IMD0033 - Probabilidade
Aula 18 - Moda

Ivanovitch Silva
Maio, 2019

Agenda

- A moda
  - Variáveis ordinais
  - Variáveis nominais
  - Variáveis discretas
- Casos especiais
  - Unimodal
  - Bimodal
  - Multimodal
- Distribuições assimétricas
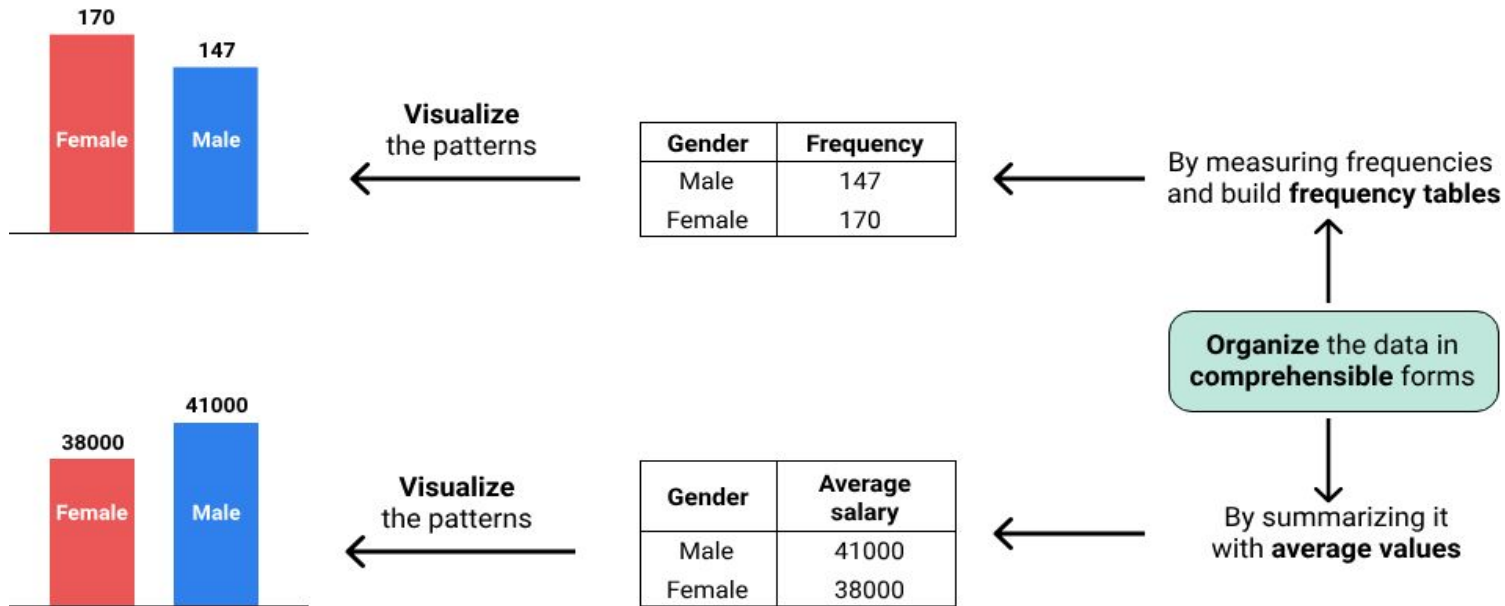- Distribuições simétricas

Atualizar o repositório

git clone https://github.com/ivanovitchm/imd0033_2019_1.git

Ou ....

git pull

PREVIOUSLY ON...

Gender	Frequency
Male	147
Female	170

Visualize the patterns

By measuring frequencies and build **frequency tables**

Organize the data in **comprehensible** forms

Gender	Average salary
Male	41000
Female	38000

Visualize the patterns

By summarizing it with **average values**

Depending on the particular characteristics of a distribution, we'll see that we can summarize it using the **mean**, the **weighted mean**, the **median**, or the **mode**.

Introduction

- Mean
    - It takes into account each value in the distribution
    - It can be defined algebraically.
    - These two properties make the mean far superior to the median.
- Median
    - It comes in handy, however, when it's not possible or appropriate to compute the mean.

The dataset

	Order	PID	MS SubClass	MS Zoning	Lot Frontage	Lot Area	Street	Alley	Lot Shape	Sale Condition	SalePrice
0	1	526301100	20	RL	141.0	131770	Pave	NaN	WD	Normal	215000
1	2	526350040	20	RH	80.0	11622	Pave	NaN	WD	Normal	105000
2	3	526351010	20	RL	81.0	14267	Pave	NaN	WD	Normal	172000
3	4	526353030	20	RL	93.0	11160	Pave	NaN	WD	Normal	244000
4	5	527105010	60	RL	74.0	13830	Pave	NaN	WD	Normal	189900

```
1  houses["Land Slope"].unique()
```

```
array(['Gtl', 'Mod', 'Sev'], dtype=object)
```

```
1  houses["Roof Style"].unique()
```

```
array(['Hip', 'Gable', 'Mansard', 'Gambrel', 'Shed', 'Flat'], dtype=object)
```

The mode for Ordinal Variables

- We call the **most frequent value** in the distribution the **mode**.
- The mode of the **Land Slope** variable is **'Gtl'**. In other words, the typical house has a gentle slope.

```
1 houses['Land Slope'].value_counts()
```

```
Gtl      2789
Mod       125
Sev        16
```

The mode for nominal variables

- The mode is also a good choice for nominal data.
- We obviously **can't compute the mean** for this variable because the values are words.
- Because the Roof Style variable is nominal, **there's also no inherent order** of the values in the distribution. We can't compute the median.

```
1  houses['Roof Style'].value_counts()
```

```
Gable       2321
Hip          551
Gambrel       22
Flat          20
Mansard       11
Shed           5
Name: Roof Style, dtype: int64
```

The mode for discrete variables

- There are some cases where computing the **mean** and the **median** is possible and correct, but the **mode** is preferred nonetheless.
- This is sometimes the case for discrete variables.

```
1  houses['Kitchen AbvGr'].value_counts().sort_index()
```

```
0       3
1    2796
2     129
3       2
```

	Value	Reporting to non-technical audiences
Mean	1.04	The average house has 1.04 kitchens.
Median	1	The average house has one kitchen.
Mode	1	The typical house has one kitchen.

The mode: special cases

There are distributions that can have more than one mode.

$[1, 1, 2, 2, 3, 3, 4, 4]$

Without a mode

$[0, 1, 1, 2, 2, 3, 3, 4, 4, 8]$

Multimodal

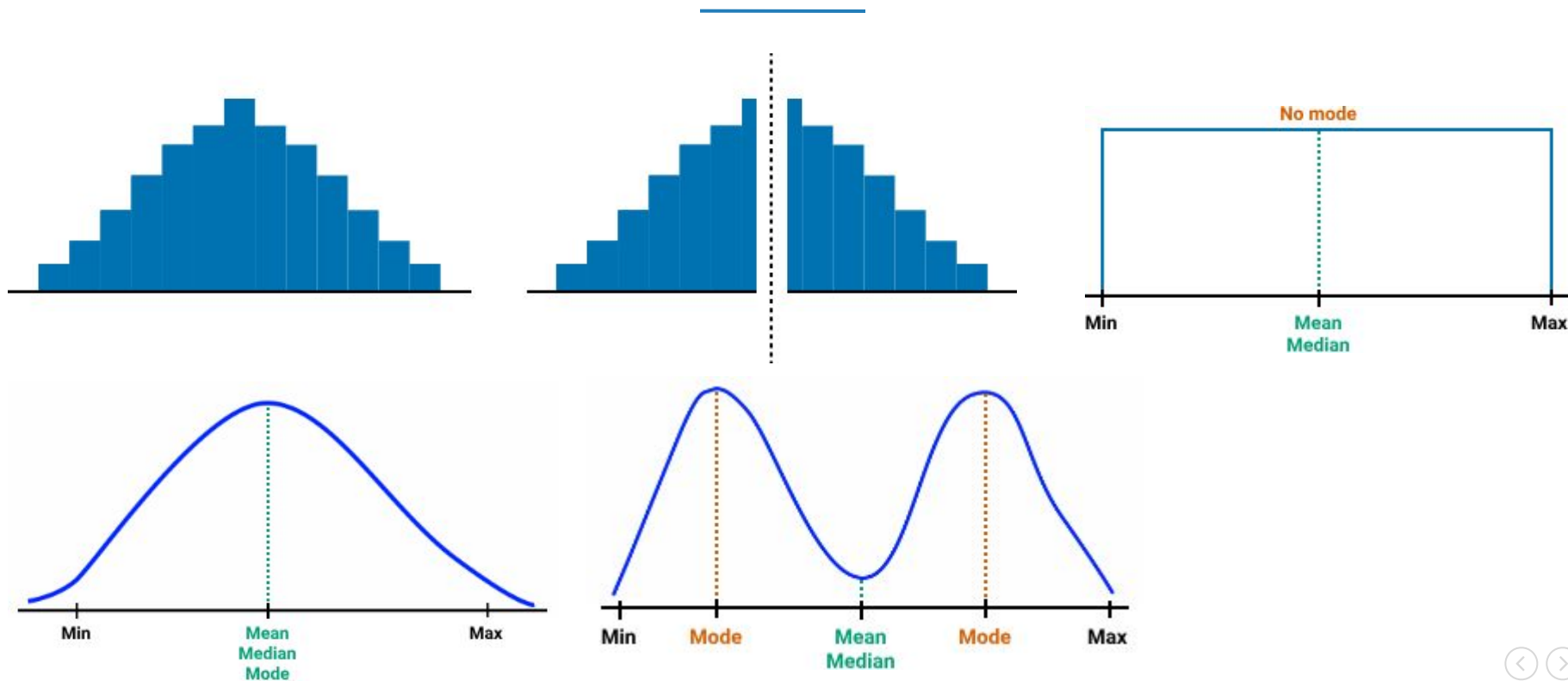$[0, 1, 1, 1, 2, 2, 2, 3]$

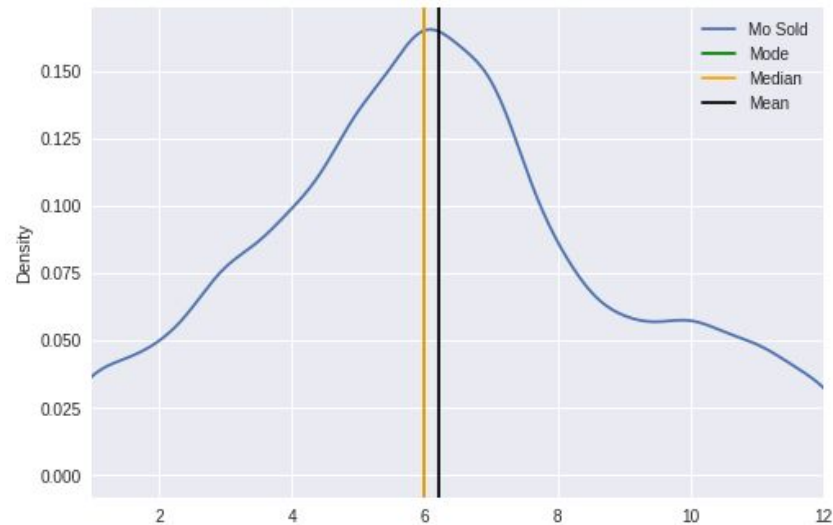Bimodal

Skewed distributions
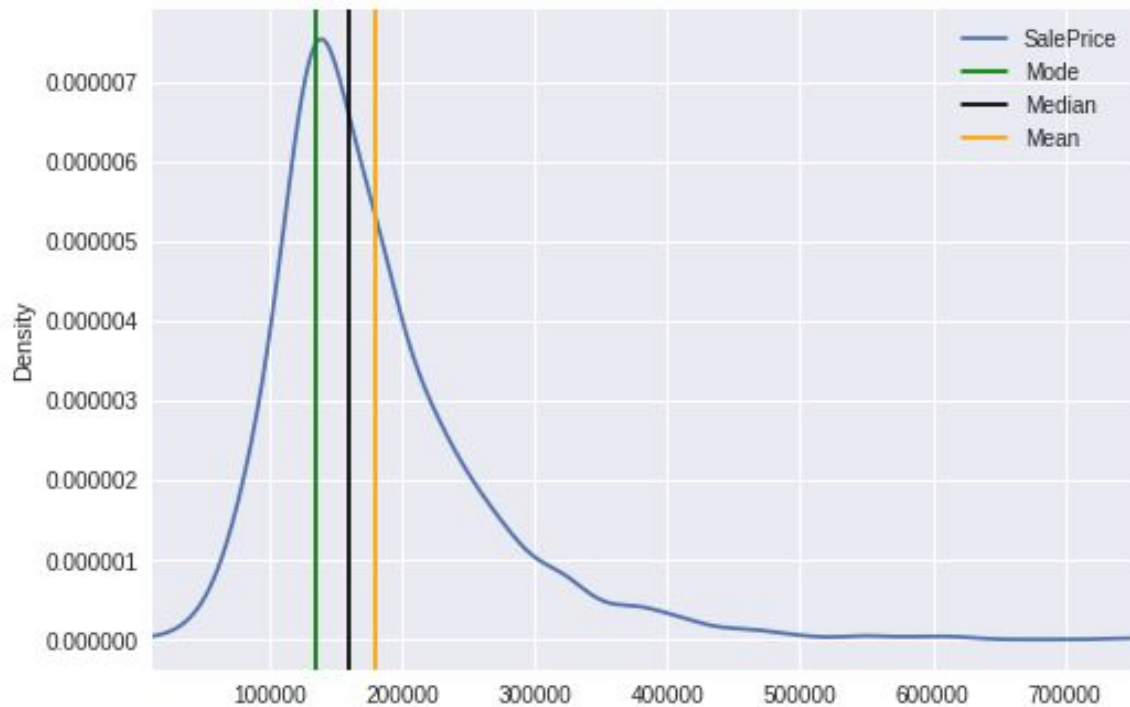
Skewed distributions
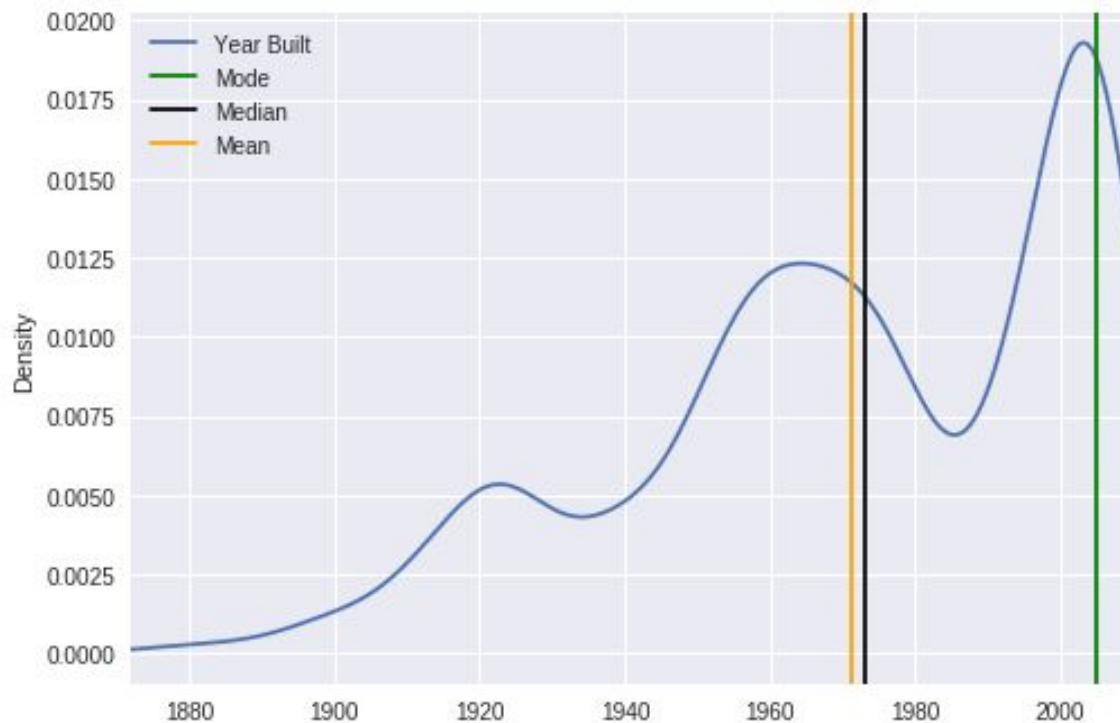
Symmetrical distribution

Symmetrical distribution


```
houses['Mo Sold'].plot.kde(xlim = [1,12])

import matplotlib.pyplot as plt
plt.axvline(houses['Mo Sold'].mode()[0], color = 'Green', label = 'Mode')
plt.axvline(houses['Mo Sold'].median(), color = 'Orange', label = 'Median')
plt.axvline(houses['Mo Sold'].mean(), color = 'Black', label = 'Mean')
plt.legend()
```

Next Steps

	Can be used for	Can't be used for	Ideal for
Mean	• Data measured on an **interval or ratio scale** • **Ordinal data** encoded with numbers (contentious)	• **Nominal data** • **Ordinal data** that is not encoded with numbers • Data points with different weights — we need to use the **weighted mean** instead	• Summarizing **numerical** distributions by taking into account **each value** in the distribution
Median	• Data measured on an **interval or ratio scale** • **Ordinal data** encoded with numbers	• **Nominal data** • **Ordinal data** that is not encoded with numbers	• Summarizing numerical distributions that have **outliers** • **Open-ended** distributions
Mode	• **Nominal data** • **Ordinal data** • Data measured on an **interval or ratio scale**	• Perfectly **uniform distributions** • **Continuous variables** (we need to create grouped frequency tables as a workaround)	• **Nominal data** • **Ordinal data** that is not encoded with numbers • **Discrete data** — when the result needs to be communicated to a non-technical audience