Laboratório 6: Inferência para Dados Categoriais

Em agosto de 2012, agências de notícias como Washington Post e o Huffington Post publicaram reportagens sobre o aumento do ateísmo na América do Norte. A fonte da reportagem foi uma pesquisa que perguntou às pessoas, "Independente de você frequentar algum culto religioso ou não, você diria que você é uma pessoa religiosa, não é uma pessoa religiosa ou é um ateu convicto?" Esse tipo de pergunta, que pede para as pessoas se classificarem de uma forma ou outra, é comum em pesquisas de opinião e gera dados categoriais. Neste laboratório vamos explorar a pesquisa sobre ateísmo e investigar o que está em jogo quando fazemos inferências sobre proporções populacionais utilizando dados categoriais.

A Pesquisa de Opinião

Para acessar o comunicado à imprensa da pesquisa de opinião, realizada pela WIN-Gallup International, clique no link abaixo:

http://www.wingia.com/web/files/richeditor/filemanager/Global_INDEX_of_Religiosity_and_Atheism_PR__6.pdf

Revise com cuidado as informações do relatório e então tente resolver as seguintes questões:

Exercício 1 No primeiro parágrafo, vários resultados importantes são relatados. Essas porcentagens parecem ser *estatísticas amostrais* (derivadas dos dados da amostra) ou *parâmetros populacionais*?

Exercício 2 O título do relatório é "Índice Global de Religiosidade e Ateísmo" ("Global Index of Religiosity and Atheism"). Para generalizar os resultados do relatório para a população humana global, o que devemos assumir a respeito do método amostral? Parece ser uma suposição razoável?

Os Dados

Preste atenção na Tabela 6 (páginas 15 e 16), que informa o tamanho da amostra e o percentual de respostas de todos os 57 países que fizeram parte da pesquisa. Mesmo sendo um formato útil para resumir os dados, basearemos nossas análises no conjunto de dados original das respostas individuais à pesquisa. Carregue esse conjunto de dados no R utilizando o seguinte comando.

```
download.file("http://www.openintro.org/stat/data/atheism.RData", destfile = "atheism.RData")
load("atheism.RData")
```

Exercício 3 A que corresponde cada linha da Tabela 6? A que corresponde cada linha do banco de dados `atheism` (ateísmo)?

Para investigar o elo entre essas duas maneiras de organizar esses dados, dê uma olhada na proporção estimada de ateus nos Estados Unidos. Perto do fim da Tabela 6, verificamos que é 5%. Devemos ser capazes de chegar ao mesmo número usando o banco de dados `atheism`.

Exercício 4 Utilizando o comando abaixo, crie um novo banco de dados denominado `us12` que contém apenas as linhas do banco de dados `atheism` associadas aos respondentes da pesquisa

realizada em 2012 nos Estados Unidos. Em seguida, calcule a proporção de respostas dos que se afirmam ateus. Ela é semelhante à porcentagem da Tabela 6? Se não, por quê?

```
us12 <- subset(atheism, atheism$nationality == "United States" & atheism$year == "2012")
```

Inferência de Proporções

Como foi sugerido pelo Exercício 1, a Tabela 6 apresenta *estatísticas*, ou seja, cálculos feitos a partir da amostra de 51.927 pessoas. O que nós gostaríamos, porém, é obter informações sobre os *parâmetros* populacionais. Você pode responder à pergunta "Qual a proporção de pessoas na amostra que informaram serem ateus?" com uma estatística; por outro lado, uma questão como "Qual a proporção de pessoas na Terra que informariam serem ateus?" é respondida com uma estimativa do parâmetro.

As ferramentas inferenciais para estimar proporções populacional são análogas àquelas utilizadas para as médias no último laboratório: o intervalo de confiança e o teste de hipótese.

> **Exercício 5** Descreva as condições para inferência necessárias para construir um intervalo de confiança de 95% para a proporção de ateus nos Estados Unido em 2012. Você está confiante de que todas as condições são atendidas?

Se as condições para inferência são razoáveis, podemos calcular o erro padrão e construir o intervalo de confiança manualmente, ou deixar que a função `inference` faça isso por nós.

```
inference(y = us12$response, est = "proportion", type = "ci", method = "theoretical",
          success = "atheist")
```

Perceba que, uma vez que o objetivo é construir uma estimativa intervalar para uma proporção, é necessário especificar o que constitui um "sucesso", que nesse caso é a resposta `atheist` (ateu).

Apesar de intervalos de confiança formais e testes de hipótese não aparecerem no relatório, sugestões de inferência aparecem no final da página 7: "Em geral, a margem de erro para pesquisas de opinião deste tipo é de $\pm 3 - 5\%$ com 95% de confiança."

> **Exercício 6** Com base nos resultados do R, qual é a margem de erro para a estimativa da proporção de ateus nos EUA em 2012?

> **Exercício 7** Utilizando a função `inference`, calcule os intervalos de confiança para a proporção de ateus em 2012 para dois outros países de sua escolha, e informe as margens de erro associadas a eles. Certifique-se de observar se as condições para inferência são atendidas. Pode ser útil primeiro criar novos conjuntos de dados para cada um dos dois países, e então usar essas conjuntos de dados junto com a função `inference` para construir os intervalos de confiança.

Como a Proporção Afeta a Margem de Erro?

Imagine que você fez um levantamento com 1000 pessoas a respeito de duas questões: você é mulher? E você é canhoto? Uma vez que ambas as proporções amostrais foram calculadas a partir de um mesmo tamanho de amostra, elas devem ter a mesma margem de erro, certo? Errado! Apesar da margem de erro mudar em relação ao tamanho da amostra, ela também é afetada pela proporção.

Lembre-se da fórmula para calcular o erro padrão: $EP = \sqrt{p(1-p)/n}$. O resultado é utilizado na fórmula para calcular a margem de erro para um intervalo de confiança de 95%: $ME = 1.96 \times EP =$

$1.96 \times \sqrt{p(1-p)/n}$. Já que a proporção populacional p se encontra na fórmula para calcular o ME, faz sentido que a margem de erro depende, de alguma forma, da proporção populacional. Podemos visualizar essa relação criando um gráfico relacionando ME com p.

O primeiro passo é criar um vetor `p`, que é uma sequência de 0 a 1 com cada número separado por 0,01. Podemos então criar um vetor para a margem de erro (`me`), associando com cada um dos palores de `p` utilizando a fórmula aproximada já conhecida ($ME = 2 \times SE$). Por fim, fazemos um gráfico com os dois vetores para revelar a relação entre eles.

```
n <- 1000
p <- seq(0, 1, 0.01)
me <- 2*sqrt(p*(1 - p)/n)
plot(me ~ p)
```

Exercício 8 Descreva a relação entre `p` e `me`.

Condição de Sucesso ou Fracasso

O livro enfatiza que você deve sempre verificar as condições antes de fazer qualquer inferência. Para inferência de proporções, a proporção amostral pode ser considerada como se distribuindo de maneira aproximadamente normal se for baseada numa amostra aleatória de observações independentes e se $np \geq 10$ e $n(1-p) \geq 10$. Essa regra geral é fácil o suficiente de ser seguida, mas deixa aberta a questão: o que há de tão especial com o número 10? A resposta mais curta é: nada. Você pode argumentar que estaríamos bem com 9 ou que deveríamos utilizar 11. O "melhor" valor para essa regra geral é, pelo menos em alguma medida, arbitrário.

Podemos investigar as relações entre n e p e a forma da distribuição amostral utilizando simulações. Para começar, simulamos o processo de retirar 5000 amostra de 1040 elementos de uma população com a verdadeira proporção de ateus igual a 0.1. Para cada uma das 5000 amostras, calculamos $\hat{p}$ e então criamos um histograma para visualizar sua distribuição.

```
p <- 0.1
n <- 1040
p_hats <- rep(0, 5000)

for(i in 1:5000){
  samp <- sample(c("atheist", "non_atheist"), n, replace = TRUE, prob = c(p, 1-p))
  p_hats[i] <- sum(samp == "atheist")/n
}

hist(p_hats, main = "p = 0.1, n = 1040", xlim = c(0, 0.18))
```

Esses comandos constroem a distribuição amostral de `p_hats` por meio do *loop* do comando `for` que já nos é familiar. Você pode pode ler o procedimentos amostral da primeira linha de código dentro do *loop* como "retire uma amostra com reposição de n elementos a partir das opções de ateu e não-ateu com probabilidades p e $1-p$, respectivamente." A segunda linha do *loop* diz "calcule a proporção de ateus nesta amostra e registre e valor." O *loop* nos permite repetir esse processo 5.000 vezes para construir uma boa representação da distribuição amostral.

Exercício 9 Descreva a distribuição amostral da proporção com $n = 1040$ e $p = 0.1$. Certifique-se de identificar seu centro, dispersão e forma.
Dica: Lembre-se que o R tem funções como `mean` para calcular estatísticas descritivas.

Exercício 10 Repita a simulação acima mais três vezes mas com diferentes tamanhos de amostra e proporções: com $n = 400$ e $p = 0.1$, $n = 1040$ e $p = 0.02$, e $n = 400$ e $p = 0.02$. Crie histogramas para as quatro distribuições e exiba-os em conjunto utilizando o comando `par(mfrow = c(2,2))`. Talvez você precise expandir a janela do gráfico para acomodar o gráfico maior. Descreva as três distribuições amostrais novas. Com base nesses gráficos limitados, como que n parece afetar a distribuição de $\hat{p}$? Como que p afeta a distribuição amostral?

Depois de terminar, você pode resetar a disposição da janela de gráfico utilizando o comando `par(mfrow = c(1,1))` ou clicando no botão "Clear All" ("Limpar Tudo") logo acima da janela de gráficos (se estiver usando o RStudio). Preste atenção pois a última opção irá apagar todos os gráficos anteriores.

Exercício 11 Se você retomar a Tabela 6, verá que a Austrália tem uma proporção amostral de 0,1 numa amostra de 1040, e que o Equador tem uma proporção amostral de 0,02 com 400 sujeitos. Vamos supor, para esse exercício, que essas estimativas pontuais são verdadeiras. Dada a forma de suas respectivas distribuições amostrais, você acha razoável efetuar inferência e informar a margem de erros, como o relatório faz?

Sua Vez

A questão sobre o ateísmo foi também feita pelo WIN-Gallup International numa pesquisa de opinião parecida realizada em 2005.[†] A Tabela 4 na página 13 do relatório resume os resultados da pesquisa de 2005 a 2012 em 29 países.

1. Responda às duas perguntas seguintes utilizando a função `inference`. Como sempre, descreva as hipóteses para qualquer teste que você realizar e esboce sobre as condições para inferência.

   (a) Há evidência convincente de que a Espanha teve uma mudança em seu índice de ateísmo entre 2005 e 2012?
   Dica: Crie um novo conjunto de dados para os respondentes da Espanha. Depois, utilize suas respostas como a primeira entrada na função `inference`, e utilize a variável `year` (ano) para definir os grupos.

   (b) Há evidência convincente de que os Estados Unidos tiveram uma mudança em seu índice de ateísmo entre 2005 e 2012?

2. Se de fato não houve nenhuma mudança no índice de ateísmo nos países listados na Tabela 4, em quantos países você esperar detectar uma mudança (com um nível de significância de 0,05) simplesmente por acaso?
   Dica: Procure no índice do livro sobre erros do Tipo 1.

3. Suponha que você foi contratado pelo governo local para estimar a proporção de residentes que participam de cultos religiosos semanalmente. De acordo com diretrizes, a estimativa deve ter uma margem de erro inferior a 1% com nível de confiança de 95%. Você não tem nenhuma noção de que valor supor para p. Quanto pessoas você teria que amostrar para garantir que você está dentro das diretrizes?
   Dica: Retome seu gráfico da relação entre p e a margem de erro. Não use o conjunto de dados para responder a essa questão.

4. Quais conceitos do livro são abordados neste laboratório? Quais conceitos, se houver algum, que não são abordados no livro? Você viu esses conceito em algum outro lugar, p.e., aulas, seções de discussão, laboratórios anteriores, ou tarefas de casa? Seja específico em sua resposta.

[†] Assumimos aqui que o tamanho das amostras permaneceram iguais.