90 <- media_amostra - margem_erro_90
limite_superior_ic_90 <- media_amostra + margem_erro_90

limite_inferior_ic_95 <- media_amostra - margem_erro_95
limite_superior_ic_95 <- media_amostra + margem_erro_95

limite_inferior_ic_99 <- media_amostra - margem_erro_99
limite_superior_ic_99 <- media_amostra + margem_erro_99
```

Os intervalos de confiança dessa estimativa da média de votos dessa amostra de n = 150 dos candidatos a vereador em Belo Horizonte nesta eleição simulada são (v. 3.6 do R):

- Com nível de confiança de 90%: entre 596 e 607 votos;
- Com nível de confiança de 95%: entre 595 e 608 votos;
- Com nível de confiança de 99%: entre 593 e 610 votos.

Em versões anteriores do R, os intervalos de confiança (com os valores dos votos arrendondados com 0 casas decimais) estimados são:

- Com nível de confiança de 90%: entre 595 e 605 votos;
- Com nível de confiança de 95%: entre 595 e 605 votos (com o arrendondamento os ICs ficam equivalentes);
- Com nível de confiança de 99%: entre 592 e 608 votos.