Universidade de São Paulo – USP
Faculdade de Filosofia, Letras e Ciências Humanas – FFLCH
Programa de Pós-Graduação em Geografia Física

Seminários de Pesquisa em Geografia

Introdução à análise descritiva de dados

Professores: Emerson Galvani

São Paulo, Inverno/Primavera de 2017

Estrutura da aula

1) Tipos de variáveis
2) Medidas de Tendência Central
(Média, moda, mediana, amplitude e valor máximo e mínimo)
3) Medidas de dispersão
(Desvio em relação à média, variância, desvio padrão e coeficiente de variação)
4) Distribuição de frequência
(Frequência absoluta e relativa, probabilidade e tempo de retorno)
5) Correlação e Regressão Linear
(Diagrama de dispersão, coeficiente de correlação e dispersão, regressão linear e teste de significância)
6) Dígitos significativos e Arredondamento de dados
7) Atividades
8) Leituras complementares

O que é estatística?

A estatística é uma ciência que se dedica **à coleta, análise e interpretação de dados.** Preocupa-se com os métodos de coleta, organização, resumo, apresentação e interpretação dos dados, assim compreender as situações e concluir com base **objetiva** sobre o conjunto de dados(Vieira, 1999).

No contexto da Geografia a **Geoestatística** é um ramo da Estatística Espacial que usa o conceito de funções aleatórias para incorporar a dependência espacial aos modelos para variáveis georreferenciadas.

A estatística nos ajuda a responder questões como?

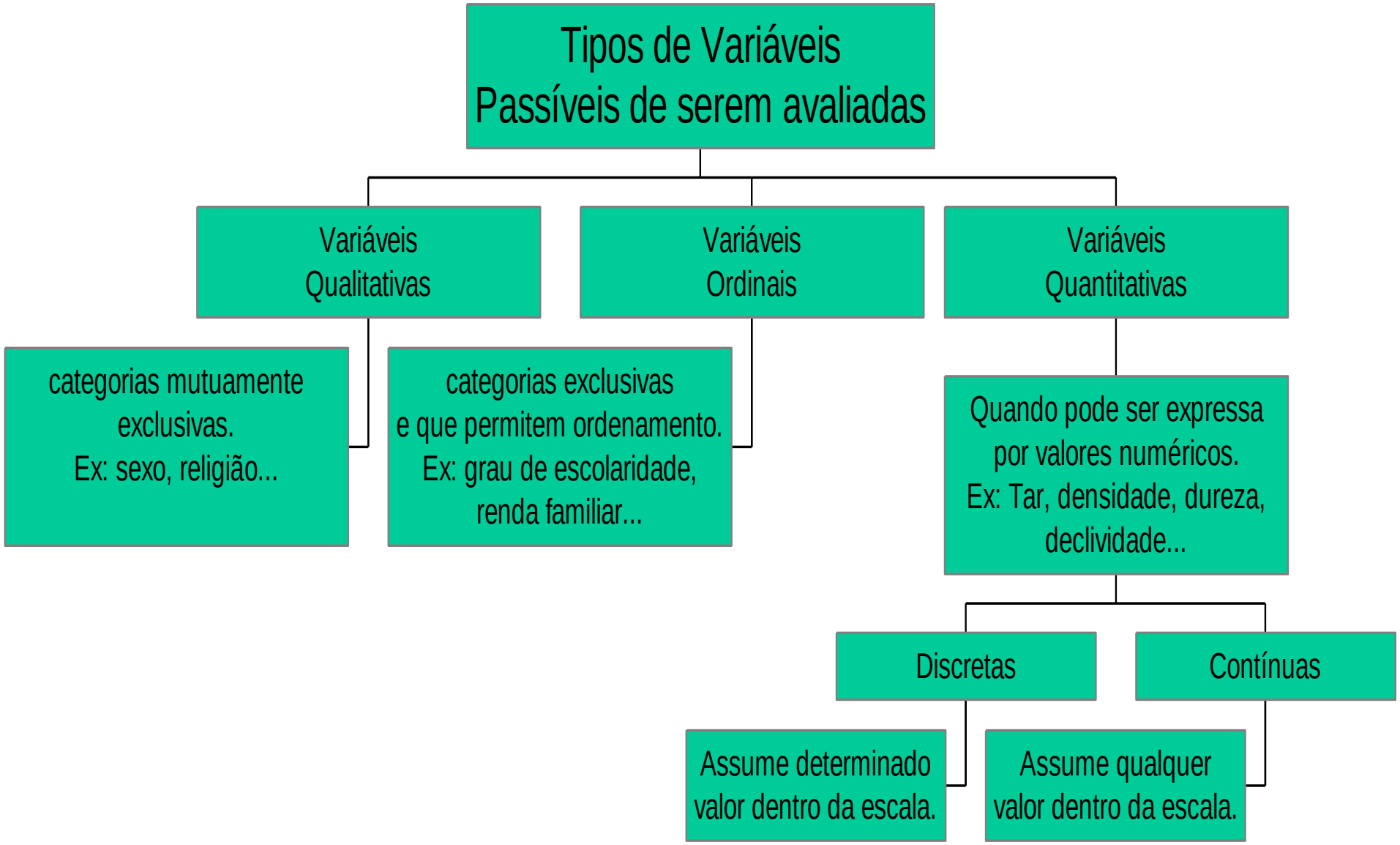
Esses dados são iguais ou diferentes?

O valor de 24,5 oC é igual ou diferente a 24,8 oC?

Textos como: “A temperatura do ar no ambiente 1 foi de 32,4oC ***ligeiramente*** superior ao ambiente 2 com 32,1 oC”.

Com a análise estatística é possível concluir com mais objetividade se o valor é igual ou diferente tornando o texto mais objetivo e conclusivo, e portanto com mais aceitação científica.

Análise de Dados em Geografia



Em Climatologia, especificamente, temos:

Temperatura do ar – contínua
(em qualquer hora do dia há uma temperatura)

Precipitação pluvial – discreta
(tem hora que tem chuva outras não)

Radiação solar onda curta – discreta
(apenas durante o dia)

Umidade do ar – contínua
(em qualquer hora do dia há um valor de umidade)

e assim por diante para outras áreas.

MEDIDAS DE TENDÊNCIA CENTRAL

1. MÉDIA (M)

Somatório de todos os elementos dividido por *n*.

$$\overline{X} = \frac{\sum_{1}^{n} X_i}{n}$$

2. MODA (MO)

É o valor que ocorre com maior freqüência.

3. MEDIANA (ME)

É o valor que ocupa posição central do conjunto de dados ordenados. A mediana pode se apresentar em alguns casos melhor que a média pois não é influenciada por valores extremos.

4. Valor Máximo, Valor Mínimo e amplitude

MEDIDAS DE TENDÊNCIA CENTRAL

Exemplo:

Dados Brutos

A	B
121	171
171	152
158	170
173	168
184	169
163	171
157	190

Ordenados

A	B
121	152
157	168
158	169
163	170
171	171
173	171
184	190

M(A)=161,0

M(B)=170,1

MO(A)= ?

MO(B)=171,0

ME(A)=163,0

ME(B)=170,0

Em séries climatológicas de 50, 100, anos de dados a mediana é utilizada para caracterizar as regiões.

MEDIDAS DE TENDÊNCIA CENTRAL

Interpretando média e mediana: Exemplo

Dados Brutos

A	B
121/300	171
171	152
158	170
173	168
184	169
163	171
157	190

Ordenados

A	B
157	152
158	168
163	169
171	170
173	171
184	171
300	190

Média(A)=187,0

Média(B)=170,1

Mediana(A)=171,0

Mediana(B)=170,0

No caso das variáveis quantitativas, quando o valor da Mediana é muito diferente da Média, é aconselhável considerar sempre a Mediana como valor de referência mais importante.

MEDIDAS DE TENDÊNCIA CENTRAL

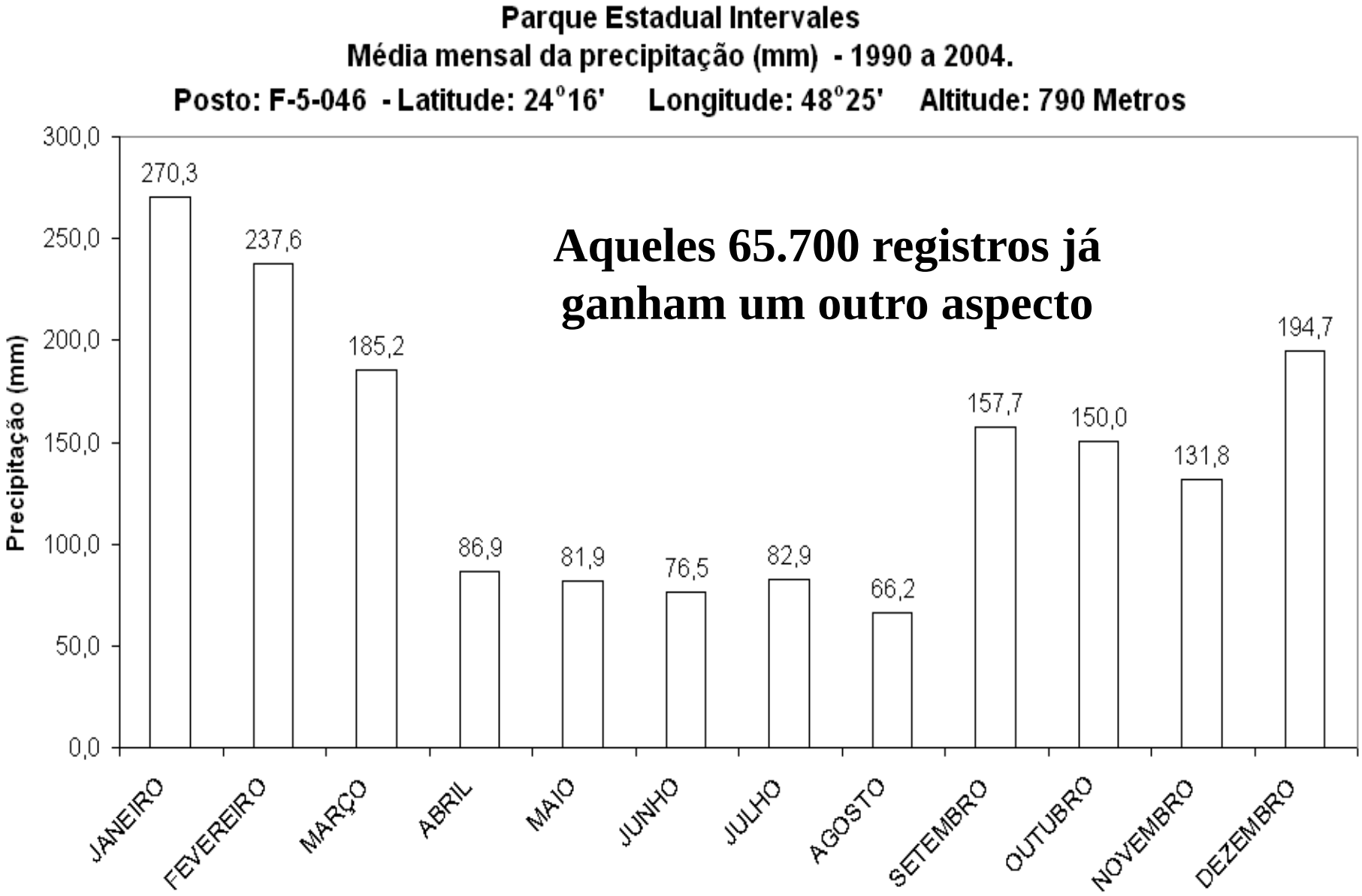
Parque estadual intervales - Precipitação diária em mm.

Série de 1990 a 2004 - Total de **65.700 registros**. O que fazer com isso???

Dia	JAN	FEV	MAR	ABR	MAIO	JUN	JUL	AGO	SET	OUT	NOV	DEZ
1	10,9	0,2	0,0	0,1	1,2	0,0	0,0	2,8	2,7	0,1	14,3	0,0
2	46,3	0,0	0,1	0,0	1,1	0,3	0,3	0,1	1,2	5,7	0,3	0,0
3	31,3	0,0	26,8	0,6	0,2	0,2	0,1	0,0	1,7	0,3	0,2	0,0
4	0,8	0,0	0,1	0,1	0,1	0,1	0,1	0,0	0,8	0,9	16,3	18,6
5	5,3	0,0	0,6	0,3	0,1	0,1	9,6	0,1	7,1	0,1	0,3	0,1
6	59,7	4,6	0,7	0,4	0,3	0,2	0,2	0,3	0,3	0,0	0,1	0,8
7	17,0	0,0	2,3	0,1	0,4	16,7	3,0	0,0	1,1	0,0	27,5	11,2

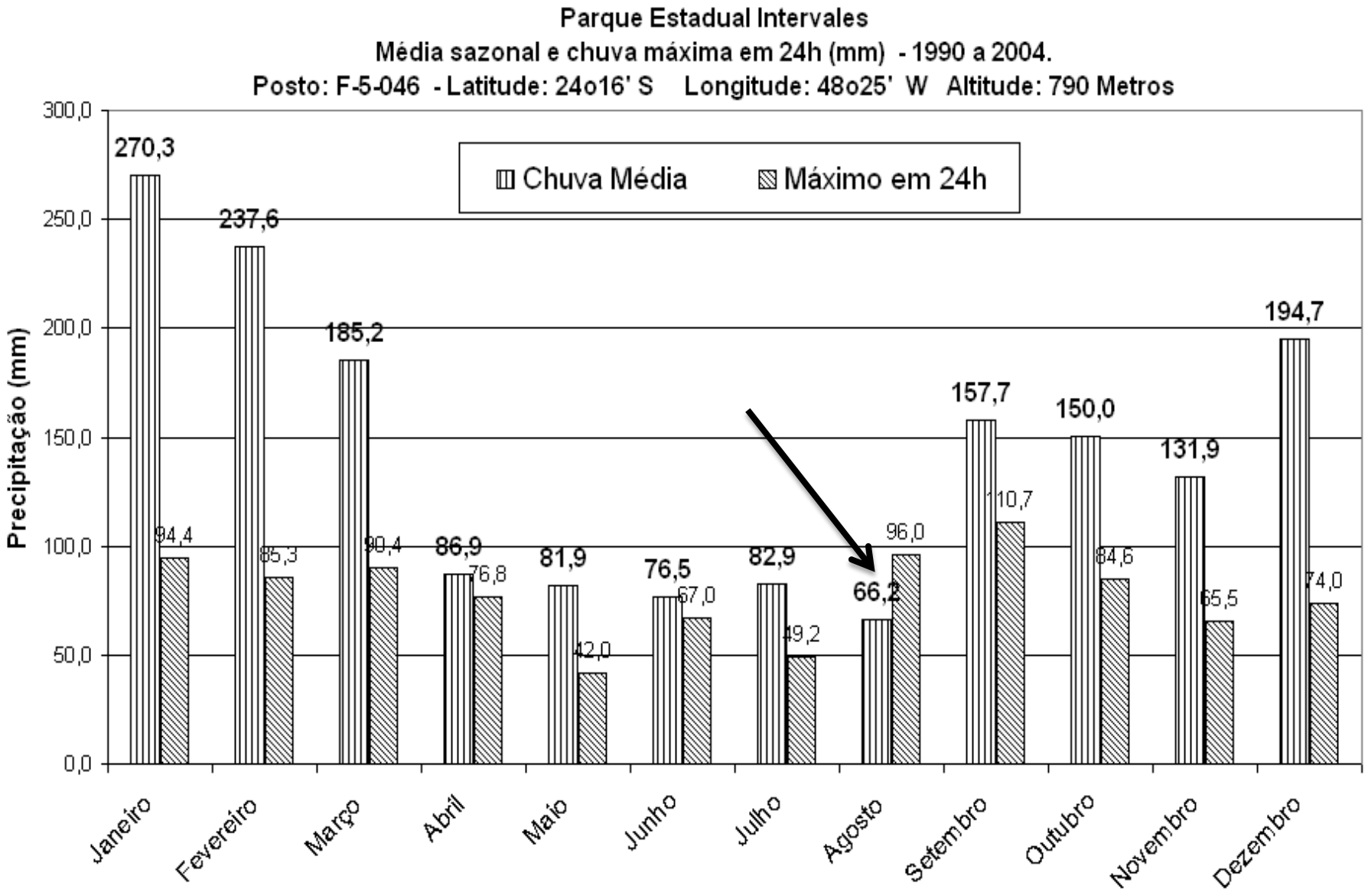
Tabela quase infinita que nos cansa só de caminhar na tela.
Utilizem recursos da **tabela dinâmica** do Excel para trabalhar com séries históricas extensas.

MEDIDAS DE TENDÊNCIA CENTRAL



Aqueles 65.700 registros já ganham um outro aspecto

MEDIDAS DE TENDÊNCIA CENTRAL

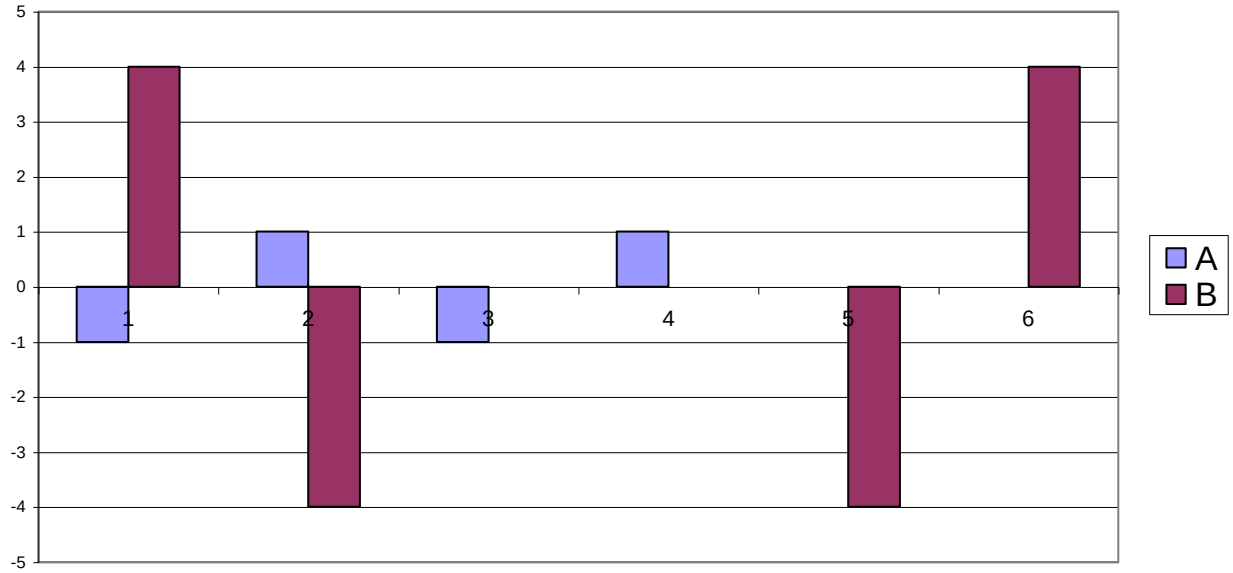


MEDIDAS DE DISPERSÃO

1. DESVIO EM RELAÇÃO A MÉDIA (DM)

É a diferença entre o valor observado é a média do conjunto.

A	B
4	9
6	1
4	5
6	5
5	1
5	9
M=5	M=5



As duas variáveis apresentam médias iguais contudo *variabilidade* de B é maior do A.

Atenção ao valor da média!!!

MEDIDAS DE DISPERSÃO

2. VARIÂNCIA DA AMOSTRA (S^2)

É a somatória dos quadrados dos desvios de cada observação em relação a média, dividido por *n-1 (grau de liberdade da amostra)*.

$$S^2 = \frac{\sum \left(X - \overline{X}\right)^2}{n-1}$$

$$S^2 = \frac{4}{6-1}$$

$$S^2 = 0{,}8$$

X	X-M	$(X-M)^2$
4	-1	1
6	1	1
4	-1	1
6	1	1
5	0	0
5	0	0
M=5		$\sum$=4

MEDIDAS DE DISPERSÃO

3. DESVIO PADRÃO (S)

É a raiz quadrada da variância.

$$S^2 = \frac{\sum\left(X - \overline{X}\right)^2}{n-1} \quad \Rightarrow \quad S = \sqrt{\frac{\sum\left(X - \overline{X}\right)^2}{n-1}}$$

4. COEFICIENTE DE VARIAÇÃO (CV)

É uma medida da variância e do desvio padrão expressa em porcentagem.

$$CV = \frac{100 * S}{\overline{X}}$$

A	B	C
4	9	9
6	1	1
4	5	1
6	5	2
5	1	8
5	9	9

$$CV(A) = 17{,}88\%$$

$$CV(B) = 71{,}56\%$$

$$CV(C) = 80{,}92\%$$

DISTRIBUIÇÃO DE FREQUENCIA

1. FREQUÊNCIA (f)

É o numero de vezes que determinado evento ocorreu.

2. FREQUÊNCIA RELATIVA (fr)

É o numero de vezes que determinado evento ocorreu (na) em relação ao número total de eventos observados (n).

$$fr = \frac{na}{n}$$

Espécie	(na)	Fr
A	32	31%
B	17	16%
C	43	41%
D	13	12%
Total	105	100%

DISTRIBUIÇÃO DE FREQUENCIA

3. PROBABILIDADE (P)

Expressa a relação entre o número de vezes que determinado evento ocorreu (na) e o número total de eventos observados (n).

$$P = fr = \frac{na}{n}$$

4. TEMPO DE RETORNO (T)

Período ou tempo de retorno é definido como o inverso da probabilidade.

$$T = \frac{1}{P = fr}$$

Espécie	(na)	Fr
A	32	31%
B	17	16%
C	43	41%
D	13	12%
Total	105	100%

$$T(A) = \frac{1}{0{,}31}$$

$$T(A) = 3{,}2$$

Ou seja, a cada 3,2 indivíduos 1 e da espécie A

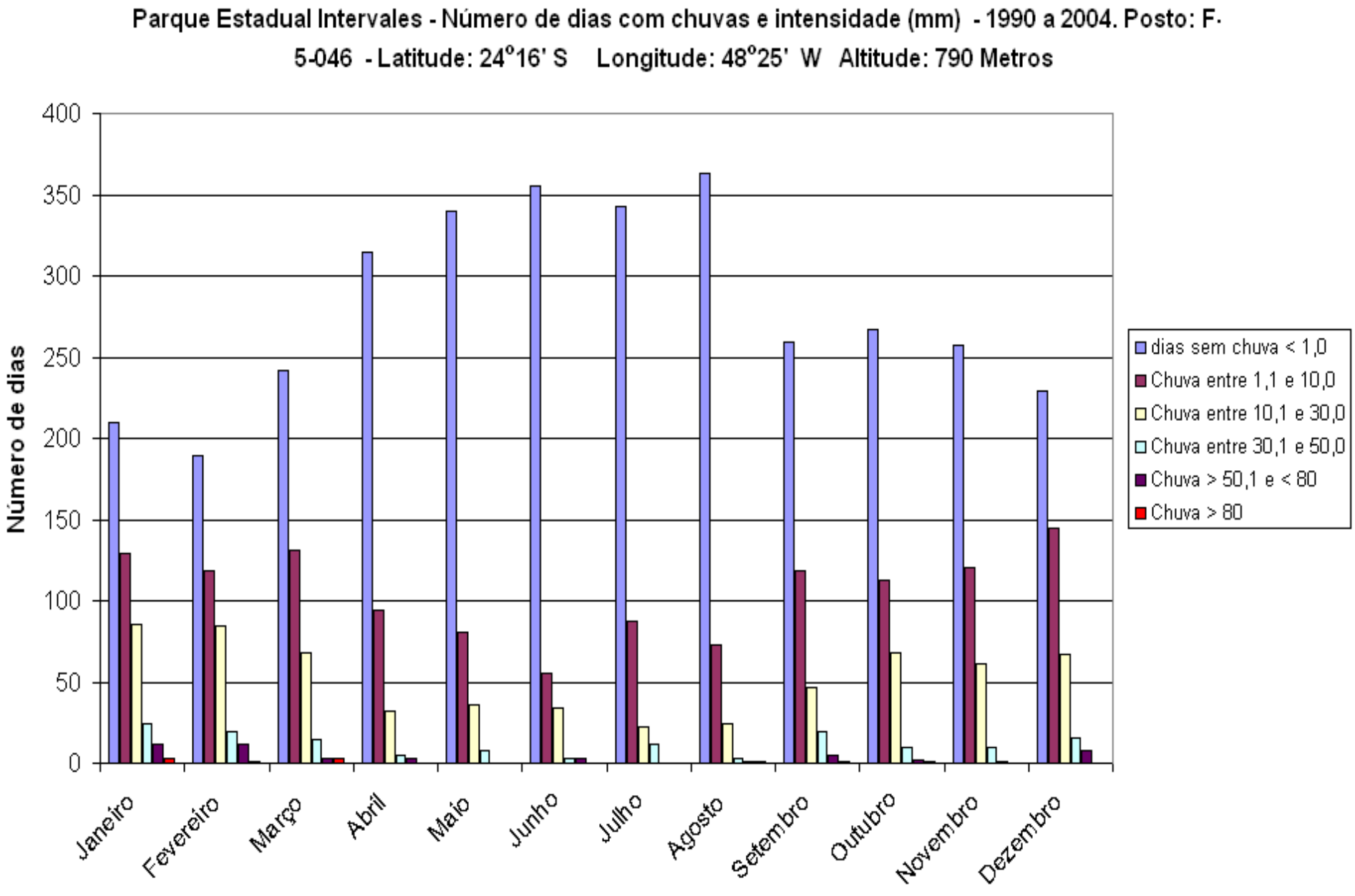
DISTRIBUIÇÃO DE FREQUÊNCIA

Parque Estadual Intervales - Precipitação diária em mm. Série de 1990 a 2004 - Total de 65.700 registros. O que fazer com isso???

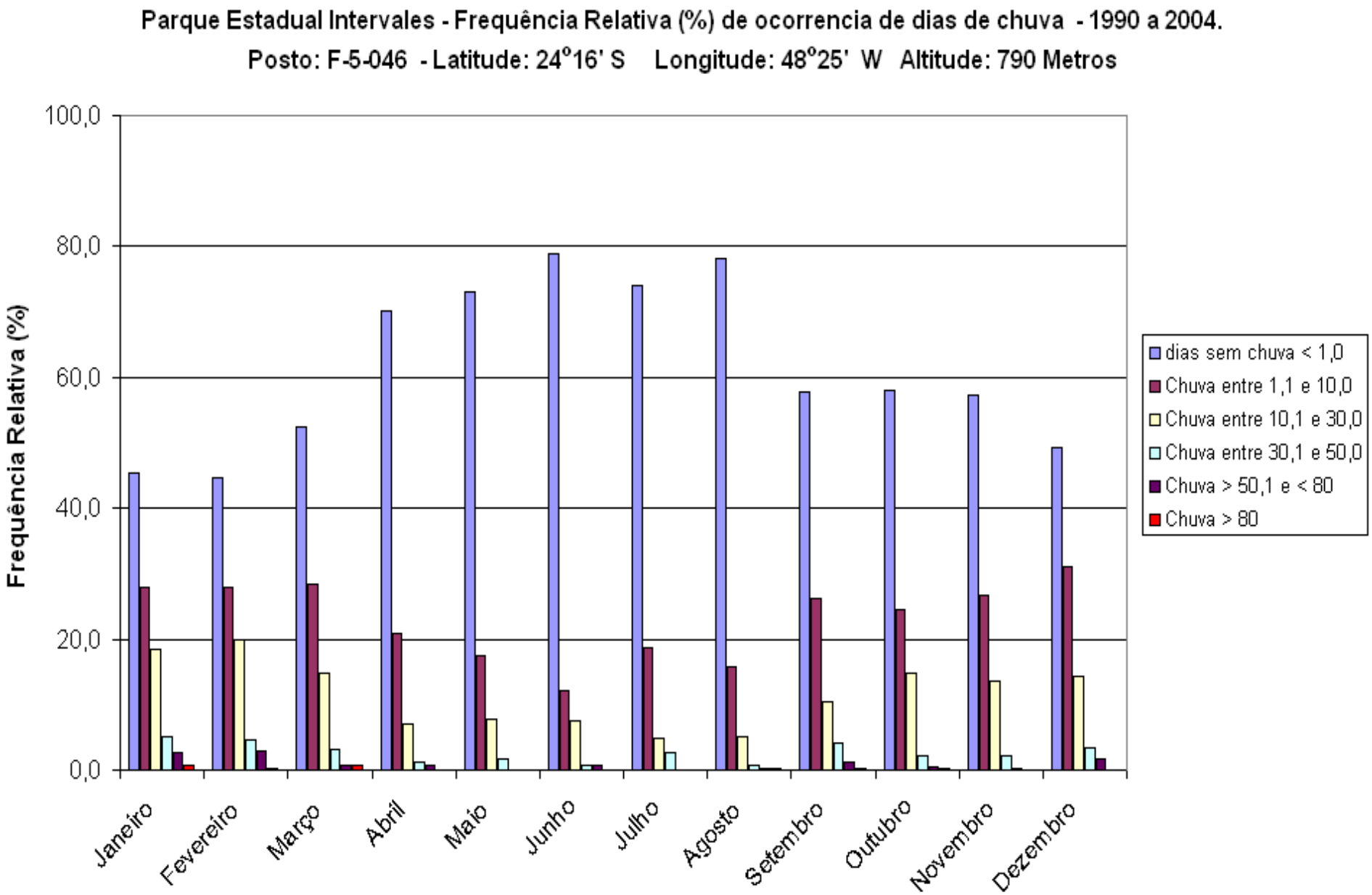
Número de dias
dias sem chuva < 1,0
Chuva entre 1,1 e 10,0
Chuva entre 10,1 e 30,0
Chuva entre 30,1 e 50,0
Chuva > 50,1 e < 80
Chuva > 80

Hierarquizar os eventos e determinar a Frequência de ocorrência desses eventos

DISTRIBUIÇÃO DE FREQUENCIA



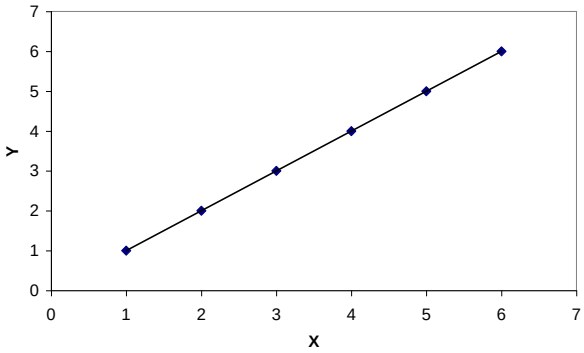
DISTRIBUIÇÃO DE FREQUENCIA



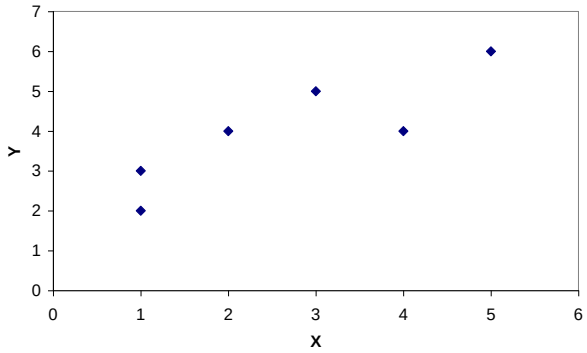
REGRESSÃO LINEAR

1. DIAGRAMA DE DISPERSÃO

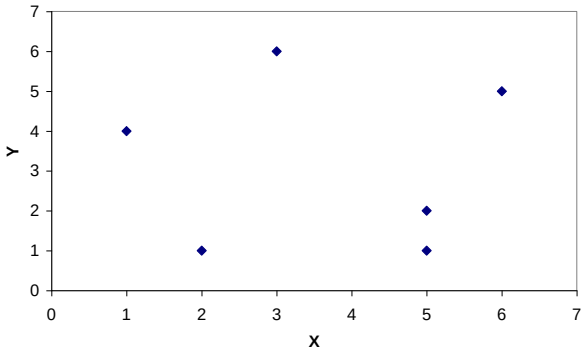
Expressa se a relação entre as variáveis é linear ou não.



BOA RELAÇÃO



RELAÇÃO RUIM



NÃO EXISTE

2. COEFICIENTE DE CORRELAÇÃO (R)

É a medida do grau de associação linear entre duas variáveis.

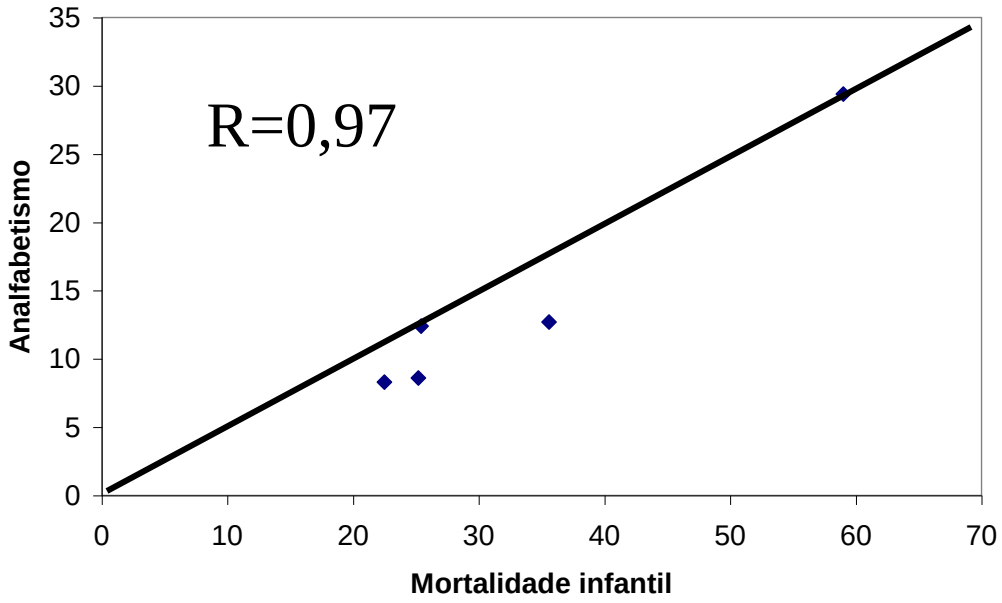
$$R = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sqrt{\left[\sum X^2 - \frac{(\sum X)^2}{n}\right] * \left[\sum Y^2 - \frac{(\sum Y)^2}{n}\right]}}$$

Parece complicado mas o Excel faz para nós.

REGRESSÃO LINEAR

2. COEFICIENTE DE CORRELAÇÃO (R)

	Mortalidade infantil	Analfa-betismo
N	35,6	12,7
NE	59,0	29,4
SE	25,2	8,6
S	22,5	8,3
CO	25,4	12,4

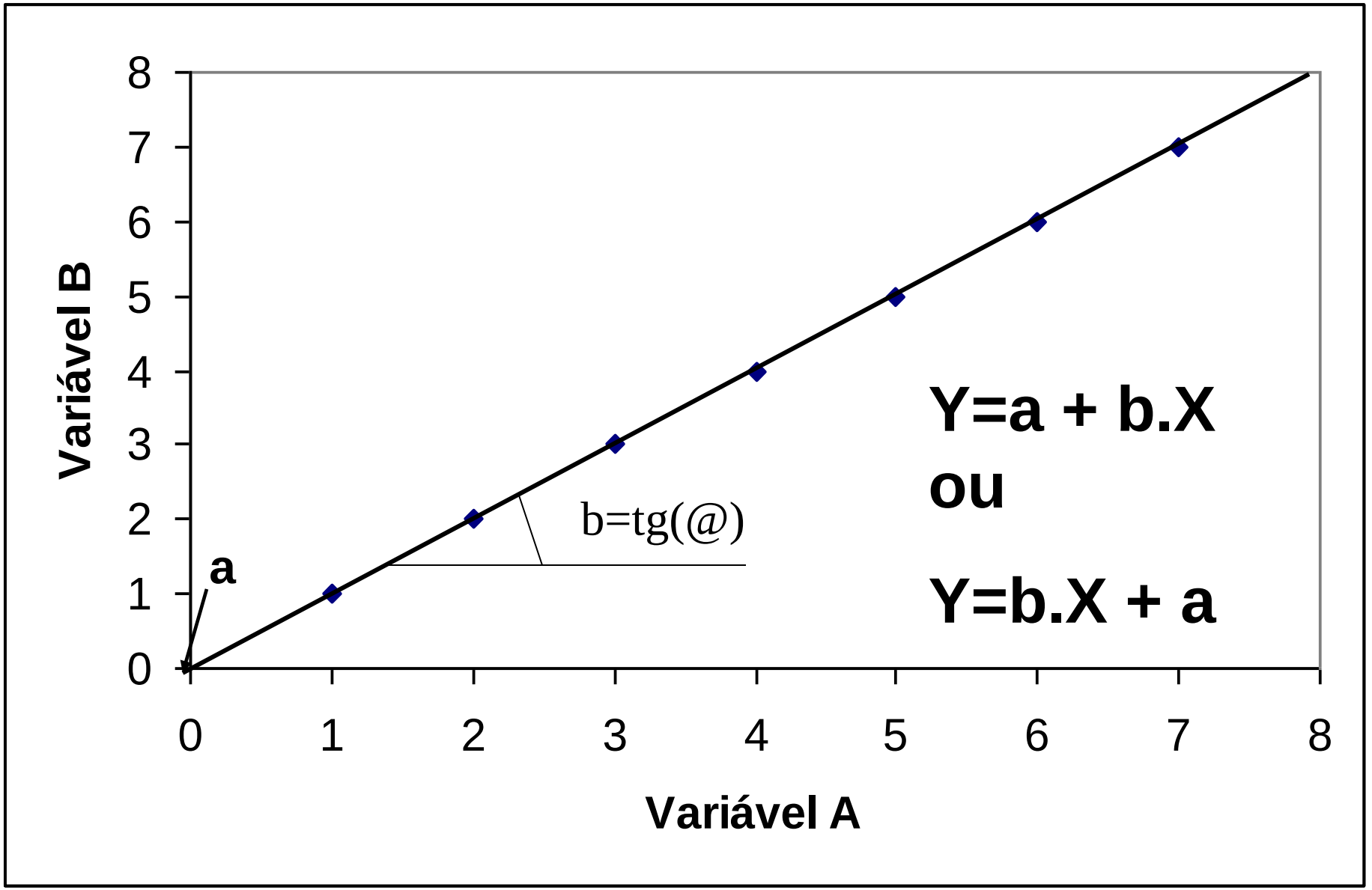


R = 0,97, ou seja, é possível explicar a relação entre analfabetismo e mortalidade infantil em até 97%.

Esse valor de R para uma amostra reduzida é significativo?

REGRESSÃO LINEAR

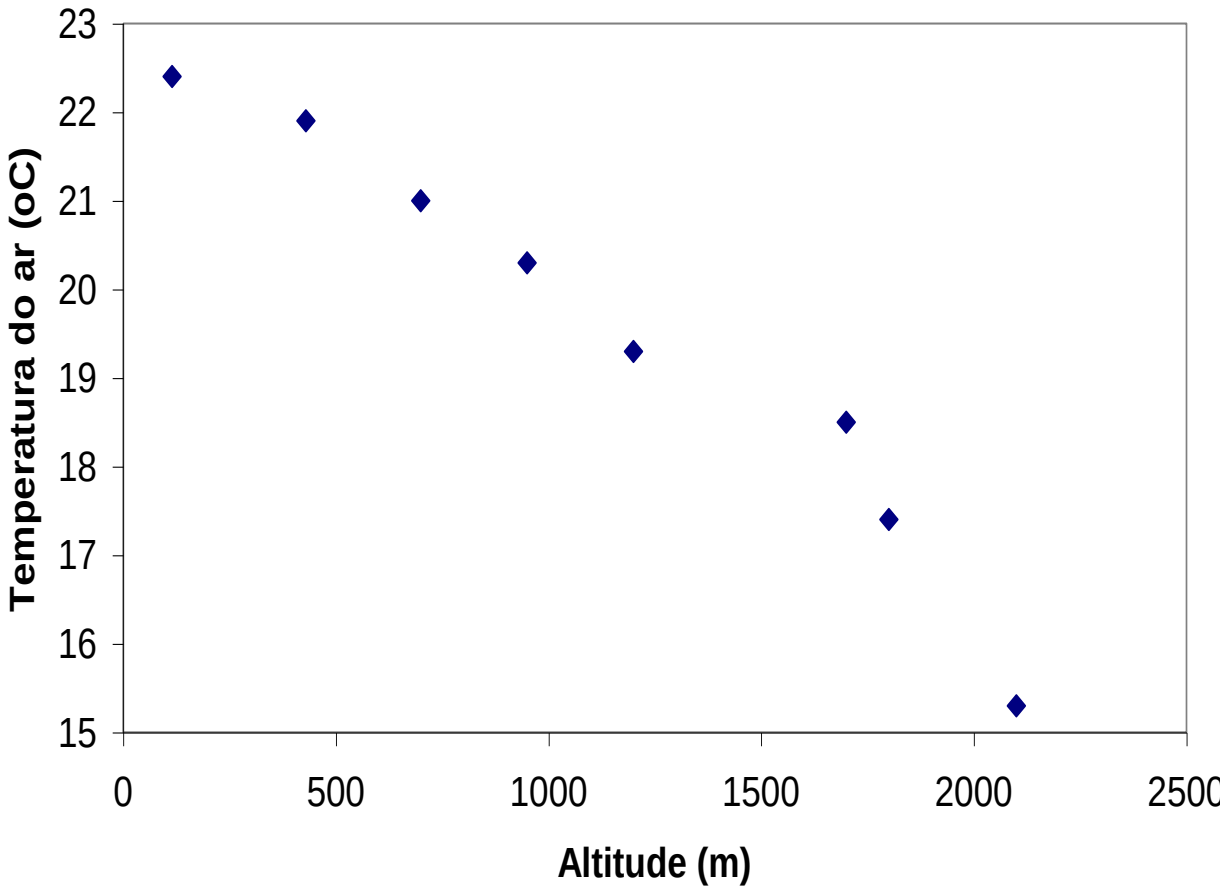
3. COEFICIENTE LINEAR *a* e *b* DA RETA DE REGRESSÃO



REGRESSÃO LINEAR

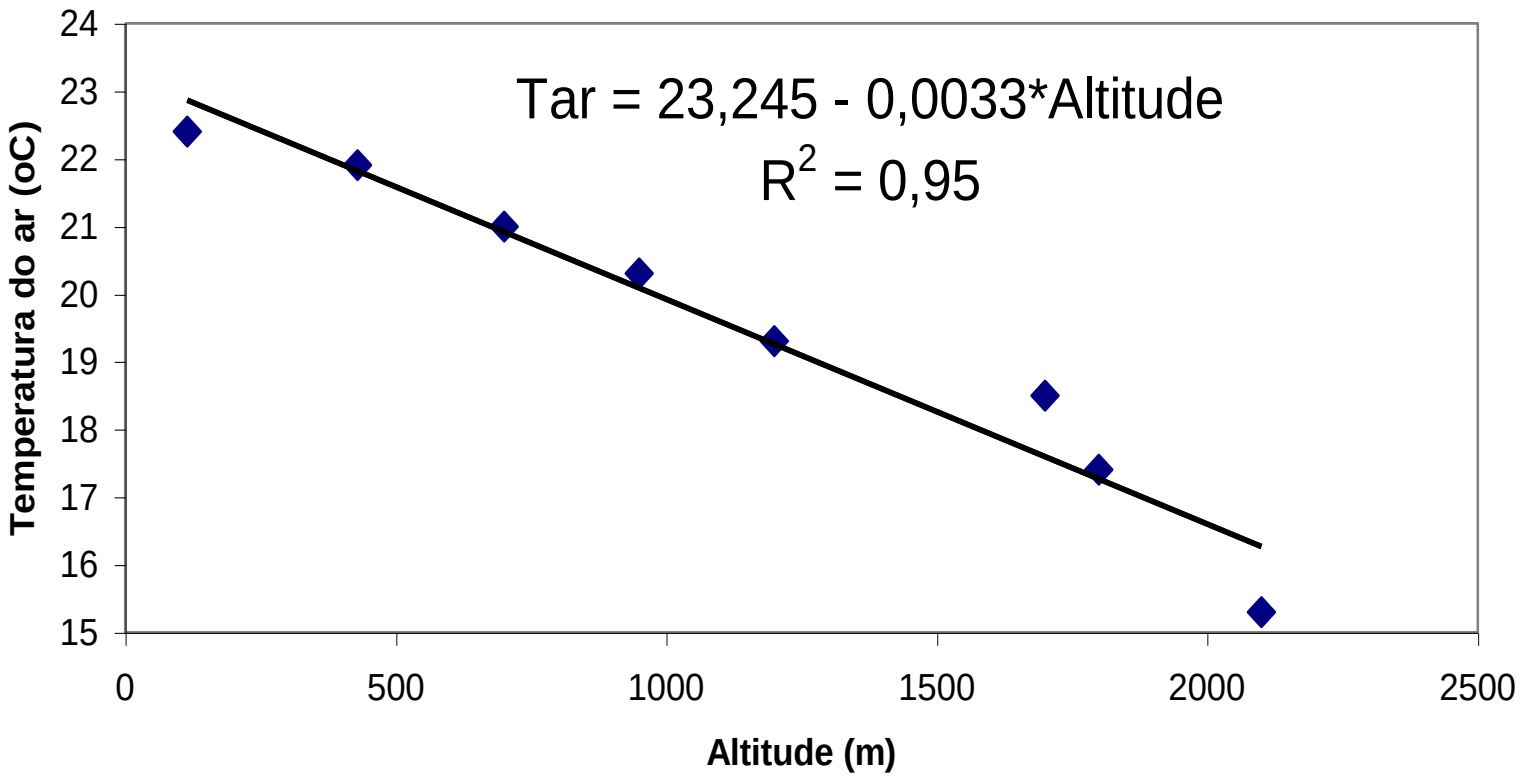
3. COEFICIENTE LINEAR *a* E *b* DA RETA DE REGRESSÃO

Altitude (m)	Tar (ºC)
2100	15,3
1800	17,4
1700	18,5
1200	19,3
950	20,3
700	21,0
430	21,9
115	22,4



REGRESSÃO LINEAR

3. COEFICIENTE LINEAR a E b DA RETA DE REGRESSÃO



Posso, então, estimar os valores de temperatura do ar para quaisquer outras altitudes nas quais não foram efetuadas medidas.

$$Tar = 25{,}3 - 0{,}0054.Altitude\ (m) \quad R^2 = 0{,}98$$

Por exemplo, a altitude **1500m** o valor de Tar, fazendo uso da equação, será de 17,2 ^{o}C.

O uso da regressão linear permite, portanto, uma redução da amostragem do trabalho de campo, permitindo maior rapidez e menor custo na obtenção dos dados. Cabe lembrar que a regressão linear só se aplica quando os elementos apresentam entre si uma relação de **dependência natural**.

Esse valor de correlação R é significativo para essa amostra?

-Cálculo da significância o valor R

Um simples exemplo de que a interpretação estatistica do coeficiente de correlação r se faz necessária está em que, se tivermos apenas dois pontos, r será $+1$ ou -1, pois quaisquer dois pontos estarão alinhados. Isso, entretanto, não quer absolutamente dizer que tenhamos um caso de correlação linear perfeita; simplesmente, a amostra é tão pequena que não se pode tirar qualquer conclusão. Vemos que a correta interpretação de um valor r calculado está diretamente ligada ao número de pontos com base no qual foi calculado.

Muitas vezes desejamos saber se um dado valor de r, combinado com o respectivo tamanho da amostra n, permite concluir, a um dado nível de significância α, que realmente existe correlação linear entre as variáveis. Testamos, então, as hipóteses

$$H_0, \quad \rho = 0,$$
$$H_1, \quad \rho \neq 0.$$

Esse teste pode ser feito através da quantidade

$$t_{n-2} = r\sqrt{\frac{n-2}{1-r^2}}, \tag{8.8}$$

que será testada como um t de Student com $n-2$ graus de liberdade[6]. O teste poderá também ser feito unilateralmente.

Após o cálculo de tn-2 com os dados da analise de correlação obtém o valor de limite para aceitação da hipótese.

Acompanhe cálculo na planilha anexo.

Como o valor calculado (26,7) é superior a tabelado (1,740) confirma-se que existe uma relação significativa entre Tar e Altitude.

Assim, no texto podemos escrever que a correlação é significativa ao nível de 95% de probabilidade pelo teste t de Student.

Sempre que o valor calculado for superior ao tabelado aceita-se a hipótese e diz que o r é significativo ao nível de probabilidade (no caso aqui 95%).

Tabela A6.5 Distribuições t de Student — valores de $t_{\nu,P}$, onde $P = P(t_\nu \geq t_{\nu,P})$

ν \ P	0,10	0,05	0,025	0,01	0,005
1	3,078	6,314	12,706	31,821	63,657
2	1,886	2,920	4,303	6,965	9,925
3	1,638	2,353	3,182	4,541	5,841
4	1,533	2,132	2,776	3,747	4,604
5	1,476	2,015	2,571	3,365	4,032
6	1,440	1,943	2,447	3,143	3,707
7	1,415	1,895	2,365	2,998	3,499
8	1,397	1,860	2,306	2,896	3,355
9	1,383	1,833	2,262	2,821	3,250
10	1,372	1,812	2,228	2,764	3,169
11	1,363	1,796	2,201	2,718	3,106
12	1,356	1,782	2,179	2,681	3,055
13	1,350	1,771	2,160	2,650	3,012
14	1,345	1,761	2,145	2,624	2,977
15	1,341	1,753	2,131	2,602	2,947
16	1,337	1,746	2,120	2,583	2,921
17	1,333	1,740	2,110	2,567	2,898
18	1,330	1,734	2,101	2,552	2,878
19	1,328	1,729	2,093	2,539	2,861
20	1,325	1,725	2,086	2,528	2,845
21	1,323	1,721	2,080	2,518	2,831
22	1,321	1,717	2,074	2,508	2,819
23	1,319	1,714	2,069	2,500	2,807
24	1,318	1,711	2,064	2,492	2,797
25	1,316	1,708	2,060	2,485	2,787
26	1,315	1,706	2,056	2,479	2,779
27	1,314	1,703	2,052	2,473	2,771
28	1,313	1,701	2,048	2,467	2,763
29	1,311	1,699	2,045	2,462	2,756
30	1,310	1,697	2,042	2,457	2,750
50	1,299	1,676	2,009	2,403	2,678
80	1,292	1,664	1,990	2,374	2,639
120	1,289	1,657	1,980	2,351	2,618
∞	1,282	1,645	1,960	2,326	2,576

Distribuição *t* de *Student* para diversos valores de probabilidade.

A maioria dos pacotes estatísticos nos fornece esses parâmetros estatísticos não sendo necessário recorrer a tabelas.

5. Dígitos significativos e Arredondamento de dados

Nos resultados, devem ser apresentados apenas os dígitos significativos, para evitar a falsa impressão de exatidão (VIEIRA, 1999). O resultado de um cálculo estatístico não deve conter mais dígitos significativos que os dados de menor precisão. Por ex:

hora	Temperatura do ar (^{o}C)
06h00min	18,2
08h00min	19,5
10h00min	20,6
12h00min	22,9
14h00min	24,5
média	21,14 = 21,1

5. Dígitos significativos e Arredondamento de dados

O arredondamento dos dados deverá seguir os seguintes critérios:

* Se você vai cortar dígitos e o resto é menor do que 5, apenas faça o corte;

* Se você vai cortar dígitos e o resto é maior do que 5, aumente o último número em uma unidade;

6. Dígitos significativos e Arredondamento de dados

* Se você vai cortar dígitos e o resto é exatamente igual a 5, a convenção é:

** Se o dígito anterior ao que vai ser cortado é par, apenas faça o corte;

** Se dígito anterior ao que vai ser cortado é impar, aumente esse dígito em uma unidade.

Esta prática faz com que, ao longo das operações, os aumentos e reduções devidos aos arredondamentos se compensem.

5. Dígitos significativos e Arredondamento de dados

Exemplo:

16,44 = 16,4

16,46 = 16,5

16,45 = 16,4

16,75 = 16,8

Frase do dia

É possível mentir usando estatísticas, mas se mente mais, e melhor, sem estatísticas. É preciso entender que as amostras podem levar a conclusões erradas. Contudo, as opiniões pessoais, sem base de dados, levam, em geral, a conclusões muito mais erradas.

Frederick Mosteller (Vieira, 1999)

Atividades

1)Para o conjunto de dados anexo determine todos os procedimentos apresentados nos slides de aula. Os dados foram coletados em trabalho de Campo de Disciplina de Estágio Supervisionado em Climatologia ao longo da estrada Caminhos do Mar.

2) Para calcular o desvio utilize P1 como referência.

3) Para os cálculos de correlação utilize os dados de temperatura do ar média em cada ponto e estabeleça a relação com a altitude (m) indicada na caderneta de campo. Calcule o tn-2 para 95% de probabilidade (0,05).

4) Para os cálculos de frequência estabeleça classes de temperatura do ar com intervalos de 1oC.

5) Organize o material em documento do editor de texto, comente os resultados.

Caderneta de campo

Ponto	Altitude (m)	Patm (mmHg)	UTM X (m)	UTM Y (m)	Descrição
P1	740	715	350.791	7.361.446	Planalto (vegetação porte baixo – Floresta Ombrófila Densa Montana). P1a: Próximo ao estacionamento.
P2	600	725	351.551	7.360.055	Alta Encosta (vegetação porte alto – Floresta Ombrófila Densa Montana). P2: Depois dos dutos d'água.
P3	500	734	351.985	7.360.262	Alta Encosta (vegetação porte alto – Floresta Ombrófila Densa Montana). P3: Calçada do Lorena, ponto com a mata mais fechada.
P4	400	743	352.166	7.360.502	Média Encosta (vegetação porte alto – Floresta Ombrófila Densa Submontana). P4: Vegetação com porte bem alto.
P5	300	751	352.746	7.360.708	Média Encosta (vegetação porte alto – Floresta Ombrófila Densa Submontana). P5: Presença de grande número de embaúbas.
P6	200	760	353.484	7.360.571	Média Encosta (vegetação porte alto – Floresta Ombrófila Densa Submontana). P6: Vegetação com porte alto.
P7	100	768	354.023	7.360.184	Baixa Encosta (vegetação porte baixo – Floresta Ombrófila Densa Submontana). P7: Ponto próximo a estrada e a Cubatão.
P8	30	770	353.910	7.360.053	Baixa Encosta (vegetação porte baixo – Floresta Ombrófila Densa Terras Baixas). P8: Próximo a base de Cubatão (estacionamento).

Consulta material

Galvani, Emerson. Estatística Descritiva em sala de aula. In: Geografia - práticas de campo, laboratório e sala de aula. São Paulo - SP: Editora Sarandi, 2011. v. 1. 528p.

Para outros métodos busque:

OLIVEIRA, Maria Rita Pelegrin de; GALVANI, Emerson. Eventos Extremos de Precipitação no Perfil Longitudinal Paraty (RJ) - Campos do Jordão (SP). Revista do Departamento de Geografia, São Paulo, p. 58-66, june 2017. ISSN 2236-2878. Disponível em: <http://www.revistas.usp.br/rdg/article/view/133419>. Acesso em: 17 sep. 2017. doi:http://dx.doi.org/10.11606/rdg.v0ispe.133419.

Grato pela atenção.