Laboratório 5: Inferência para Dados Numéricos

Nascimentos na Carolina do Norte

Em 2004, o estado da Carolina do Norte, Estado Unidos, disponibilizou um grande conjunto de dados contendo informações sobre os nascimentos registrados no estado. Esse conjunto de dados é útil para pesquisadores que estudam a relação entre hábitos e práticas de gestantes e o nascimento de seus filhos. Nós trabalharemos com uma amostra aleatória das observações deste conjunto de dados.

Análise Exploratória

Carregue o conjunto de dados `nc` em seu espaço de trabalho.

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

load("nc.RData")
```

Temos dados de 13 variáveis diferentes, algumas categoriais e outras numéricas. Cada variável representa as seguintes informações:

`fage`	idade do pai em anos.
`mage`	idade da mãe em anos.
`mature`	maioridade da mãe.
`weeks`	duração da gestação em semanas.
`premie`	se o nascimento é classificado como prematuro ou a termo.
`visits`	número de visitas hospitalares durante a gravidez.
`marital`	se a mãe estava `casada` (`married`) ou `solteira` (`not married`) no momento do nascimento.
`gained`	peso ganho pela mãe durante a gravidez, em libras.
`weight`	peso do bebê no nascimento, em libras.
`lowbirthweight`	se o bebê foi classificado como tendo baixo peso ao nascer (`low`) ou não (`not low`).
`gender`	sexo do bebê, `feminino` (`female`) ou `masculino` (`male`).
`habit`	se a mãe é `não-fumante` (`nonsmoker`) ou `fumante` (`smoker`).
`whitemom`	se a mãe é `branca` (`white`) ou `não-branca` (`not white`).

Exercício 1 Quais são os casos neste conjunto de dados? Há quantos casos em nossa amostra?

Como um primeiro passo na análise, devemos levar em consideração alguns sumários dos dados. Isso pode ser feito utilizando o comando `summary` ("sumário"):

```
summary(nc)
```

Enquanto você confere os sumários das variáveis, considere quais variáveis são categoriais e quais são numéricas. Para as variáveis numéricas, há algum caso atípico, um *outlier*? Se você não tem certeza ou quer dar uma olhada mais aprofundada nos dados, crie um gráfico.

Considere a possibilidade de uma relação entre o hábito de fumar da mãe e o peso de seu bebê. Criar um gráfico com os dados é uma etapa útil porque nos ajuda a visualizar tendências rapidamente, identificar associações fortes, e elaborar questões de pesquisa.

Exercício 2 Crie um gráfico de caixas lado-a-lado das variáveis `habit` (`hábito`) e `weight` (`peso`). O que o gráfico revela sobre a relação entre essas duas variáveis?

O gráfico de caixas permite comparar as medianas das distribuições, mas podemos também comparar as médias das distribuições utilizando a seguinte função para dividir a variável `weight` nos grupos definidos pela variável `habit`, e então calcular a média de cada um utilizando a função `mean`.

```
by(nc$weight, nc$habit, mean)
```

Há uma diferença evidente, mas essa diferença é estatisticamente significante? Para responder a essa questão, vamos realizar um teste de hipótese.

Inferência

Exercício 3 Verifique se as condições necessárias para realizar a inferência são atendidas. Perceba que você precisará obter o tamanho das amostras para verificar as condições. Você pode calcular o tamanho dos grupos utilizando o mesmo comando `by` utilizado acima, mas substituindo a função `mean` pela função `length`.

Exercício 4 Escreva as hipóteses para testar se a média dos pesos dos bebês que nasceram de mães fumantes é diferente daqueles que nasceram de mães não fumantes.

Em seguida, utilizaremos uma nova função, `inference`, que será utilizada para realizar os testes de hipótese e para construir os intervalos de confiança.

```
inference(y = nc$weight, x = nc$habit, est = "mean", type = "ht", null = 0,
          alternative = "twosided", method = "theoretical")
```

Vamos com calma para analisar cada argumento desta função personalizada.

- O primeiro argumento é `y`, que é a variável resposta na qual estamos interessados: `nc$weight` (peso).
- O segundo argumento é a variável explicativa, `x`, que é a variável que divide os dados em dois grupos, fumantes e não fumantes: `nc$habit`.
- O terceiro argumento, `est`, é o parâmetro no qual estamos interessados: `"mean"` (média) (há outras opções: `"median"` (mediana), ou `"proportion"` (proporção)).
- Em seguida decidimos sobre o tipo de inferência que queremos (`type`): um teste de hipótese (`"ht"`) ou um intervalo de confiança (`"ci"`).
- Quando realizamos um teste de hipótese, também precisamos informar o valor nulo (`null`), que neste caso é `0`, já que a hipótese nula supõe que as duas médias populacionais são iguais uma a outra.
- A hipótese alternativa (`alternative`) pode ser `"less"` (menor), `"greater"` (maior), ou `"twosided"` (bi-caudal).
- Por fim, o método (`method`) de inferência pode ser `"theoretical"` (teórico) ou `"simulation"` (baseado em simulações).

Exercício 5 Mude o argumento `type` (tipo) para `"ci"` para construir e registrar um intervalo de confiança para a diferença entre os pesos dos bebê que nasceram de mães fumantes e não fumantes.

Por padrão, a função utilizada informa um intervalo para a diferença ($\mu_{não-fumante} - \mu_{fumante}$) (a diferença entre médias dos dois grupos). Podemos mudar facilmente essa ordem utilizando o argumento `order` (ordem):

```
inference(y = nc$weight, x = nc$habit, est = "mean", type = "ci", null = 0,
          alternative = "twosided", method = "theoretical",
          order = c("smoker","nonsmoker"))
```

Sua Vez

1. Calcule o intervalo de confiança de 95% para a duração média das gravidezes (`weeks`) e o interprete no contexto do conjunto de dados. Perceba que, uma vez que você está realizando uma inferência sobre um único parâmetro populacional, não há nenhuma variáveis explanatória, e portanto você pode omitir a variável `x` da função.
2. Calcule um novo intervalo de confiança para o mesmo parâmetro com nível de confiança de 90%. Você pode mudar o nível de confiança adicionando um novo argumento à função: `conflevel =0.90`.
3. Realize um teste de hipótese para avaliar se o a média do peso ganho pelas mães mais jovens é diferente da média de peso ganho pelas mães mais velhas.
4. Agora, um tarefa não-inferencial: determine o ponto de corte da idade das mães jovens e maduras. Utilize um método da sua escolha, e explique como seu método funciona.
5. Escolha um par de variáveis, sendo uma numérica e outra categorial, e desenvolva um pergunta de pesquisa para avaliar a relação entre essas variáveis. Formule a questão de maneira que ela possa ser respondida utilizando um teste de hipótese e/ou um intervalo de confiança. Responda a sua questão utilizando a função `inference`, informe os resultados estatísticos, e também elabora uma explicação em linguagem simples.
6. 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.