ME414 - Estatística para Experimentalistas

Estatística Descritiva: Tabelas de Frequencia e Gráficos

Prof. Carlos Trucíos ctrucios@unicamp.br ctruciosm.github.io

Instituto de Matemática, Estatística e Computação Científica, Universidade Estadual de Campinas

Aula 2

Métodos Tabulares

Gráficos para variáveis qualitativas

Tabelas e Gráficos para duas variáveis

Para conhecermos a historia que os dados tem para nos contar é necessário primeiro explorar os dados e termos uma visão geral do panorama.

- Para conhecermos a historia que os dados tem para nos contar é necessário primeiro explorar os dados e termos uma visão geral do panorama.
- ▶ Isto é feito com tabelas, gráficos, medidas resumo, etc.

- Para conhecermos a historia que os dados tem para nos contar é necessário primeiro explorar os dados e termos uma visão geral do panorama.
- Isto é feito com tabelas, gráficos, medidas resumo, etc.
- Esse processo é conhecido como Análise Exploratória de Dados (EDA em Ingles) e nos ajuda a identificar alguns modelos que serão usados posteriormente ou a definir melhor as hipóteses a serem testadas.

- Para conhecermos a historia que os dados tem para nos contar é necessário primeiro explorar os dados e termos uma visão geral do panorama.
- Isto é feito com tabelas, gráficos, medidas resumo, etc.
- Esse processo é conhecido como Análise Exploratória de Dados (EDA em Ingles) e nos ajuda a identificar alguns modelos que serão usados posteriormente ou a definir melhor as hipóteses a serem testadas.
- A forma como exploramos os dados depende do tipo de variável e escala de medida.

Suponha que estamos trabalhando com um *dataset* de 1000 observações e entre as variáveis contidas, temos a variável *Cor de olhos* (em que 0 significa negros, 1 significa marrom, 2 significa azul, 3 significa verde e 4 significa outras cores).

Suponha que estamos trabalhando com um *dataset* de 1000 observações e entre as variáveis contidas, temos a variável *Cor de olhos* (em que 0 significa negros, 1 significa marrom, 2 significa azul, 3 significa verde e 4 significa outras cores).

► Faria sentido calcular a média para essa variável? Porque?

Suponha que estamos trabalhando com um *dataset* de 1000 observações e entre as variáveis contidas, temos a variável *Cor de olhos* (em que 0 significa negros, 1 significa marrom, 2 significa azul, 3 significa verde e 4 significa outras cores).

- ► Faria sentido calcular a média para essa variável? Porque?
- ▶ O que poderiamos fazer para termos uma ideia de como se comportam os dados?

Suponha que estamos trabalhando com um *dataset* de 1000 observações e entre as variáveis contidas, temos a variável *Cor de olhos* (em que 0 significa negros, 1 significa marrom, 2 significa azul, 3 significa verde e 4 significa outras cores).

- ► Faria sentido calcular a média para essa variável? Porque?
- O que poderiamos fazer para termos uma ideia de como se comportam os dados?
- Muitas vezes o unico que podemos fazer é contar

Tabelas de Frequência

As tabelas de frequência são utilizadas quando as variáveis são categóricas¹ elas mostram a frequência (a quantidade de vezes) com que cada categória/classe da variável aparece.

Carlos Trucíos (UNICAMP) | ME414 |

¹Contudo, variáveis numéricas podem ser categorizadas

Tabelas de Frequência

As tabelas de frequência são utilizadas quando as variáveis são categóricas¹ elas mostram a frequência (a quantidade de vezes) com que cada categória/classe da variável aparece.

Exemplo

Tabela 1: Distribuição de Frequências dos passagéiros do Titanic segundo sexo.

Sexo	Frequência		
female	466		
male	843		

¹Contudo, variáveis numéricas podem ser categorizadas

► A frequência (*i.e,* o número de observações em cada classe) é também chamada de **frequência absoluta**

- ► A frequência (*i.e,* o número de observações em cada classe) é também chamada de **frequência absoluta**
- ▶ Veja que cada observação pertence unicamente a uma classe

- ▶ A frequência (*i.e,* o número de observações em cada classe) é também chamada de **frequência absoluta**
- ▶ Veja que cada observação pertence unicamente a uma classe
- Muitas vezes, além da frequência absoluta, estamos também interessados na proporção de observações em cada classe, isso é chamado de frequência relativa

- ▶ A frequência (*i.e,* o número de observações em cada classe) é também chamada de **frequência absoluta**
- ▶ Veja que cada observação pertence unicamente a uma classe
- Muitas vezes, além da frequência absoluta, estamos também interessados na proporção de observações em cada classe, isso é chamado de frequência relativa
- Quando reportamos a porcentagem isto é conhecido como frequência percentual = frequência relativa ×100

- ▶ A frequência (*i.e,* o número de observações em cada classe) é também chamada de **frequência absoluta**
- ▶ Veja que cada observação pertence unicamente a uma classe
- Muitas vezes, além da frequência absoluta, estamos também interessados na proporção de observações em cada classe, isso é chamado de frequência relativa
- Quando reportamos a porcentagem isto é conhecido como frequência percentual = frequência relativa ×100

- ▶ A frequência (*i.e,* o número de observações em cada classe) é também chamada de **frequência absoluta**
- ▶ Veja que cada observação pertence unicamente a uma classe
- Muitas vezes, além da frequência absoluta, estamos também interessados na proporção de observações em cada classe, isso é chamado de frequência relativa
- Quando reportamos a porcentagem isto é conhecido como frequência percentual = frequência relativa ×100

Frequência relativa

Frequência relativa classe
$$i = \frac{\text{Frequência absoluta da classe } i}{\text{Total de observações}}$$

Exemplo

Tabela 2: Distribuição de Frequências dos passagéiros do Titanic segundo sexo.

Sexo	Freq. absoluta	Freq. relativa	Freq. percentual
female	466	0.356	35.6
male	843	0.644	64.4

Exemplo

Tabela 2: Distribuição de Frequências dos passagéiros do Titanic segundo sexo.

Sexo	Freq. absoluta	Freq. relativa	Freq. percentual
female	466	0.356	35.6
male	843	0.644	64.4

► Muitas vezes, também é desejável apresentar as frequência acumuladas (que são obtidas somandos as frequências até a *i*-éssima classe)

Exemplo

Tabela 3: Distribuição de Frequências dos passagéiros do Titanic segundo sexo.

Sexo	Freq. absoluta	Freq. relativa	Freq. percentual	Freq. percentual acumulada
female	466	0.36	35.6	35.6
male	843	0.64	64.4	100.0

Também podemos fazer tabelas de frequência com varáveis quantitativas, mas, obviamente, não faremos com que cada observação diferente seja uma classe, **categorizamos** os dados!.

Também podemos fazer tabelas de frequência com varáveis quantitativas, mas, obviamente, não faremos com que cada observação diferente seja uma classe, **categorizamos** os dados!.

Como categorizar? (sem a ajuda de um especialista)

determinar o número de classes (nem muitas nem poucas),

Também podemos fazer tabelas de frequência com varáveis quantitativas, mas, obviamente, não faremos com que cada observação diferente seja uma classe, **categorizamos** os dados!.

Como categorizar? (sem a ajuda de um especialista)

- determinar o número de classes (nem muitas nem poucas),
- determinar a amplitude de cada classe (por padrão usamos a mesma amplitude em todas as classes),

$$\mbox{Amplitude} = \frac{\mbox{Maior valor} - \mbox{Menor valor}}{\mbox{N\'umero de classes}}$$

Também podemos fazer tabelas de frequência com varáveis quantitativas, mas, obviamente, não faremos com que cada observação diferente seja uma classe, **categorizamos** os dados!.

Como categorizar? (sem a ajuda de um especialista)

- determinar o número de classes (nem muitas nem poucas),
- determinar a amplitude de cada classe (por padrão usamos a mesma amplitude em todas as classes),

$$\mbox{Amplitude} = \frac{\mbox{Maior valor} - \mbox{Menor valor}}{\mbox{N\'umero de classes}}$$

 Determinar os limites da classe (uma observação deve pertencer a uma e somente a uma classe)

Tabela 4: Distribuição de Frequências dos passagéiros do Titanic segundo idade (categorizada).

	Freq. absoluta	Freq. percentual	Freq. percentual acumulada
(0,16]	134	10.2	10.2
(16,32]	524	40.0	50.3
(32,48]	269	20.6	70.8
(48,64]	106	8.1	78.9
(64,80]	13	1.0	79.9
NA	263	20.1	100.0
Total	1309	100.0	100.0

Tabela 4: Distribuição de Frequências dos passagéiros do Titanic segundo idade (categorizada).

	Freq. absoluta	Freq. percentual	Freq. percentual acumulada
(0,16]	134	10.2	10.2
(16,32]	524	40.0	50.3
(32,48]	269	20.6	70.8
(48,64]	106	8.1	78.9
(64,80]	13	1.0	79.9
NA	263	20.1	100.0
Total	1309	100.0	100.0

- ▶ [,] significa inclui,
- ► (,), <, > significa não inclui.

Tabela 5: Distribuição de Frequências dos passagéiros do Titanic segundo idade (categorizada).

	Freq. abso- luta	Freq. percen- tual	Freq. percentual acumulada	Freq. percentual (-NA)	Freq. percentual acumulada (-NA)
(0,16]	134	10.2	10.2	12.8	12.8
(16,32]	524	40.0	50.3	50.1	62.9
(32,48]	269	20.6	70.8	25.7	88.6
(48,64]	106	8.1	78.9	10.1	98.8
(64,80]	13	1.0	79.9	1.2	100.0
NA	263	20.1	100.0	0.0	100.0
Total	1309	100.0	100.0	100.0	100.0

► Mas... conhecendo uma das frequências, conhecemos todas as outras através de simples operações!. Por que não utilizar, por exemplo, apenas a freq. relativa e deixar o *leitor* fazer a conta?

- Mas... conhecendo uma das frequências, conhecemos todas as outras através de simples operações!. Por que não utilizar, por exemplo, apenas a freq. relativa e deixar o *leitor* fazer a conta?
- ▶ A ideia é apresentar os resultados da forma mais simples e informativa para o *leitor*, que ele não precise fazer contas, que ele veja a Tabela e saiba tudo aquilo que precisa saber.

"Uma imagem vale mais que mil palavras". Confúsio

"Uma imagem vale mais que mil palavras". Confúsio

Os gráficos devem ser de tal forma que apenas *batendo o olho* tenhamos toda a informação que precissamos. Por isso, eles devem ser **informatívos** e **fáceis de entender**.

"Uma imagem vale mais que mil palavras". Confúsio

Os gráficos devem ser de tal forma que apenas *batendo o olho* tenhamos toda a informação que precissamos. Por isso, eles devem ser **informatívos** e **fáceis de entender**.

Assim como as tabelas, os gráficos nos ajudam a ter um panorama geral dos dados.

"Uma imagem vale mais que mil palavras". Confúsio

Os gráficos devem ser de tal forma que apenas *batendo o olho* tenhamos toda a informação que precissamos. Por isso, eles devem ser **informatívos** e **fáceis de entender**.

- Assim como as tabelas, os gráficos nos ajudam a ter um panorama geral dos dados.
- Existem diferentes tipos de gráficos e eles dependem do tipo de variável com que estamos trabalhando.

Gráfico de setores (ou Pizza/Pie)

Gráfico para descrever variáveis qualitativas

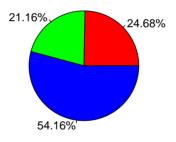
- Gráfico para descrever variáveis qualitativas
- Apresenta a distribuição de frequências

- Gráfico para descrever variáveis qualitativas
- Apresenta a distribuição de frequências
- Útil quando temos poucas classes.

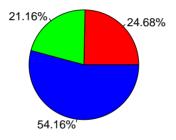
- Gráfico para descrever variáveis qualitativas
- Apresenta a distribuição de frequências
- Útil quando temos poucas classes.
- Uma prática comúm quando temos muitas classes, é agrupar todas as classes com poucas frequências em uma nova categoria chamada outros.

- Gráfico para descrever variáveis qualitativas
- Apresenta a distribuição de frequências
- Útil quando temos poucas classes.
- Uma prática comúm quando temos muitas classes, é agrupar todas as classes com poucas frequências em uma nova categoria chamada outros.
- ▶ No seguinte slide, veremos um exemplo

Distribuição dos passagéiros do Titanic segundo a classe da passagem



Distribuição dos passagéiros do Titanic segundo a classe da passagem



O que você achou do gráfico?

Distribuição dos passagéiros do Titanic segundo a classe da passagem

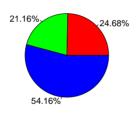


Distribuição dos passagéiros do Titanic segundo a classe da passagem



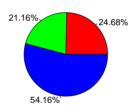
O que você achou do gráfico?

Distribuição dos passagéiros do Titanic segundo a classe da passagem





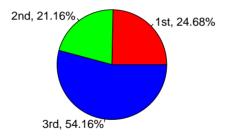
Distribuição dos passagéiros do Titanic segundo a classe da passagem



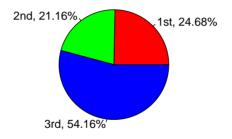
1st2nd3rd

O que você achou do gráfico?

Distribuição dos passagéiros do Titanic segundo a classe da passagem



Distribuição dos passagéiros do Titanic segundo a classe da passagem



O que você achou do gráfico?

Os gráficos devem ser de tal forma que apenas *batendo o olho* tenhamos toda a informação que precissamos. Por isso, eles devem ser **informatívos** e **fáceis de entender**.

Gráfico de barras

Gráfico para descrever variáveis qualitativas.

- Gráfico para descrever variáveis qualitativas.
- Apresenta a distribuição de frequências.

- Gráfico para descrever variáveis qualitativas.
- Apresenta a distribuição de frequências.
- Útil quando tempos muitas ou poucas classes.

- Gráfico para descrever variáveis qualitativas.
- Apresenta a distribuição de frequências.
- Útil quando tempos muitas ou poucas classes.
- ► A prática comúm de agrupar classes com poucas frequências em uma nova classe chamada de "outros", se mantém.

- Gráfico para descrever variáveis qualitativas.
- Apresenta a distribuição de frequências.
- Útil quando tempos muitas ou poucas classes.
- ▶ A prática comúm de agrupar classes com poucas frequências em uma nova classe chamada de "outros", se mantém.
- ► Em um dos eixos especificamos as classes (grupos/categorias) e no outro eixo a frequência.

Exemplos

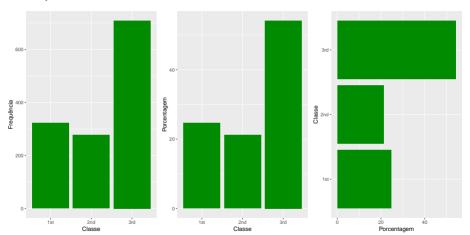


Figura 1: Distribuição dos passagéiros segundo classe da passagem

Gráfico de Pareto

Chamado também de diagrama de Pareto. É parecido ao gráfico de barras, mas contem algumas informações adicionais:

Gráfico de Pareto

Chamado também de diagrama de Pareto. É parecido ao gráfico de barras, mas contem algumas informações adicionais:

Ordena as barras de maior frequência a menor frequência

Gráfico de Pareto

Chamado também de diagrama de Pareto. É parecido ao gráfico de barras, mas contem algumas informações adicionais:

- Ordena as barras de maior frequência a menor frequência
- As barras representam a frequência absoluta

Gráfico de Pareto

Chamado também de diagrama de Pareto. É parecido ao gráfico de barras, mas contem algumas informações adicionais:

- Ordena as barras de maior frequência a menor frequência
- As barras representam a frequência absoluta
- ▶ É criada uma curva (um polígono) para representar a frequência percentual acumulada.

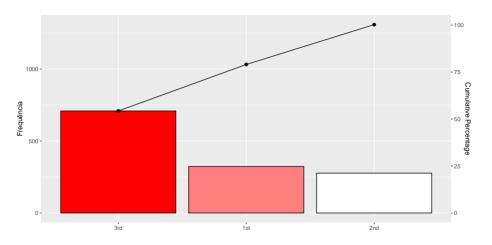


Figura 2: Gráfico de Pareto: Distribuição dos diamantes segundo a qualidade do corte

Tabelas e Gráficos para duas variáveis

Às vezes, estamos interessados em saber mais do que apenas a analise de uma única variável por separado pode nos dizer. Nesses casos, as tabelas ou gráficos antes apresentados não nos ajudam.

Às vezes, estamos interessados em saber mais do que apenas a analise de uma única variável por separado pode nos dizer. Nesses casos, as tabelas ou gráficos antes apresentados não nos ajudam.

Tabelas de contingência

As Tabelas de contingência (chamadas também Tabelas de dupla entrada ou tabelas cruzadas) são uteis para sintetizar a informação conjunta de duas variáveis, de forma que nos permita ter um melhor panorama dos nossos dados.

Às vezes, estamos interessados em saber mais do que apenas a analise de uma única variável por separado pode nos dizer. Nesses casos, as tabelas ou gráficos antes apresentados não nos ajudam.

Tabelas de contingência

- As Tabelas de contingência (chamadas também Tabelas de dupla entrada ou tabelas cruzadas) são uteis para sintetizar a informação conjunta de duas variáveis, de forma que nos permita ter um melhor panorama dos nossos dados.
- Nas Tabelas de contingência teremos, na vertical, a distribuição das classes de uma das variáveis, e no horizontal, a distribuição das classes da outra variável.

Às vezes, estamos interessados em saber mais do que apenas a analise de uma única variável por separado pode nos dizer. Nesses casos, as tabelas ou gráficos antes apresentados não nos ajudam.

Tabelas de contingência

- As Tabelas de contingência (chamadas também Tabelas de dupla entrada ou tabelas cruzadas) são uteis para sintetizar a informação conjunta de duas variáveis, de forma que nos permita ter um melhor panorama dos nossos dados.
- Nas Tabelas de contingência teremos, na vertical, a distribuição das classes de uma das variáveis, e no horizontal, a distribuição das classes da outra variável.
- ▶ Além disso, poderemos ver a distribuição **conjunta** das variáveis (*i.e,* conheceremos o comportamento da intereção entre as duas variáveis).

Tabela 6: Distribuição dos passagéiros do Titanic segundo classe da passagem e sexo

	Feminino	Masculino	Total
1st	144	179	323
2nd	106	171	277
3rd	216	493	709
Total	466	843	1309

Tabela 6: Distribuição dos passagéiros do Titanic segundo classe da passagem e sexo

	Feminino	Masculino	Total
1st	144	179	323
2nd	106	171	277
3rd	216	493	709
Total	466	843	1309

Como representaríamos isso em um gráfico?

Gráficos

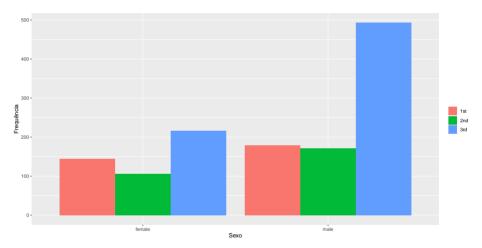


Figura 3: Distribuição dos passagéiros do Titanic segundo classe da passagem e sexo.

Gráficos

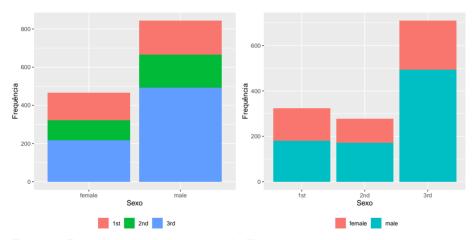


Figura 4: Distribuição dos passagéiros do Titanic segundo classe da passagem e sexo.

Tabelas de contingência e Gráficos

Tabela 7: Distribuição dos passagéiros (em %) do Titanic segundo classe da passagem e sexo

	Feminino	Masculino	Total
1st	11.0	13.67	24.68
2nd	8.1	13.06	21.16
3rd	16.5	37.66	54.16
Total	35.6	64.40	100.00

Tabela 7: Distribuição dos passagéiros (em %) do Titanic segundo classe da passagem e sexo

	Feminino	Masculino	Total
1st	11.0	13.67	24.68
2nd	8.1	13.06	21.16
3rd	16.5	37.66	54.16
Total	35.6	64.40	100.00

▶ Aprox. 54% dos passagéiros eram da terceira classe.

Tabela 7: Distribuição dos passagéiros (em %) do Titanic segundo classe da passagem e sexo

	Feminino	Masculino	Total
1st	11.0	13.67	24.68
2nd	8.1	13.06	21.16
3rd	16.5	37.66	54.16
Total	35.6	64.40	100.00

- ▶ Aprox. 54% dos passagéiros eram da terceira classe.
- ▶ Aprox. 64% dos passagéiros eram do sexo masculino.

Tabela 7: Distribuição dos passagéiros (em %) do Titanic segundo classe da passagem e sexo

	Feminino	Masculino	Total
1st	11.0	13.67	24.68
2nd	8.1	13.06	21.16
3rd	16.5	37.66	54.16
Total	35.6	64.40	100.00

- ▶ Aprox. 54% dos passagéiros eram da terceira classe.
- ▶ Aprox. 64% dos passagéiros eram do sexo masculino.
- ► Aprox. 8% dos passagéiros eram do sexo feminino e viajaram em segunda classe.

Às vezes, podemos montar as tabelas e interpretá-las em função de uma das variáveis. Por exemplo, podemos interpretar as porcentagens por linha.

Às vezes, podemos montar as tabelas e interpretá-las em função de uma das variáveis. Por exemplo, podemos interpretar as porcentagens por linha.

Tabela 8: Distribuição dos passagéiros do Titanic segundo classe da passagem e sexo (% segundo a classe da passagem)

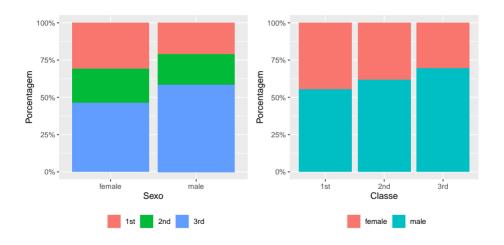
	Feminino	Masculino	Total
1st	44.58	55.42	100
2nd	38.27	61.73	100
3rd	30.47	69.53	100

Ou podemos interpretar as porcentagens por coluna.

Ou podemos interpretar as porcentagens por coluna.

Tabela 9: Distribuição dos passagéiros do Titanic segundo classe da passagem e sexo (% segundo o sexo)

Feminino	Masculino
30.90	21.23
22.75	20.28
46.35	58.48
100.00	100.00
	30.90 22.75 46.35



► É possivel estender as ideias das tabelas e gráficos de duas variáveis, para mais de duas dimensões.

- ▶ É possivel estender as ideias das tabelas e gráficos de duas variáveis, para mais de duas dimensões.
- Escolher entre tabelas e gráficos; frequências absolutas, relativas ou percentuais; ou típo de gráfico a utilizar são detalhes que podem causar dor de cabeça no princípio, mas a regra é simples "O que for mais fácil de entender e interpretar é o que devemos reportar".

- ▶ É possivel estender as ideias das tabelas e gráficos de duas variáveis, para mais de duas dimensões.
- Escolher entre tabelas e gráficos; frequências absolutas, relativas ou percentuais; ou típo de gráfico a utilizar são detalhes que podem causar dor de cabeça no princípio, mas a regra é simples "O que for mais fácil de entender e interpretar é o que devemos reportar".

- ▶ É possivel estender as ideias das tabelas e gráficos de duas variáveis, para mais de duas dimensões.
- Escolher entre tabelas e gráficos; frequências absolutas, relativas ou percentuais; ou típo de gráfico a utilizar são detalhes que podem causar dor de cabeça no princípio, mas a regra é simples "O que for mais fácil de entender e interpretar é o que devemos reportar".

Leituras recomendadas

- Anderson, D. R; Sweeney, D. J.; e Williams, T. A. (2008). Estatística Aplicada à Administração e Economia. 2ed. Cengage Learning. Cap 2
- ► Freund, J. E.; Perles, B. M. (2014). *Modern elementary statistics*. 12ed. Pearson College Division. **Chapter 1**