



DEPARTAMENTO DE ESTATÍSTICA

17 maio 2024

## **Prova 1 - Questão Bônus**

Prof. Maria Teresa Leão Costa

Aluno: Bruno Gondim Toledo

Matrícula: 15/0167636

Análise de Dados Categorizados

1º/2024

## 1) Especifique as variáveis estudadas classificando-as quanto ao tipo e escala de mensuração utilizada.

Existem duas variáveis neste estudo, sendo elas Renda e Felicidade. Ambas estão categorizadas de forma ordinal. Ambas as variáveis podem ser vistas como explicativa, mas neste caso, me parece mais apropriado a variável resposta ser a Felicidade, enquanto que a variável renda seria a variável explicativa. As categorias em que foram ordinalizadas são nominais; não confundir com variáveis nominais (as variáveis SÃO qualitativas ordinais); apenas os rótulos não são como Likert ou algo do tipo, que dêem alguma ideia da magnitude da diferença entre as categorias.

## 2) Determine Frequencias relativas adequadas que permitam analisar descritivamente os resultados encontrados na amostra tendo em vista a pergunta a ser respondida pela pesquisa. Não se esqueça de representar graficamente os dados para ilustrar sua argumentação.

A variável Renda pode ser utilizada para responder a variável Felicidade? A tabela abaixo mostra a Frequencia de felicidade por renda. A tabela seguinte mostra a Frequencia relativa de felicidade por renda.

### Frequencias absolutas

| Renda           | Pouco_Feliz | Feliz | Muito_Feliz | Total |
|-----------------|-------------|-------|-------------|-------|
| Abaixo da Media | 21          | 159   | 110         | 290   |
| Media           | 53          | 372   | 221         | 646   |
| Acima da Media  | 94          | 249   | 83          | 426   |
| Total           | 168         | 780   | 414         | 1362  |

### Frequencias relativas

#### Frequências relativas de linha

|             | Abaixo da Media | Media | Acima da Media | Total |
|-------------|-----------------|-------|----------------|-------|
| Pouco_Feliz | 0.072           | 0.548 | 0.379          | 1     |
| Feliz       | 0.082           | 0.576 | 0.342          | 1     |
| Muito_Feliz | 0.221           | 0.585 | 0.195          | 1     |

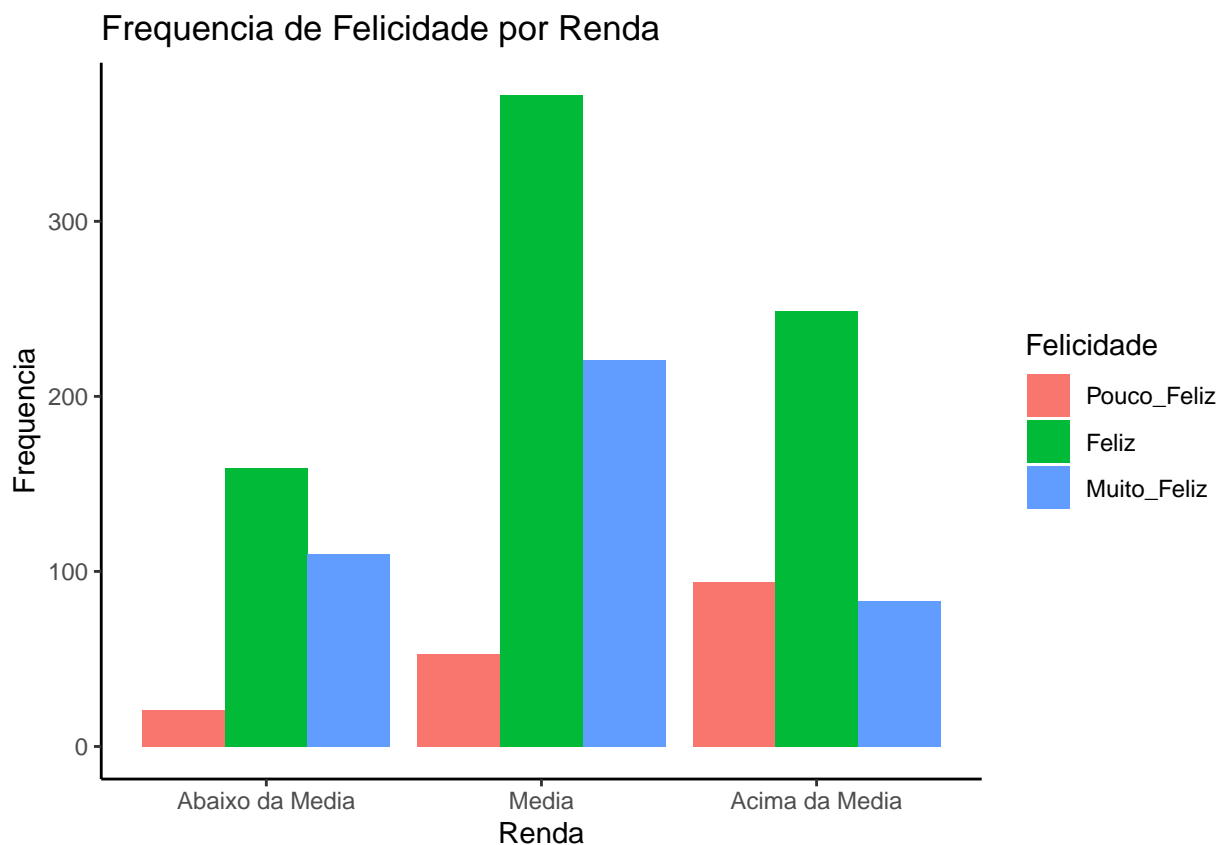
#### Frequências relativas de coluna

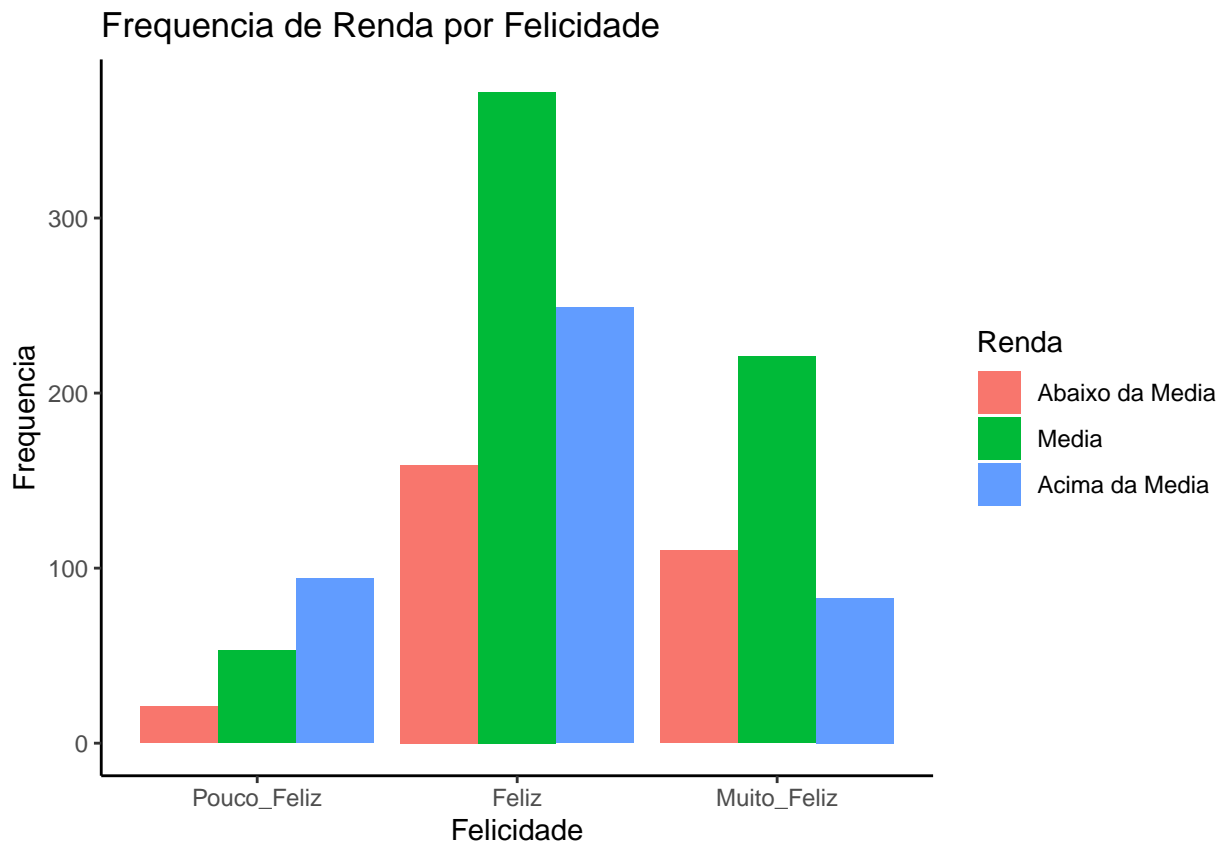
|             | Abaixo da Media | Media | Acima da Media |
|-------------|-----------------|-------|----------------|
| Pouco_Feliz | 0.125           | 0.204 | 0.266          |
| Feliz       | 0.315           | 0.477 | 0.534          |
| Muito_Feliz | 0.560           | 0.319 | 0.200          |
| Total       | 1.000           | 1.000 | 1.000          |

### Frequências relativas globais

| Renda           | Pouco_Feliz | Feliz | Muito_Feliz | Total |
|-----------------|-------------|-------|-------------|-------|
| Abaixo da Media | 0.015       | 0.117 | 0.081       | 0.213 |
| Media           | 0.039       | 0.273 | 0.162       | 0.474 |
| Acima da Media  | 0.069       | 0.183 | 0.061       | 0.313 |
| Total           | 0.123       | 0.573 | 0.304       | 1.000 |

Os gráficos abaixo mostram a Frequencia de felicidade por renda e a Frequencia de renda por felicidade, respectivamente.





Pelas tabelas e gráficos, tendemos a acreditar que existe uma associação inversa de felicidade com renda. Aparentemente, quanto maior a renda, menor a felicidade. Nota-se um aumento da proporção de Pouca felicidade em relação à Muita felicidade conforme a renda cresce. Em todos os casos, “Feliz” (Media) se destaca como a categoria com maior Frequencia. Para o caso de pessoas com renda Abaixo da Media, a proporção de Feliz e Muito feliz são bem mais próximas que nas demais categorias de renda. Para pessoas de renda Acima da Media, a lógica inverte e a proporção de Pouco feliz é a que mais se aproxima da categoria com maior Frequencia absoluta (Feliz).

Portanto, não aparenta haver associação positiva entre aumento de renda e aumento de felicidade por esta análise exploratória preliminar.

**3) A equipe responsável pela análise dos dados optou por utilizar o Teste Qui-quadrado de Independência para verificar a hipótese de associação entre felicidade e renda. Os resultados apresentados a seguir foram obtidos durante a aplicação da técnica escolhida pela equipe e incluem as Frequencias observadas e esperadas das células e os resíduos padronizados para testar a independência.**

O valor da estatística do teste obtido foi de  $\chi^2 = 73,4$ .

**a) Utilizando o SAS reproduza a análise realizada pela equipe e interprete os resultados, indicando a conclusão obtida através do teste utilizado.**

Obs: Não consegui incluir os resíduos padronizados na mesma tabela

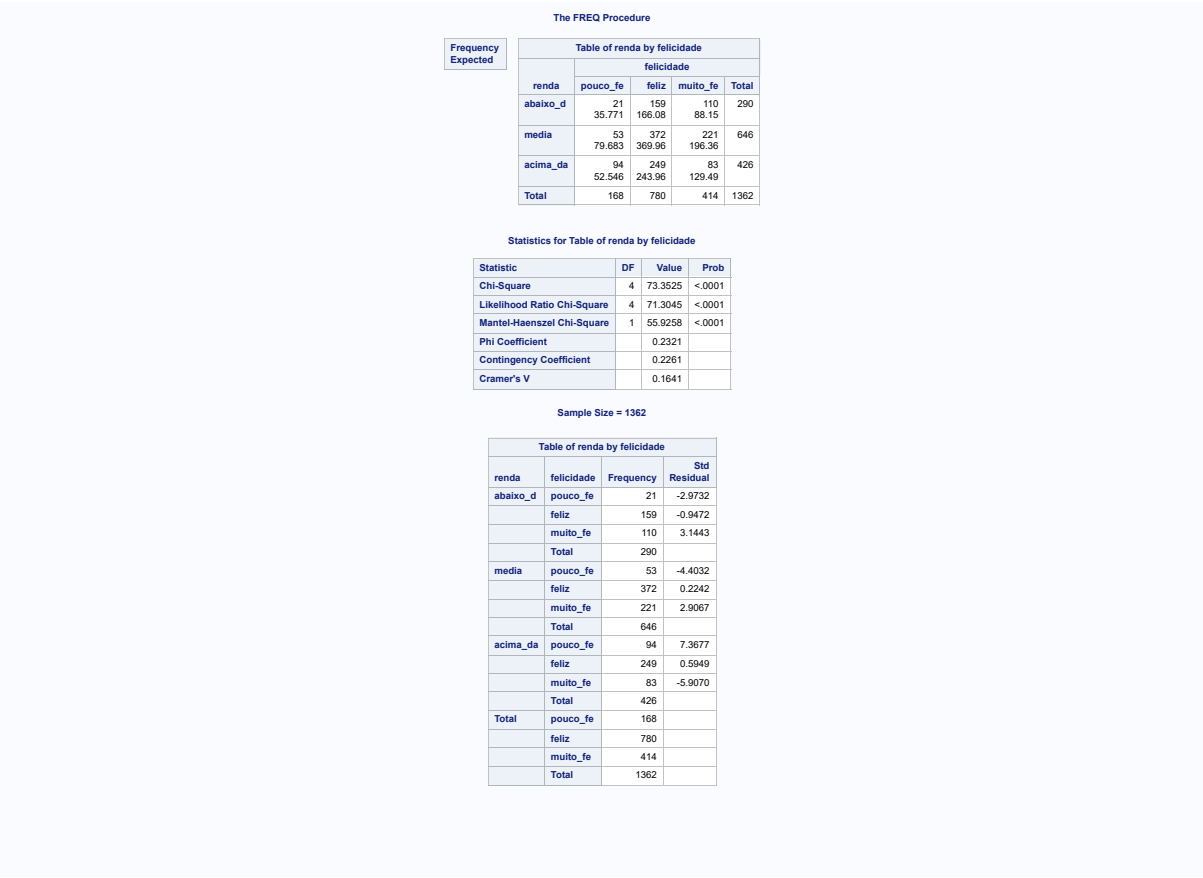


Figure 1: Saídas do SAS

Código utilizado:

```
data df;
input renda $ felicidade $ freq;
datalines;
abaixo_da_media pouco_feliz 21
abaixo_da_media feliz 159
abaixo_da_media muito_feliz 110
media pouco_feliz 53
media feliz 372
media muito_feliz 221
acima_da_media pouco_feliz 94
acima_da_media feliz 249
acima_da_media muito_feliz 83
;
run;

proc freq data=df order=data;
weight freq;
tables renda*felicidade/nopercent nocol expected norow chisq;
tables renda*felicidade/nopercent nocol norow CROSSTAB(STDRES);
run;
```

Quanto à análise:

Hipóteses de teste:

$H_0$ ) Não existe associação entre felicidade e renda.

$H_1$ )  $\exists$  associação entre felicidade e renda.

Estatística do teste:

$$\chi^2 = \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}} \sim \chi^2_{(i-1)(j-1)g.l. = 4 \text{ g.l.}} = 73,3525$$

Para qualquer valor de significância  $\alpha$ , rejeitamos a hipótese nula com p-valor  $\mathbb{P}(\chi^2_4 \geq 73,352) < 0,0001$ , indicando que existe associação entre felicidade e renda.

## **b) Interprete os resíduos padronizados nas células dos cantos tendo:**

### **i) Frequências 21 e 83,**

Nas células especificadas, o resíduo padronizado calculado foi de -2,9732 e -5,9070 respectivamente. Valores negativos indicam falta de ajuste, no sentido de a Frequência observada ser menor que a Frequência esperada. Portanto para as células especificadas, a Frequência observada é menor que a Frequência esperada.

### **ii) Frequências 110 e 94.**

Nessas células, o resíduo padronizado calculado foi de 3,1443 e 7,3677 respectivamente. Neste caso, a interpretação também é pela indicação da falta de ajuste do valor observado em relação ao valor esperado. Valores positivos indicam que a Frequência observada é maior que a Frequência esperada. Portanto neste caso, a Frequência observada é maior que a Frequência esperada.

### c) Indique qual a deficiência da utilização do teste aplicado para estes dados?

A hipótese foi formulada para indicar alguma dependência em relação às variáveis do estudo, qualquer que fossem. O teste conseguiu indicar o que observamos com o resíduo padronizado, ou seja, nos para altos níveis de felicidade e baixo nível de renda, a Frequência observada é maior que a Frequência esperada. O inverso vale para altos níveis de renda e baixa felicidade. Entretanto, o teste de qui-quadrado de independência não é o mais adequado para este tipo de análise, pois as variáveis são ordinais e não nominais. O teste de qui-quadrado é mais adequado para variáveis nominais. Portanto, o qui-quadrado não consegue captar com tanta sensibilidade a relação entre as variáveis que o estudo exigiria.

### 4) Indique um teste alternativo de independência que seja mais adequado para estes dados, justificando sua escolha.

Poderíamos utilizar diversos testes para esta abordagem. Por exemplo, um mais adequado e trabalhado na disciplina seria o coeficiente de correlação momento-produto de Pearson. Claro que, para este, teríamos que definir *scores* coerentes com a magnitude das variáveis, o que exigiria um pouco mais de conhecimento do problema do que simplesmente a tabela de contingência. Poderíamos também utilizar o teste de correlação de Spearman, que não exige tanta formalidade na definição dos *scores*. Poderíamos também pensar em agregar categorias (o que novamente exigiria um conhecimento maior do problema, das variáveis e características do estudo) para aplicar testes mais sensíveis à tabelas 2x2, como o teste exato de Fisher ou o teste de McNemar. É ainda plausível calcular a razão de chances (odds ratio) para avaliar a associação entre as variáveis, e depois testar se a razão de chances é diferente de 1 com a estatística de breslow-day ou teste de Wald. Há ainda o teste de Goodman-Kruskal, que é uma generalização do teste de McNemar para tabelas maiores que 2x2.

### 5) Aplique o teste indicado e interprete os resultados obtidos.

Irei aplicar os testes de McNemar, gama de Goodman-Kruskal e estatística  $G^2$  de razão de verossimilhança para testar a associação entre as variáveis.

Hipóteses:  $H_0$  Não existe associação entre felicidade e renda (variáveis independentes).  $H_1 \exists$  associação entre felicidade e renda.

```
##
## McNemar's Chi-squared test
##
## data:  cofre
## McNemar's chi-squared = 55.923, df = 3, p-value = 4.363e-12

##
## Log likelihood ratio (G-test) test of independence without correction
##
## data:  cofre
## G = 71.305, X-squared df = 4, p-value = 1.199e-14

##
## Goodman-Kruskal's gamma for ordinal categorical data
##
## data:  cofre
## Z = -7.5803, p-value = 3.447e-14
## 95 percent confidence interval:
## -0.3818329 -0.2297177
## sample estimates:
## Goodman-Kruskal's gamma
## -0.3057753
```

Todos os três testes rejeitam a hipótese nula de independência entre felicidade e renda. Portanto, a associação entre felicidade e renda é estatisticamente significativa.

Como o teste de Goodman-Kruskal é uma generalização do teste de McNemar, já era esperado que fornecessem o mesmo valor. E o teste de razão de verossimilhança é uma generalização do teste do qui-quadrado, portanto a interpretação dos resultados é a mesma para os três testes.

## **6) Elabore um parágrafo sobre as conclusões obtidas no estudo.**

Os resultados do experimento apresentam um paradigma. Era esperado encontrar um aumento de felicidade quanto renda também cresce. Entretanto, é sabido que existe uma correlação negativa aparente entre estas variáveis. Estamos diante de um clássico Paradoxo de Simpsons. Índices globais cautelosamente elaborados colocam países de renda alta no topo de felicidade. Porém, para observar estes resultados, é necessário ponderar outras covariáveis, pois simplesmente perguntando as pessoas o seu nível de felicidade, chegamos a uma situação como a apresentada neste estudo. A felicidade é uma característica latente que não pode ser diretamente medida, necessitando de técnicas elaboradas na área de psicometria para chegar a uma medida mais fidedigna. Renda é outra variável difícil de ser medida, também geralmente sendo estimada à partir de covariáveis, como consumo de energia, PIB per capita, etc.

Não me parece adequado aplicar os resultados deste estudo para uma inferência, visto que a análise foi feita de forma superficial e sem considerar outras variáveis que poderiam influenciar a relação entre felicidade e renda. Ainda assim, os resultados são interessantes e podem ser utilizados para direcionar futuras pesquisas sobre o tema, demonstrando a necessidade de estudar bem instrumentos que meçam características latentes e estimem renda de forma fidedigna. É claro tanto felicidade quanto renda são conceitos relativos, devendo ser ponderados por expectativa de vida, acesso a serviços básicos, e também a cultura local dos entrevistados. Nem todas as culturas, em especial as chamadas “não ocidentais”, tem o mesmo paradigma nosso de prosperidade; isto é, a vida como um constante crescimento econômico. Mas também é sabido que não se pode inferir sobre a felicidade de um povo que não tem saneamento básico, acesso a serviços, etc. Portanto, a relação entre felicidade e renda é complexa e deve ser estudada com cuidado.