

AULAS 28 E 29

Construindo e compreendendo tabelas

Ernesto F. L. Amaral

22 e 28 de junho de 2012
Metodologia (DCP 033)

Fonte:

Babbie, Earl. 1999. “Métodos de Pesquisas de *Survey*”. Belo Horizonte: Editora UFMG. pp.337-361.

ANÁLISE MULTIVARIADA

- A maioria das análises de *survey* utilizam análise multivariada para explorar os dados disponíveis.
- Esse termo se refere ao exame simultâneo de diversas variáveis.
- A análise das associações simultâneas entre idade, escolaridade e preconceito é um exemplo de análise multivariada.
- Essa análise utiliza diferentes técnicas estatísticas para desenvolver inferências descritivas e causais.
- A análise multivariada pode ser realizada com o estudo de tabelas simples, chamadas de tabelas de contingência ou tabulações cruzadas.
- Mesmo as análises univariada e bivariada permitem explorar de diversas formas o banco de dados.

ANÁLISE UNIVARIADA

- Análise univariada é o exame da distribuição de casos de apenas uma variável de cada vez.
- O formato mais básico para apresentar dados univariados é relatar todos os casos individuais, listando o atributo (categoria) de cada caso estudado na variável em questão.
- Ao invés de apresentar a lista com todos os dados, podemos distribuir os dados em tabelas de frequência sem perder qualquer detalhe.
- Podemos apresentar distribuições de frequências de dados agrupados (dados marginais):
 - Neste caso, temos menos dados para examinar e interpretar, mas não podemos reproduzir todos os dados originais.
 - Esses dados marginais podem ser apresentados em valores absolutos (brutos) ou porcentagens.

CASOS SEM RESPOSTA

- No caso de faltarem dados para alguns casos do banco, podemos:
 - Apresentar as porcentagens sobre o número total de respondentes, relatando os que não responderam como porcentagens do total.
 - Usar o número de pessoas que responderam à pergunta como a base sobre a qual computar as porcentagens.
- A escolha da base (denominador) depende do propósito da análise.
- Se objetivo for comparar a distribuição da amostra com dados da população, provavelmente omitiremos os casos “sem resposta” (“missings”) da análise.
- Como “sem resposta” em geral não é uma categoria significativa, sua presença entre as categorias de base confunde a comparação com dados da população.

Tabela 1. Ilustração de análise univariada, Brasil - 2010.

Idades de Executivos de Empresas	Distribuição Percentual
Abaixo de 35	9,0
36-45	21,0
46-55	45,0
56-65	19,0
66 ou mais	6,0
Total absoluto	433
Sem dados (valor absoluto)	18

Fonte: Dados hipotéticos, 2010.

TENDÊNCIA CENTRAL

- Além de informar dados marginais, é possível apresentar dados na forma de **médias** resumidas ou medidas de **tendência central**:
 - **Moda**: atributo mais freqüente, agrupado ou não agrupado.
 - **Média aritmética**: soma dos dados de uma variável, dividida pelo total de observações no banco.
 - **Mediana**: é o atributo do meio na distribuição, a qual deve estar ordenada pelos atributos observados.
- Geralmente os estudos mostram as médias e medianas.
- As médias sofrem efeito de valores extremos (“outliers”).
- Se entrevistamos 31 adolescentes de 13 a 19 anos, qual é a idade deles em geral?
 - Podemos informar a moda, média e/ou mediana.

Tabela 2. Distribuição de adolescentes por idade.

Idade	Freq.	Moda (mais freqüente)	Média (média aritmética)	Mediana (ponto médio)
13	3	---	$13 \times 3 = 39$	1-3
14	4	---	$14 \times 4 = 56$	4-7
15	6	---	$15 \times 6 = 90$	8-13
16	8	---	$16 \times 8 = 128$	14-21
17	4	---	$17 \times 4 = 68$	22-25
18	3	---	$18 \times 3 = 54$	26-28
19	3	---	$19 \times 3 = 54$	29-31
Total	31	16	492 (total) / 31 (casos) = 15,87	16 ou...

Fonte: Dados hipotéticos.

CÁLCULO DA MEDIANA

- O entrevistado do meio é o número 16, já que existem quinze pessoas mais jovens e quinze mais velhos.
- O grupo com o 16º indivíduo é composto pelos indivíduos: 14º, 15º, 16º, 17º, 18º, 19º, 20º e 21º.
- Vamos considerar que esses jovens estão distribuídos igualmente na idade de 16 anos, já que não sabemos suas idades exatas.
- Eles terão intervalos de 0,125 anos ($1/8$) entre um e outro, se posicionando exatamente no meio do intervalo (0,0625):
 - O 14º indivíduo tem 16,0625 anos ($16 + 0,0625$).
 - O 15º indivíduo tem 16,1875 anos ($16,0625 + 0,125$).
 - O 16º indivíduo tem 16,3125 anos ($16,1875 + 0,125$).
- Portanto, a idade mediana do grupo é de 16,3125 anos.

DISPERSÃO

- As medidas de tendência central oferecem a vantagem de apresentar os dados brutos de uma forma mais simples.
- Um único número pode representar todos os dados coletados sobre uma variável.
- Porém, não podemos reconstruir os dados originais a partir da média ou mediana.
- Podemos compensar essa debilidade pela informação sobre dispersão das respostas.
- A medida mais simples da dispersão é a **amplitude** que é a distância entre o valor mais alto e o mais baixo.

DESVIO PADRÃO

- Uma medida mais sofisticada de dispersão é o **desvio padrão**, que é uma medida de variação dos valores em torno da média.
- Há o desvio padrão populacional (σ) e o desvio padrão amostral (s).

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$$

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

- A divisão por $n - 1$ aparece quando desejamos que o desvio padrão amostral (s) seja um estimador não tendencioso do desvio padrão populacional (σ).
- Em amostras aleatórias: 68% das observações estão dentro do intervalo de 1 desvio padrão para cima e para baixo da média; 95% entre 2 desvios padrão; e 99,9% entre 3 desvios padrão.

VARIÁVEIS CONTÍNUAS E DISCRETAS

- Algumas das medidas propostas acima não são apropriadas para todas variáveis, já que existem variáveis contínuas e discretas:
 - Idade é uma variável contínua de razão.
 - Sexo é uma variável discreta.
- Medianas devem ser calculadas apenas para **dados de intervalo**.
- Médias devem ser calculadas apenas para **dados de razão**.
- Para **dados nominais e ordinais**, podemos apresentar números brutos ou porcentagens marginais.
- A moda é uma análise correta para todos tipos de dados, mas algumas vezes não será muito informativa.

DESCRIÇÕES DE SUBGRUPOS

- Análises univariadas servem para **descrever** a amostra do *survey* e, por extensão, a população da qual foi extraída.
- Análises bivariadas e multivariadas objetivam temas **explicativos** (relações causais).
- Há o **caso intermediário** de descrição de subgrupos, em que podemos examinar separadamente as respostas à uma questão por grupos da amostra.
- Ao computar as descrições estratificadas, os passos são dados independentemente para cada subgrupo.
- Cada grupo é submetido a uma análise univariada simples.
- Distribuições de frequência para subgrupos são chamadas de **marginais estratificadas**.
- O principal objetivo é de comparar os grupos, com o pressuposto de que a variável de estratificação terá algum efeito causal (**explicação**) sobre a variável de descrição.

FUNDINDO CATEGORIAS DE RESPOSTAS

- Podemos construir tabelas que possuem grande quantidade de informações de diferentes grupos, o que dificulta perceber algum padrão significativo.
- Esse tipo de problema pode ocorrer quando pequenas porcentagens de entrevistados selecionam alguma categoria de resposta.
- Nestes casos, devemos combinar ou fundir as categorias extremas ou semelhantes.
- O ideal é somar as frequências brutas e recalculas as porcentagens para as categorias combinadas.

**Tabela 3. Atitude em relação às Nações Unidas:
 “Como a ONU está resolvendo os problemas
 que ela tem que enfrentar?”
 (Distribuição percentual)**

Avaliação	Alemanha Ocidental	Inglaterra	França	Japão	Estados Unidos
Muito bom	2,0	7,0	2,0	1,0	5,0
Bom	46,0	39,0	45,0	11,0	46,0
Ruim	21,0	28,0	22,0	43,0	27,0
Muito ruim	6,0	9,0	3,0	5,0	13,0
Não sei	25,0	17,0	28,0	40,0	9,0

Fonte: 5-Nation Survey Finds Hope for U.N. New York Times, p.6, 26 June 1985.

**Tabela 4. Atitude em relação às Nações Unidas:
 “Como a ONU está resolvendo os problemas
 que ela tem que enfrentar?”
 (Distribuição percentual combinada)**

Avaliação	Alemanha Ocidental	Inglaterra	França	Japão	Estados Unidos
Trabalho bom ou melhor	48,0	46,0	47,0	12,0	51,0
Trabalho ruim ou pior	27,0	37,0	25,0	48,0	40,0
Não sei	25,0	17,0	28,0	40,0	9,0

Fonte: 5-Nation Survey Finds Hope for U.N. New York Times, p.6, 26 June 1985.

LIDANDO COM OS “NÃO SEI”

- Em pesquisas de *survey*, geralmente é bom dar às pessoas a opção de dizer “não sei”, quando se pede suas opiniões sobre certos assuntos.
- Porcentagens substanciais respondendo “não sei” podem confundir os resultados de uma tabela.
- Mesmo sem os dados absolutos, há um modo fácil de calcular as porcentagens excluindo os “não sei”:
 - 1) Calcular proporção dos que apresentaram uma opinião.
 - 2) Dividir o percentual observado pela proporção acima.
- Por exemplo, 48% dos alemães disseram que a ONU está resolvendo problemas de forma boa ou melhor; 27% de forma ruim ou pior; e 25% não opinaram:
 - 1) Proporção com opinião: $(100\% - 25\%) / 100 = 0,75$.
 - 2) Percentual de avaliação positiva: $48\% / 0,75 = 64\%$.
 - 3) Percentual de avaliação negativa: $27\% / 0,75 = 36\%$.

**Tabela 5. Atitude em relação às Nações Unidas:
 “Como a ONU está resolvendo os problemas
 que ela tem que enfrentar?”
 (Distribuição percentual combinada, omitindo “não sei”)**

Avaliação	Alemanha Ocidental	Inglaterra	França	Japão	Estados Unidos
Trabalho bom ou melhor	64,0	55,0	65,0	20,0	56,0
Trabalho ruim ou pior	36,0	45,0	35,0	80,0	44,0

Fonte: 5-Nation Survey Finds Hope for U.N. New York Times, p.6, 26 June 1985.

QUAL É A VERSÃO CERTA?

- Qual versão das tabelas anteriores seria a correta?
 - Distribuição original.
 - Distribuição combinada.
 - Distribuição combinada, omitindo os “não sei”.
- A melhor versão depende do objetivo da análise e de interpretação dos dados.
- Se não for essencial distinguir entre “muito bom” e “bom”, faz sentido utilizar distribuição combinada.
- O fato de uma grande porcentagem de japoneses não ter opinado pode ser uma descoberta importante, o que nos faria continuar com categoria “não sei”.
- Se quisermos saber como pessoas votariam uma questão, excluiríamos os “não sei”.
- Em geral, é correto informar os dados de ambas formas: com e sem os “não sei”.

ANÁLISE BIVARIADA ≠ DESCRIÇÃO DE SUBGRUPOS

- A análise bivariada explicativa é, basicamente, a mesma coisa que descrição de subgrupos, com certas restrições.
- Nas descrições de subgrupos, temos liberdade para escolher a variável de estratificação e descrever cada subgrupo nos termos de outra variável.
- Os dados a seguir são exemplos de descrições de subgrupos.

Tabela 6. “Você concorda ou discorda da proposição de que homens e mulheres devem ser tratados igualmente em todos os aspectos?” (Distribuição percentual)

Opinião	Homens	Mulheres
Concordam	63,0	75,0
Discordam	37,0	25,0
Total absoluto	400	400

Fonte: Dados hipotéticos.

Tabela 7. “Você concorda ou discorda da proposição de que homens e mulheres devem ser tratados igualmente em todos os aspectos?” (Distribuição percentual)

Opinião	Concordam	Discordam
Homens	46,0	60,0
Mulheres	54,0	40,0
Total absoluto	552	248

Fonte: Dados hipotéticos.

ANÁLISE BIVARIADA

- Numa análise bivariada explicativa, somente a primeira tabela faria sentido:
 - Geralmente, mulheres têm status inferior na sociedade americana; portanto, elas apoiariam mais a proposta de igualdade entre sexos.
 - O sexo do respondente afeta sua resposta ao item do questionário; mulheres têm maior probabilidade de aprovar do que os homens.
 - Se os respondentes homens e mulheres forem descritos separadamente em termos de suas respostas, uma porcentagem maior de mulheres do que de homens aprovaria.

PRIMEIRA TABELA ≠ SEGUNDA TABELA

- A primeira tabela divide a amostra em dois subgrupos (homens e mulheres) e descreve as atitudes dos dois separadamente:
 - As porcentagens são comparadas e vemos que mulheres têm maior probabilidade de aprovar do que homens.
- Na segunda tabela, a lógica seria de que as atitudes relacionadas à igualdade sexual afetam o sexo:
 - Aprovar a igualdade sexual tende a tornar a pessoa mais mulher do que homem, o que é um raciocínio ilógico.
 - O sexo dos respondentes é predeterminado bem antes de se formarem atitudes sobre igualdade sexual.
 - Porém, a segunda tabela seria legítima para descrição de subgrupos.

VARIÁVEIS INDEPENDENTES E DEPENDENTES

- Nosso objetivo é de explicar valores da variável dependente (y), baseado nos valores da variável independente (x).
- A variável independente causa a variável dependente, com uma determinada chance (probabilidade).
- No exemplo anterior, sexo (variável independente) causa atitudes sobre igualdade sexual (variável dependente).
- Algumas vezes, é difícil determinar qual variável é dependente e qual é independente.
- De todo modo, toda tabela bivariada explicativa implicitamente designa uma variável independente e uma dependente.

DETERMINANDO INDEPENDENTES E DEPENDENTES

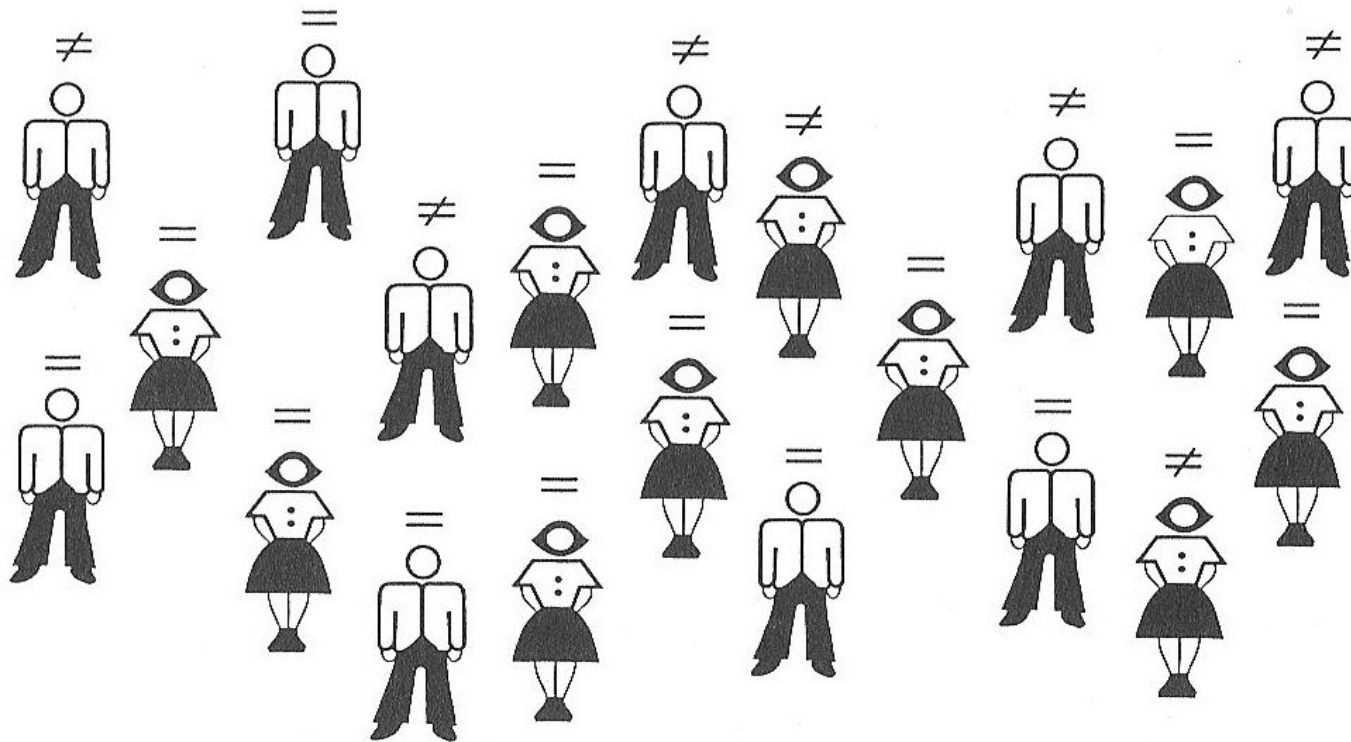
- Sempre que houver uma clara ordem temporal relacionando as duas variáveis:
 - A variável cujos valores são determinados antes é a variável independente.
 - A variável cujos valores são determinados depois é sempre a variável dependente.
- Esse ponto está ligado à idéia de que a ciência (assim como a pesquisa de *survey*) é lógica.
- Uma implicação disso é que duas variáveis ocorrendo ao mesmo tempo não podem ser ligadas causalmente.
- O sexo e a raça de uma pessoa não podem ser analisados de forma explicativa.
- Em situações em que a ordem temporal das variáveis não é clara, a designação de variáveis independentes e dependentes precisa ser feita em bases lógicas.

CONSTRUÇÃO DE TABELAS

- Antes de tudo, é preciso designar uma variável como dependente e outra como independente.
- A construção de tabelas segue estes passos:
 - A amostra é dividida em valores ou categorias da variável independente.
 - Cada subgrupo é descrito em termos dos valores ou categorias das variáveis dependentes.
 - A tabela é lida comparando-se os subgrupos da variável independente em termos de um valor da dependente.

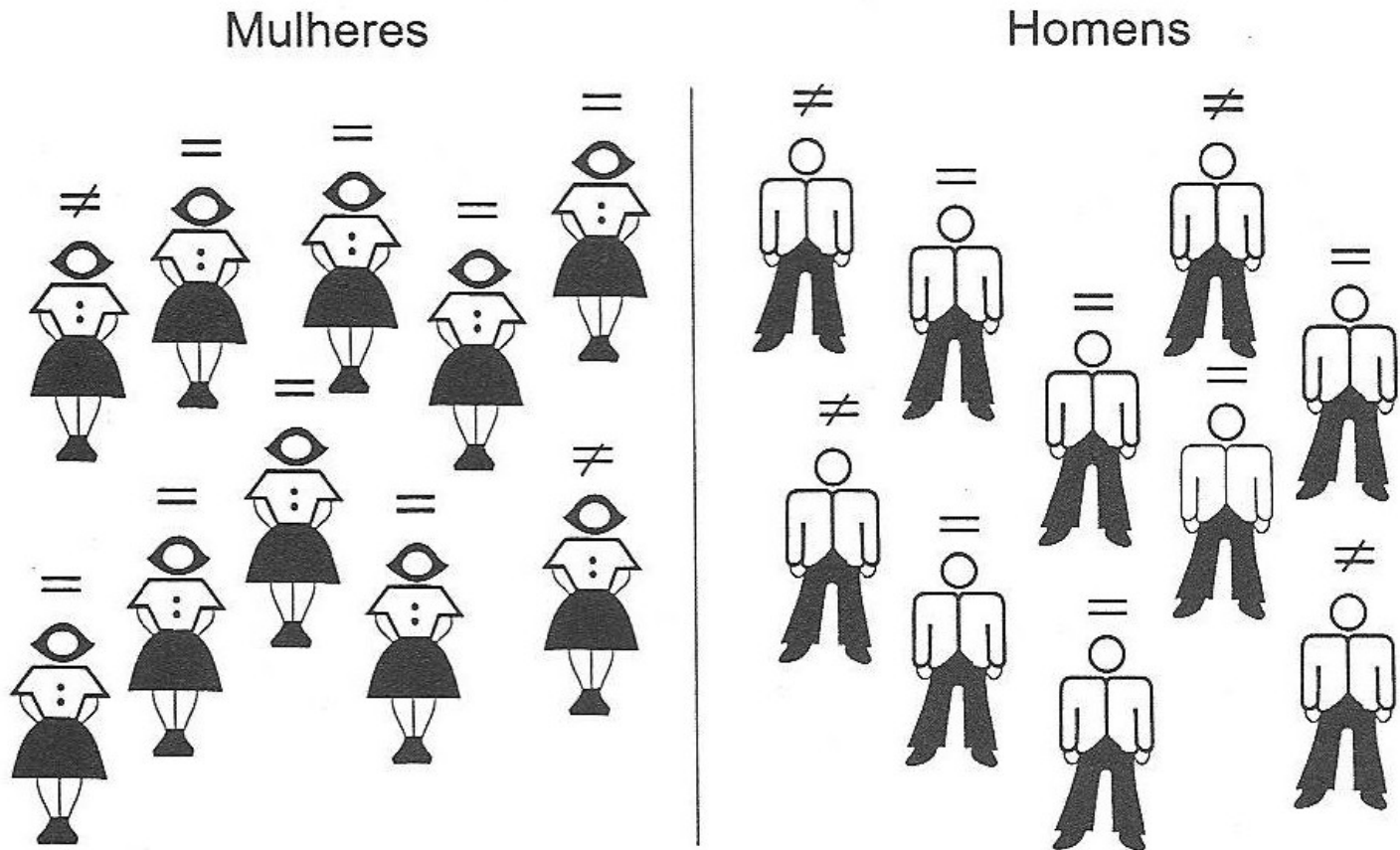
PERCENTUANDO UMA TABELA

A. Alguns homens e mulheres que ou são a favor (sinal de igual) da igualdade sexual, ou são contra (\neq)



PERCENTUANDO UMA TABELA

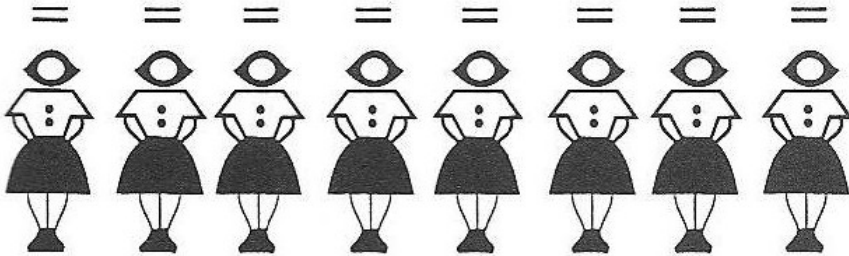
B. Separar homens e mulheres (a variável independente)



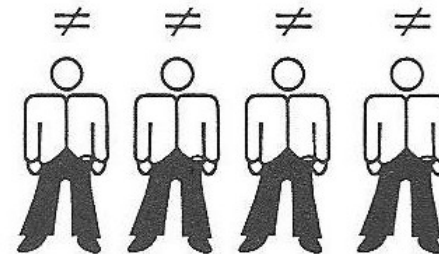
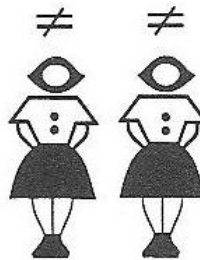
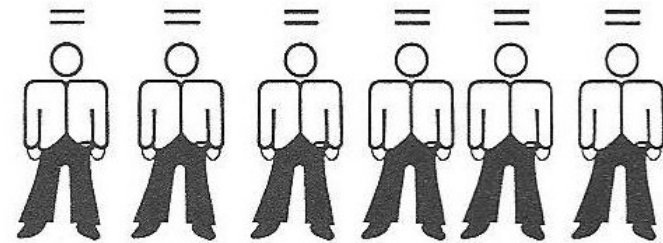
PERCENTUANDO UMA TABELA

C. Em cada grupo do gênero, separar os que são a favor da igualdade sexual dos que são contra (variável dependente)

Mulheres



Homens

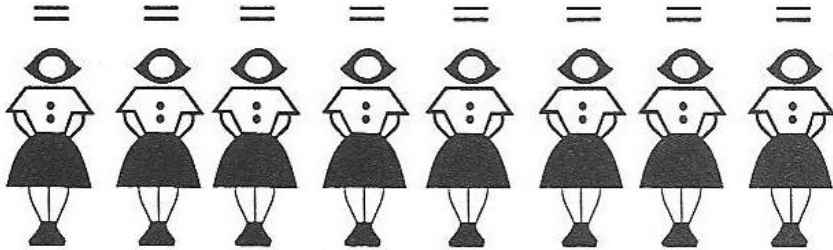
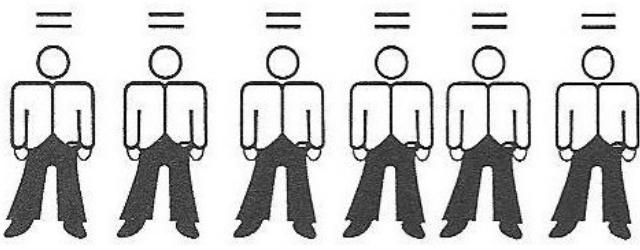
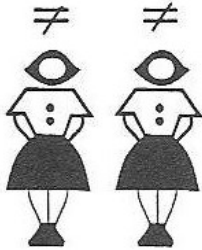
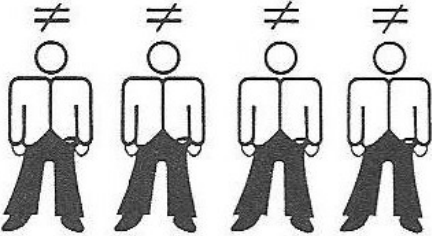


PERCENTUANDO UMA TABELA

D. Conte os números em cada célula da tabela

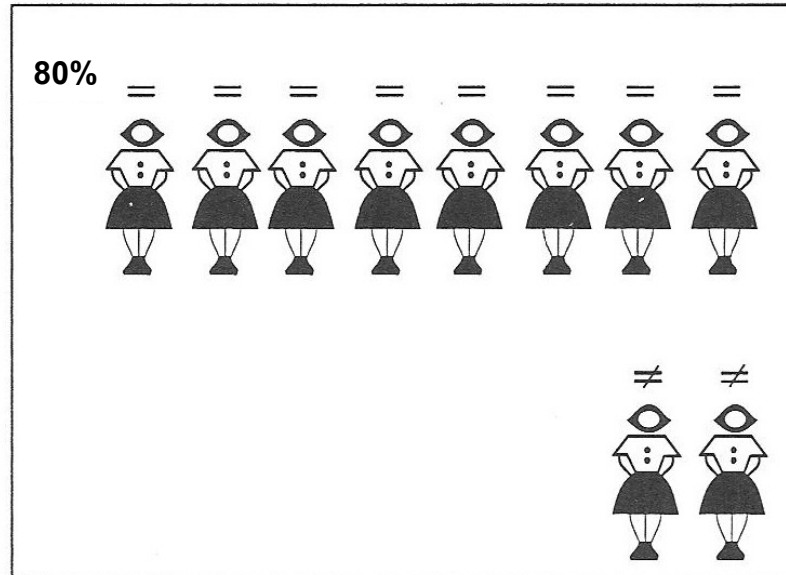
Mulheres

Homens

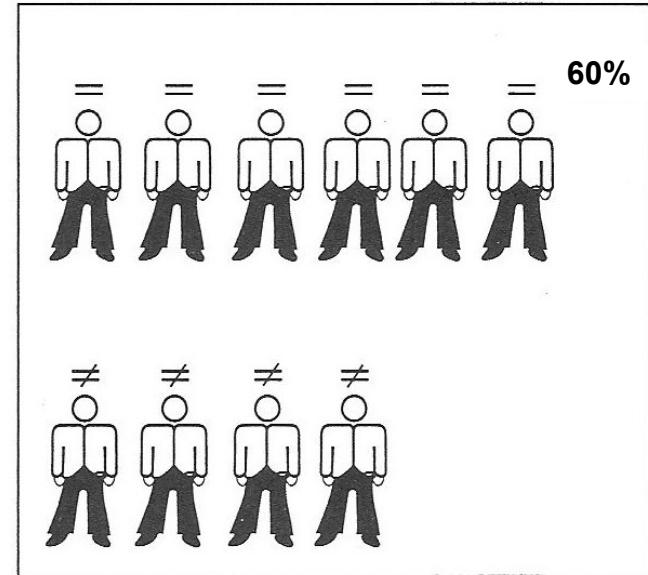
<p>8</p> 	 <p>6</p>
<p>2</p> 	 <p>4</p>

PERCENTUANDO UMA TABELA

E. Qual a porcentagem das mulheres é a favor da igualdade?



F. Qual a porcentagem dos homens é a favor da igualdade?



G. Conclusões

Enquanto a maioria dos homens e mulheres são a favor da igualdade sexual, mulheres tem maior chances de o ser.

Portanto sexo parece ser uma das causas da atitude em relação a igualdade sexual.

	Mulheres	Homens
A favor da igualdade	80 %	60 %
Contra a igualdade	20 %	40 %
Total	100 %	100 %

FORMATO DE TABELAS BIVARIADAS

- Não há formato de apresentação padronizado para tabelas de porcentagens.
- Porém, podemos seguir algumas diretrizes:
 - Tabelas devem ter cabeçalhos ou títulos descrevendo seu conteúdo.
 - O conteúdo original das perguntas do questionário deve ser apresentado na própria tabela ou no texto.
 - Os valores ou categorias de cada variável devem ser claramente indicados.
 - Quando são apresentadas porcentagens, deve-se indicar a base (denominador) em que foram computadas.
 - É redundante apresentar todos os números brutos.
 - Informar na tabela se há respondentes “sem resposta”.

ANÁLISE MULTIVARIADA

- Tabelas multivariadas podem ser construídas com base numa descrição mais complicada de subgrupo.
- Em vez de explicar a variável dependente com uma variável independente (como na análise bivariada), temos mais de uma variável independente.
- O primeiro passo é dividir a amostra total em subgrupos baseados nos vários valores das duas variáveis independentes simultaneamente.
- Depois, os vários subgrupos são descritos nos termos da variável dependente, e se fazem as comparações.

Tabela 8. “Você aprova ou desaprova a proposição geral de que homens e mulheres devem ser tratados igualmente em todos os aspectos?” (Distribuição percentual)

Opinião	Mulheres		Homens	
	Abaixo de 30	30 e acima	Abaixo de 30	30 e acima
Aprovam	90,0	60,0	78,0	48,0
Desaprovam	10,0	40,0	22,0	52,0
Total absoluto	200	200	200	200
Sem resposta (valor absoluto)	2	3	10	2

Fonte: Dados hipotéticos.

SIMPLIFICANDO

Tabela 9. “Você aprova ou desaprova a proposição geral de que homens e mulheres devem ser tratados igualmente em todos os aspectos?” (Distribuição percentual)

Porcentagem que Concorda	Mulheres	Homens
Abaixo de 30	90,0 (200)	78,0 (200)
30 e acima	60,0 (200)	48,0 (200)

Fonte: Dados hipotéticos.

Obs.: Números entre parênteses indicam casos sobre os quais as porcentagens são baseadas.

ANÁLISE MULTIVARIADA

- Tabelas parecem muito simples para merecer discussão extensa.
- Porém, elas são complexas, sendo freqüentemente mal-construídas e mal-interpretadas.
- Vimos análises univariada, descrição de subgrupos, explicativa bivariada e multivariada.
- Lembrem-se de:
 - Dividir a amostra em subgrupos baseados nos valores da variável independente.
 - Descrever cada grupo com base nos valores da variável dependente.
 - Comparar subgrupos da variável independente em relação ao valor da variável dependente.
 - Obedecer regra básica: percentue na coluna e interprete na linha; ou percentue na linha e interprete na coluna.