Estatística Básica e Análise de Dados para Diagnóstico, Monitoramento e Avaliação de Programas Sociais

Introdução ao R e Testes de Hipóteses

Gilvan Guedes
Departamento de Demografia - UFMG

Gabriel Assunção
Departamento de Estatística - UFMG

Escola Nacional de Administração Pública
Ministério do Desenvolvimento Social

Brasília, Distrito Federal, Brasil

13 de janeiro de 2016

Sumário

1 Testes Paramétricos - continuação

1.1 Teste para a Proporção de uma População

Suponha que tenhamos uma amostra $X_1, X_2, \ldots, X_n$ de distribuição Normal, com média p desconhecida e variância $p(1-p)$ conhecida e desejamos testar as seguintes hipóteses:

$$\begin{cases} H_0 : p = p_0 \\ H_A : p \neq p_0 \end{cases} \tag{1}$$

A estatística de teste é dada por:

$$z_0 = \frac{(\hat{p} - p_0)}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

A região crítica para um teste bilateral com um nível α de significância, a região crítica assume a seguinte forma:

$$RC = \{z \in \mathbb{R} : z > z_{(\alpha/2)} | z < -z_{(\alpha/2)}\}$$

onde $z_{(\frac{\alpha}{2})}$ é encontrado na tabela da distribuição Normal padrão e é tal que:

$$P(Z > z_{(\frac{\alpha}{2})}) = \frac{\alpha}{2}$$

O p-valor do teste é dado por:

$$\text{p-valor} = 2P(Z > |z_0|\,|H_0)$$

Caso a hipótese de interesse seja unilateral, ou seja, quando temos as seguintes hipóteses:

$$\begin{cases} H_0 : p \leq p_0 \\ H_A : p > p_0 \end{cases} \tag{2}$$

ou

$$\begin{cases} H_0 : p \geq p_0 \\ H_A : p < p_0 \end{cases} \tag{3}$$

A região crítica será, respectivamente, dada por:

$$\text{p-valor} = P(Z > z_0|H_0)$$

ou

$$\text{p-valor} = P(Z < -z_0|H_0)$$

e o p-valor será dado por:

$$\text{p-valor} = P(Z > |z_0|\,|H_0)$$

Para implementar os testes no R, basta executar os seguintes comandos:

```
> # Teste para a proporcao de uma populacao com variancia conhecida
>
> set.seed(11) # Definindo a semente de numeros aleatorios
> #=================================================================
> # Gera uma amostra de 1000 observacoes da distribuicao de Bernoulli
> #com media p = 0.3
>
> X <- rbinom(1,1000,0.3)
> #=================================================================
> # Teste Bicaudal com Significancia = 5%
> prop.test(X, # objeto
+           1000, # tamanho do vetor
+           p=0.1, # probabilidade teorica
+           alternative='two.sided', # tipo da Hipotese Alternativa
+           conf.level=0.95, # nivel de confianca
+           correct=TRUE) # correcao para populacao finita

        1-sample proportions test with continuity correction

data:  X out of 1000, null probability 0.1
X-squared = 497.025, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.1
95 percent confidence interval:
 0.2835641 0.3418997
sample estimates:
    p
0.312

> #=================================================================
> # Teste Unicaudal com Significancia = 5%
>
> prop.test(X, # objeto
+           1000, # tamanho do vetor
+           p=0.1, # probabilidade teorica
+           alternative='less', # tipo da Hipotese Alternativa (p < p_0)
+           conf.level=0.95, # nivel de confianca
+           correct=TRUE) # correcao para populacao finita

        1-sample proportions test with continuity correction

data:  X out of 1000, null probability 0.1
X-squared = 497.025, df = 1, p-value = 1
alternative hypothesis: true p is less than 0.1
95 percent confidence interval:
 0.0000000 0.3370882
sample estimates:
    p
0.312
```

```
> #==================================================================
> prop.test(X, # objeto
+           1000, # tamanho do vetor
+           p=0.1, # probabilidade teorica
+           alternative='greater', # tipo da Hipotese Alternativa (p > p_0)
+           conf.level=0.95, # nivel de confianca
+           correct=TRUE) # correcao para populacao finita

        1-sample proportions test with continuity correction

data:  X out of 1000, null probability 0.1
X-squared = 497.025, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is greater than 0.1
95 percent confidence interval:
 0.2879473 1.0000000
sample estimates:
    p
0.312

>
```

1.2 Teste para Diferença de Proporções

Sejam $X_1, X_2, \ldots, X_n$ uma amostra aleatória simples selecionada na população X e seja $Y_1, Y_2, \ldots, Y_n$ uma amostra aleatória simples selecionada na população Y. Admita que a população X seja independente da população Y. Assuma também que as duas populações sejam distribuídas segundo a Distribuição de Bernoulli com médias p_X e p_Y e variâncias populacionais $p_X(1-p_X)$ e $p_Y(1-p_Y)$, ambas desconhecidas.

O teste terá a seguinte estrutura:

$$\begin{cases} H_0 : p_X = p_Y \\ H_A : p_X \neq p_Y \end{cases} \tag{4}$$

Como não conhecemos os valores de p_X e p_Y, vamos realizar uma estimação ponderada entre esses valores para determinar a variância do teste. O valor do desvio-padrão seria dado por:

$$\sigma = \sqrt{\frac{p_X(1-p_X)}{n_X} + \frac{p_Y(1-p_Y)}{n_Y}} = \sqrt{\frac{p(1-p)}{n_X} + \frac{p(1-p)}{n_Y}}$$

No entanto, como não conhecemos o valor de p, precisamos estimá-lo com os valores amostrais. A fórmula do valor estimado será:

$$\hat{p} = \frac{n_X \hat{p}_X + n_Y \hat{p}_Y}{n_X + n_Y}$$

A estatística de teste para as hipóteses ($H_A : p_X = p_Y$ e $H_A : p_X \neq p_Y$) é dada por:

$$Z_0 = \frac{(\hat{p}_X - \hat{p}_Y)}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_X} + \frac{\hat{p}(1-\hat{p})}{n_Y}}} \sim N(0,1)$$

A região crítica para um teste bilateral com um nível α de significância, assume a seguinte forma:

$$RC = \{z \in \mathbb{R} : z_0 > z_{(\alpha/2)} | z_0 < -z_{(\alpha/2)}\}$$

onde $z_{(\frac{\alpha}{2})}$ é encontrado na tabela da distribuição Normal padrão e é tal que:

$$P(Z > z_{(\frac{\alpha}{2})}) = \frac{\alpha}{2}$$

O p-valor do teste é dado por:

$$\text{p-valor} = 2P(Z > |z_0|\,|H_0)$$

Caso a hipótese de interesse seja unilateral, ou seja, quando temos as seguintes hipóteses:

$$\begin{cases} H_0 : p_X \leq p_Y \\ H_A : p_X > p_Y \end{cases} \tag{5}$$

ou

$$\begin{cases} H_0 : p_X \geq p_Y \\ H_A : p_X < p_Y \end{cases} \tag{6}$$

A região crítica será, respectivamente, dada por:

$$\text{p-valor} = P(Z > z_0|H_0)$$

ou

$$\text{p-valor} = P(Z < -z_0|H_0)$$

e o p-valor será dado por:

$$\text{p-valor} = P(Z > |z_0|\,|H_0)$$

Para implementar os testes no R, siga os comandos abaixo:

```
> set.seed(11) # Definindo a semente de numeros aleatorios
> #=================================================================
> # Gera duas amostras independentes de 1000 e 1500 observacoes da distribuicao de B
>
> # X com media p = 0.3
>
> X <- rbinom(1,1000,0.3)
> # Y com media p = 0.5
>
> Y <- rbinom(1,1500,0.5)
> # Aplicando o teste:
> #=================================================================
> # Teste Bicaudal com Significancia = 5%
> prop.test(X,Y, # objetos
+           alternative='two.sided', # tipo da Hipotese Alternativa
+           conf.level=0.95, # nivel de confianca
+           correct=TRUE) # correcao para populacao finita
```

```
        1-sample proportions test with continuity correction

data:  X out of Y, null probability 0.5
X-squared = 30.0887, df = 1, p-value = 4.127e-08
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.3665265 0.4365207
sample estimates:
        p
0.4010283

> #=================================================================
> # Teste Unicaudal com Significancia = 5%
>
> prop.test(X,Y, # objeto
+           alternative='less', # tipo da Hipotese Alternativa (p_X < p_Y)
+           conf.level=0.95, # nivel de confianca
+           correct=TRUE) # correcao para populacao finita

        1-sample proportions test with continuity correction

data:  X out of Y, null probability 0.5
X-squared = 30.0887, df = 1, p-value = 2.064e-08
alternative hypothesis: true p is less than 0.5
95 percent confidence interval:
 0.0000000 0.4308732
sample estimates:
        p
0.4010283

> #=================================================================
> prop.test(X,Y, # objeto
+           alternative='greater', # tipo da Hipotese Alternativa (p_X > p_Y)
+           conf.level=0.95, # nivel de confianca
+           correct=TRUE) # correcao para populacao finita

        1-sample proportions test with continuity correction

data:  X out of Y, null probability 0.5
X-squared = 30.0887, df = 1, p-value = 1
alternative hypothesis: true p is greater than 0.5
95 percent confidence interval:
 0.3718845 1.0000000
sample estimates:
        p
0.4010283
```