ANÁLISE DESCRITIVA

ICB

Airlane P. Alencar

Banco de dados

- http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets
- Diabetes
- These data are courtesy of Dr John Schorling, Department of Medicine, University of Virginia School of Medicine.
- 19 variables on 403 subjects from 1046 subjects who were interviewed in a study to understand the prevalence of obesity, diabetes, and other cardiovascular risk factors in central Virginia for African Americans.
- According to Dr John Hong, Diabetes Mellitus Type II (adult onset diabetes) is associated most strongly with obesity. The waist/hip ratio may be a predictor in diabetes and heart disease. DM II is also associated with hypertension - they may both be part of "Syndrome X". The 403 subjects were the ones who were actually screened for diabetes. Glycosolated hemoglobin > 7.0 is usually taken as a positive diagnosis of diabetes.
- Willems JP, Saunders JT, DE Hunt, JB Schorling: Prevalence of coronary heart disease risk factors among rural blacks: A community-based study. *Southern Medical Journal* 90:814-820; 1997
  Schorling JB, Roach J, Siegel M, Baturka N, Hunt DE, Guterbock TM, Stewart HL: A trial of church-based smoking cessation interventions for rural African Americans. *Preventive Medicine* 26:92-101; 1997

Diabetes

Name	Labels	Units	Levels	Storage	NAs	
id	Subject ID			double		0
chol	Total Cholesterol			double		1
stab.glu	Stabilized Glucose			double		0
hdl	High Density Lipoprotein			double		1
ratio	Cholesterol/HDL Ratio			double		1
glyhb	Glycosolated Hemoglobin			double		13
location			2	integer		0
age		years		double		0
gender			2	integer		0
height		inches		double		5
weight		pounds		double		1
frame			3	integer		12
bp.1s	First Systolic Blood Pressure			double		5
bp.1d	First Diastolic Blood Pressure			double		5
bp.2s	Second Systolic Blood Pressure			double		262
bp.2d	Second Diastolic Blood Pressure			double		262
waist		inches		double		2
hip		inches		double		2
time.ppn	Postprandial Time when Labs were Drawn	minutes		double		3

Tipos de variáveis

- Quantitativas: numéricas
  - Discretas: número de filhos
  - Contínuas: renda, concentração de alguma substância
- Qualitativas ou categóricas
  - Nominal: estado civil, sexo, presença de diabetes (s/n)
  - Ordinal: grau de instrução, dor forte, moderada ou fraca

Medidas resumo para as variáveis quantitativas

- Média

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{y_1 + \ldots + y_n}{n}$$

- Mediana (md): metade das observações são menores que md e metade são maiores que md

  Ordeno todos os valores e encontro a observação central

  Ex: 2, 2, 4, 5, 8 => Md=4 pega valor na posição (1+5)/2=3

  1, 3, 5, 7, 7, 9 => Md=(5+7)/2=6 posição (1+6)/2=3,5

- Moda: é o valor mais frequente na amostra

Percentis

- O valor px tal que x% das observações são menores que ele é denominado percentil x%.
- O percentil 25%, também chamado de primeiro quartil, deixa ¼ das observações abaixo dele. P25=Q1= (79+81)/2=80
- ¼ dos pacientes tem glicose abaixo de 80.
- A mediana é o percentil 50% =88.

i	glicose	ordenado
1	82	75
2	97	75
3	92	76
4	93	78
5	90	79
6	94	81
7	92	82
8	75	82
9	87	83
10	89	87
11	82	89
12	128	90
13	75	92
14	79	92
15	76	93
16	83	94
17	78	97
18	112	112
19	81	128
20	206	206

Medidas de variabilidade

- Variância amostral

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \frac{(y_1 - \bar{y})^2 + \ldots + (y_n - \bar{y})^2}{n-1}$$

  - Nos dados dos 20 pacientes, a glicose média é 94,55 e a variância amostral é 854,26.
- Desvio Padrão= raiz quadrada da variância
  - dp= s= 29.23

Medidas de variabilidade

- Amplitude= Máximo – Mínimo
- Amplitude interquartílica= Q3-Q1
  - Amplitude= 206-75
  - Amplitude interquartílica=93,5-80=13,5

i	glicose	ordenado
1	82	75
2	97	75
3	92	76
4	93	78
5	90	79
6	94	81
7	92	82
8	75	82
9	87	83
10	89	87
11	82	89
12	128	90
13	75	92
14	79	92
15	76	93
16	83	94
17	78	97
18	112	112
19	81	128
20	206	206

Desvio e Erro Padrão

- Desvio Padrão

$$s = \sqrt{s^2}$$

- Erro Padrão

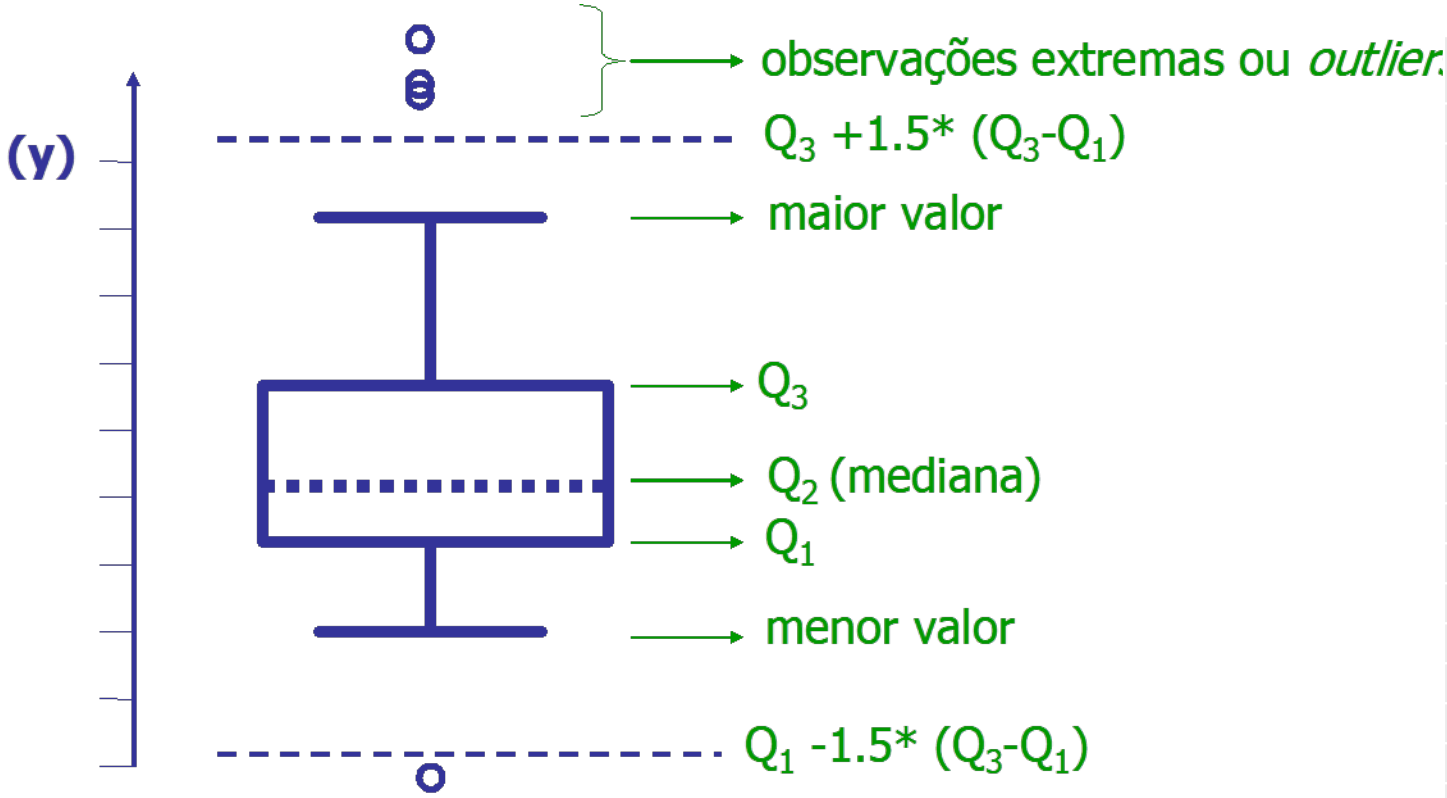
$$ep = \frac{s}{\sqrt{n}}$$

- **Qual a diferença?**

  O erro padrão corresponde ao desvio padrão da média amostral.

- Ex: dp= s= 29.23 e o ep=6,5

Boxplot



i	glicose	ordenado
1	82	75
2	97	75
3	92	76
4	93	78
5	90	79
6	94	81
7	92	82
8	75	82
9	87	83
10	89	87
11	82	89
12	128	90
13	75	92
14	79	92
15	76	93
16	83	94
17	78	97
18	112	112
19	81	128
20	206	206

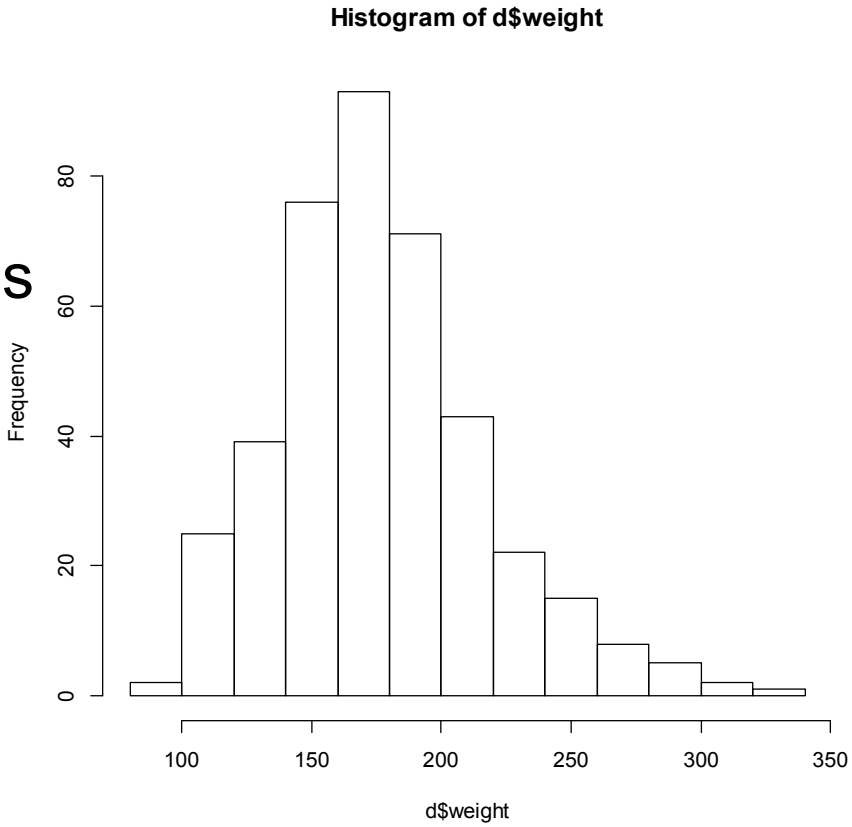
- Q1=80; Q2=Md=88; Q3=93,5
- Vamos usar o R também.

Histograma

- Gráfico de colunas justapostas
- Definir Número de intervalos
  - Sugestão: Sturges: $k=1+\ln_2 n$
  - Altera até ficar adequado
- Construir tabela de frequências
- Ex: Diabetes- Peso

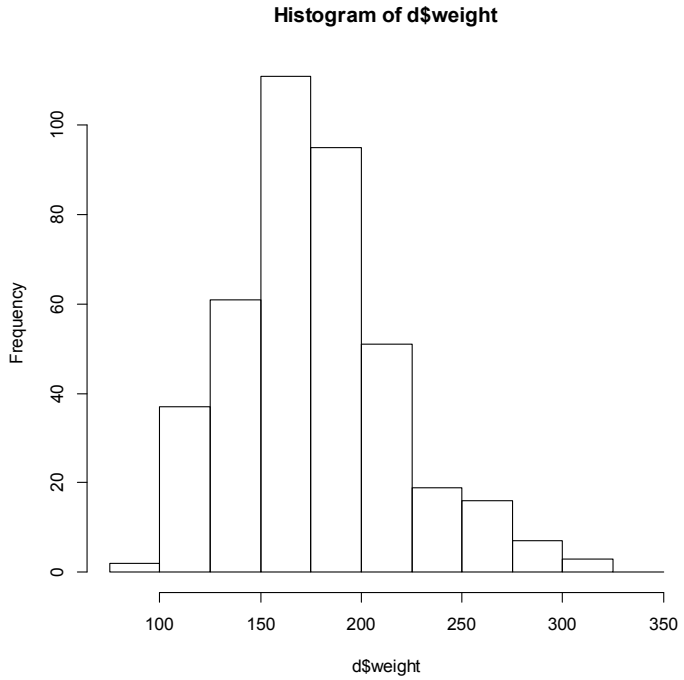
  n=402 k=9.7

  No R, hist()



Histograma

- Posso definir os valores dos intervalos
- hist(d$weight, breaks=seq(75,350,25))



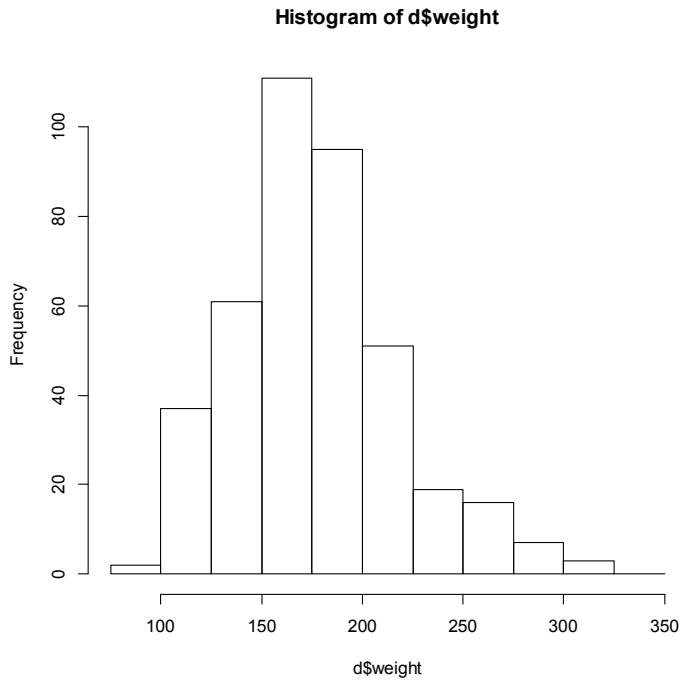
- Contagens

2	25	39	76	93	71	43	22	15	8	5	2	1
2	27	66	142	235	306	349	371	386	394	399	401	402

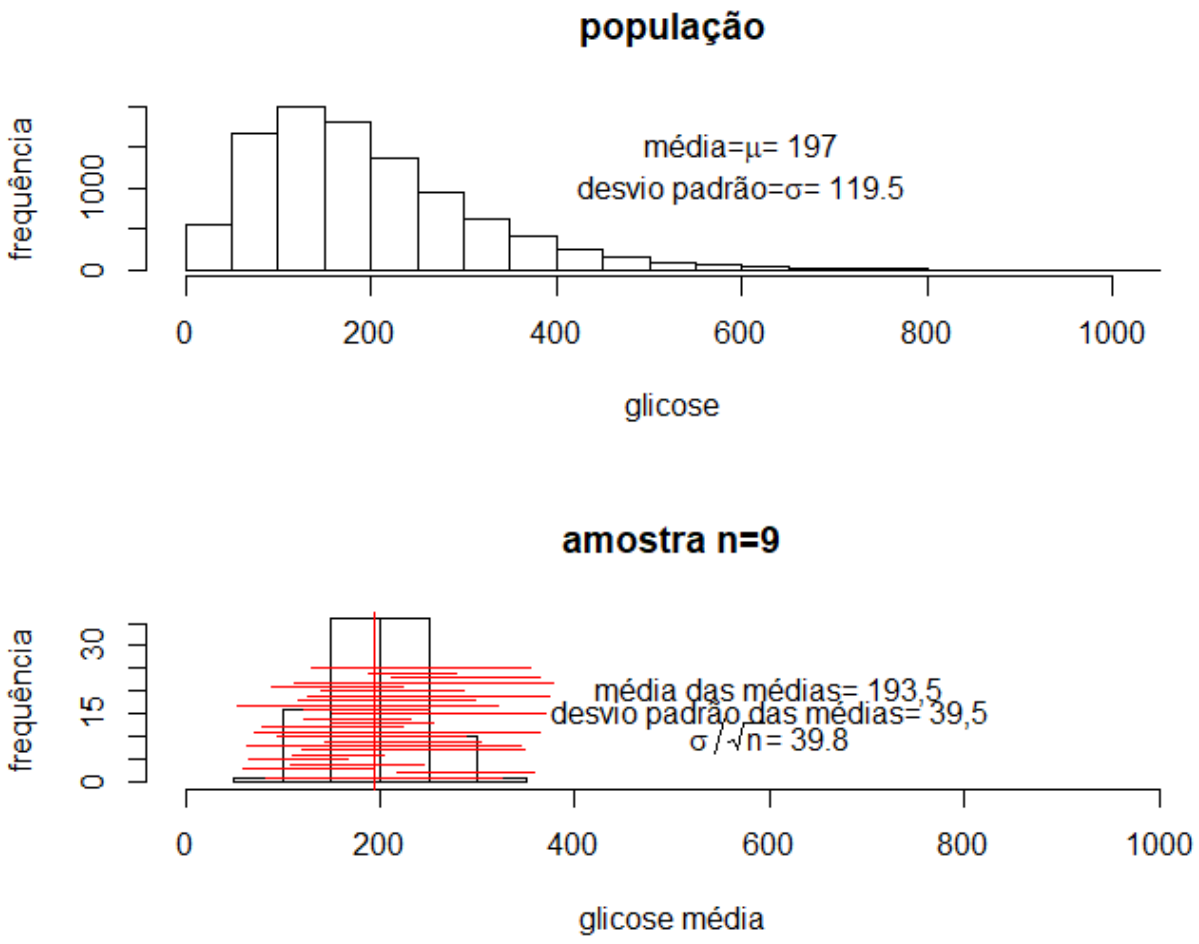
Cálculo de quantis em histograma

- Tabela

Lim.Inf	75	100	125	150	175	200	225	250	275	300	325	350	375
freq.	2	25	39	76	93	71	43	22	15	8	5	2	1
freq.acum.	2	27	66	142	235	306	349	371	386	394	399	401	402



Dados glicose população de N=10000



https://www.ime.usp.br/~lane/home/MAE0261/medias.r

Tabelas

- Vale a pena fazer tabelas e gráficos dinâmicos na planilha eletrônica.

- Frequências absolutas e relativas

Programa R

- www.r-project.org

- Instalar library(Rcmdr)

- Usar os dados de glicose para fazer tabelas de frequências, boxplots e histogramas.

Alguns comandos

- d= read.csv("C:/2017/MAE0261/dados/diabetes.csv", sep=";")
- names(d)

- hist(d$stab.glu)

- summary(d$stab.glu)

- hist(d$stab.glu, breaks=seq(0,400,25), main="Histograma", xlab="Glicose", ylab="Frequência")
- boxplot(stab.glu ~ gender, data=d, ylab="glicose")