


IMD0033 - Probabilidade

Aula 14 - Tipos de escalas e medidas de tendência central

Ivanovitch Silva
Outubro, 2017



Agenda

- Estudo de caso: salários, EAD, dados abertos UFRN
- Tipos de escalas
- Histogramas
- Simetria e curtose
- Moda, média e mediana

Atualizar o repositório

```
git clone https://github.com/ivanovitchm/IMD0033_Probabilidade.git
```

Ou

```
git pull
```

Tipos de escalas

Escalas com intervalos equidistantes



$60\text{km/h} - 50\text{km/h} = 10\text{km/h}$

$50\text{km/h} - 40\text{km/h} = 10\text{km/h}$

Escalas logarítmicas

M	A	A amplitude cresce por um fator 10 a cada magnitude
0	A_0	A_0
1	$A_1 = 10A_0$	$A_1 = 10A_0$
2	$A_2 = 100A_0$	$A_2 = 10A_1$
3	$A_3 = 1000A_0$	$A_3 = 10A_2$
4	$A_4 = 10000A_0$	$A_4 = 10A_3$
5	$A_5 = 100000A_0$	$A_5 = 10A_4$
6	$A_6 = 1000000A_0$	$A_6 = 10A_5$
7	$A_7 = 10000000A_0$	$A_7 = 10A_6$
8	$A_8 = 100000000A_0$	$A_8 = 10A_7$
9	$A_9 = 1000000000A_0$	$A_9 = 10A_8$

$$3 - 2 = 900$$

$$2 - 1 = 90$$

Os tipos de escalas influenciam as propriedades estatísticas

Escalas discretas e contínuas



Caminhada de um caracol por dia
(em polegadas): 2.10, 2.33, 3.45



Número de carros no estacionamento
por dia: 30, 50, 101

Variáveis discretas podem ter propriedades estatísticas contínuas.

Qual a média de carros no estacionamento?

Entendendo o ponto inicial das escalas

Algumas escalas utilizam o **zero** de diferentes formas.



Zero carros no estacionamento é realmente zero carros

Escala	Zero absoluto
Kelvin	0 K
Rankine	0 Ra
Celsius	-273,15 °C
Fahrenheit	-459,67 °F
Réaumur	-218,52 °Ré

A ausência de calor (zero absoluto) em escalas de temperatura não é em Zero

Se hoje o estacionamento possui 20 carros, e ontem haviam 10 carros, eu posso afirmar que a quantidade de carros dobrou?

Se a temperatura em Natal hoje foi de 40°C e ontem a temperatura alcançou 20°C , eu posso afirmar que a temperatura dobrou?

Escalas ordinais

Em escalas ordinais os itens são ordenados com um ranking.

Quantas horas por dia você estuda?

[nada, pouco, normal, muito, absurdamente]

[0, 1, 2, 3, 4]

Escalas categóricas

Escala que organiza os itens em categoria (não existe um ordem de quem é maior ou menor).

Qual o seu gênero?

- Masculino
- Feminino
- Prefiro não mencionar

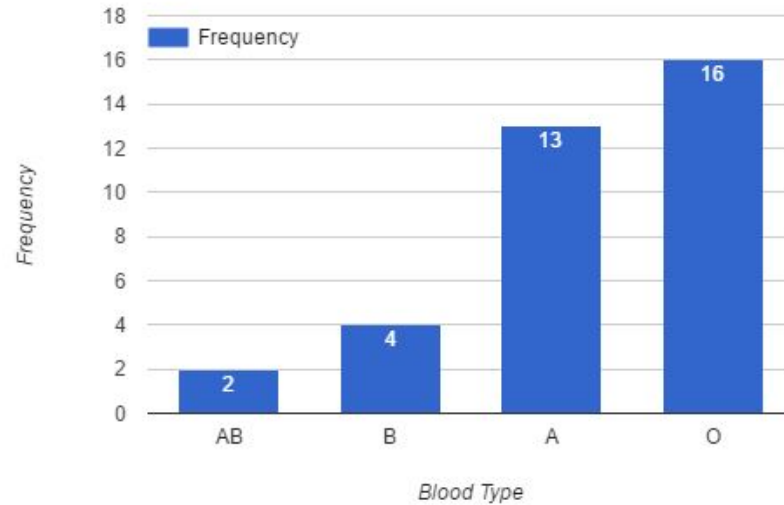
Medidas de tendência central

Histograma

Blood Types

O	O	A	O	O
A	A	B	A	O
O	O	O	AB	O
A	O	A	O	A
B	A	O	A	B
AB	O	A	O	B
A	A	O	A	O

Histogram

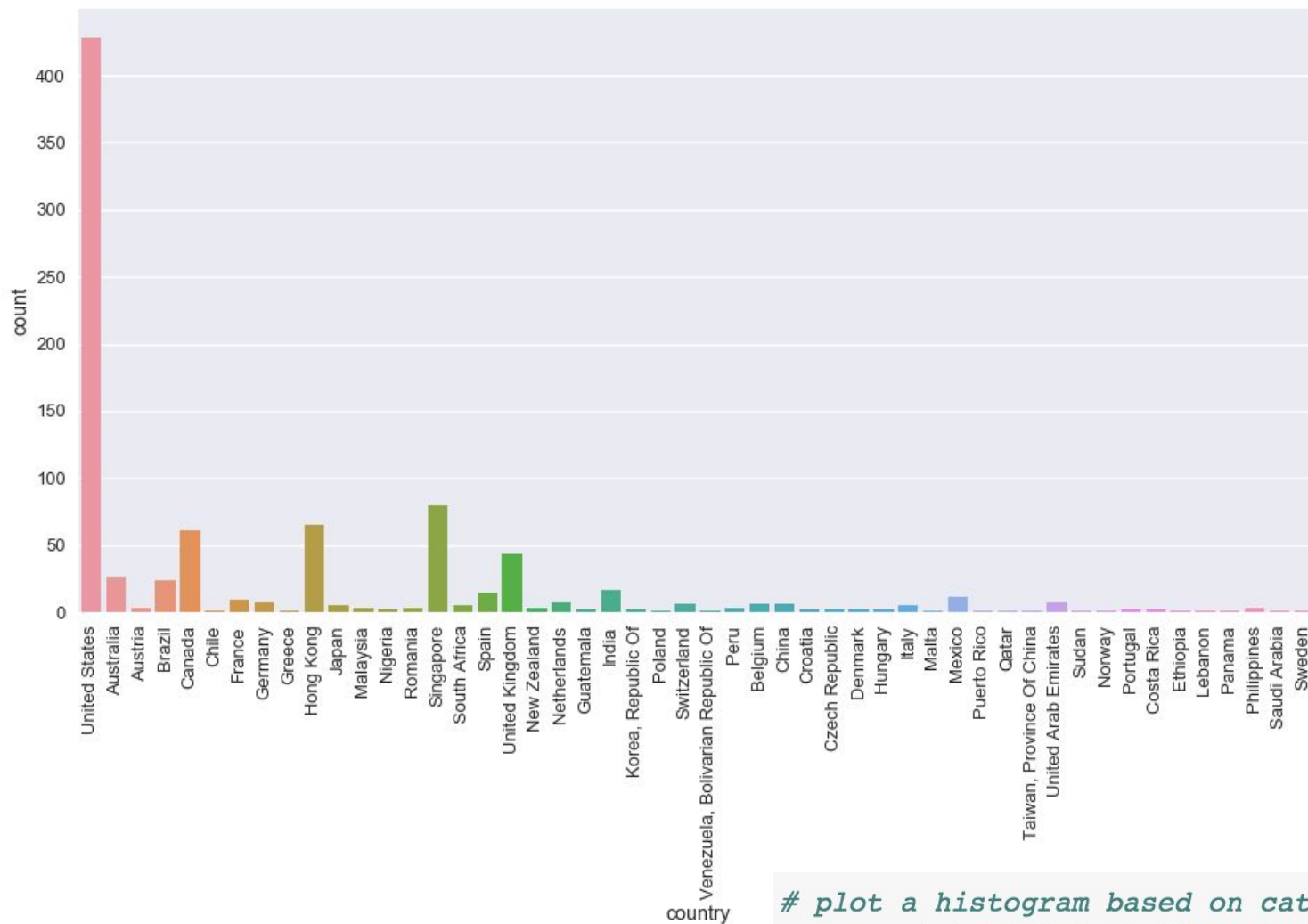


Histograma de variáveis categóricas

Studying about frequency

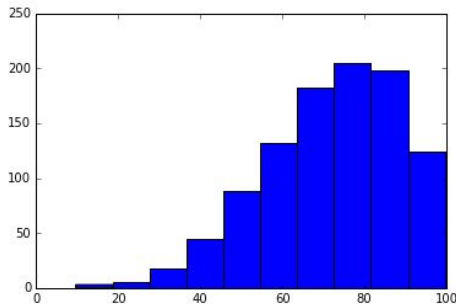
	id	industry	country	city
0	10001	Agriculture	United States	Davis
1	10002	Arts & Education	Australia	Perth
2	10003	Arts & Education	Austria	Lieboch
3	10004	Arts & Education	Brazil	São Paulo
4	10005	Arts & Education	Canada	Georgetown
5	10006	Arts & Education	Canada	Hamilton
6	10007	Arts & Education	Canada	Milton
7	10008	Arts & Education	Canada	Mississauga
8	10009	Arts & Education	Canada	Regina
9	10010	Arts & Education	Canada	Toronto



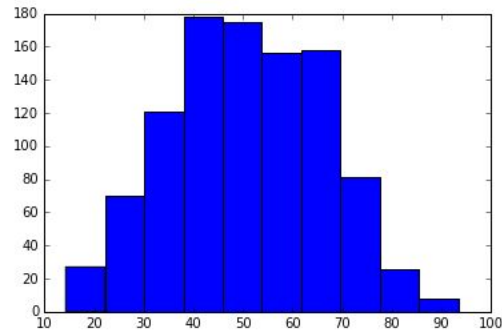


plot a histogram based on categorical variables
 sns.countplot(students['country'])

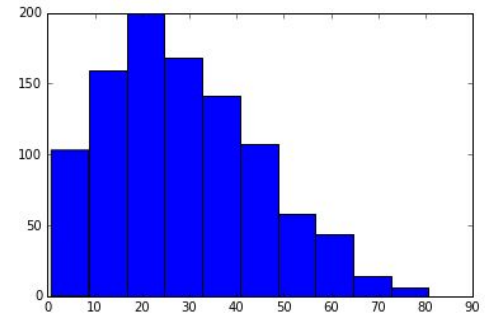
Assimetria (skewness)



Assimetria Negativa
Skew < 0



Simetria
Skew = 0



Assimetria Positiva
Skew > 0

Assimetria (skewness)

```
# We can test how skewed a distribution is using the skew function.  
# A positive value means positive skew,  
# a negative value means negative skew, and close to zero means no skew.  
from scipy.stats import skew
```

```
skewness = skew(students['country'].value_counts())  
skewness
```

```
6.109103958619247
```

Curtose (kurtosis)

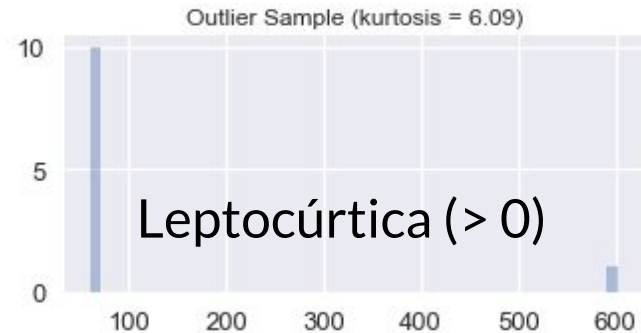
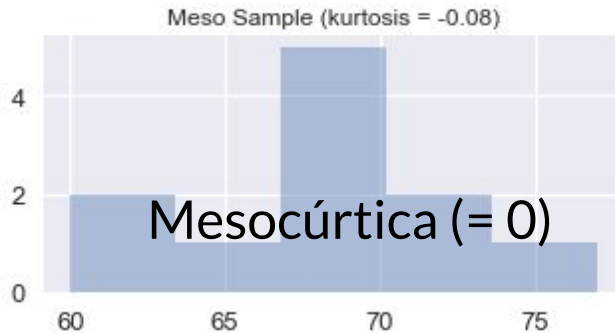
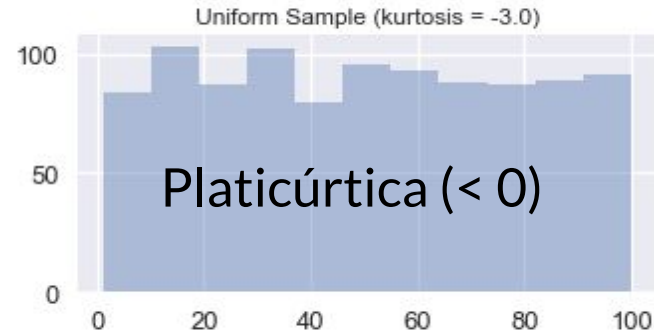
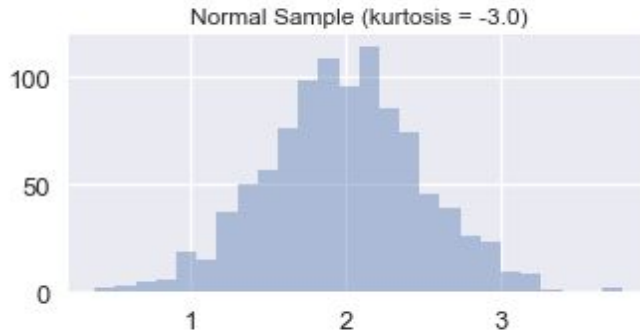
Curtose é uma medida de dispersão que caracteriza o "achatamento" da curva da função de distribuição.

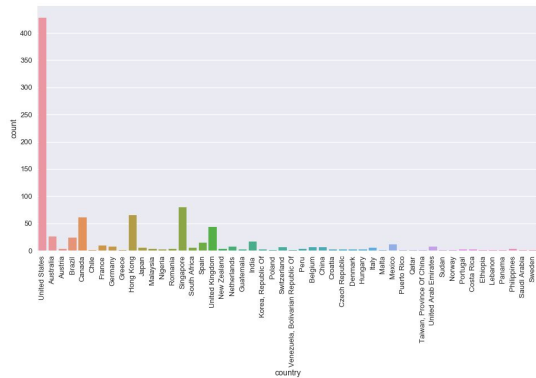
O valor do curtose é relacionado com a cauda da distribuição e não ao valor mais comum (pico).

Um alto valor para a métrica indica a presença de pontos fora da curva (outliers).

Curtose (kurtosis)

Checking for Outliers with Kurtosis





```
# We can measure kurtosis with the kurtosis function.
# Negative values indicate platykurtic distributions, positive values indicate
# leptokurtic distributions, and values near 0 are mesokurtic.

# platykurtic (< 0) = produces fewer and less extreme outliers than does the normal distribution
# leptokurtic (> 0) = produces more outliers than the normal distribution

from scipy.stats import kurtosis

kurtosiness = kurtosis(students['country'].value_counts())
kurtosiness
```

37.93760879517695



PESQUISAR DADOS

Ex.: cursos



Etiquetas mais comuns

projetos de pesquisa

pesquisadores

graduação

GRUPOS



Comunicados e Documentos



Contratos e Convênios



Despesas e Orçamento



Ensino



Extensão



Institucional



Materiais



Patrimônio



Pesquisa



Pessoas



Processos

TURMAS

Relação de turmas dos cursos de nível médio, técnico, graduação e pós-graduação da UFRN

CSV PDF

COMPONENTES CURRICULARES

Relação de componentes curriculares oferecidos pela UFRN nas modalidades de ensino presencial, à distância e semi-presencial

CSV PDF

MATRÍCULAS EM COMPONENTES

Relação das matrículas em componentes dos cursos da instituição.

CSV PDF

DOCENTES

Relação de docentes da UFRN

CSV PDF

CURSOS DE PÓS GRADUAÇÃO

Relação dos cursos de pós Graduação da UFRN

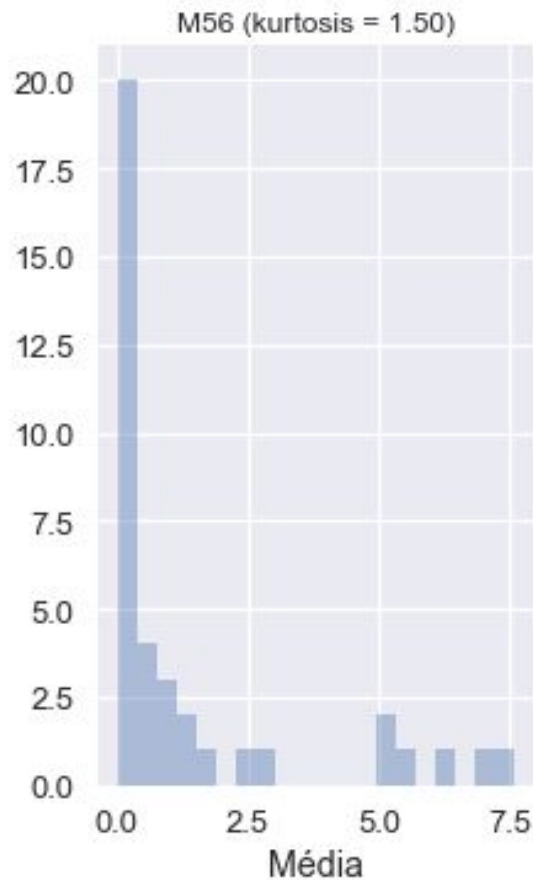
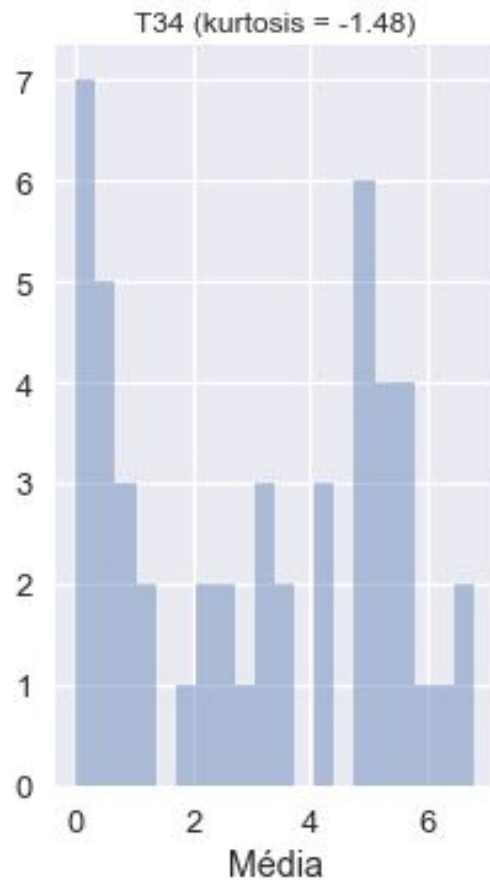
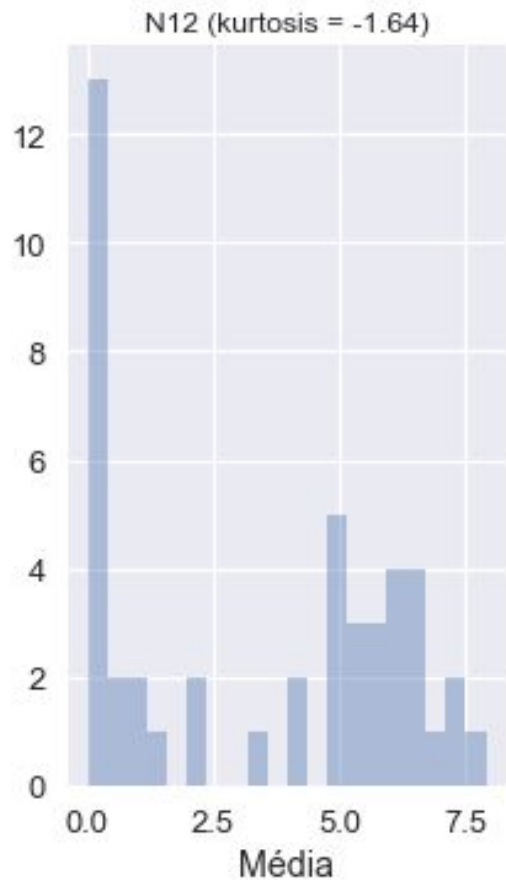
CSV PDF

PROGRAMAS DE PÓS-GRADUAÇÃO

Relação de programas de graduação oferecidos pela UFRN

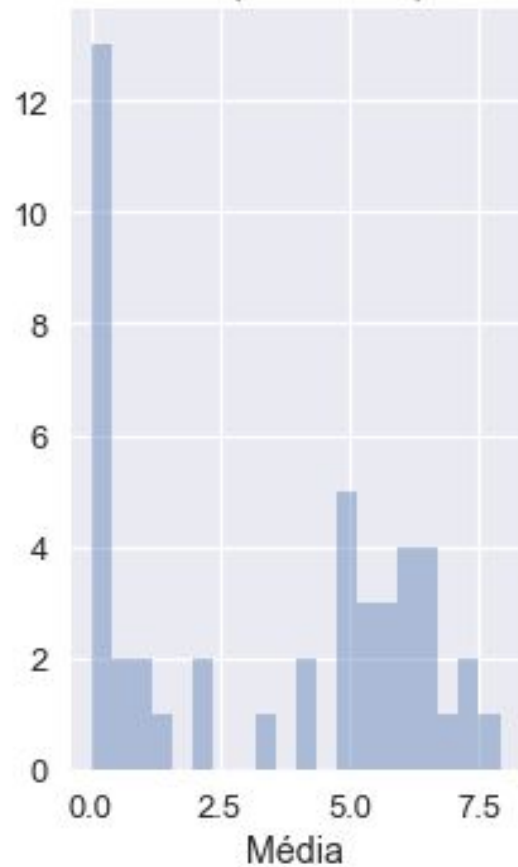
IMD0028 - Fundamentos
Matemáticos para
Computação I
2017.1
03 turmas

Checking for Outliers with Kurtosis

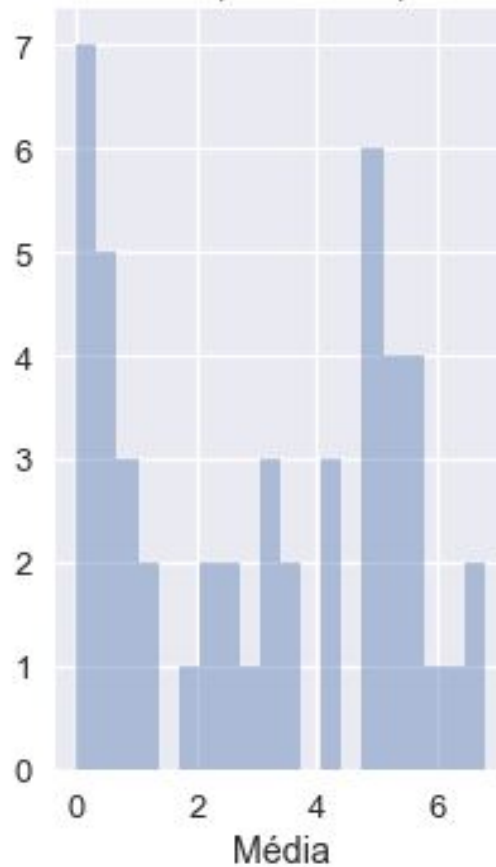


Skew Analysis

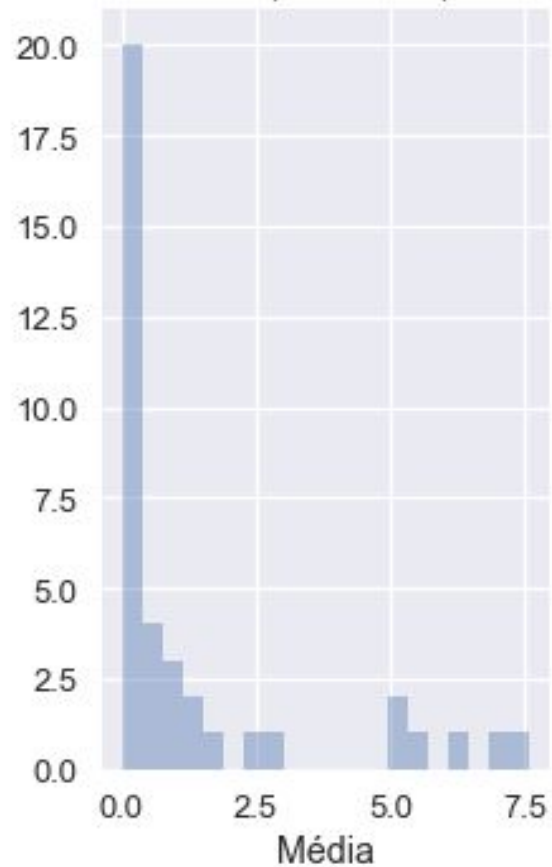
N12 (skew = -0.11)



T34 (skew = -0.06)



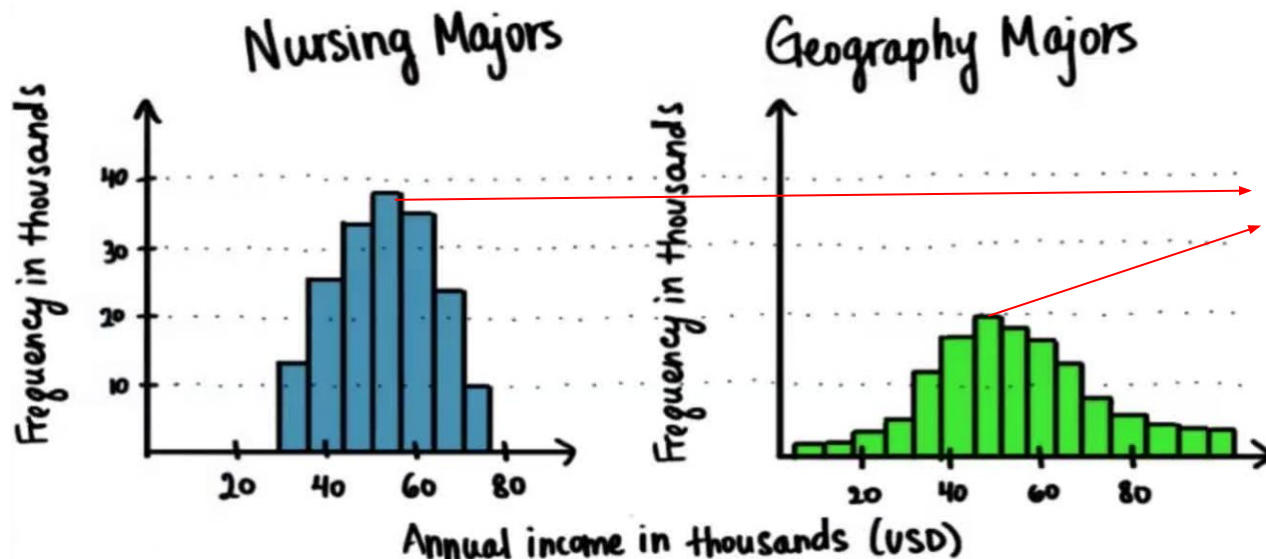
M56 (skew = 1.70)



Moda vs Média vs Mediana

Tendência Central

Mode



**Moda do
conjunto
de dados**

Aproximadamente, qual é a faixa de salário mais comum para a enfermagem e geografia?

Quiz: moda

2 5 5 9 8 3 2 10 1

Qual a moda?

Propriedades da Moda

V/F

Todos os valores da base de dados afetam a moda

A moda será a mesma para várias amostras da população

Existe uma equação para a moda

Média

Enfermagem	Geografia
\$58.350	\$48.670
\$63.120	\$57.320
\$44.640	\$38.150
\$56.389	\$41.290
\$72.250	\$53.160

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Propriedades da Média

- ☐ Todos os valores da distribuição afetam a média
- ☐ A média pode ser descrita por uma fórmula
- ☐ Amostras da população podem ter a mesma média
- ☐ A média das amostras podem ser utilizadas para se fazer previsões sobre a população
- ☐ A média não irá mudar caso valores extremos sejam adicionados a base de dados

Média e Pontos fora da Curva

Nursing	Geography
\$58.350	\$48.670
\$63.120	\$57.320
\$44.640	\$38.150
\$56.389	\$41.290
\$72.250	\$53.160

A média para os formandos em geografia na UNC era \$100.000 na década de 80.

O que aconteceu?



\$500.000

Mediana

Nursing	Geography
\$58.350	\$48.670
\$63.120	\$57.320
\$44.640	\$38.150
\$56.389	\$41.290
\$72.250	\$53.160

1. Ordene a base de dados
2. Elimine os extremos até que reste apenas um elemento.

Mediana e Pontos fora da Curva

Nursing	Geography
\$58.350	\$48.670
\$63.120	\$57.320
\$44.640	\$38.150
\$56.389	\$41.290
\$72.250	\$53.160



\$500.000

TURMAS

Relação de turmas dos cursos de nível médio, técnico, graduação e pós-graduação da UFRN

CSV PDF

COMPONENTES CURRICULARES

Relação de componentes curriculares oferecidos pela UFRN nas modalidades de ensino presencial, à distância e semi-presencial

CSV PDF

MATRÍCULAS EM COMPONENTES

Relação das matrículas em componentes dos cursos da instituição.

CSV PDF

DOCENTES

Relação de docentes da UFRN

CSV PDF

CURSOS DE PÓS GRADUAÇÃO

Relação dos cursos de pós Graduação da UFRN

CSV PDF

PROGRAMAS DE PÓS-GRADUAÇÃO

Relação de programas de graduação oferecidos pela UFRN

IMD0028 - Fundamentos
Matemáticos para
Computação I
2017.1
03 turmas

```
print(M56[ 'Média' ].mean() )  
print(M56[ 'Média' ].mode() )  
print(M56[ 'Média' ].median() )
```

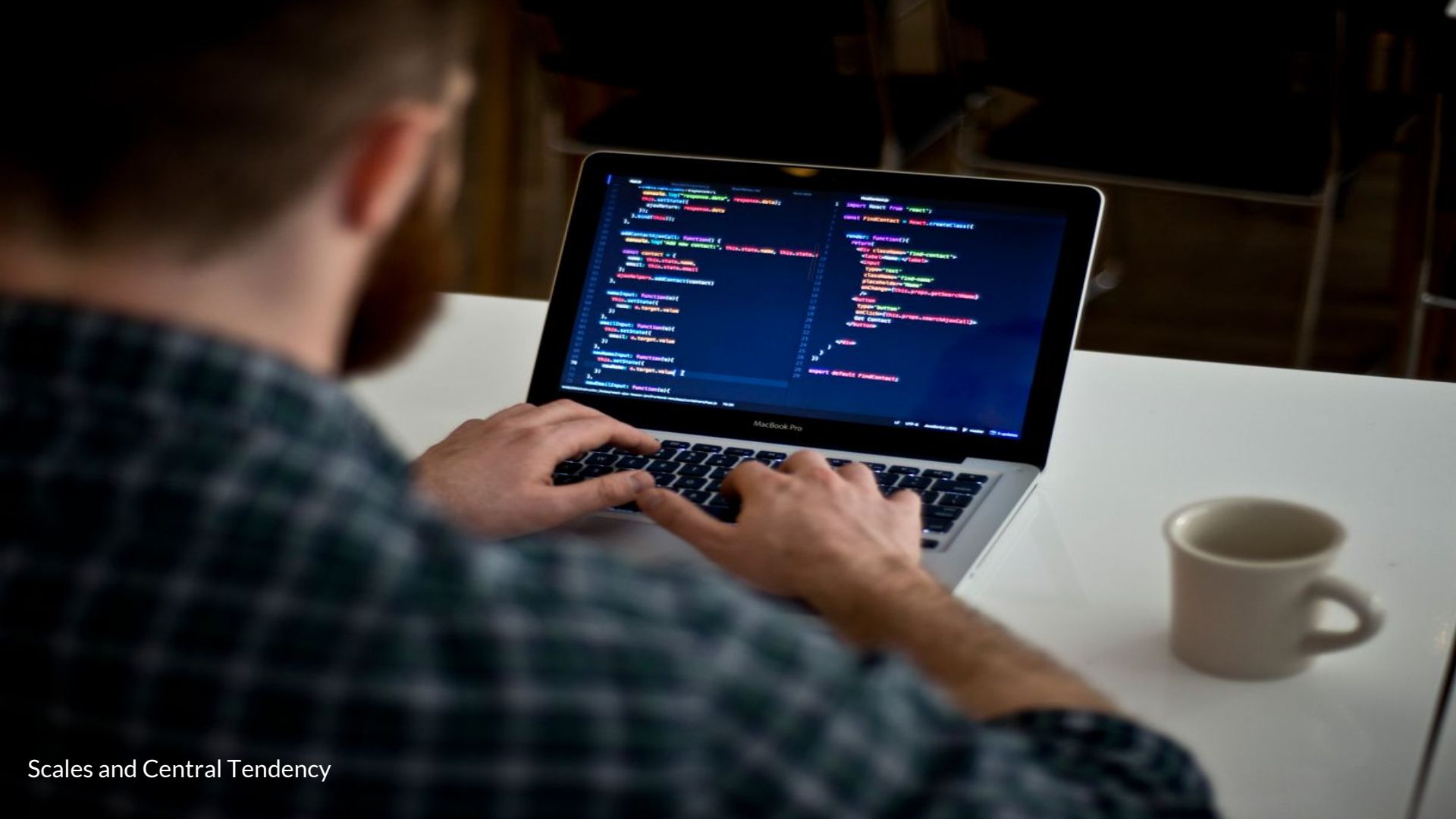
```
1.407894736842105  
0      0.0  
dtype: float64  
0.3
```

Mean (red) vs Median (green)



```
# Plot the mean in red
ax1.axvline(N12['Média'].mean(), color="r")
ax2.axvline(T34['Média'].mean(), color="r")
ax3.axvline(M56['Média'].mean(), color="r")

# Plot the median in green
ax1.axvline(N12['Média'].median(), color="g")
ax2.axvline(T34['Média'].median(), color="g")
ax3.axvline(M56['Média'].median(), color="g")
```



Scales and Central Tendency