DESCRIÇÃO DE UMA VARIÁVEL CATEGÓRICA

DESCRIÇÃO DE UMA VARIÁVEL CATEGORICA

Uma variável categórica (nominal, qualitativa) registra variações que são distribuídas entre grupos, categorias ou fatores (sim/não; mulher/homem; esquerda/centro/direita; europeu/americano/ africano/asiático) que não se podem utilizar em cálculos aritméticos de modo direto. Nesta secção apresentamos as técnicas de estatística descritiva para o estudo destas variáveis.

1. Antes da Análise: pergunta e dados

Pergunta de investigação

Uma linha importante da literatura sobre perfis ministeriais tem-se focado na expertise dos ministr@s, particularmente nos contextos de crise económica. Um dos argumentos dominantes é que individuos com expertise técnica são considerados mais competentes para a função governativa do que aqueles sem expertise. Porém, a investigação tem-se focado em países grandes ou democracias consolidadas, não havendo muitos estudos sobre países pequenos com democracias recentes. Interessa-nos estudar quão especialistas sos ministros nestes democracias.

> *Quanta expertise tem os ministros de governo nas democracias novas de países pequenos?*

Dados

Para responder à pergunta de investigação formulada, utilizaremos os dados da base "Base_tutorial_1". Como vimos na introdução, esta é uma base fictícia que tem como unidade de análise ministros e ministras de governo de uma democracia nova e pequena, com 16 observações e X variáveis. Interessa-nos examinar a variável '*expert*', que regista se o indivíduo observado tem experiência (ou não) na área da pasta ministerial ocupada.

Com fins didáticos, criamos mais duas variáveis para duas amostras diferentes.

```
exp_a2 <- B$expert
exp_a2[5] <- NA
```

A variável '*exp_a2*' simula a mesma medição para uma segunda amostra.

No R, as variáveis categóricas são chamadas 'factor'. Para a realização das computações, muitas vezes as variáveis 'character' são igualmente válidas, mas podem existir constrangimentos ao uso de certas funções.

Controlamos a classe das variáveis

```
class(B$expert)
class(exp_a2)
```

```
[1]"factor"
```

```
[1] "factor"
```

Controlamos pela presença de valores em falta

```
table (is.na(B$expert))
table (is.na(exp_a2))
```

```
FALSE
   16
FALSE  TRUE
   15     1
```

O resultado indica que não há (é falso que existam) valores em falta para a variável da primeira amostra. Por outro lado, existe um valore em falta para a variável da amostra 2.

Identificamos a extensão da variável ou o número de observações (N) das duas amostras

```
length(B$expert)
length(exp_a2)
```

```
[1] 16
[1] 16
```

2. Dados 'crus' (desagregados)

O '**print**' da variável devolve-nos a distribuição dos seus valores tal como aparecem na base de dados, conforme foram introduzidos. Segue a ordem das linhas da base de dados, correspondendo muitas vezes à cronologia da recolha dos dados.

```
print (B$expert)
```

```
[1]  non-expert  expert      expert     non-expert  non-expert
[6]  expert      expert      expert     non-expert  expert
[11] expert      non-expert  non-expert non-expert  non-expert
[16] non-expert
Levels: expert non-expert
```

O resultado indica que a variável tem duas categorias (levels) e compreende 16 indivíduos (N=16),. Vemos que o primeiro indivíduo era um não-especialista, ou segundo um especialista, e o último um não-especialista, por exemplo.

Como já vimos, olhar para os dados desagregados da-nos uma visualização muito básica dos dados no seu estado natural, pouco prático sobretudo à medida que a extensão da variável cresce. Contudo

este acesso ao 'estado natural' dos dados pode ser um modo rápido de detetar anomalias tais como a presença de algum valor não esperado, o excesso de NAs, ou o enviesamento na recolha (se, por exemplo, uma mesma categoria se encontrar excessivamente agrupada).

3. Análise

Para a análise de uma variável categórica, não podemos aplicar as mesmas ferramentas que aplicamos para a análise de uma variável numérica, tais como médias e medianas. Contudo, o objetivo básico da descrição estatística mantém-se: caracterizar a distribuição dos valores da variável através da agregação dos dados.

A reação mais intuitiva, e acertada, para a análise de uma variável categórica é contar a quantidade de observações que se agrupam por categoria, ou seja, quantos ministros são especialistas e quantos não são.
Este cálculo é viabilizado pela ferramenta estatística mais básica para a análise de uma categórica, isto é, as tabelas de frequência.

Tabela de frequência absoluta

Estas tabelas resumem as frequências absolutas de cada categoria: quantos casos em cada categoria.

```
t1 <- table (B$expert)
t1
```

```
non-expert expert
9            7
```

Do total dos indivíduos observados, vemos que 9 são não especialistas e 7 são especialistas.

Mas nestas situações, referirmo-nos a número absolutos continua a ser pouco informativo. Por exemplo, dizermos que há 7 especialistas num total de 16 ministros/as não é o mesmo que dizermos que há 7 especialistas num conjunto de 70. No primeiro caso, estamos perante uma percentagem de 43,75% de ministros/as que são especialistas (7 ÷ 16 × 100), enquanto no segundo exemplo estamos perante uma percentagem de 10% (7 ÷ 70 × 100). Torna-se assim relevante olhar para as proporções.

Tabela de frequência relativa

A frequência relativa ou proporção corresponde basicamente ao rácio entre a frequência absoluta e o total de casos observados:

```
prop.table(t1 )
```

```
non-expert     expert
0.5625         0.4375
```

Tipicamente a frequência relativa é transformada em percentagem:

```
prop.table(t1 )*100
```

```
non-expert     expert
56.25          43.75
```

Arredondar, por exemplo a uma casa decimal, costuma a ser muito prático:

```
round (prop.table(t1 )*100, digits = 1)
```

```
non-expert     expert
56.2           43.8
```

Assim como o também é útil visualizar o número total de observações:

```
addmargins(t1)
```

```
non-expert     expert    Sum
  9            7         16
```

A tabela completa:

```
addmargins(t1)

round (addmargins(prop.table(t1 )*100),1)
```

```
non-expert     expert    Sum
  9            7         16

non-expert     expert    Sum
 56.2          43.8      100.0
```

Esta tabela completa dá-nos as frequências absolutas por categoria (i.e., o número de casos especialistas e não-especialistas), o número total de observações, as percentagens arredondas a uma casa decimal (tendo por base as frequências relativas) e o total (i.e., 100%).

Gráfico de barras (`barplot`)

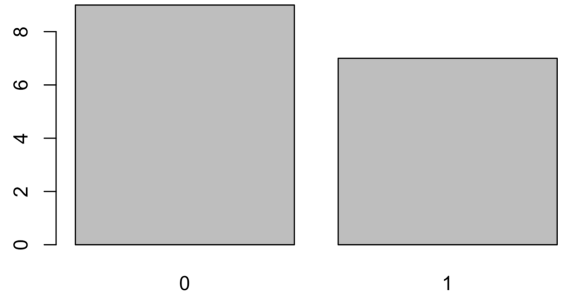
Os gráficos de barras é a ferramenta adequada para a representação gráfica de dados categóricos. [1]

As barras podem estar na vertical ou na horizontal, sendo que um dos eixos identifica as categorias e o outro o valor que sinaliza a quantidade associada à categoria (frequências absolutas e relativas), respondendo normalmente à pergunta de "quantos?" em cada categoria.

O uso de gráfico de barras pode tornar-se problemático quando há um grande número de categorias. Há quem sugira que para variáveis nominais as barras devem ser horizontais, ao passo que para variáveis ordinais (categóricas ordinais) as barras devem ser verticais e apresentadas de forma ordenada. (Daniela. cita?)

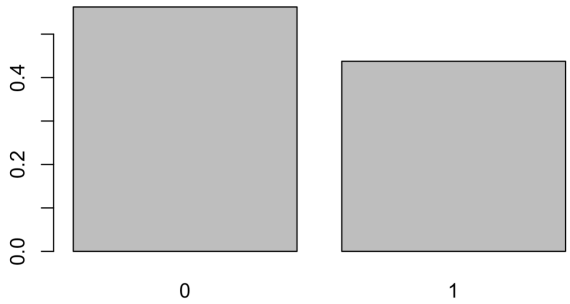
Um gráfico de barras com frequências:

```
barplot (t1 )
```



O mesmo gráfico de barras com proporções:

```
barplot (prop.table(t1))
```



[1] É muito importante pensar qual a forma mais eficaz, clara e transparente de representar os nossos dados. Neste link estão muito boas práticas quanto à forma adequada de se usar cada tipo de gráfico: https://depictdatastudio.com/charts/ [Daniela]

Gráfico de área ou mosaico

Outras representações visuais incluem os gráficos de área, também chamados de gráficos de mosaico. Estes gráficos permitem uma rápida perceção da porção relativa de cada categoria no seu todo.

Cada coluna representa uma categoria da variável, entanto que a largura indica a densidade ou proporção de casos incluídos em cada uma das categorias

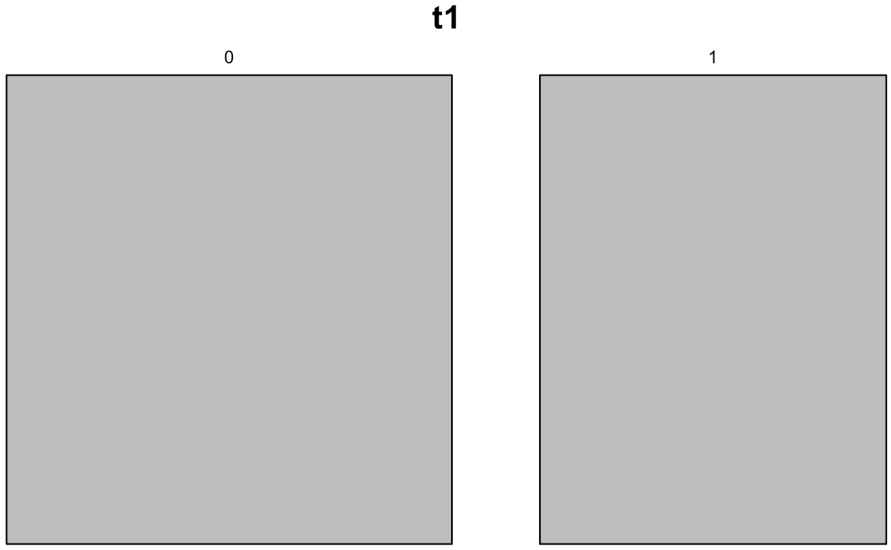
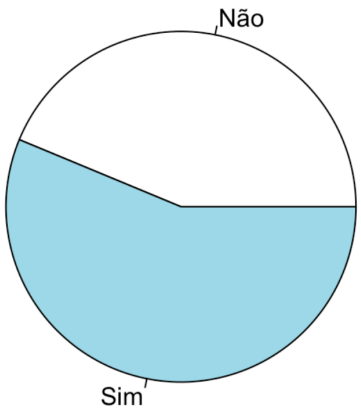


Gráfico circulares (pie chart)

Embora sejam muito frequentes, os gráficos circulares são vistos como limitados em muitas disciplinas, pela sua dificuldade para interpretar variáveis com mais de duas categorias.

```
pie (t1, labels = c("Não", "Sim"))
```



4. EXERCÍCIOS

Secção PANORAMA

1) Identificar e escolher 5 artigos vinculados a um tópico do seu interesse, que utilizem análise estatística.
2) Para cada um dos 5 artigos escolhidos, identificar que tipo de análise aplicam (descritivo, explicativo) e que técnicas e ferramentas (coeficinetes, regressões, boxplots, etc.).
3) Para 2 dos 5 artigos escolhidos,
   a) identificar e descrever a base de dados utilizados (unidade de observação, número de observações, variáveis) e
   b) reconstruir o codebook incluindo pelo menos 5 variáveis.

Secção UMA QUANTITATIVA

4) Para 1 dos 5 artigos escolhidos, identificar a descrição de uma variável numérica. Transcrever essa descrição.

Secção UMA QUALITATIVA

5) Para 1 dos 5 artigos escolhidos, identificar a descrição de uma variável qualitativa. Transcrever essa descrição.

Secção ANALISE BIVARIADA

6) .....

5. Referências

OpenIntro

6. APÊNDICE (R)

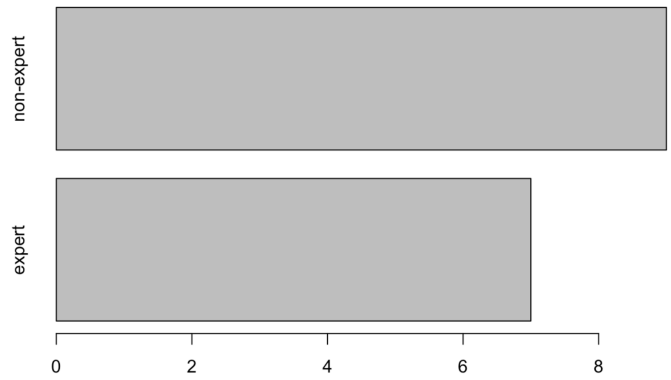
Gráficos com Pacote Base

Gráfico de Barras

http://www.sthda.com/english/wiki/bar-plots-r-base-graphs

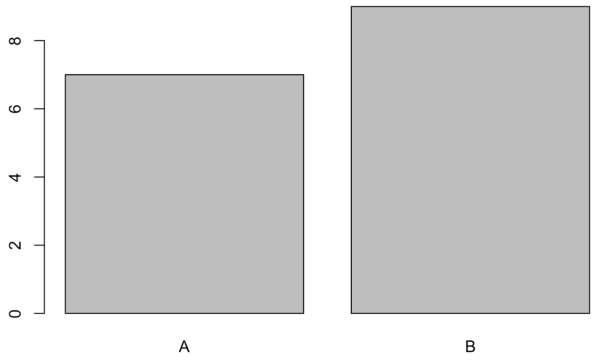
Posicionamento horizontal:

```
barplot(t1 , horiz = TRUE)
```



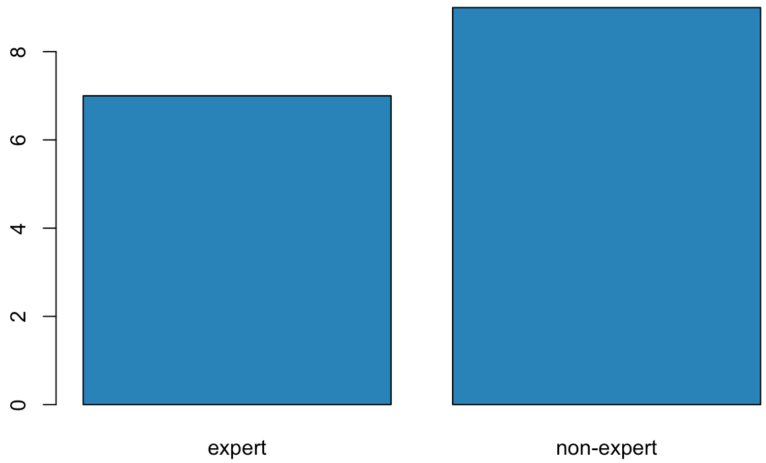
Nomeação das categorias:

```
barplot(t1, names.arg = c("A", "B"))
```



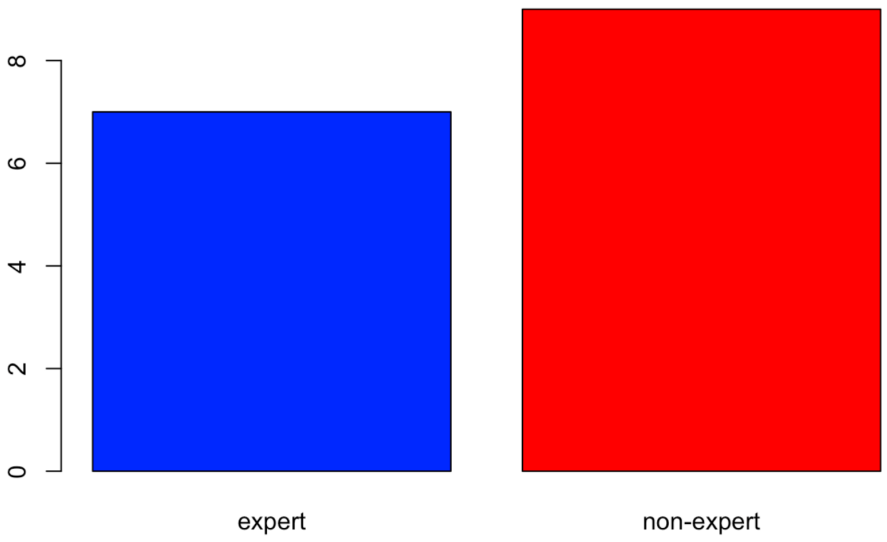
Cor interna

```
barplot(t1, col = "steelblue")
```



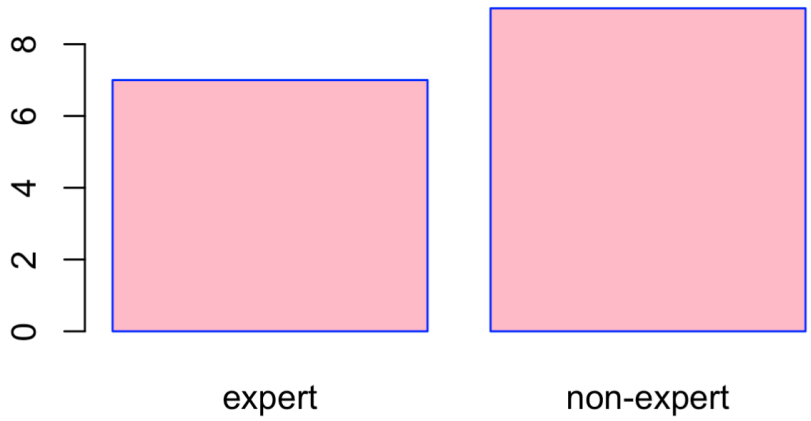
Cor interna por categoria

```
barplot(t1, col = c("blue", "red"))
```



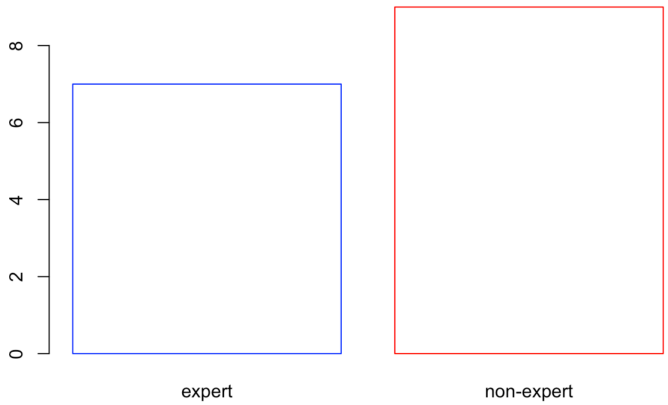
Cor interna e das bordas:

```
barplot(t1,
        col = "pink",
        border = "blue") # Uma unica cor
```



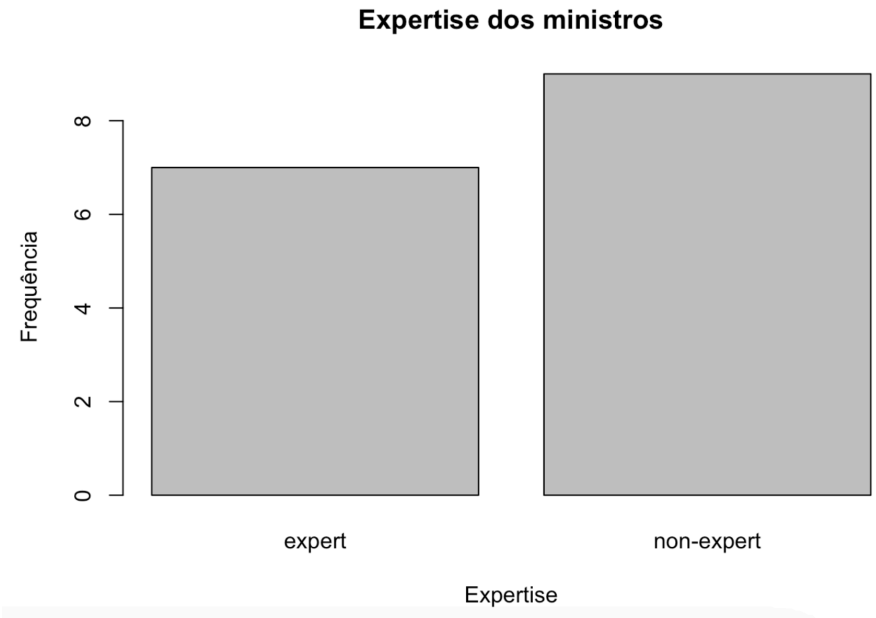
Cor das bordas, por categoria

```
barplot(t1,
        col = "grey",
        border = c("blue", "red")) # Cor por categoria
```



Título e etiquetas

```
barplot(t1,
        main = "Expertise dos ministros",
        xlab = "Expertise",
        ylab = "Frequência")
```



Stacked bar plots (PARA BIVARIADA)

```
barplot(VADeaths,
        col = c("lightblue", "mistyrose", "lightcyan",
                "lavender", "cornsilk"),
        legend = rownames(VADeaths))
```