
FIAP
MBA

MBA em DATA SCIENCE & ARTIFICIAL INTELLIGENCE

STATISTICS WITH R

profregina.bernal@fiap.com.br
reginabernal@terra.com.br

Dra. Regina Tomie Ivata Bernal

Cientista de Dados na área da Saúde

Formação Acadêmica:

- Estatístico - UFSCar
- Mestre em Saúde Pública – FSP/USP
- Doutor em Ciências – Epidemiologia - FSP/USP

Atividades Profissionais:

- Professora de pós-graduação na FIAP
- Consultora externa da SVS/MS
- Cientista de Dados em Saúde

DETECÇÃO DE OUTLIERS

Detecção de dados suspeitos - "Outlier"

- Dado incorreto
- População diferente
- Dado correto – Evento raro

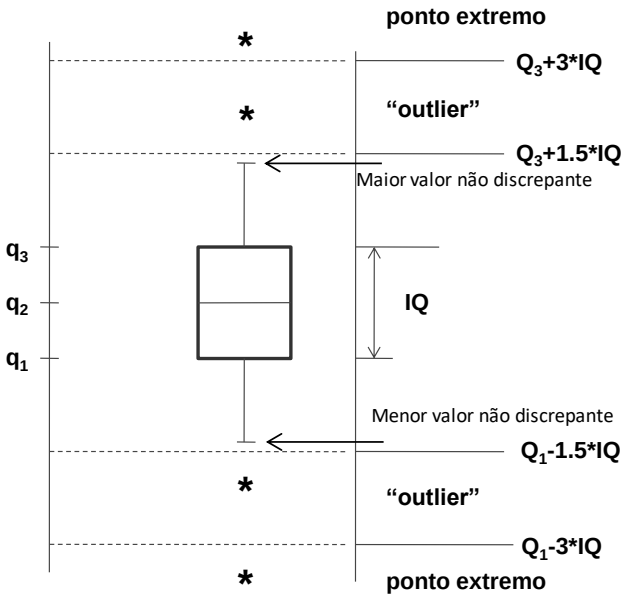
Detecção de dados suspeitos - "Outlier"

- Representação Gráfica na Análise dos Dados

O Box Plot (desenho esquemático) informa medidas de posição, dispersão, assimetria, caudas e dados atípicos (outliers). A posição central é dada pela mediana e a dispersão pela amplitude inter-quartílica. As medidas de posição q1, q2 e q3 informam a assimetria da distribuição. Os comprimentos das caudas são dados pelas linhas que vão do retângulo aos valores distantes e pelos valores atípicos.

Detecção de dados suspeitos - "Outlier"

- Representação Gráfica na Análise dos Dados



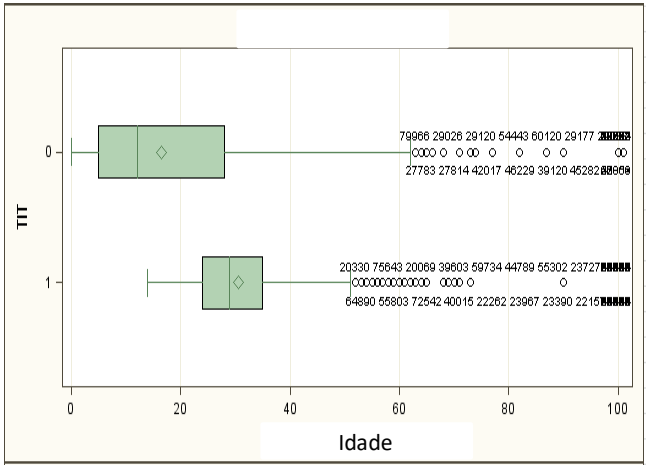
Legenda:

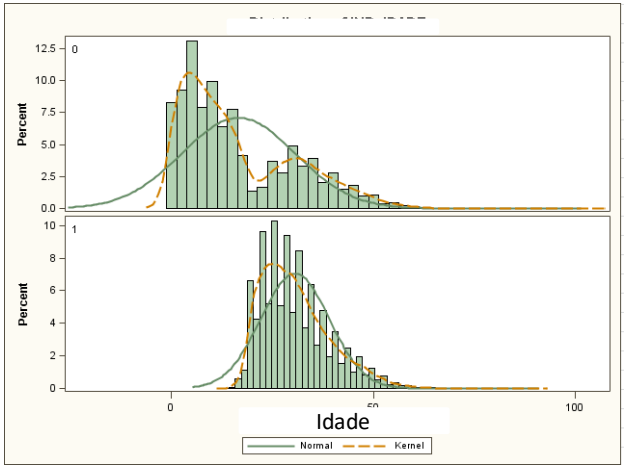
Q1 = quartil 1
Q2 = quartil 2 = mediana
Q3 = quartil 3
IQ = interquartil

Detecção de dados suspeitos - "Outlier"

Exemplo

- Exemplo





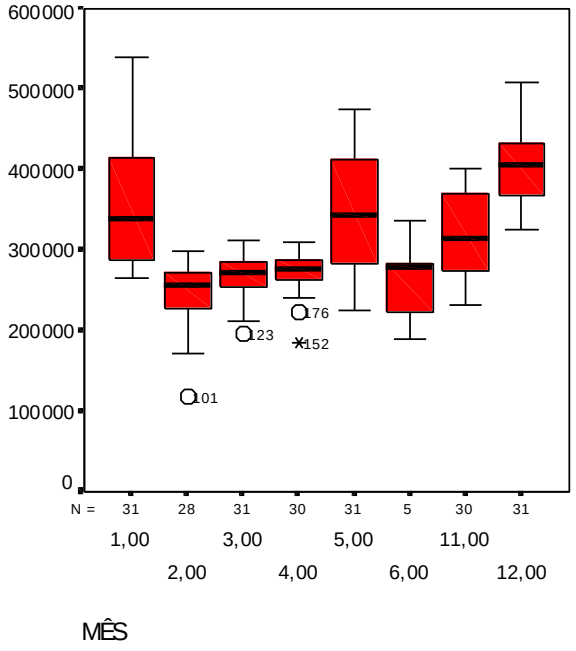
Aplicação

Detecção de dados suspeitos - "Outlier"

Exemplo

Gráfico Box-Plot

Exemplo: "Total de unidades vendidas por produto – Campanha 1 a 12 de 2016

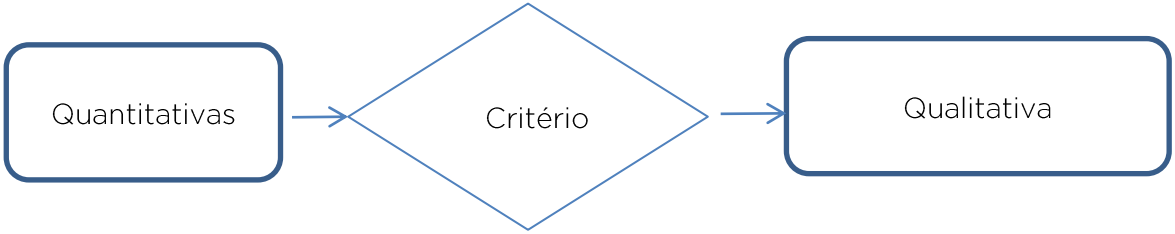


Exercitando!!!!!

Base
Cadastro

TABELAS DE FREQUÊNCIAS

Transformando variáveis quantitativas em qualitativas



Exemplo:

Anos de estudo

Critério		Grau instrução
0	→	Analfabeto
[1 - 9]	→	Fundamental
[10 – 12]	→	Médio
>= 13	→	Superior

Transformando variáveis quantitativas em qualitativas

Exemplo: Quantas classes serão necessárias para representar a despesa anual?

Fórmula de Sturges ➡ K = 1 + 3,322 * log10(n)

K = número de classes

Despesa	N	%	%ac
[0 - 265)	26740	44,0	44,0
[265 - 530)	10939	18,0	62,0
[530 - 795)	4862	8,0	70,0
[795 - 1060)	4254	7,0	77,0
[1060 - 1325)	3646	6,0	83,0
[1325 - 1590)	3039	5,0	88,0
[1590 - 1855)	2431	4,0	92,0
[1855 - 2120)	1823	3,0	95,0
[2120 - 2385)	1215	2,0	97,0
[2385 - 2650)	608	1,0	98,0
[2650 - 2915)	243	0,4	98,4
[2915 - 3180)	243	0,4	98,8
[3180 - 3445)	182	0,3	99,1
[3445 - 3710)	182	0,3	99,4
[3170 - 3975)	122	0,2	99,6
[3975 - 4240)	122	0,2	99,8
[4240 - 4505)	122	0,2	100,0
Total	60773	100,0	

Medidas resumo da despesa anual

Mean	Std Dev	Minimum	Maximum	Mode	Range	Sum	N
265,22	537,55	0	4491,19	0	4491,19	16118247,5	60773

K = 1 + 3,3 * log (60773) = 16,78 ~ 17

$$Intervalo = \frac{(Máximo - Mínimo)}{K} = \frac{4491{,}19}{17} = 264{,}18 \cong 265$$

Distribuição de Frequência

O número de vezes que ocorreram valores em cada classe ou valores chama-se frequência absoluta. O conjunto das ocorrências, com correspondentes frequências absolutas (FA) e relativas (FR), define a distribuição de frequências da variável. Conhecer o comportamento da variável.

Distribuição etária dos trabalhadores da Empresa XXX, 01/05/2019

Faixa etária	Frequency	Percent	Cumulative Frequency	Cumulative Percent
00 - 17	19052	33,8	19052	33,8
18 - 29	16143	28,6	35195	62,4
30 - 39	13710	24,3	48905	86,7
40 - 49	5773	10,2	54678	96,9
50 - 59	1559	2,8	56237	99,7
60 - 69	174	0,3	56411	100,0
Acima 69	13	0,0	56424	100,0
Total	56424	100,0	.	.

Exercitando!!!!!

Base
Bike Sharing

GRÁFICOS

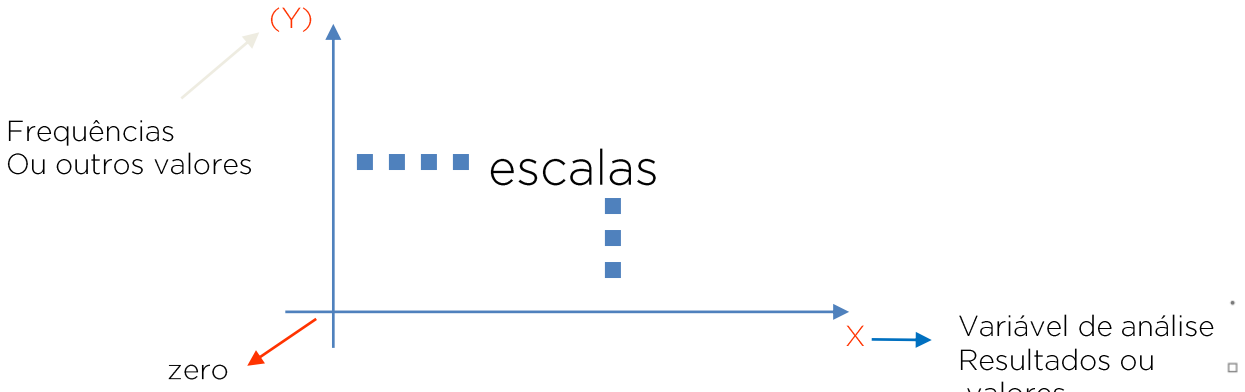
Apresentação Gráfica dos Dados

As regras básicas de elaboração de um gráfico são:

- simplicidade
- clareza
- veracidade

Apresentação Gráfica dos Dados

➢ EIXOS CARTESIANOS



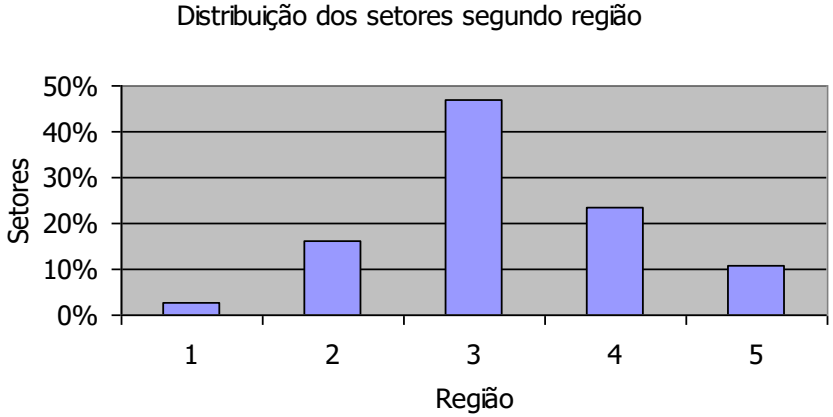
Apresentação Gráfica dos Dados

Variáveis qualitativas ou discretas

a) Colunas

Um gráfico de colunas ilustra comparações entre itens. As categorias são organizadas na horizontal e os valores são distribuídos na vertical.

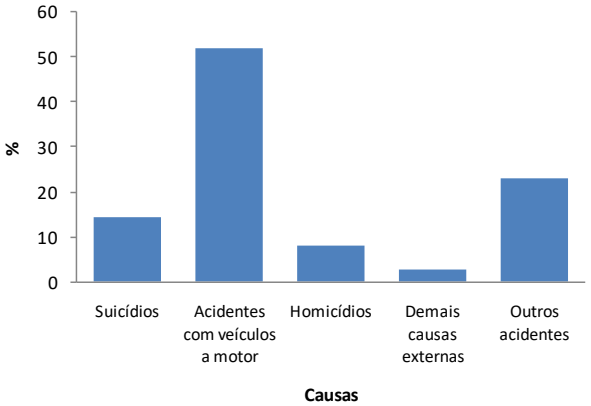
Exemplo:



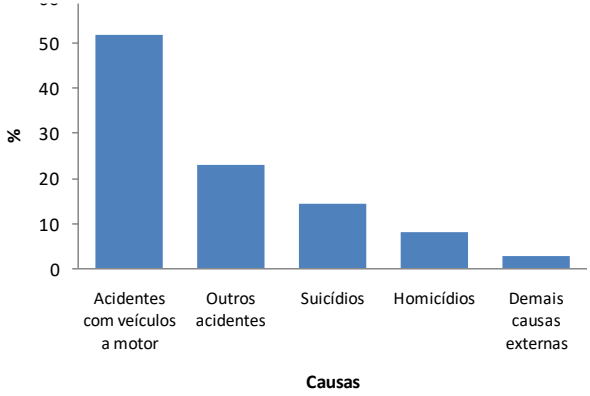
Apresentação Gráfica dos Dados

Variáveis qualitativas

Causa	%
Suicídios	14.2
Acidentes com veículos a motor	52.1
Homicídios	8.1
Demais causas externas	2.6
Outros acidentes	23
Total	100.0



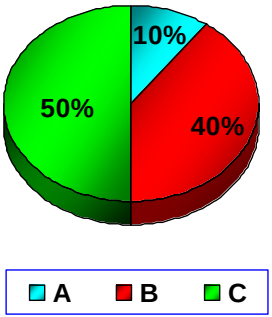
Causa	%
Acidentes com veículos a motor	52.1
Outros acidentes	23
Suicídios	14.2
Homicídios	8.1
Demais causas externas	2.6
Total	100.0



Apresentação Gráfica dos Dados

b) Setores ou pizza

Um gráfico de pizza mostra o tamanho proporcional de itens que constituem uma série de dados para a soma dos itens. A freqüência relativa (%) transformada em graus mediante o calculo proporcional.



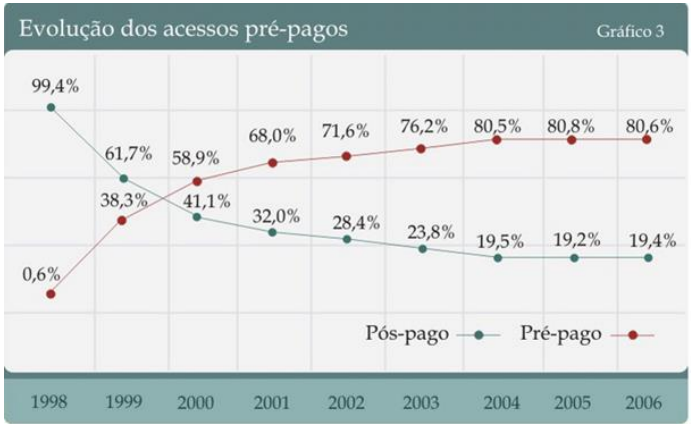
100	360
50	X

$$X = \frac{360 \cdot 50}{100} = 180$$

Apresentação Gráfica dos Dados

c) Linha

Um gráfico de linha mostra tendências nos dados em intervalos iguais.

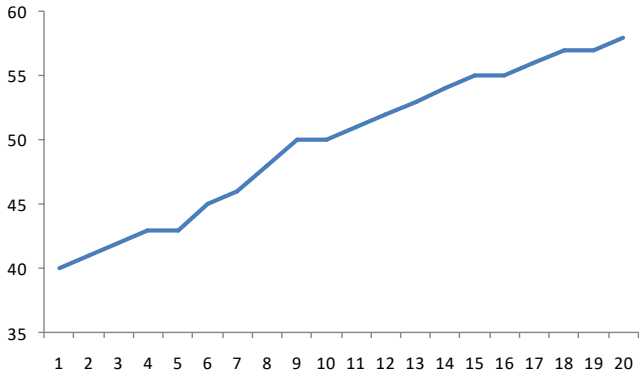


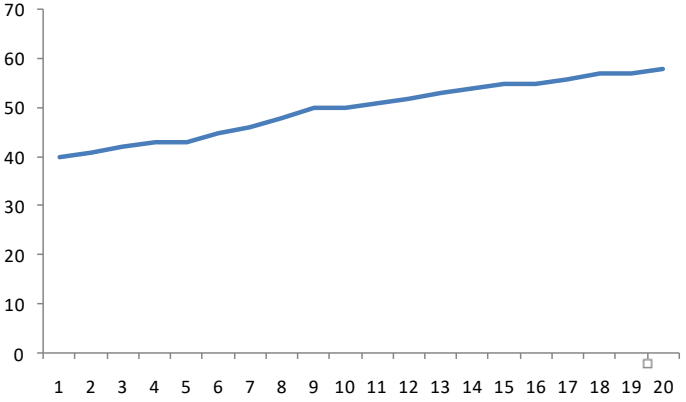
Fonte: Relatório Anual da Anatel, 2007.

O gráfico está adequado?

Apresentação Gráfica dos Dados

c) Linha





Qual gráfico está adequado?

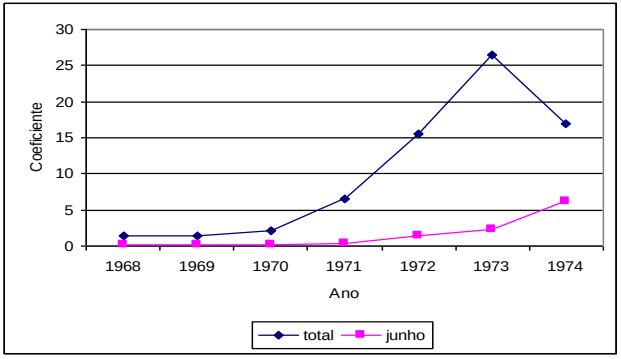
Apresentação Gráfica dos Dados

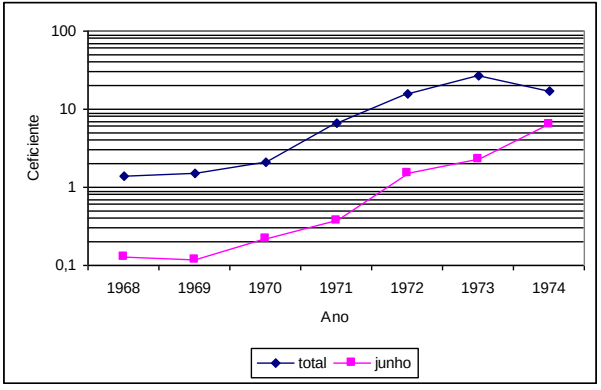
Tabela 2.4-Coeficientes de mortalidade (por 100.000 hab.) por meningite meningocócica no Município de São Paulo, no período de 1968 a 1974 observados durante todo o ano (total) e mês de junho de cada ano.

Ano	Total	Junho
1968	1,4	0,13
1969	1,5	0,12
1970	2,1	0,22
1971	6,6	0,37
1972	15,6	1,49
1973	26,5	2,24
1974	17,0	6,26

FONTE: *Rev. Saúde Públ., 10*: 1-16, 1976

Figura 1- Coeficientes de mortalidade (por 100.000 hab.) por meningite meningocócica no Município de São Paulo, no período de 1968 a 1974 observados durante todo o ano (total) e mês de junho de cada ano.





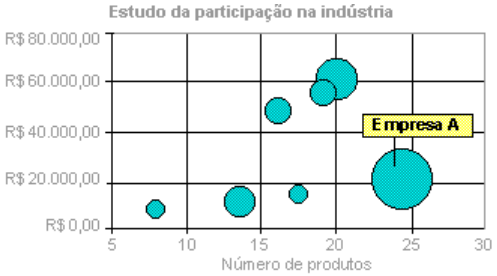
Apresentação Gráfica dos Dados

e) Bolhas

Um gráfico de bolhas é um tipo de gráfico xy (dispersão). O tamanho do marcador de dados indica o valor de uma terceira variável.

Exemplo:

Nº de produtos	Vendas	Partic. no mercado %
14	R$ 11.200,00	13
20	R$ 60.000,00	23
18	R$ 14.400,00	5
Valores X	Valores Y	Tamanho da bolha



O gráfico nesse exemplo mostra que a Empresa A tem a maioria dos produtos e a maior fatia do mercado, mas não necessariamente as melhores vendas.

Apresentação Gráfica dos Dados

Variáveis contínuas

a) Histograma

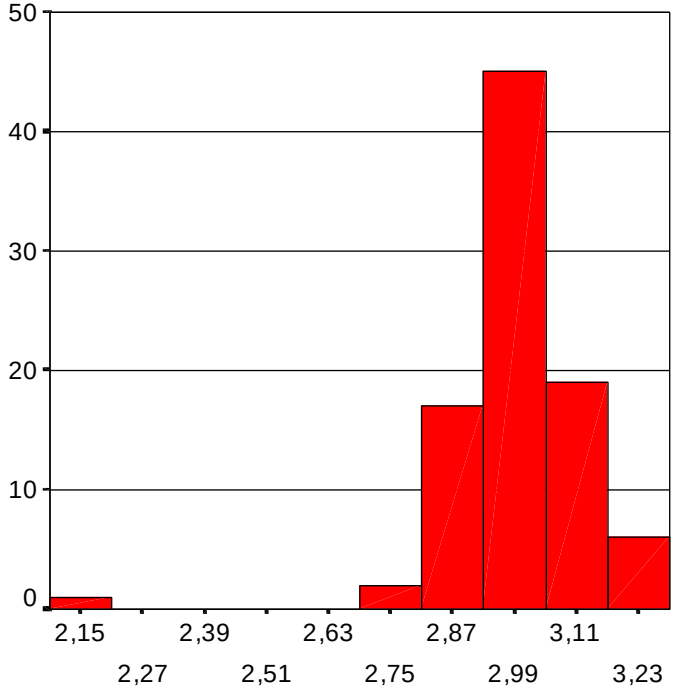
O histograma é formado por retângulos cujas áreas representam freqüências dos intervalos de suas classes. Esta apresentação é indicada para séries contínuas, e portanto não há espaço entre as barras.

Apresentação Gráfica dos Dados

Histograma

Exemplo: Preço médio (net price) do produto A (em reais)

Classes	Freqüência	Freqüência Relativa	Ponto Médio
2,09|---- 2,21	1	0,01	2,15
2,21|---- 2,33	0	0,00	2,27
2,33|---- 2,45	0	0,00	2,39
2,45|---- 2,57	0	0,00	2,51
2,57|---- 2,69	0	0,00	2,63
2,69|---- 2,81	2	0,02	2,75
2,81|---- 2,93	19	0,21	2,87
2,93|---- 3,05	45	0,50	2,99
3,05|---- 3,17	17	0,19	3,11
3,17|---- 3,29	6	0,07	3,23
Total	90	1,00	



Apresentação Gráfica dos Dados

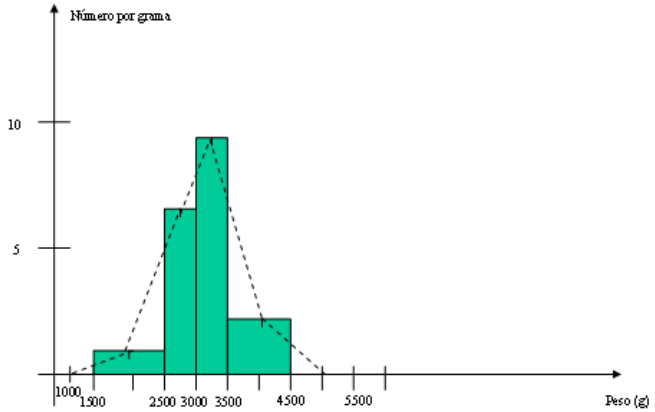
Tabela 1 – Nascidos vivos segundo peso ao nascer

Peso ao nascer (g)	Nº	Amplitude (g)	h´(frequência por gramas)
1.500 |— 2.500	1.200	1000	1,2
2.500 |— 3.000	3.600	500	7,2
3.000 |— 3.500	4.800	500	9,6
3.500 |— 4.500	2.400	1000	2,4
Total	12.000		

Tabela de frequência com amplitudes desiguais.

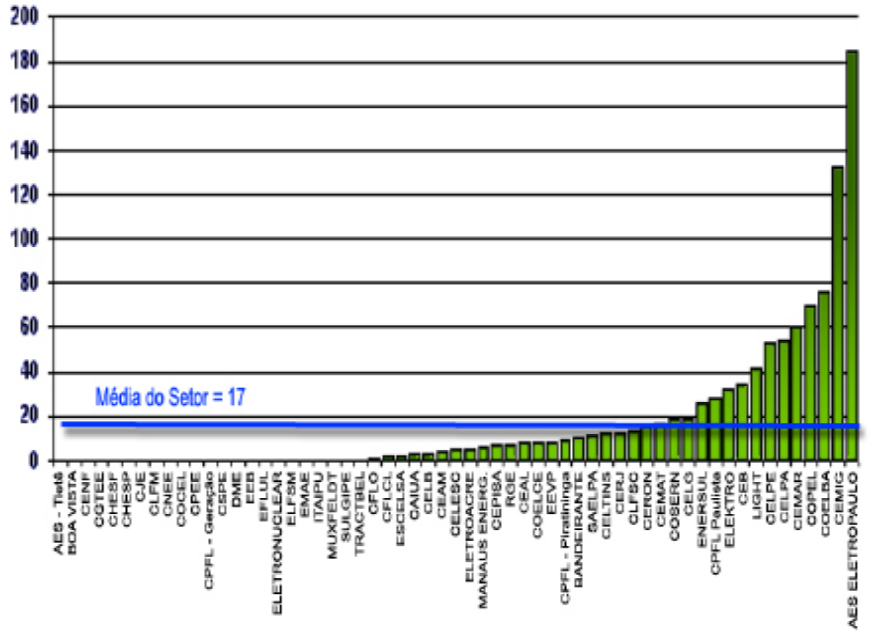
Frequência ajustada $\longrightarrow h' = \frac{N^{o}}{Amplitude}$

Figura 2.9 - Distribuição de nascidos vivos segundo peso ao nascer. Maternidade X, 1999.



FONTE: SAME, Maternidade X

Apresentação Gráfica dos Dados



O gráfico mostra que a AES Eletropaulo tem maior número de acidentes. Você concorda com esse resultado?

Apresentação Gráfica dos Dados

- Variáveis qualitativas ou quantitativas discretas

Colunas: Um gráfico de colunas ilustra comparações entre itens. As categorias são organizadas na horizontal e os valores são distribuídos na vertical.

Setor ou pizza: Um gráfico de pizza mostra o tamanho proporcional de itens que constituem uma série de dados para a soma dos itens.
A frequência relativa (%) transformada em graus mediante o calculo proporcional.

Bolhas: Um gráfico de bolhas é um tipo de gráfico dispersão (x,y). O tamanho do marcador de dados indica o valor de uma terceira variável.

Apresentação Gráfica dos Dados

- Variáveis quantitativas contínuas

  Histograma: É formado por retângulos cujas áreas representam frequências dos intervalos de suas classes. Esta apresentação é indicada para séries contínuas, e portanto não há espaço entre as barras.

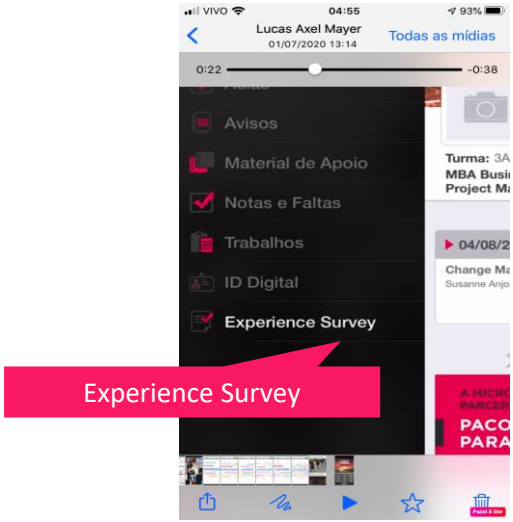
Exercitando!!!!!

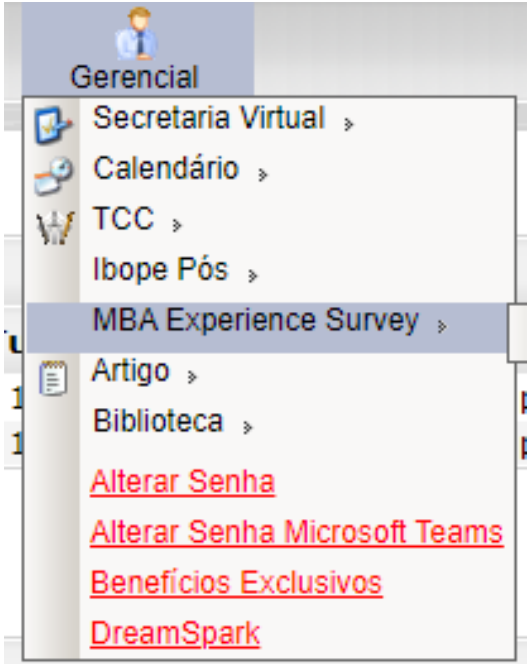
Base
Bike Sharing

O que você achou da aula de hoje?

Pelo aplicativo da FIAP

(Entrar no FIAPP, e no menu clicar em Experience Survey)





OBRIGADA

 / Regina T. I. Bernal

FIAP


FIAP