Descrição de dados

Jairo Nicolau

Variáveis e planilhas

Variável

Uma variável é um atributo ou característica de um indivíduo ou objeto.

O termo *variável* dá ênfase ao fato de que os valores dos dados *variam*.

Planilha de dados

renda_indiv_faixa[10] 2

	id	municipio	distrito	distrito1	subdistrito	subdistrito1	setor	bairro	rural_urbano	estado	idade	i
1	1	LORENA (	5	Lorena	.	3.52721e+10	5	Olaria -	Urbana	S„o Paul	36	
2	2	COSMORAM	5	Cosmorama	.	3.51290e+10	2	Centro	Urbana	S„o Paul	42	
3	3	SOROCABA	5	Sorocaba	.	3.55221e+10	229	Vila Bar	Urbana	S„o Paul	33	
4	4	SOROCABA	5	Sorocaba	.	3.55221e+10	229	Vila Bar	Urbana	S„o Paul	56	
5	5	ARARAQUA	5	Araraquara	.	3.50321e+10	130	Santana	Urbana	S„o Paul	60	
6	6	BARIRI (	5	Bariri	.	3.50520e+10	5	Centro	Urbana	S„o Paul	51	
7	7	COSMORAM	5	Cosmorama	.	3.51290e+10	11	Rural	Rural	S„o Paul	24	
8	8	COSMORAM	5	Cosmorama	.	3.51290e+10	11	Rural	Rural	S„o Paul	32	
9	9	BARIRI (	5	Bariri	.	3.50520e+10	5	Centro	Urbana	S„o Paul	82	
10	10	SAO JOAO	5	Sao Joao da	.	3.54910e+10	53	Progress	Urbana	S„o Paul	56	
11	11	ARARAQUA	5	Araraquara	.	3.50321e+10	130	Santana	Urbana	S„o Paul	44	
12	12	PEDREIRA	5	Pedreira	.	3.53711e+10	3	Vila Mon	Urbana	S„o Paul	45	
13	13	CAMPINAS	5	Campinas	.	3.50950e+10	832	Jardim I	Urbana	S„o Paul	60	
14	14	BERTIOGA	5	Bertioga	.	3.50636e+10	28	Jardim I	Urbana	S„o Paul	33	
15	15	TATUI (S	5	Tatui	.	3.55400e+10	9	Vila S„o	Urbana	S„o Paul	25	
16	16	RANCHARI	5	Rancharia	.	3.54221e+10	14	Jardim P	Urbana	S„o Paul	61	
17	17	TATUI (S	5	Tatui	.	3.55400e+10	21	Vila Dr.	Urbana	S„o Paul	47	
18	18	ARARAQUA	5	Araraquara	.	3.50321e+10	2	Centro	Urbana	S„o Paul	37	
19	19	RIBEIRAO	5	Ribeirao Pre	.	3.54340e+10	199	Subsetor	Urbana	S„o Paul	38	
20	20	BARIRI (	5	Bariri	.	3.50520e+10	1	Centro	Urbana	S„o Paul	50	
21	21	CAMPINAS	5	Campinas	.	3.50950e+10	645	Jardim I	Urbana	S„o Paul	32	
22	22	BERTIOGA	5	Bertioga	.	3.50636e+10	31	Jardim I	Urbana	S„o Paul	43	
23	23	BERTIOGA	5	Bertioga	.	3.50636e+10	31	Jardim I	Urbana	S„o Paul	57	
24	24	PEDREIRA	5	Pedreira	.	3.53711e+10	31	Jardim T	Urbana	S„o Paul	25	
25	25	PEDREIRA	5	Pedreira	.	3.53711e+10	31	Jardim T	Urbana	S„o Paul	20	
26	26	LORENA (	5	Lorena		3.52721e+10	12	Centro	Urbana	S„o Paul	22	

Vars: 83 Obs: 2,000 Filter: Off

Tipos de variáveis

- Numérica (quantitativa)
- Categórica

Variável numérica (ou quantitativa)

As observações portam valores numéricos que são passíveis de operação matemática (soma, divisão, multiplicação)

- idade
- número de filhos
- renda per capita
- anos de estudo

Variável categórica

Cada observação pertence a uma categoria em conjunto de categorias:

sexo: masculino, feminino.

religião: católica, evangélica, espírita e outras

tipo de moradia: apartamento, casa e outras

crença na vida após a morte: sim, não

Problema

Identifique se cada uma da variáveis é categórica ou quantitativa:

a) número de crianças em uma creche.
b) domicílio eleitoral.
c) estado civil.
d) distância (em quilômetros) de deslocamento até a escola.
e) código postal.

Problema

Identifique se cada uma da variáveis é categórica ou quantitativa:

a) número de crianças em uma creche. *numérica*
b) domicílio eleitoral. *categórica*
c) estado civil. *categórica*
d) distância (em quilômetros) de deslocamento até a escola. *numérica*
e) código postal. *categórica*

Regras para fazer uma planilha

- Nomes das variáveis: simples, sem acento e usando underline ou maiúscula/minúscula.

- O label (segmento) da variável categórica deve ser digitada igual.

- Números sem pontos e vírgulas

- Atenção para data: melhor o formato: 2018/03/21

Responder ao seguinte questionário

- sexo

- altura

- tempo médio de deslocamento até o CPDOC

- cor do cabelo

- time de futebol

- votou na ultima eleição

Medidas de tendência central: média e mediana

Onde está o centro da distribuição

Se você tiver que sugerir um único número para resumir uma variável numérica, qual você sugeriria?

Mediana

- Ordene as observações:

- Se o n é ímpar, a mediana é a observação central:
  78, 91, 94, 98, 99, 101, 103, 105, 114.
  mediana = 99.

- Se o n é par, divida as duas observações centrais:
  78, 91, 94, 98, 99, 101, 103, 105, 114, 121.
  mediana = 100.

Média

$$\bar{y} = \frac{Total}{n} = \frac{\sum y}{n}$$

- 78, 91, 94, 98, 99, 101, 103, 105, 114.

média = 98

Mediana ou média?

Se acrescentamos 1200 (um valor extremo) na distribuição:

- 78, 91, 94, 98, 99, 101, 103, 105, 114, 1200.
  mediana = 100.

- 78, 91, 94, 98, 99, 101, 103, 105, 114, 1200.
  média = 208

Mediana ou média?

- A média é sensível a valores extremos.
- Como uma distribuição assimétrica, a mediana é melhor opção.
- Com uma distribuição simétrica, a média representa bem os dados.

Stephen Jay Gould

- A mediana não é a mensagem

- https://coelhoprecambriano.blogspot.com/2020/07/a-mediana-nao-e-mensagem-por-stephen.html

Medidas de posição relativa

Quantis

O quantil é definido como segmento de tamanho igual de uma determinada população. Uma das métricas mais comuns em análise estatística, a mediana, é na verdade apenas o resultado da divisão de uma população em dois quantis.

Um quintil é um dos 4 valores que dividem os dados em 5 partes iguais, cada uma sendo 1/5 (20%). Uma população dividida em tercis tem 3 partes iguais, enquanto uma dividida em quartis tem 4 partes.

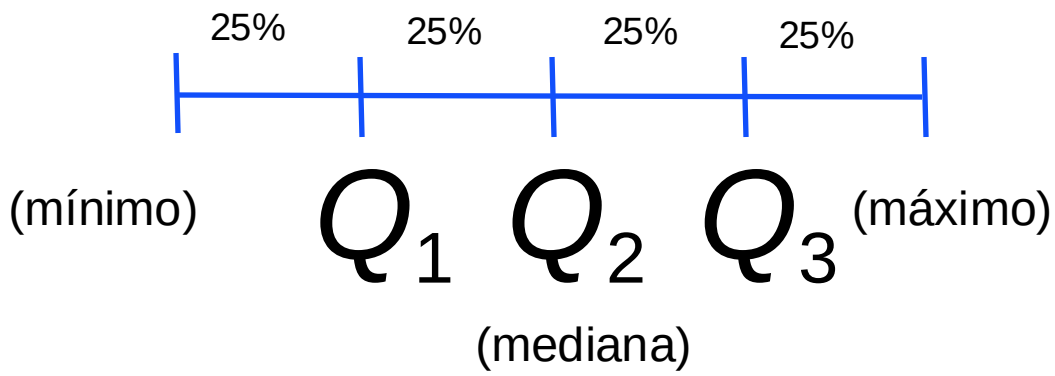
Os quantis são mensurados de formas de diferentes, mas sempre parte de uma distribuição de dados ordenado do menor para o maior valor, ou vice-versa.

Quartis

Um quartil divide os dados em três pontos – um quartil inferior, a mediana e o quartil superior – para formar quatro grupos do conjunto de dados.

Assim como a mediana divide os dados na metade, de modo que 50% da medida fique abaixo da mediana e 50% acima dela, o quartil divide os dados em quartos para que 25% das medidas sejam menores que o quartil inferior, 50 % sejam menores que a mediana e 75% sejam menores do que o quartil superior.

Quartis



Quartis

Na distribuição: 78, 91, 94, 98, 99, 101, 103, 105, 114, 117, 121

O 94 é o quartil inferior, também conhecido como Q1

O 101 é a mediana: metade dos números estão abaixo dele, metade acima

O 114 é o quartil superior, também conhecido como o Q3

78, 91, **94**, 98, 99, **101**, 103, 105, **114**, 117, 121

Observe que 25% dos casos estão abaixo de 94; 50% dos casos estão abaixo de 101; 75% dos casos estão abaixo de 114

Quintil

- Um **quintil** é um valor estatístico de um conjunto de dados que representa 20% de uma determinada população.
- Portanto, o primeiro **quintil** representa o quinto mais baixo dos dados (1% até 20%); o segundo **quintil** representa o segundo quinto (> 20% até 40%) e assim por diante.

Qual é o percentual de renda que cada quintil recebe no Brasil?

1. Ordene as famílias (ou indivíduos) do menor para o maior.

2. Calcule os quatro valores que dividirão a distribuição em 5 faixas

3. Some a renda de todos os indivíduos de um determinado quintil e calcule quanto esse
   resultado representa do percentual total

4. Em 2015, por exemplo, os 20% mais pobres ficavam com 3,6% da renda; enquanto os 20%
   mais ricos ficavam com 56% da renda

A distribuição de renda no Brasil

https://ourworldindata.org/grapher/income-shares-by-quintile-pip?country=~BRA

Decil

Se dividirmos uma distribuição em dez segmentos temos os **decis**. O procedimento é o mesmo: o primeiro **decil** representa o décimo mais baixo dos dados (1% até 10%); o **decil** o seguinte representa o segundo décimo (> 10% até 20% dos dados).

Imagine uma turma em que os alunos tenham tirado diferentes notas entre 20 e 100. Digamos que o valor do primeiro **decil** seja 35;

Isso significa que os alunos que tiraram até 35 estão no primeiro **decil**. Se a nota que demarca o **decil** superior é 90, isso significa que os alunos que estão nessa faixa tiraram pelo menos 90.

Desigualdade nos Estados Unidos

https://www.youtube.com/watch?v=QPKKQnijnsM&ab_channel=politizane

2.3.Medidas de dirpersão

desvio padrão e intervalo interquartil (IQR)

Medidas de dispersão

Desvio padrão

- O desvio padrão de um conjunto de dados, expresso pela letra *s,* é uma medida que expressa quanto os valores desviam da média.

- Fórmula:

$$s = \sqrt{\frac{\Sigma(x - \overline{x})^2}{n - 1}}$$

Como calcular o desvio padrão

Salário de quatro indivíduos:

1792 1666 1362 1614 1460 1867 1439

$$\bar{X} = \frac{1792 + 1666 + 1362 + 1614 + 1460 + 1867 + 1439}{7}$$

$$= \frac{11{,}200}{7}$$

$$= 1600$$

Como calcular o desvio padrão

observações x_i	Desvios $x_i - \bar{x}$	Desvio ao Quadrado $(x_i - \bar{x})^2$
1792	1792–1600 = 192	$(192)^2$ = 36,864
1666	1666 –1600 = 66	$(66)^2$ = 4,356
1362	1362 –1600 = -238	$(-238)^2$ = 56,644
1614	1614 –1600 = 14	$(14)^2$ = 196
1460	1460 –1600 = -140	$(-140)^2$ = 19,600
1867	1867 –1600 = 267	$(267)^2$ = 71,289
1439	1439 –1600 = -161	$(-161)^2$ = 25,921
	soma = 0	soma = 214,870

$$s^2 = \frac{214,870}{7-1} = 35,811.67$$

$$s = \sqrt{35,811.67} = 189.24$$

Desvio padrão: propriedades

- O valor do desvio padrão é sempre positivo (nunca negativo).
- O valor do desvio padrão aumenta dramaticamente com a inclusão de um ou mais valores extremos (outliers).
- A unidade do desvio padrão é a mesma unidade dos dados originais.

O intervalo interquartil (IQR)

- Os quartis mais baixos e mais altos são os 25^{th} e 75^{th} percentis dos dados.

- O intervalo interquartil = $Q_3 - Q_1$

- O IQR indica qual "território" a metade central dos dados cobrem.

Exemplo: intervalo interquartil (IQR)

- O intervalo interquartil da distribuição:

Max	9.0
Q3	7.6
Median	7.0
Q1	6.6
Min	3.7

- $Q_3 - Q_1 = 1.0$
- Observamos que 50% dos dados estão entre 7.6 e 6.6; portanto próximos da mediana.

“Pares” de medidas

- Média + desvio padrão
- Mediana + intervalo interquartil.

Taxas e Indicadores

Indicador

- Um número que sumariza informação quantitativa de um determinado fenômeno.

  http://data.worldbank.org/indicator

Taxas por população

- Por que dividir um número absoluto por 100, 1.000, 10.000, 1000.000?

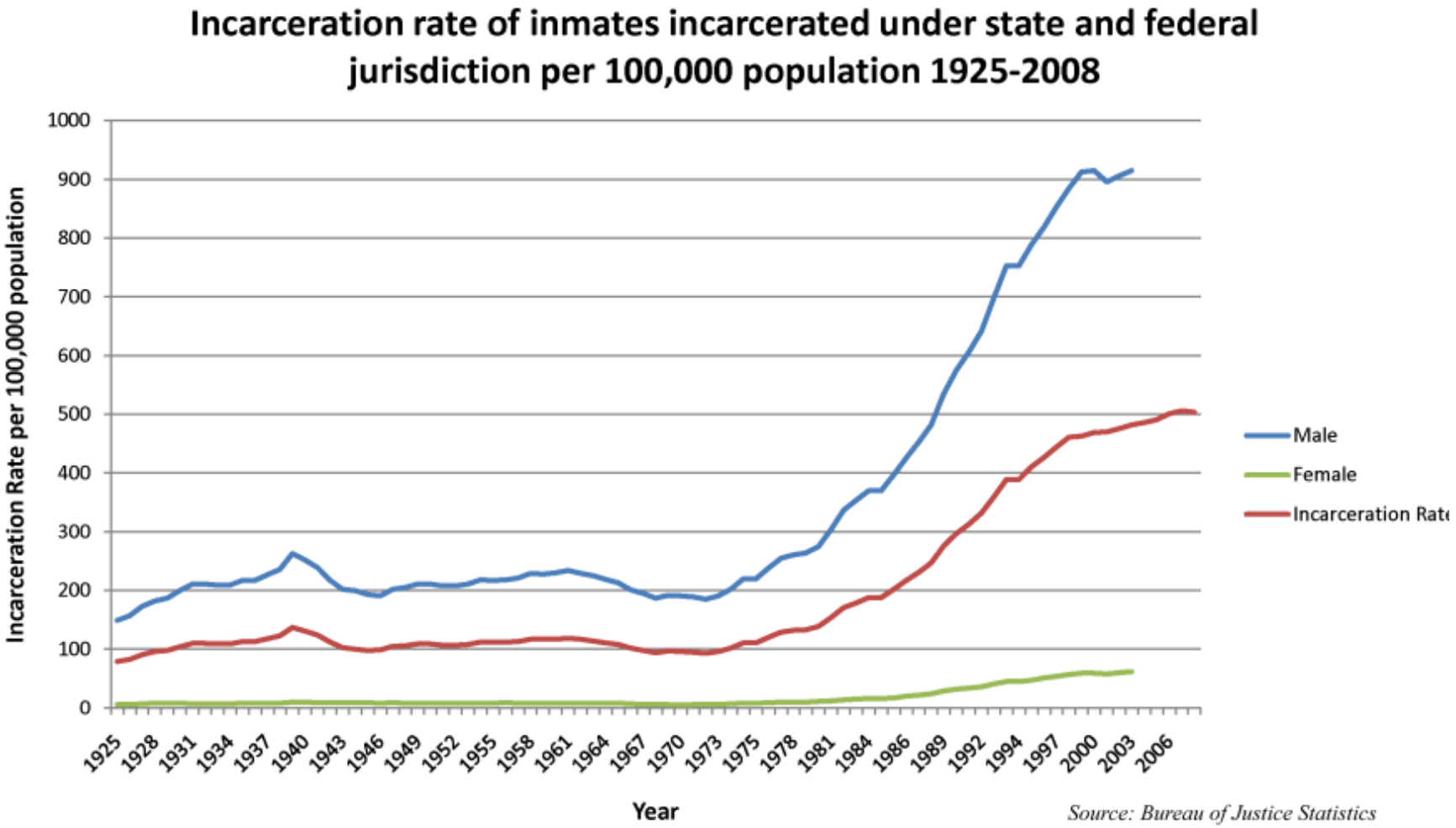
- Para garantir a comparação no tempo e no espaço, sempre relacionando o número de episódios sobre o tamanho da população.

Taxas por população

- Qual é a cidade mais violenta do Brasil?
- Qual estado tem mais casos de AIDS?
- O número de presos está diminuindo nos Estados Unidos?

Taxas por população

- Qual é a cidade mais violenta do Brasil? Simões Filho (BA)= 139 por 100 mil/ Brasil= 27 por 100 mil.
- Qual estado tem mais casos de AIDS? Rio Grande do Sul = 40 casos por 100 mil/ Brasil= 20 por 100 mil.
- O número de presos está diminuindo nos Estados Unidos? Não


Incarceration rate of inmates incarcerated under state and federal jurisdiction per 100,000 population 1925-2008
Incarceration Rate per 100,000 population
1000
900
800
700
600
500
400
300
200
100
0
1925
1928
1931
1934
1937
1940
1943
1946
1949
1952
1955
1958
1961
1964
1967
1970
1973
1975
1978
1981
1984
1986
1989
1992
1994
1997
2000
2003
2006
Year
Male
Female
Incarceration Rate
Source: Bureau of Justice Statistics

Como calcular a taxa de homicidio por 100 mil?

Estado do Rio de Janeiro

- 2012: 4.081
- 2013: 4.761

- 2013: 4.761/16.621.238 = 0,002864 x 100.000 = 28,64

- 2012: 4.081/16.431.019 = 0,002248 x 100.000 = 24,83

Como calcular a taxa de homicidio por 100 mil?

Estado do Rio de Janeiro

- 2012: 4.081
- 2013: 4.761

Crescimento de 16,6 %

- 2013 = 28,64
- 2012 = 24,83

Crescimento de 15,3%

Indicadores para cientistas políticos

- Índice fragmentação de Rae (F).
- Número efetivo de partidos (N).
- Índice de desproporcionalidade Gallagher.
- Índice de Rice.
- Índice de volatilidade.
- Taxa de comparecimento.
- Taxa de brancos e nulos.

Índice de rice

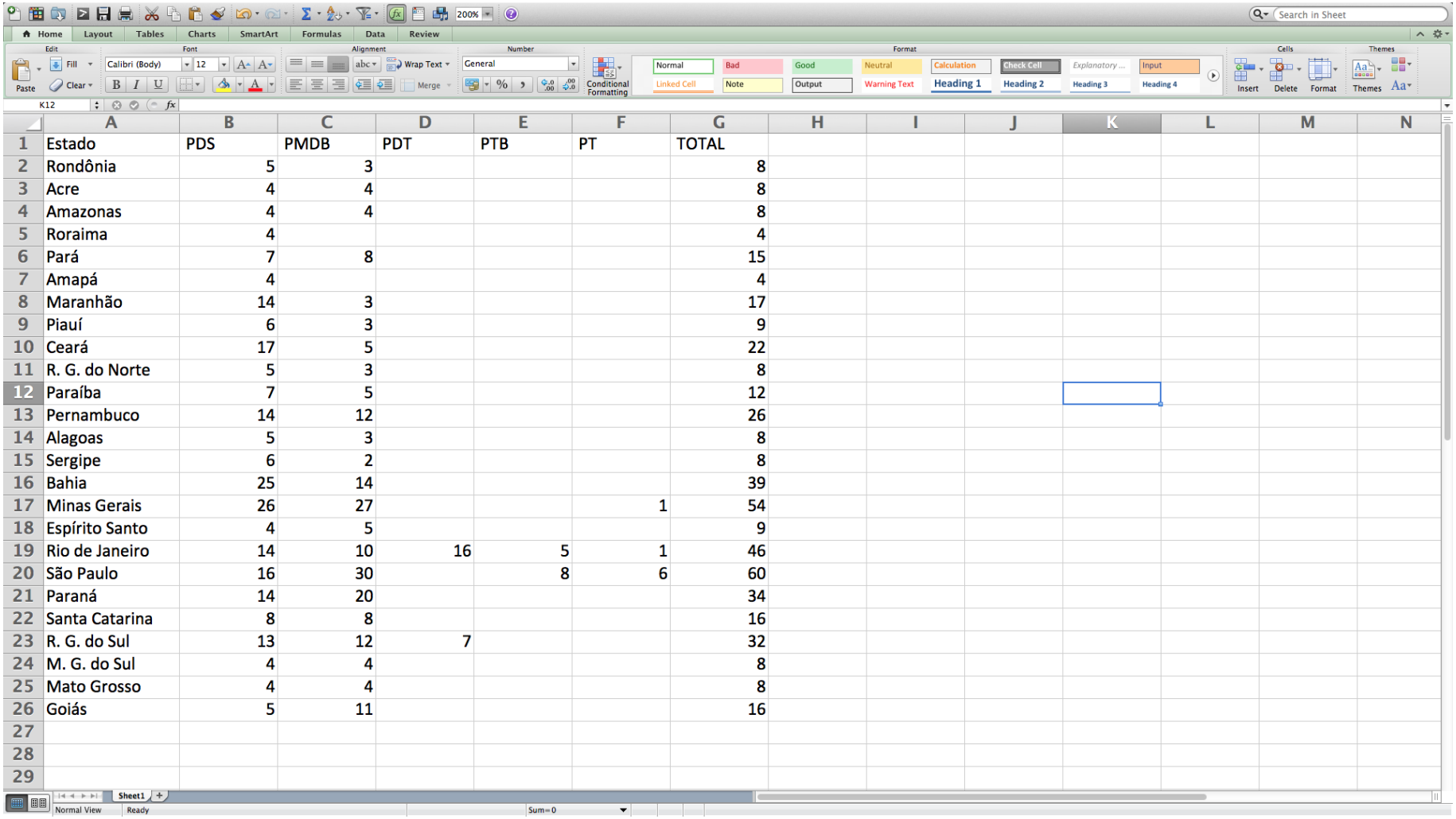
- Utilizado para medir votações nominais na Câmara dos Deputados:

% lado majoritário - % lado minoritário

- Em uma votação em que 80% da bancada votou sim e 20%votou não:

  Rice = 80 - 20 = 60

Que estado tem maior dipersão na representação?



	A	B	C	D	E	F	G
1	Estado	PDS	PMDB	PDT	PTB	PT	TOTAL
2	Rondônia	5	3				8
3	Acre	4	4				8
4	Amazonas	4	4				8
5	Roraima	4					4
6	Pará	7	8				15
7	Amapá	4					4
8	Maranhão	14	3				17
9	Piauí	6	3				9
10	Ceará	17	5				22
11	R. G. do Norte	5	3				8
12	Paraíba	7	5				12
13	Pernambuco	14	12				26
14	Alagoas	5	3				8
15	Sergipe	6	2				8
16	Bahia	25	14				39
17	Minas Gerais	26	27			1	54
18	Espírito Santo	4	5				9
19	Rio de Janeiro	14	10	16	5	1	46
20	São Paulo	16	30		8	6	60
21	Paraná	14	20				34
22	Santa Catarina	8	8				16
23	R. G. do Sul	13	12	7			32
24	M. G. do Sul	4	4				8
25	Mato Grosso	4	4				8
26	Goiás	5	11				16
27							
28							
29							

Número efetivo de partidos (N)

- Utilizado para medir a dispersão (ou concentração) de uma distribuição nas eleições (votos) ou no Legislativo (cadeiras).

$$N = \frac{1}{\sum_{i=1}^{n} p_i^2}$$

- Um partido com 70% e outro com 30%, qual é o índice?

Número efetivo de partidos (N)

- Um partido com 70% e outro com 30%, qual é o índice?

0,70 x 0,70 = 0,49

0,30 x 0,30 = 0,09

Somatório = 0,58

1÷ 0,58 = 1,72

Exercicio

Calcular o N nas eleições para Câmara dos Deputados em 1982 dos seguintes estados:

- São Paulo
- Rio de Janeiro
- Acre

Aula 4: Explorando variáveis quantitativas com gráficos

Z-score

- Comparamos dados individuais em relação à média utilizando o z-score.
- Simbolizado pela letra

$$z = \frac{observação - média}{desvio\ padrão}$$

Benefícios da Padronização

- Valores padronizados são convertidos das unidades originais para a unidade estatística de desvio padrão da média.
- Assim, podemos comparar valores que são medidos em diferentes escalas, com diferentes unidades e extraídos de diferentes populações.

Exemplo de Padronização: corrida de 800 metros

Média: 137 segundos
Desvio padrão: 5 segundos

- Corredor A: 129 s

  (129-137) / 5 = -8/5 **z = -1.6**

- Corredor B: 140 s

  (140-137)/5 = 3/5 **z = 0.6**

Exemplo de Padronização: salto em distância

Média: 6 metros
Desvio padrão: 30 cm

- Saltador A: 6.60

  (6.60 – 6.00) / 30 **z = 2.0**

- Saltador B: 5.84

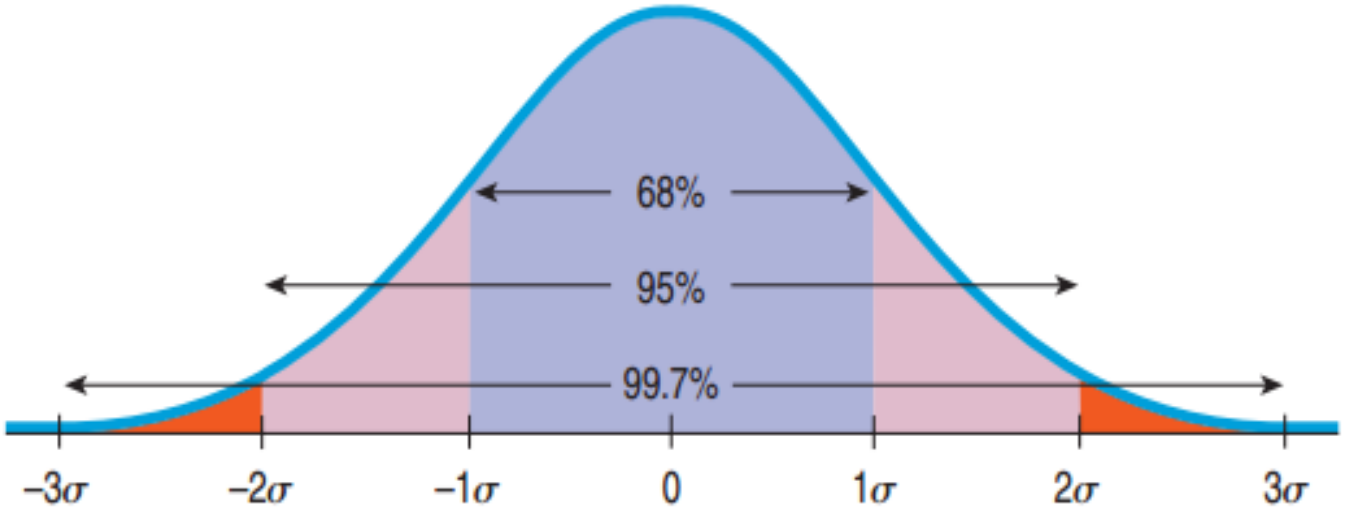
  (5.84 – 6.00) / 30 **z = - 0.4**

Padronização com o z-score

- Um *z-score* dá uma indicação de quão incomum é um valor, na medida em que ele mostra quanto o valor dista da média.
- Um *z-score* negativo diz que o valor está *abaixo* da média, enquanto um *z-score* positivo mostra que o valor está acima da média.
- Quanto maior o *z-score*(negativo ou positivo), mas ele é “incomum”.

A regra 68-95-99.7

- 68% dos valores estão em 1 desvio padrão da média.
- 95% dos valores estão a 2 desvios padrão da média.
- 99.7% dos valores estão a 3 desvios padrão da média.



A regra 68-95-99.7