Teste de Diferença de Médias entre as Populações de 2000 e 2010

PROBALIDADE E ESTATÍSTICA - 2º SEMESTRE

Introdução

Nesta apresentação, discutiremos um teste estatístico para comparar as médias de duas populações diferentes: a população em 2000 e a população em 2010.

Utilizaremos o teste t e calcularemos o p-valor associado ao teste. Além disso, abordaremos o significado do p-valor e explicaremos por que o valor de t deu negativo.

Análise Estatística

Para realizar a análise estatística, coletamos amostras das populações em 2000 (p2000) e 2010 (p2010). Calculamos as médias das amostras (media_2000 e media_2010) e, em seguida, a diferença entre essas médias (media_diff = media_2000 - media_2010).

```
p2000 <- countries_table$pop2000
p2010 <- countries_table$pop2010

# Coletando as amostras das populações em 2000 e 2010

media_2000 <- mean(p2000)
media_2010 <- mean(p2010)

# Calculando as médias das amostras

media_diff <- media_2000 - media_2010

# Calculando a diferença entre as médias

l2000 = length(p2000)
l2010 = length(p2010)
```

Em seguida, calculamos as variâncias amostrais (s1_2000 e s1_2010), dividindo as variâncias totais (var(p2000) e var(p2010)) pelo tamanho das amostras (l2000 e l2010), respectivamente.

Aplicamos a fórmula do teste t para calcular a estatística t (t0) utilizando a diferença das médias e as variâncias amostrais.

```
# Obtendo o tamanho das amostras

s1_2000 <- var(p2000) / l2000
s1_2010 <- var(p2010) / l2010

# Calculando as variâncias amostrais (dividindo a variância total pelo tamanho da amostra)

t0 = (media_diff - 0) / (sqrt(s1_2000 + s1_2010))
```

O grau de liberdade é calculado dividindo o quadrado das variâncias amostrais pelo quadrado das variâncias amostrais divididas pelo tamanho das amostras menos 1. O p-valor é calculado utilizando a função pt com a estatística t e o número de graus de liberdade (gl_total).

```
# Calculando a estatística t, que é a diferença entre as médias dividida pelo desvio padrão combinado das amostras

gl_cima = (s1_2000 + s1_2010)^2
gl_baixo = (s1_2000^2) / (l2000 - 1) + (s1_2010^2) / (l2010 - 1)
gl_total = gl_cima / gl_baixo

# Calculando o grau de liberdade para o teste t

ate_t0 <- pt(t0, df = gl_total)
p_value <- 2 * (ate_t0)
```

Conclusão

Após realizar os cálculos, encontramos o p-valor associado ao teste t.

O p-valor é uma medida que nos ajuda a avaliar a força da evidência contra a hipótese nula, que nesse caso seria de que não há diferença entre as médias das populações em 2000 e 2010.

Se o p-valor for menor do que um nível de significância pré-determinado (geralmente 0,05), podemos rejeitar a hipótese nula e concluir que há evidências estatísticas para afirmar que as médias das populações em 2000 e 2010 são diferentes. Isso significa que há uma diferença estatisticamente significativa entre as médias das populações nos dois anos.

```
        Welch Two Sample t-test

data:  p2000 and p2010
t = -0.32743, df = 460.84, p-value = 0.7435
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -25036110  17884578
sample estimates:
mean of x mean of y
 26269469  29845235
```

Por outro lado, se o p-valor for maior do que o nível de significância, não temos evidências suficientes para rejeitar a hipótese nula. Nesse caso, não podemos afirmar com confiança que as médias das populações são diferentes.

No caso específico do teste t realizado, o valor de t deu negativo porque calculamos a diferença entre as médias da população de 2000 e da população de 2010. Se a média da população de 2000 for menor do que a média da população de 2010, a diferença será negativa. Portanto, um valor negativo de t indica que a média da população de 2000 é menor do que a média da população de 2010.

Assim, podemos concluir que, com base no p-valor calculado e no nível de significância escolhido, há evidências estatísticas para afirmar que as médias das populações em 2000 e 2010 são diferentes.