Universidade de Brasília

DEPARTAMENTO DE ESTATÍSTICA

28 abril 2024

Atividade 1

Prof. Maria Tereza Leão Costa
Aluno: Bruno Gondim Toledo
Matrícula: 15/0167636

Análise de Dados Categorizados
1º/2024

A partir do conjunto de dados abaixo

consumo	ausente	presente
0	17066	48
<1	14464	38
1-2	788	5
3-5	126	1
>=6	37	1

Deseja-se testar a hipótese de associação entre consumo de álcool e presença de malformação.

H_0) Não existe associação entre consumo de álcool e presença de malformação.

H_1) c.c.

Nível de significância: $\alpha = 0,05$.

Para tal, podemos utilizar o teste do qui-quadrado de independência. O resultado do teste é:

```
##
##  Pearson's Chi-squared test
##
## data:  data
## X-squared = 12.082, df = 4, p-value = 0.01675
```

Que rejeita a hipótese nula de independência entre as variáveis. Entretando, os valores n_{ij} não são todos ≥ 5, que é um pré-requisito para confiabilidade do teste.

Podemos também calcular a *Odds Ratio* para este conjunto:

```
##                  odds ratio with 95% C.I.
## Consumo de Álcool   estimate     lower     upper
##              0    1.0000000        NA        NA
##              <1   0.9349104 0.6066157  1.430732
##              1-2  2.3206663 0.7926580  5.322638
##              3-5  3.2192267 0.1373687 14.677729
##              >=6 10.9263547 0.4594948 51.703348
```

Deste, existe um indicativo de maior chance de mal formação congênita para os grupos de maior consumo de álcool.

Para uma análise mais aprofundada, podemos calcular o coeficiente de correlação de Pearson ρ entre o consumo de álcool e a presença de malformação.

Como as variáveis são categorizadas, será necessário associar um *score* para cada categoria de consumo de álcool. Feito isso, poderemos calcular a estatística $(n-1)\rho^2 = M^2 \sim \chi^2_1 \equiv \sqrt{M^2} = M \sim N(0,1)$.

Definida a estatística de teste, e com o mesmo conjunto de hipóteses definidos para a análise de independência, resta testar para diferentes valores de *scores* arbitrários.

Scores

Scores 1,2,3,4 e 5

Valor de ρ = 0.0074908;

Valor de M^2 = 1.8277596;

P-valor = 0.1763924.

Scores 2,4,6,8 e 10

Valor de ρ = 0.0074908;

Valor de M^2 = 1.8277596;

P-valor = 0.1763924.

Scores 10,20,30,40 e 50

Valor de ρ = 0.0074908;

Valor de M^2 = 1.8277596;

P-valor = 0.1763924.

Daqui, é possível observar que os três valores são idênticos para os *scores* selecionados, o que indica o postulado de que o valor do *score* não é o que influencia as estatísticas, mas sim a diferença entre eles.

Scores 1,2,4,7 e 12

Valor de ρ = 0.0132754;

Valor de M^2 = 5.7405554;

P-valor = 0.0165775.

Desta, que seria mais próximo aos pontos médios de cada categoria, as estatísticas parecem mais coerentes com a hipótese.

Scores 5,4,3,2 e 1

Valor de ρ = -0.0074908;

Valor de M^2 = 1.8277596;

P-valor = 0.1763924.

Deste, vemos que invertendo a ordem dos scores, tanto a estatística M^2 quanto o P-valor são idênticos que os 3 primeiros casos - Apenas o ρ calculado vêm com o sinal invertido, porém o valor absoluto é o mesmo.

Extra: Exponencial do ponto médio:

$[0, e^{0,5}, e^{1,5}, e^4, e^7]$

Valor de ρ = 0.0156111;

Valor de M^2 = 7.9382136;

P-valor = 0.0048402.

CONCLUSÃO: A escolha do *score* não influencia a estatística de teste, mas sim a diferença entre eles. A escolha de um *score* mais próximo ao ponto médio da categoria parece ser mais coerente com a hipótese de correlação. Entretanto, como se trata de uma escolha arbitrária, estes testes não podem ser considerados conclusivos.