Relatório 2

Gabriela D'Agostini Contreiras Rodrigues e Pietra Giulianna de Andrade Cretella

05 julho, 2024

O processo de manipulação dos dados iniciou com a limpeza e preparação do ambiente, carregando os pacotes necessários usando o pacote *pacman* e importação da base com a função *rio*.

Questão 1 (26 junho)

- **A) Identifique os *outliers* das variáveis de interesse (ESEB e LAPOP)**

Outliers são valores que estão muito distantes da maioria dos dados em uma distribuição numérica. Eles são geralmente identificados em variáveis quantitativas, que têm uma escala numérica, como altura, peso ou renda. Para variáveis qualitativas (ou categóricas), como "casamento", "aborto", "adoção" ou respostas a perguntas de sim/não, o conceito de "distância" não se aplica da mesma forma.

As variáveis qualitativas descrevem categorias ou atributos que não têm uma ordem ou escala numérica. Como não há uma escala contínua, não há um "valor extremo" que possa ser identificado como outlier. Por exemplo, a categoria "sim" em uma pergunta de sim/não não pode ser considerada extrema em relação à categoria "não".

Além disso, as medidas usadas para identificar outliers em dados quantitativos, como a média e o desvio padrão, não são aplicáveis a dados qualitativos. Em dados qualitativos, trabalhamos com contagens de categorias, não com distribuições contínuas. Uma categoria que ocorre com pouca frequência não é necessariamente um outlier, mas apenas uma categoria menos comum.

Seleção de variáveis utilizadas no "Relatório 1" como *issue positions* da dimensão liberal/fundamentalista

1. Casamento homoafetivo
2. Adoção por casais de minorias de gênero
3. Aborto

- **B) Identifique questões semelhantes em duas bases e realize um teste de diferença de médias (para amostras independentes).**

Testes de diferença de média são usados para comparar as médias de duas ou mais amostras e verificar se há uma diferença significativa entre elas. Esses testes são apropriados para variáveis quantitativas, que têm valores numéricos contínuos e permitem o cálculo de uma média. Por outro lado, variáveis qualitativas (ou categóricas) não possuem valores numéricos que possam ser somados ou divididos para calcular uma média. Como não há uma escala numérica, não podemos calcular uma média ou comparar médias entre grupos. Para variáveis qualitativas, como "casamento", "aborto", "adoção" ou respostas a perguntas de sim/não, fizemos uma análise visual através de gráficos de distribuição de respostas que permitem uma compreensão rápida e intuitiva de como as respostas estão distribuídas entre as categorias.

Após observar que a distribuição das respostas não segue uma distribuição normal, criamos tabelas de frequência para as variáveis "casamento" e "adoção" para ESEB e LAPOP. Essas tabelas não apenas proporcionam uma compreensão mais clara da distribuição dos dados, mas também garantem a realização de análises estatísticas robustas e significativas. Em seguida, conduzi o teste não paramétrico de Wilcoxon Signed-Rank para amostras independentes. Os resultados dos testes indicaram a rejeição da hipótese nula (H0, $p < 0,05$), sugerindo evidências estatísticas significativas que sustentam a conclusão de que há diferenças significativas na distribuição das opiniões sobre casamento homoafetivo e adoção ao longo dos últimos anos nas duas bases.

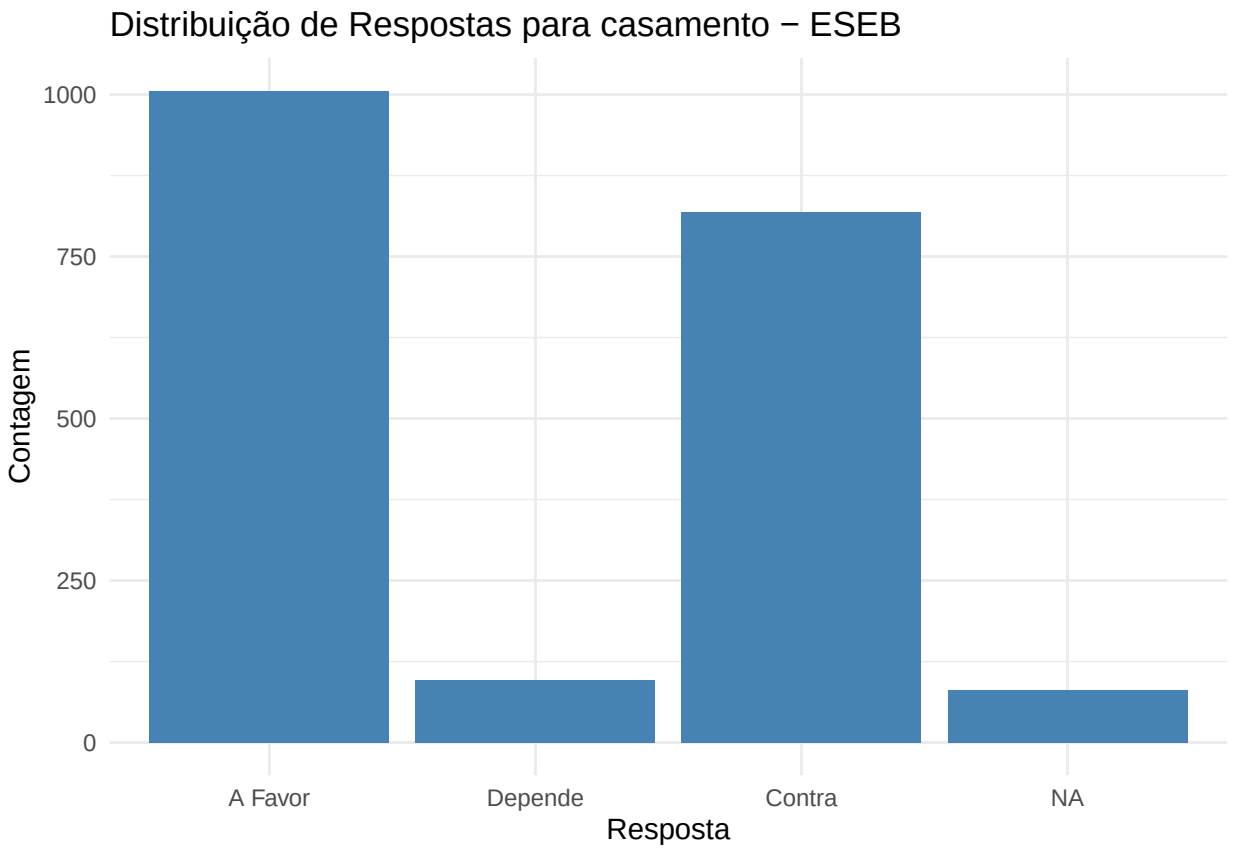


Figura 1: FONTE: ESEB, 2022

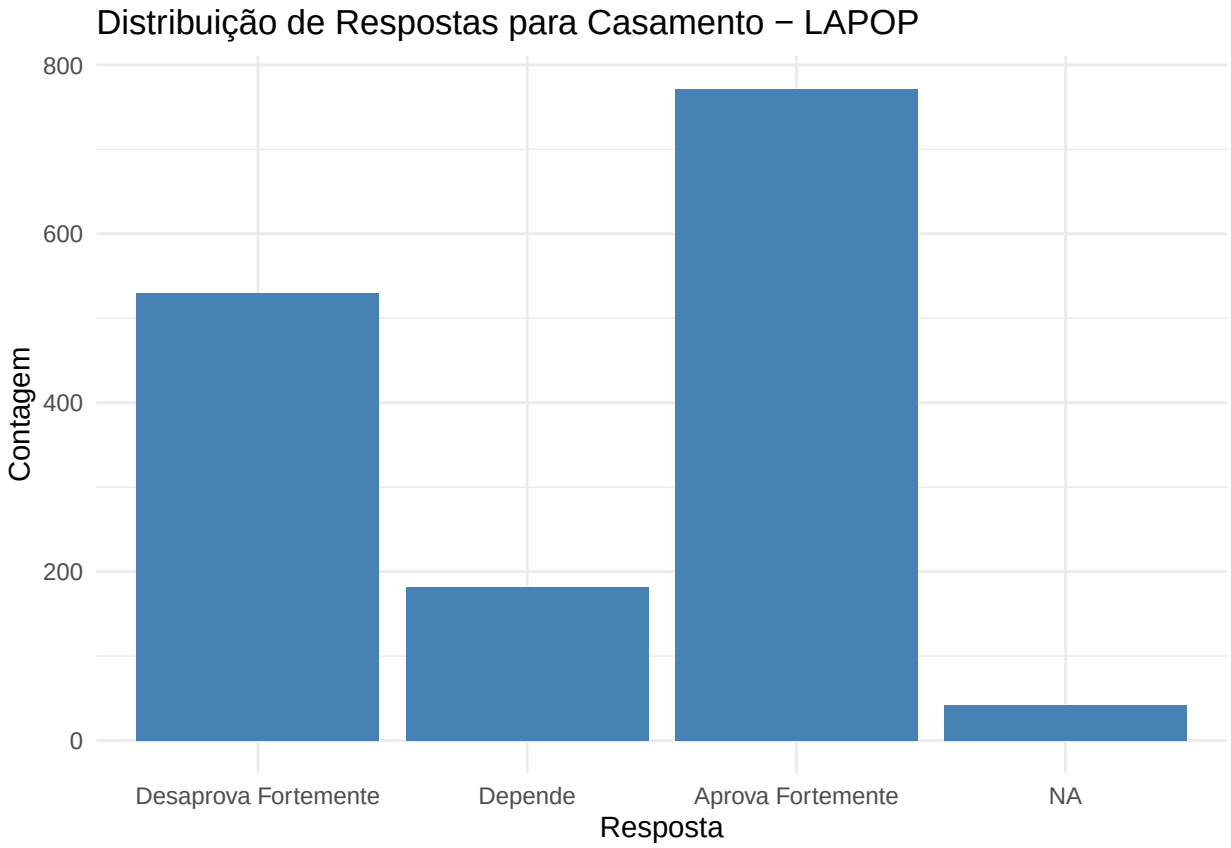


Figura 2: FONTE: LAPOP, 2023.

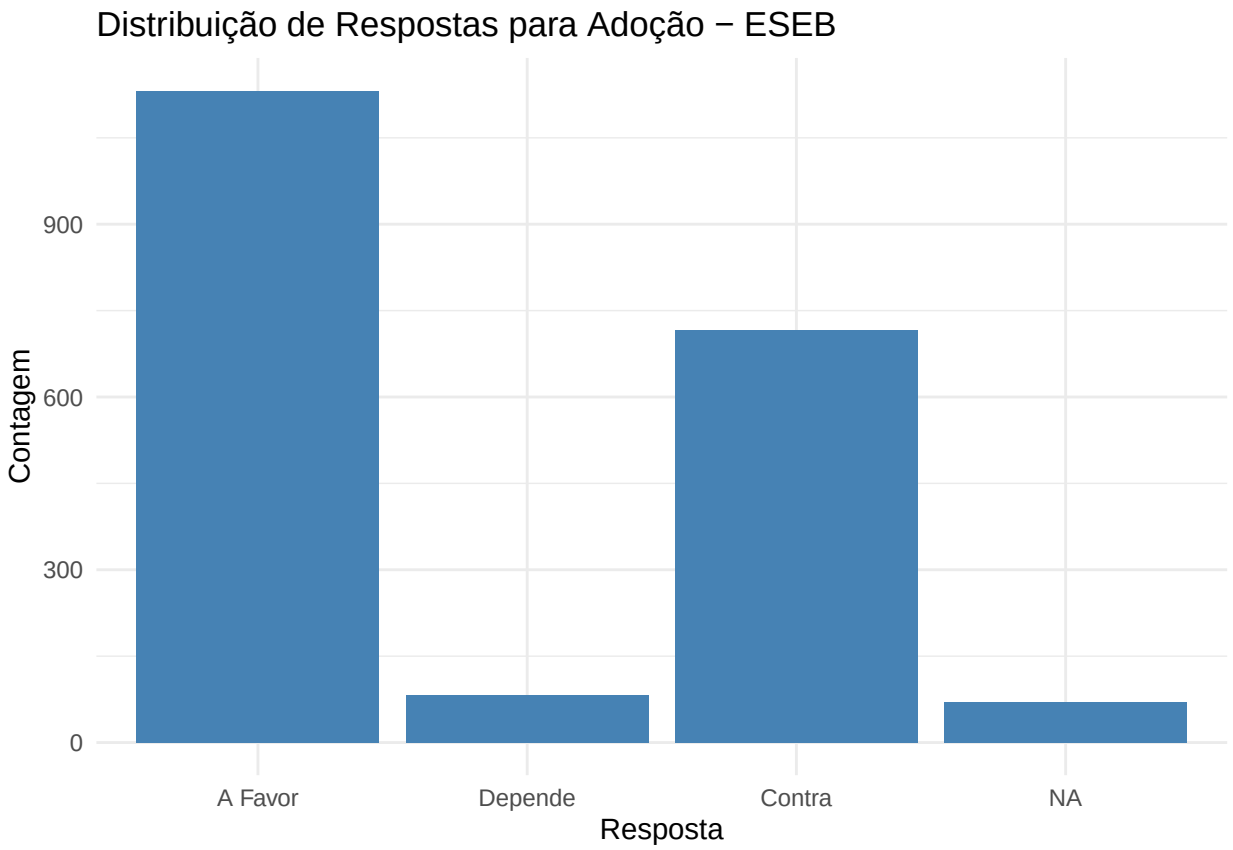


Figura 3: FONTE: ESEB, 2022

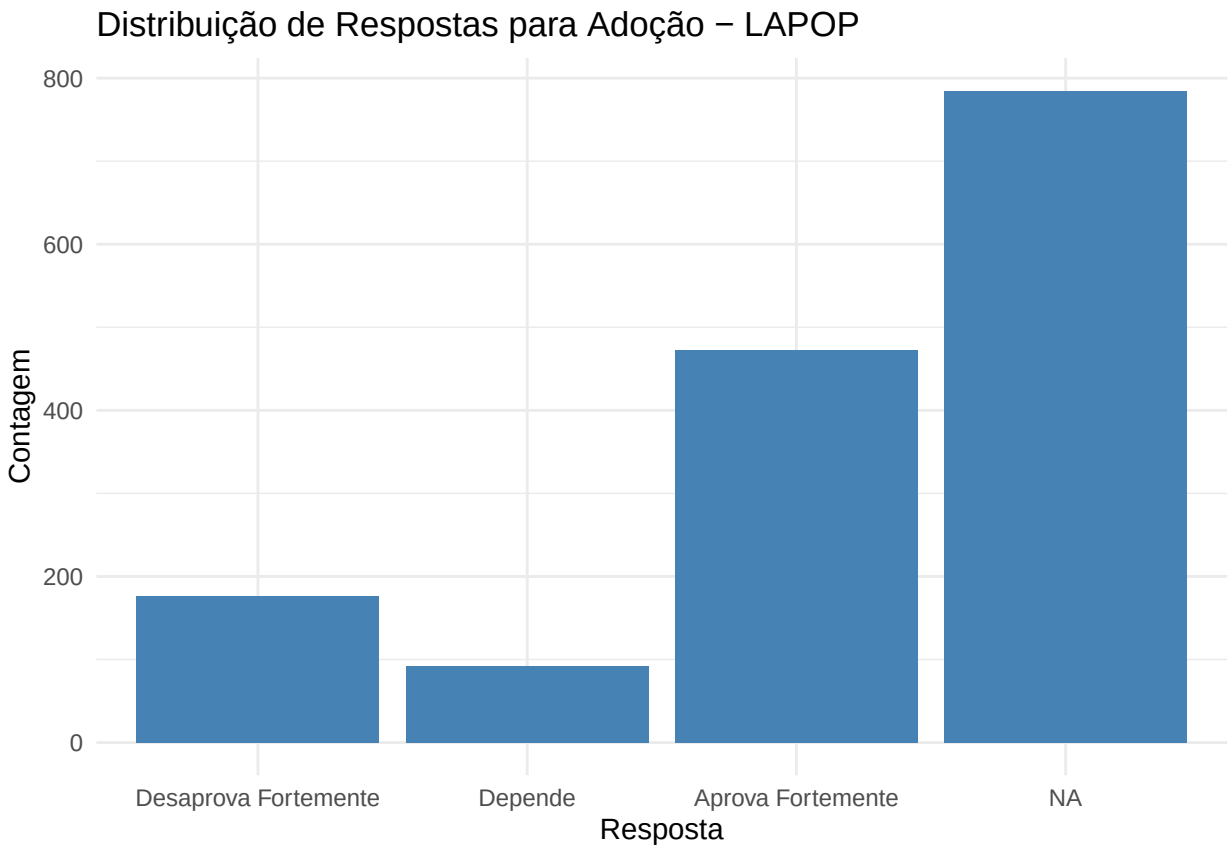


Figura 4: FONTE: base LAPOP, 2023.

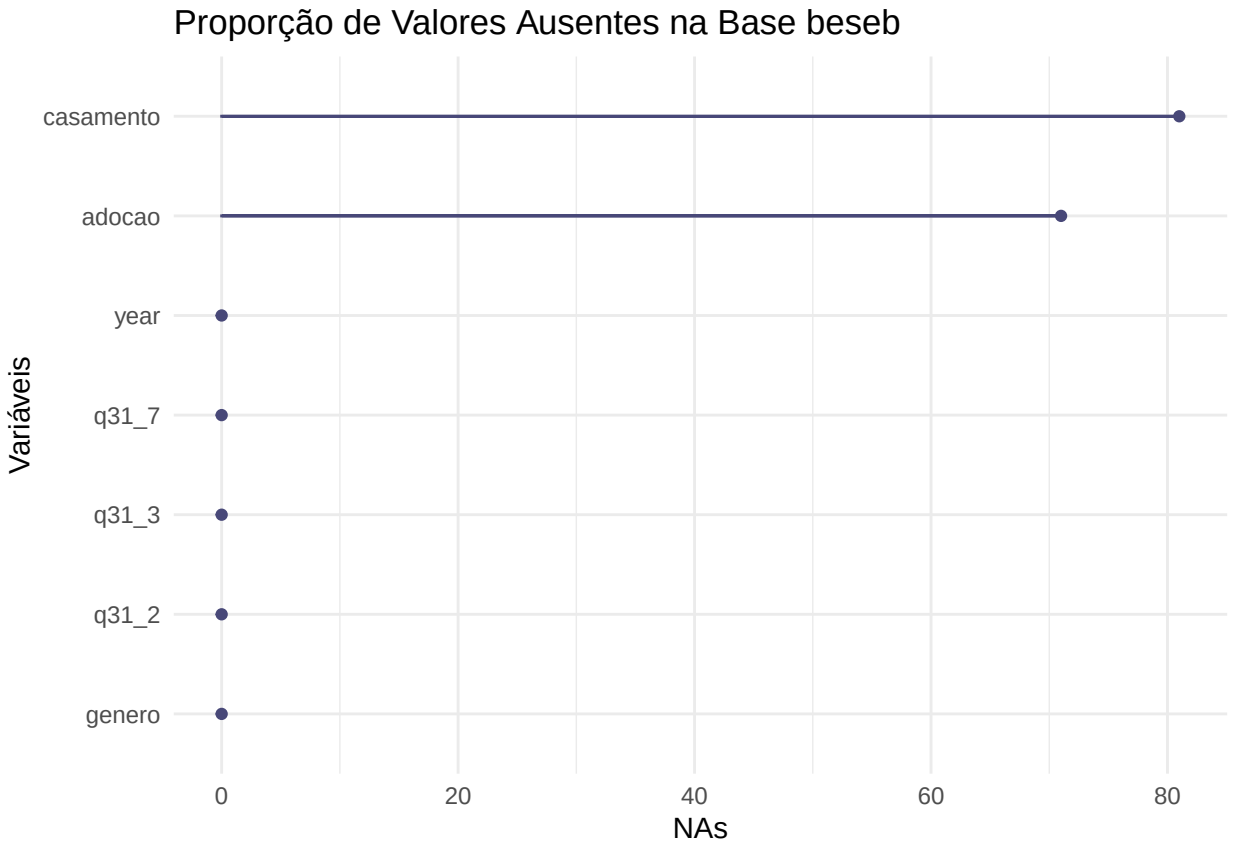
Questão 2 (27 junho)

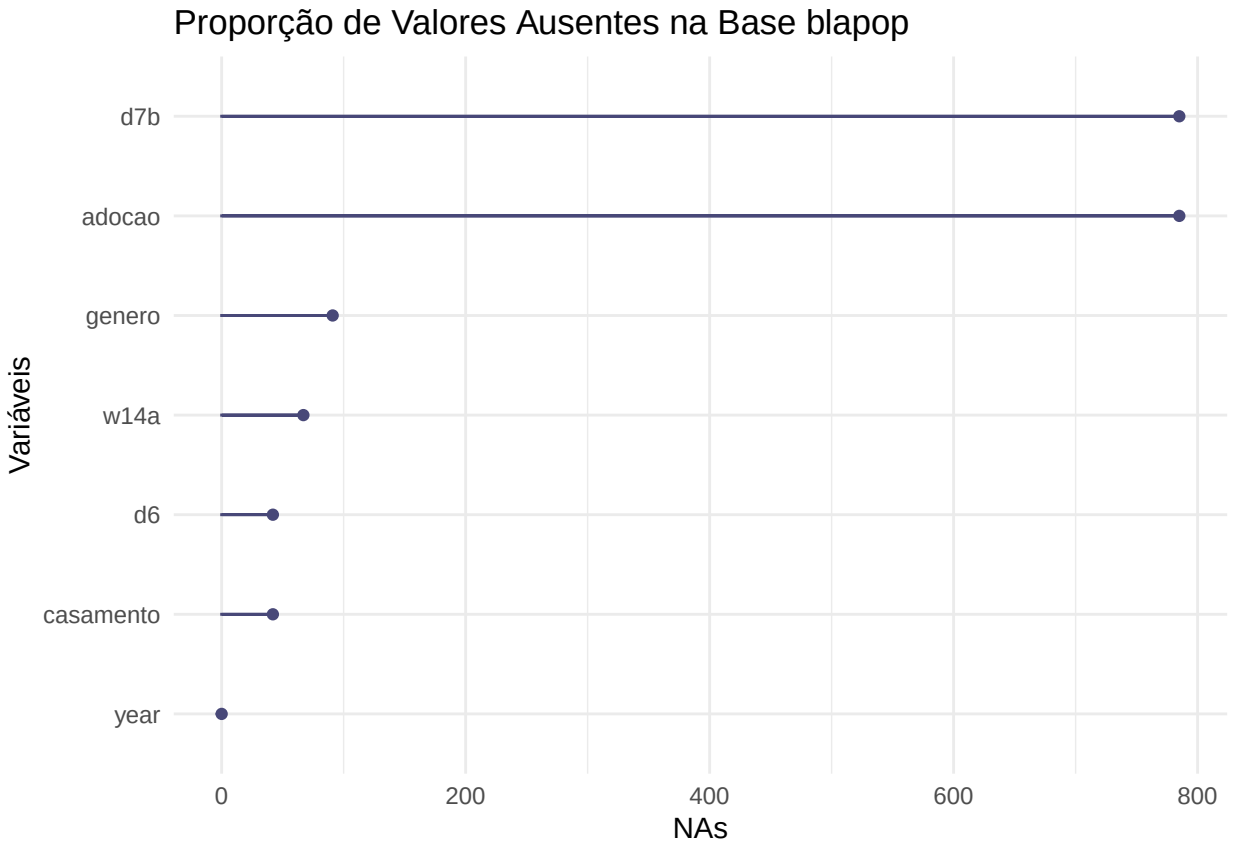
Analise e proponha soluções para os casos ausentes (NAs).

Utilizando o pacote *naniar*, verificamos a presença de valores ausentes (NA) nas bases de dados "beseb" e "blapop". A alta proporção de valores ausentes na variáveis pode ter várias razões possíveis: Não resposta ou omissão pelos respondentes: Algumas pessoas podem optar por não responder perguntas específicas, especialmente se consideram o tema sensível ou controverso. Problemas na coleta de dados: Erros durante a coleta de dados podem levar à ausência de informações. Isso pode ocorrer por falhas na instrução aos entrevistadores, problemas técnicos na coleta de dados ou até mesmo pela natureza complexa da pergunta. Desconhecimento ou falta de informação: Os respondentes podem não estar familiarizados com o conceito de "cotas" ou não ter informações suficientes para responder adequadamente. A alta proporção de valores ausentes pode comprometer a validade e a robustez das análises que utilizam essas variáveis.

variable	n_miss	pct_miss
casamento	81	4.05
adocao	71	3.55
year	0	0
q31_2	0	0
q31_3	0	0
q31_7	0	0
genero	0	0

variable	n_miss	pct_miss
d7b	785	51.4
adocao	785	51.4
genero	91	5.96
w14a	67	4.39
d6	42	2.75
casamento	42	2.75
year	0	0


Proporção de Valores Ausentes na Base beseb
casamento
adocao
year
q31_7
q31_3
q31_2
genero
Variáveis
0
20
40
60
80
NAs



- **Teste Qui-quadrado de um critério**

Para investigar a associação entre diferentes opiniões sobre o casamento e adoção separadamente, utilizei o teste de proporção *prop.test*. Primeiro, transformei a variável qualitativa "casamento" e em variáveis dummy usando o pacote *fastDummies*. Isso criou variáveis binárias (0 ou 1) para cada categoria da variável original. Utilizei a função *table* para criar uma tabela de contingência que mostra a frequência conjunta das respostas nas categorias específicas de interesse da variável "casamento", sendo elas a relação entre "A favor"-"Contra" e "A favor" - "Depende" no conjunto de dados ESEB. e "Desaprova Fortemente" - "Aprova Fortemente" e "Aprova Fortemente"-"Depende" no conjunto de dados LAPOP. Com base nos resultados do teste de proporção indica que há evidências estatísticas suficientes para rejeitar a hipótese nula (H0, $p < 0,05$) de que as proporções são iguais, sugerindo uma associação significativa entre as categorias comparadas das variáveis "casamento" e "adoção".

Para investigar possíveis diferenças na distribuição das variáveis "adocao" e "casamento" entre duas bases de dados distintas, ESEB e LAPOP, utilizei também utilizei o teste χ^2. Primeiramente, calculei as frequências de cada categoria da variável "adocao" em cada amostra. Combinei então essas tabelas de contingência em uma única tabela para análise

comparativa. Após preparar a tabela combinada, realizei o teste χ^2 para determinar se as diferenças nas frequências observadas são estatisticamente significativas. O valor p extremamente baixo $p < 0,05$ indica que há evidências estatísticas suficientes para rejeitar a hipótese nula (H0, $p < 0,05$) de independência entre as amostras para as variáveis "adocao" e "casamento".

- **Teste Qui-quadrado de dois critérios**

Testamos se a preferência do entrevistado sobre casamento homoafetivo e adoção entre casais homoafetivos variam de acordo com o "gênero". O resultado indica que há uma associação estatisticamente significativa entre as variáveis "casamento" e "gênero" em ambas as bases, com um valor p muito baixo $p < 0,05$, $p < 0,05$, indicando que as diferenças nas frequências observadas não são atribuíveis ao acaso. Importante mencionar que o teste da base LAPOP gerou uma mensagem de aviso que pode ser devido à quantidade limitada de casos na categoria "Não se identifica" na variável "gênero".

O resultado indica que há uma associação estatisticamente significativa entre as variáveis "casamento" e "gênero" em ambas as bases, com um valor p muito baixo $p < 0,05$, $p < 0,05$, indicando que as diferenças nas frequências observadas não são atribuíveis ao acaso. Importante mencionar que o teste da base LAPOP gerou uma mensagem de aviso que pode ser devido à quantidade limitada de casos na categoria "Não se identifica" na variável "gênero".

Por fim, aplicamos um teste de Fisher, para pequenas quantidades. Esse valor p extremamente baixo $p < 0,05$ indica que há uma associação estatisticamente significativa entre as variáveis "casamento" e "gênero" A hipótese alternativa de duas caudas sugere que a associação pode ser tanto positiva quanto negativa. Esses resultados reforçam a evidência de que as opiniões sobre "casamento" variam significativamente com base na identificação de gênero utilizando um teste robusto para situações com baixas frequências esperadas.

Referência: **Levin, J.; Fox, J. A. Estatística para Ciências Humanas. São Paulo: Pentice Hall, 2004. 9ª edição.** Na biblioteca: ISBN : 9788581430812