



Insper

Ciência dos dados

Use esses slides para obter algumas dicas,

Mapa Mental

Uma única variável

Qualitativa

- `.value_counts()` → tabela com frequências absolutas
- `.value_counts(True)` → tabela com frequências relativas
- `.dtypes` → mostra tipo de cada variável de um *dataframe*

Quantitativa

- `.describe()` → mostra medidas resumo de uma variável quantitativa
 - *Medidas de posição:*
 - Mínimo, máximo
 - Média, Mediana
 - Percentil 25%: separa os 25% menores valores dos demais 75% maiores valores
 - Percentil 75%: separa os 75% menores valores dos demais 25% maiores valores
 - *Medidas de heterogeneidade:*
 - Desvio padrão
- `pd.cut()` → transforma variável quantitativa em qualitativa
- Visualizar: HISTOGRAMA

Relação entre duas variáveis

Qualitativa

`pd.crosstab()` → tabela cruzada com frequências

Qualitativa

Quantitativa

`.groupby(by=VarQuali)` → separa *dataframe* nas categorias existentes na variável qualitativa

Quantitativa

- Visualizar: *scatter plot* (Gráfico de dispersão)
- Mensurar associação entre quantitativas:
 - Covariância: APENAS Sinal
 - Correlação: Sinal e Intensidade

```
pd.crosstab(df.VarQuali1, df.VarQuali2, normalize = ?)
```

- `normalize = False` → tabela cruzada com frequências absolutas (contagens) de duas variáveis **qualitativas**
- `normalize = True` → tabela cruzada com frequências relativas pelo **total geral**
- `normalize = 'index'` → tabela cruzada com frequências relativas por **total de linha** cada categoria da variável qualitativa `VarQuali1`
- `normalize = 'columns'` → tabela cruzada com frequências relativas por **total de coluna** cada categoria da variável qualitativa `VarQuali2`

Obs.: `df` → assumo aqui o nome do *dataframe* no seu código.

```
pd.crosstab(dados.PLANO, dados.EC, normalize = ?)
```

- **normalize = True** → Qual % de clientes nas combinações das categorias de PLANO e EC?
- **normalize = 'index'** → Dentre de cada categoria de PLANO, qual % de clientes nas categorias EC?
- **normalize = 'columns'** → Dentro de cada categoria de EC, qual % de clientes nas categorias de PLANO?

	Casado	Solteiro	Outros
A	?	?	?
B	?	?	?

A escolha do **normalize** depende do questionamento feito ao problema!

Covariância vs Correlação

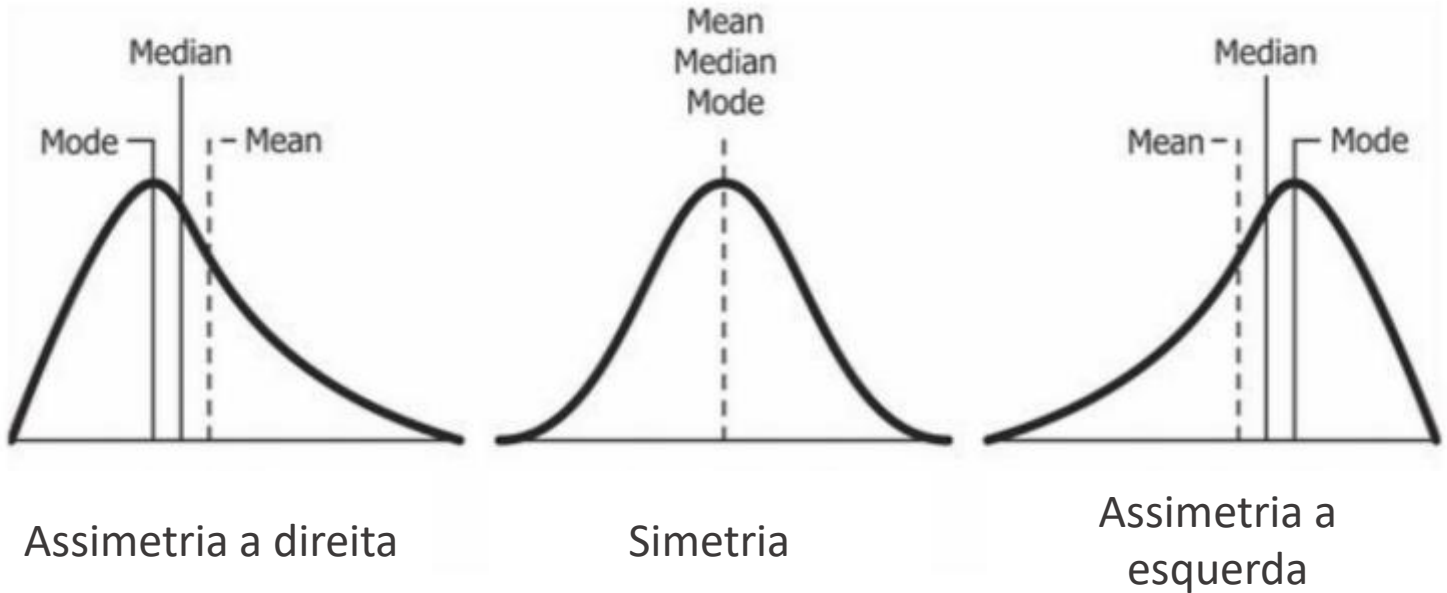
- `df.VarQuanti1.cov(df.VarQuanti2)` → calcula **covariância** entre duas variáveis **quantitativas**. Interpreta apenas SINAL da associação linear.
- `df.VarQuanti1.corr(df.VarQuanti2)` → calcula **correlação** entre duas variáveis **quantitativas**. Interpreta SINAL e GRAU da associação linear.

$$|Corr(X, Y)| < 0,3: \textit{fraca}$$

- Regra de bolso: $0,3 \leq |Corr(X, Y)| < 0,7: \textit{moderada}$
 $|Corr(X, Y)| \geq 0,7: \textit{forte}$

Obs.: `df` → assumo aqui o nome do *dataframe* no seu código.

Formato do Histograma vs Medidas de posição



Fonte: <https://en.wikipedia.org/wiki/Skewness>