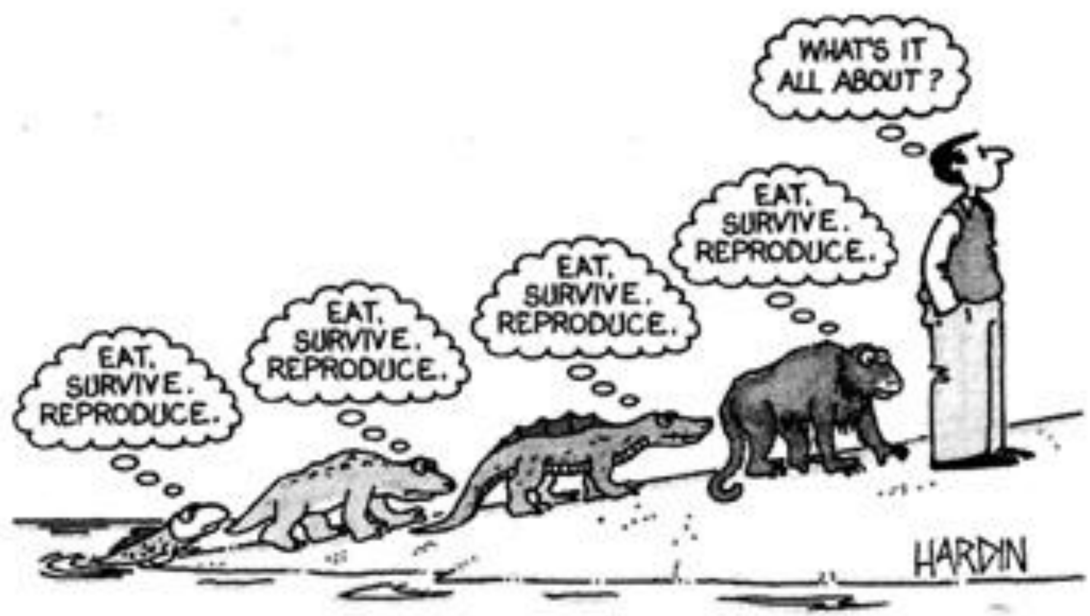
Noções de Estatística I

Renato Vicente
EACH-USP/2009

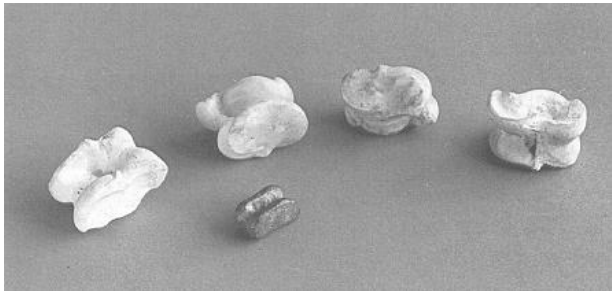
Perguntas

1. Como surgiu a idéia?
2. O que é uma distribuição estatística?
3. Como utilizo as distribuições estatísticas?
4. Como comparo populações?
5. Resumindo



Uma breve história da Estatística

Primórdios: Jogos e Divinação



Coleções de ossos da pata de cães, carneiros ou cabras (chamados *astragali*, pelos gregos) , são encontrados em sítios arqueológicos do Neolítico até a Idade Média.

Em grego o termo para dado ou cubo é *kubos.* Em árabe o termo para dado cúbico e para astrágalo é *kab,* o que sugere que os dados modernos tenham derivado dos astrágalos antigos. Na figura de fundo vermelho à direita vemos dados egípcios de 3000 anos (no Museu do Louvre).

Primórdios: Jogos e Divinação

Jogos com componentes aleatórios são comuns nos sítios arqueológicos antigos (esquerda e centro – Senet 1000 aC) e (direita - Ur 2500 aC)

Primeiros Estudos Teóricos: Cardano

Girolamo Cardano (1501-1576) era um intelectual renascentista viciado em jogos. Em um relato escreveu: “O resultado foi que após vinte jogadas recuperei minhas roupas, meus anéis e colar para o menino. ”

Escreveu um livro sobre jogos de azar (*Liber de Ludo Aleae)* relacionando

+	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Na figura a chance da soma de dois dados totalizar 10 é de 3/36.

Primeiros Estudos Teóricos: Galileu

Galileu Galilei (1564-1642) foi contratado em 1613 para ser matemático oficial de Cosimo de Medici II e Primeiro Matemático da Universidade de Pisa. Logo foi chamado a responder questões relacionadas a jogos com dados. Em particular:

Por que 10 e 11 aparecem mais freqüentemente do que 9 e 12 em arremessos de 3 dados?

Para responder a questão publicou *Sopra le Scoperte dei Dadi*

Na primeira frase do artigo escreveu: "*O fato de que em jogos de dados certos números são mais vantajosos do que outros tem uma razão óbvia, i.e., que alguns números são obtidos mais facilmente e mais frequentemente do que outros, o que depende destes poderem ser obtidos através de uma variedade maior de números*"

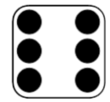
Primeiros Estudos Teóricos: Galileu

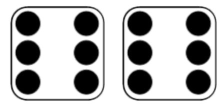
1																
3																
6	10		9		8		7		6		5		4		3	
10	631	6	621	6	611	3	511	3	411	3	311	3	211	3	111	1
15	622	3	531	6	521	6	421	6	321	6	221	3				
21	541	6	522	3	431	6	331	3	222	1						
25	532	6	441	3	422	3	322	3								
27	442	3	432	6	332	3										
108	433	3	333	1												
108		27		25		21		15		10		6		3		1
216																

A tabela acima (reproduzida do artigo original de Galileu) mostra contagens para o número de vezes que uma particular soma de três dados. Assim, por exemplo, 10 pode ser obtido através de (631,622,541,532,442,433) em um total de 27 formas diferentes

Pascal e Fermat: Nasce a Teoria de Probabilidades

O nobre francês chevalier de Méré era um jogador inveterado havia feito as seguintes apostas, acreditando-as boas e equivalentes:

APOSTA 1:Aposta em ao menos um  em 4 arremessos de um dado.

APOSTA 2: Aposta em ao menos  24 arremessos de dois dados.

Apesar de deduzir que as chances dos dois eventos seriam as mesmas (2/3) chevalier de Méré perdia consistentemente com a aposta 2 e, assim, pediu a Pascal uma explicação. Pascal então enviou uma carta a Fermat, iniciando a correspondência que deu origem a teoria de probabilidades

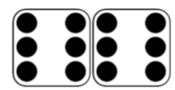
Fermat (1601-1665)

Pascal (1623-1662)

Pascal e Fermat: Nasce a Teoria de Probabilidades

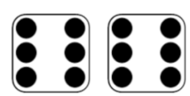
RACIOCíNIO (ERRADO) DO JOGADOR:

APOSTA 1: Prob(pelo menos 1  em 4 jogadas) = $4 \times (1/6) = 2/3$

APOSTA 2: Prob(pelo menos  em 24 jogadas) = $24 \times (1/36) = 2/3$

Fermat e Pascal:

APOSTA 1: Prob(de NENHUM  em 4 jogadas) = $(5/6)^4 = 0{,}482$

APOSTA 2: Prob(de NENHUM  em 4 jogadas) = $(35/36) = 0{,}509$

Assim a APOSTA 1 tem probabilidade de perda de 48,2% , sendo a melhor aposta.

A Teoria de Probabilidades amadurece: Huygens, Bernoulli e De Moivre

http://www.stat.sc.edu/~west/javahtml/CLT.html

http://www.math.csusb.edu/faculty/stanton/m262/intro_prob_models/intro_prob_models.html

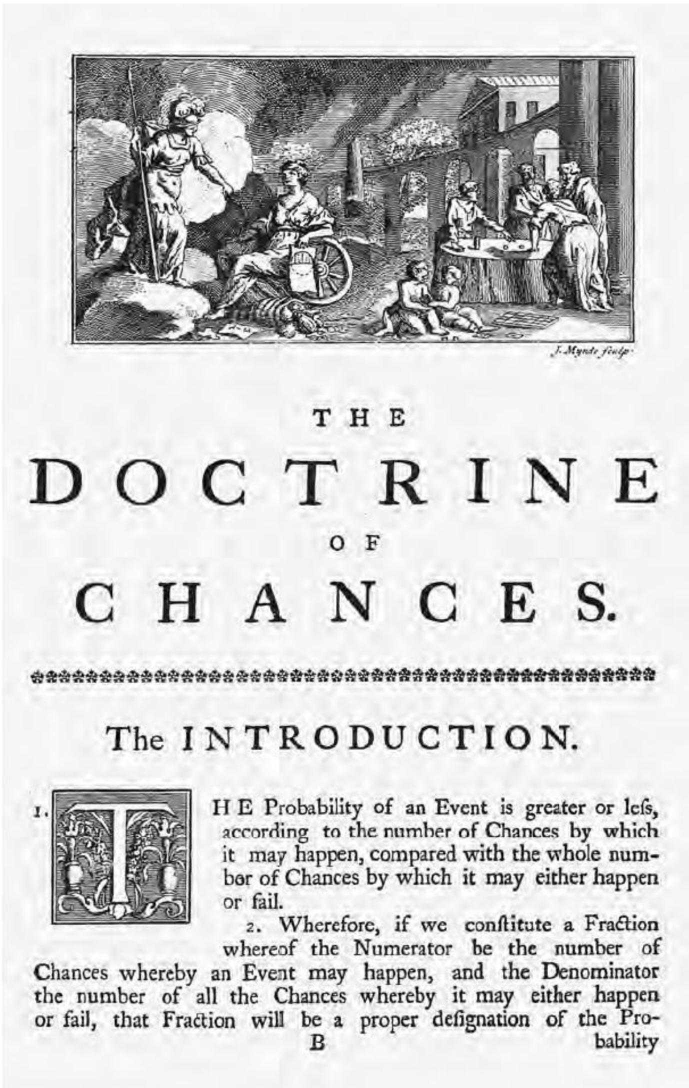
Huygens (1629-1695)

Jacob Bernoulli (1654-1705)

Abraham De Moivre (1667-1754)

Apesar de ter surgido no contexto de Jogos de Azar, a partir do trabalho de Huygens ,de Bernoulli e de De Moivre a teoria de probabilidades passou a ser entendida como uma teoria matemática para eventos aleatórios.

A Teoria de Probabilidades amadurece: Huygens, Bernoulli e De Moivre



THE

DOCTRINE

OF

CHANCES.

The INTRODUCTION.

1. THE Probability of an Event is greater or lefs, according to the number of Chances by which it may happen, compared with the whole number of Chances by which it may either happen or fail.

2. Wherefore, if we conftitute a Fraction whereof the Numerator be the number of Chances whereby an Event may happen, and the Denominator the number of all the Chances whereby it may either happen or fail, that Fraction will be a proper defignation of the Probability

B

http://www-stat.stanford.edu/~naras/jsm/NormalDensity/NormalDensity.html

http://stat.wvu.edu/SRS/Modules/Normal/normal.html

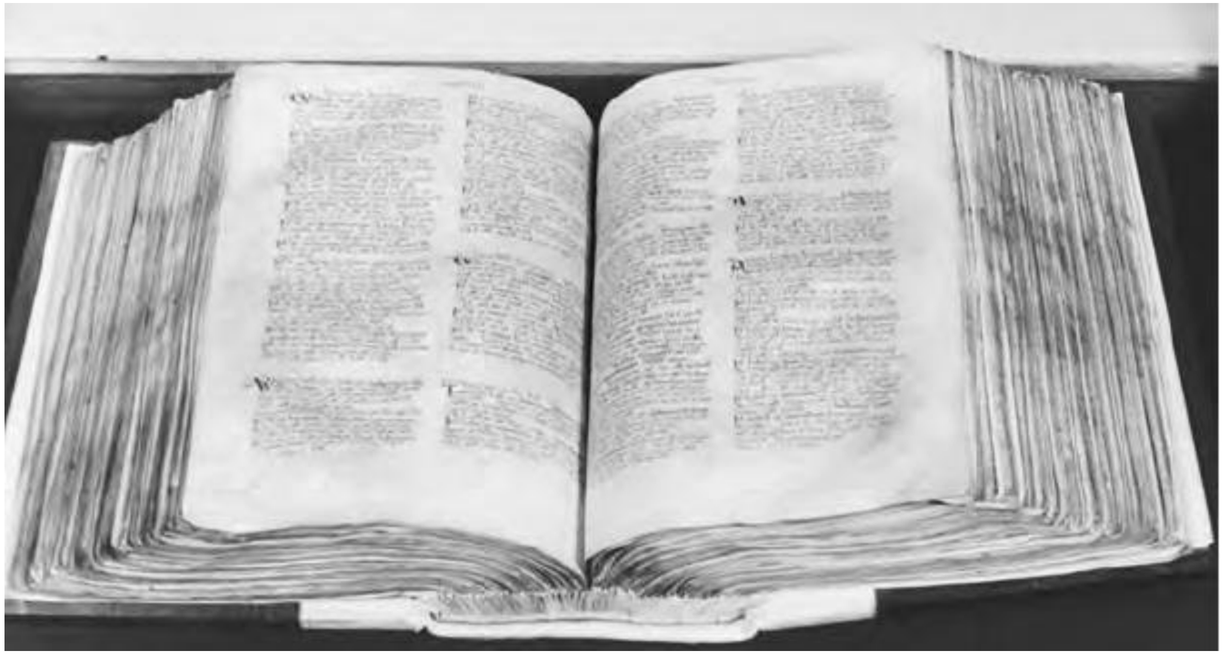
Primeira página do livro de De Moivre sobre Probabilidades. **Amigo de Newton e Halley**, De Moivre foi o primeiro a descrever a curva Normal (hoje chamada Gaussiana).

Nascimento da Demografia: Ibn Khaldun

Ibn Khaldun (1332-1406), pensador do mundo islamico medieval, produziu o primeiro trabalho teórico sobre dados demográficos de que se tem registro.

Observando dados demográficos propôs uma   teoria para a dinâmica da história baseada na idéia da *asabiyah* (solidaridade social).

Voltando para 1066: Nascimento da Demografia



Livro de Winchester: O rei normando, William o Conquistador, ordenou, após sua vitória sobre o rei anglo-saxão Harold em 1066, que um levantamento extremamente detalhado do que havia nas terras inglesas conquistadas. O rei morreu em 1087 sem nunca fazer uso do livro.

Nascimento da Estatística: John Graunt

CAPTAIN JOHN GRAUNT


A generall Bill for this present year, ending the 19 of December 1665. according to the Report made to the KING'S most Excellent Majesty.

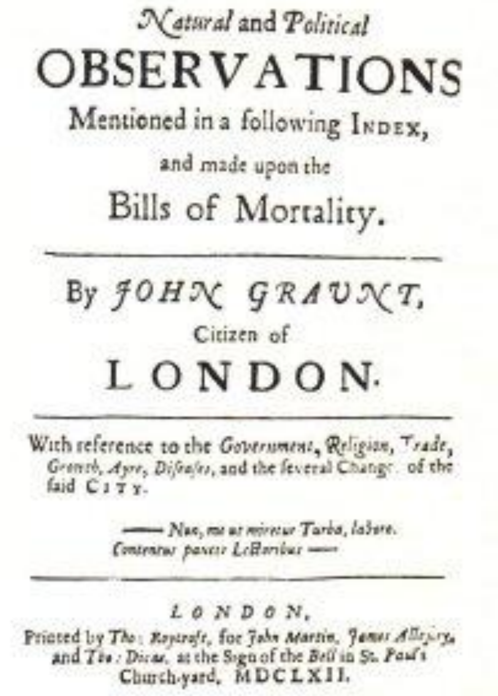
By the Company of Parish Clerks of London, &c.

The Diseases and Casualties this year.

The Total of all the Christnings — 9967

The Total of all the Burials this year — 97306

Whereof, of the Plague — 68596


Natural and Political

OBSERVATIONS

Mentioned in a following INDEX,

and made upon the

Bills of Mortality.

By JOHN GRAUNT,

Citizen of

LONDON.

With reference to the Government, Religion, Trade, Growth, Ayre, Diseases, and the several Changes of the said CITY.

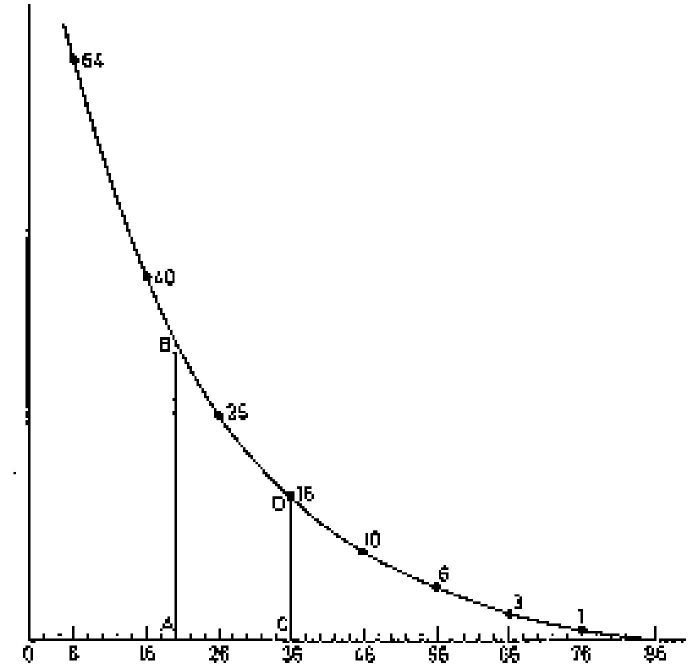
—— Non, me ut miretur Turba, laboro. Contentus paucis Lectoribus ——

LONDON,

Printed by Tho: Roycroft, for John Martin, James Allestry, and Tho: Dicas, at the Sign of the Bell in St. Paul's Church-yard, MDCLXII.

John Graunt (1620-1674) um lojista londrino que decidiu estudar sistematicamente a documentação sobre mortes e nascimentos registradas em Londres por um período de 57 anos. Os dados compilados por Graunt foram utilizados por Edmund Halley para o primeiro cálculo de expectativa de vida e subsequentes aplicações aos seguros de vida. **Halley era amigo de Newton e De Moivre.**

A Estatística encontra a Probabilidade



Christiaan Huygens(1620-1674) utilizou os dados de John Graunt para construir o primeiro gráfico representando a expectativa de vida de pessoas de uma certa idade, estes cálculos foram utilizados para venda se seguros.

O que é uma distribuição estatística?

Experimento de Cavendish (1797)

Em 1797 e 1798 Henry Cavendish realizou uma série de experimentos cujo objetivo era medir a densidade da Terra.

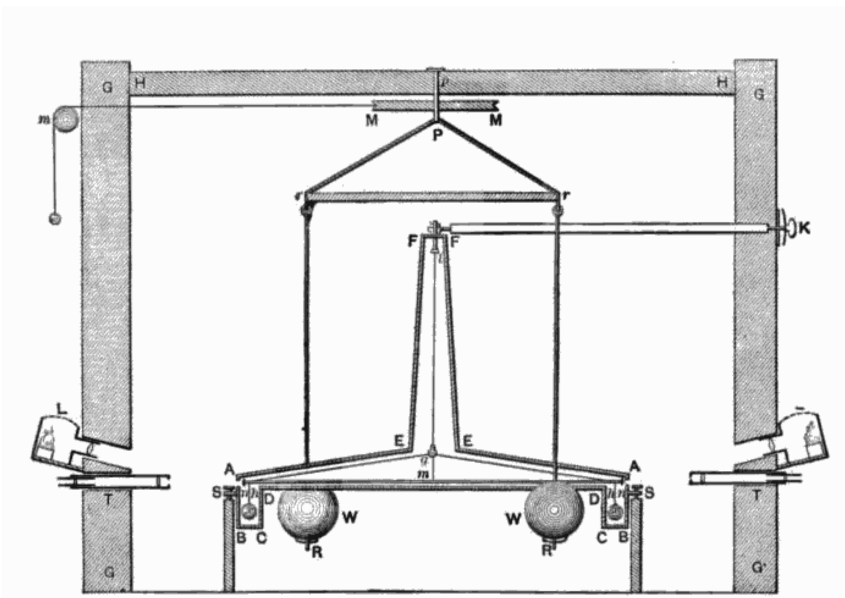


Fig. 1

Os resultados de 29 medidas da densidade em termos da densidade da água foram:

5.50	5.61	4.88
5.07	5.26	5.55
5.36	5.29	5.58
5.65	5.57	5.53
5.62	5.29	5.44
5.34	5.79	5.10
5.27	5.39	5.42
5.47	5.63	5.34
5.46	5.30	5.75
5.68	5.85	

Diagrama de Folhas e Ramos

Os resultados de 29 medidas da densidade em termos da densidade da água foram:

5.50	5.61	4.88
5.07	5.26	5.55
5.36	5.29	5.58
5.65	5.57	5.53
5.62	5.29	5.44
5.34	5.79	5.10
5.27	5.39	5.42
5.47	5.63	5.34
5.46	5.30	5.75
5.68	5.85	

As folhas contem o último dígito e os ramos os restantes em seqüência (mesmo que alguns ramos fiquem sem folhas). Se os números tiverem muitos algarismos significativos deve-se arredondá-los.

48	8
49	
50	7
51	0
52	6799
53	04469
54	2467
55	03578
56	12358
57	59
58	5

Diagrama de Folhas e Ramos

```
> cavendish <- scan("C:/Documents and Settings/Renato
Vicente/Desktop/Alesp/data/cavendish.dat")
Read 29 items
> cavendish
 [1] 5.50 5.61 4.88 5.07 5.26 5.55 5.36 5.29 5.58 5.65 5.57 5.53 5.62 5.29 5.44
5.34 5.79 5.10 5.27 5.39 5.42 5.47 5.63 5.34 5.46 5.30 5.75 5.68 5.85
> stem(cavendish)

  The decimal point is 1 digit(s) to the left of the |

  48 | 8
  49 |
  50 | 7
  51 | 0
  52 | 6799
  53 | 04469
  54 | 2467
  55 | 03578
  56 | 12358
  57 | 59
  58 | 5

>
```

Tabela de Frequências

Os resultados de 29 medidas da densidade em termos da densidade da água foram:

5.50	5.61	4.88
5.07	5.26	5.55
5.36	5.29	5.58
5.65	5.57	5.53
5.62	5.29	5.44
5.34	5.79	5.10
5.27	5.39	5.42
5.47	5.63	5.34
5.46	5.30	5.75
5.68	5.85	

Intervalo	n	f
4.8 |- 5.0	1	0.035
5.0 |- 5.2	2	0.07
5.2 |- 5.4	9	0.31
5.4 |- 5.6	9	0.31
5.6 |- 5.8	7	0.24
5.8 |- 6.0	1	0.035
Total	29	1.00

Tabela de Frequências

```
> factor(cut(cavendish, breaks = 4.8+0.2*(0:6))) -> intcavendish
> intcavendish
 [1] (5.4,5.6] (5.6,5.8] (4.8,5]   (5,5.2]   (5.2,5.4] (5.4,5.6] (5.2,5.4]
(5.2,5.4] (5.4,5.6] (5.6,5.8] (5.4,5.6] (5.4,5.6] (5.6,5.8] (5.2,5.4]
(5.4,5.6]
[16] (5.2,5.4] (5.6,5.8] (5,5.2]   (5.2,5.4] (5.2,5.4] (5.4,5.6] (5.4,5.6]
(5.6,5.8] (5.2,5.4] (5.4,5.6] (5.2,5.4] (5.6,5.8] (5.6,5.8] (5.8,6]
Levels: (4.8,5] (5,5.2] (5.2,5.4] (5.4,5.6] (5.6,5.8] (5.8,6]
> table(intcavendish)
intcavendish
  (4.8,5]   (5,5.2] (5.2,5.4] (5.4,5.6] (5.6,5.8]   (5.8,6]
        1         2         9         9         7         1
>
```

Tabela de Frequências e Histograma

Intervalo	n	f
4.8 |- 5.0	1	0.035
5.0 |- 5.2	2	0.07
5.2 |- 5.4	9	0.31
5.4 |- 5.6	9	0.31
5.6 |- 5.8	7	0.24
5.8 |- 6.0	1	0.035
Total	29	1.00

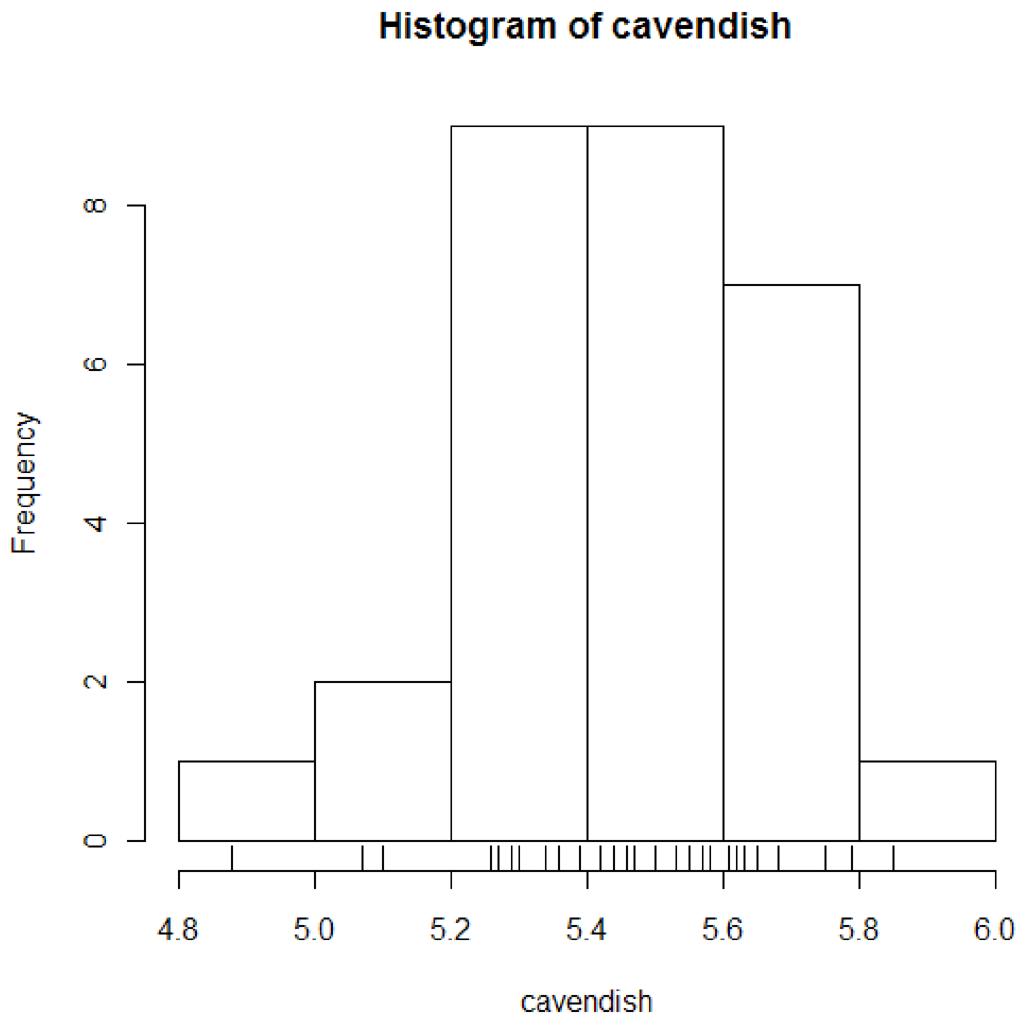
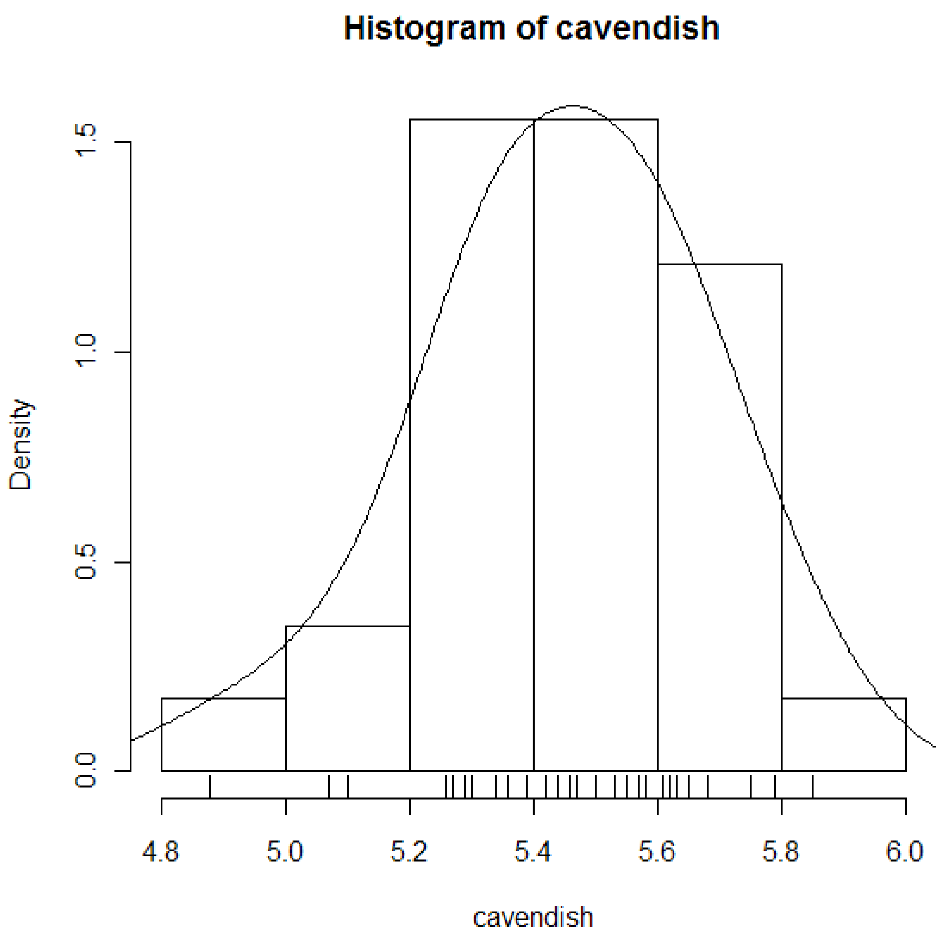


Tabela de Frequências e Histograma

```
> hist(cavendish,seq(4.8,6.0,0.2),prob=TRUE)
> rug(cavendish)
> lines(density(cavendish,bw="SJ"))
>
```



Mas afinal qual é a densidade da Terra?

Medidas Resumo1: Sumário de 5 números

Rol dos dados

```
4.88 5.07 5.10 5.26 5.27 5.29 5.29 5.30 5.34 5.34
5.36 5.39 5.42 5.44 5.46 5.47 5.50 5.53 5.55 5.57
5.58 5.61 5.62 5.63 5.65 5.68 5.75 5.79 5.85
```

4.88 5.07 5.10 5.26 5.27 5.29 5.29 5.30 5.34 5.34 5.36 5.39 5.42 5.44
5.46
5.47 5.50 5.53 5.55 5.57 5.58 5.61 5.62 5.63 5.65 5.68 5.75 5.79 **5.85**

Min = 4,88 **Max= 5.85**
Mediana (29*50 %=14 dos dados estão abaixo da mediana) = 5.46
Quartil 1 (29*25 %=7 dos dados estão abaixo de Q1) = 5.30
Quartil 3 (29*75 %= 21 dos dados estão abaixo de Q2) = 5.61

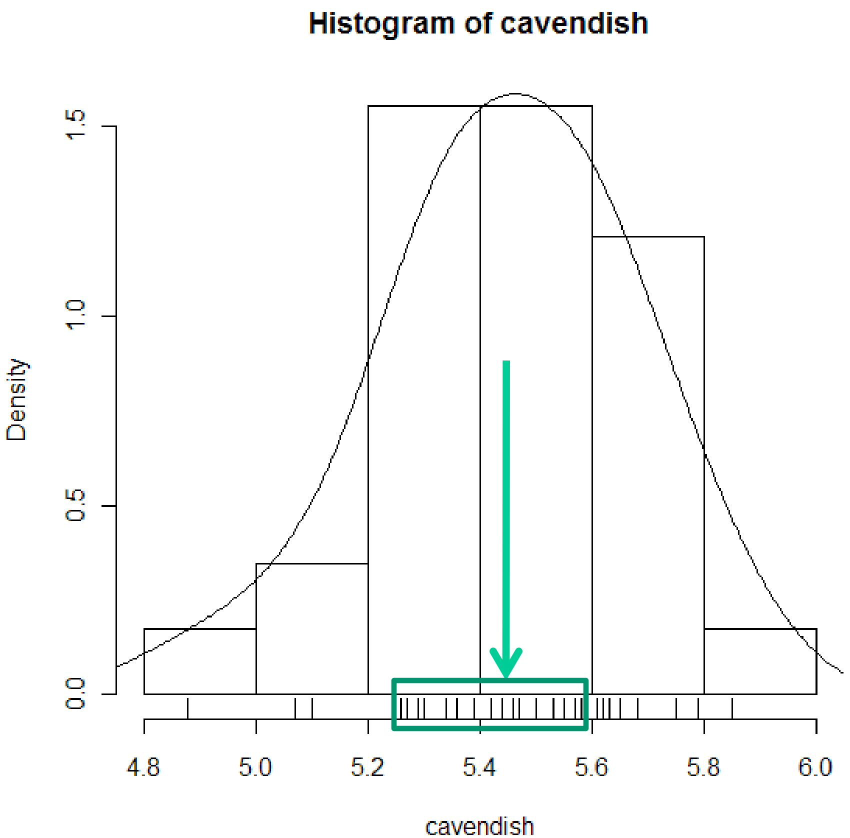
```
> summary(cavendish)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.880   5.300   5.460   5.448   5.610   5.850
>
```

Posição e Dispersão 1: Mediana e Distância Interquartis

```
> summary(cavendish)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  4.880   5.300   5.460   5.448   5.610   5.850 
> IQR(cavendish)
[1] 0.31
> median(cavendish)
[1] 5.46
>
```

$$\begin{aligned} \text{IQR} &= \text{Q3-Q1} \\ &= 5{,}61\text{-}5{,}30{=}0{,}31 \end{aligned}$$

50 % dos dados estão em uma caixa de tamanho IQR em torno da mediana. Esta caixa não precisa ser simétrica.

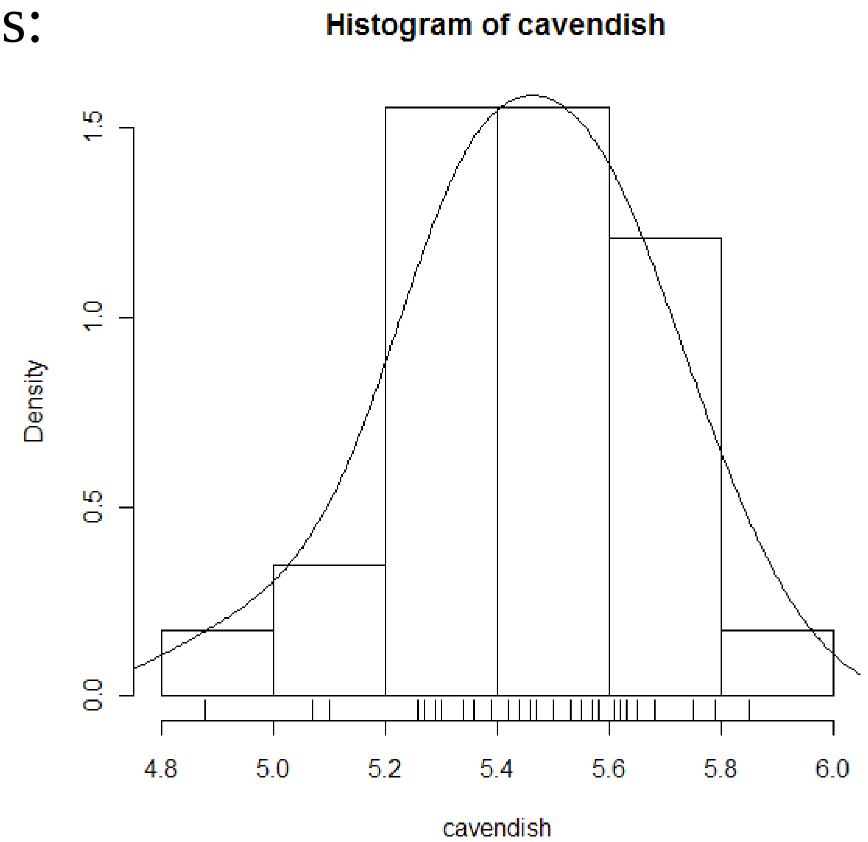


Que tamanho de intervalo utilizar para construir um histograma ?

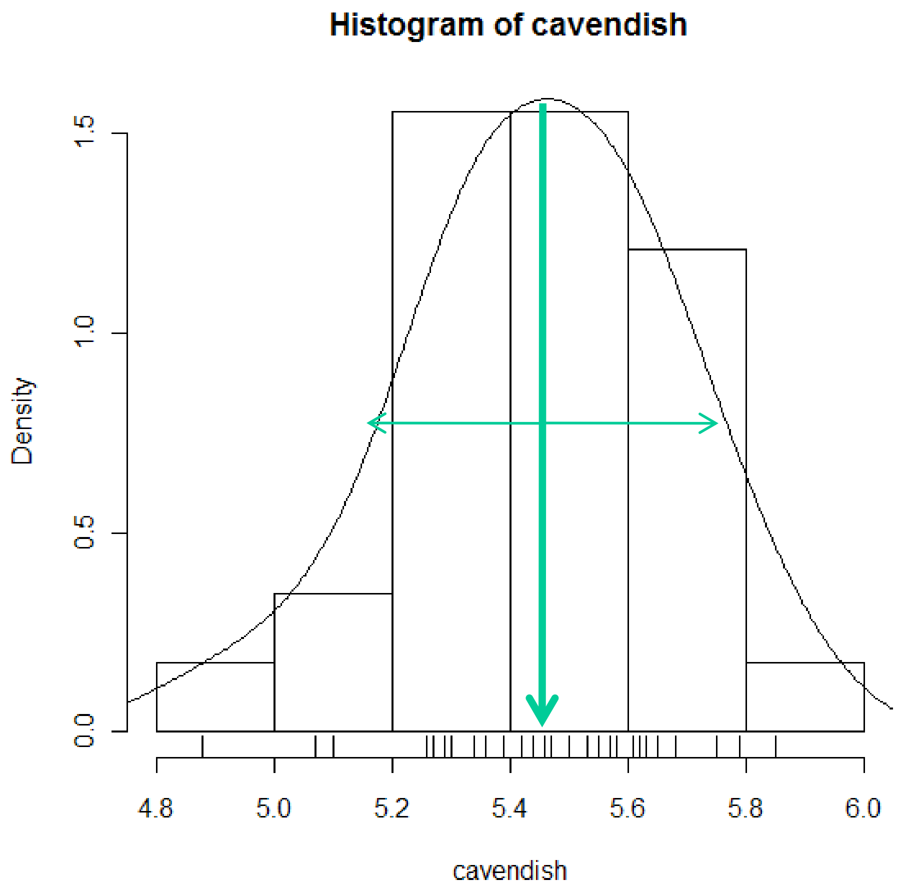
Critério de Freedman-Diaconis:

$$h = 2\frac{\mathrm{IQR}(x)}{n^{1/3}}$$

$\boxed{\begin{array}{l} \mathrm{h} = 2 \times 0.31 \;/\; (29)^{1/3} \\ = 0.2 \end{array}}$



Posição e Dispersão 2: Moda e Largura a Meia Altura (LMA)



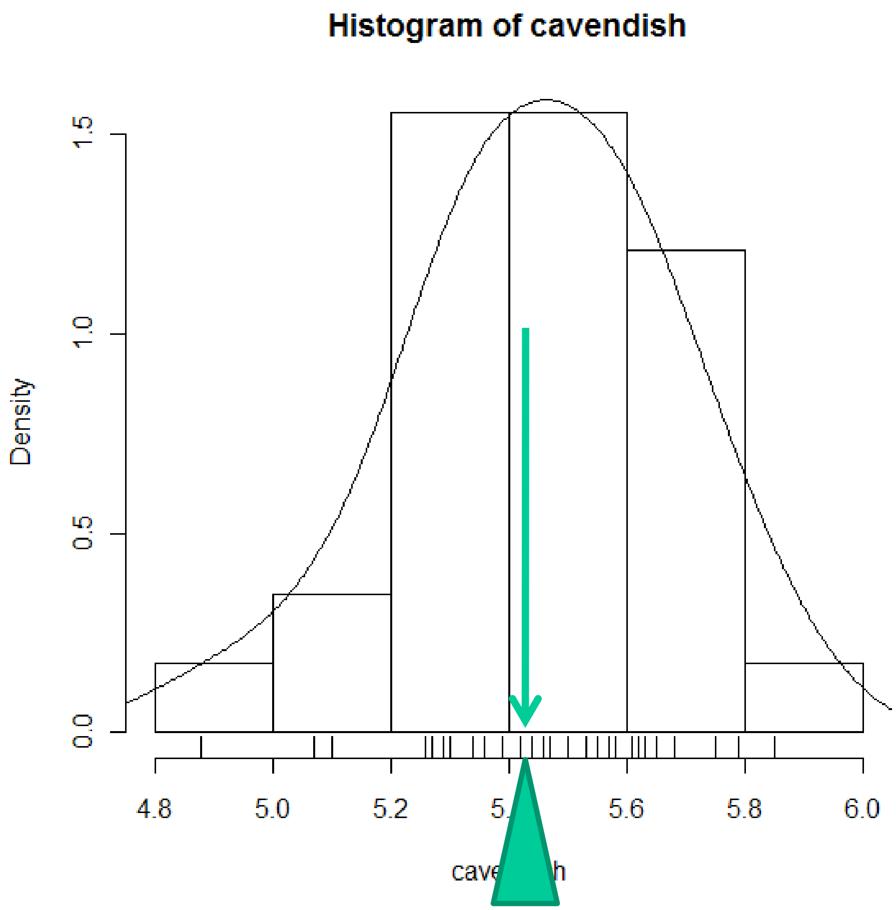
MODA = Valor (ou valores) mais freqüentes.

LMA = Largura da distribuição entre valores de densidade iguais a MAX/2.

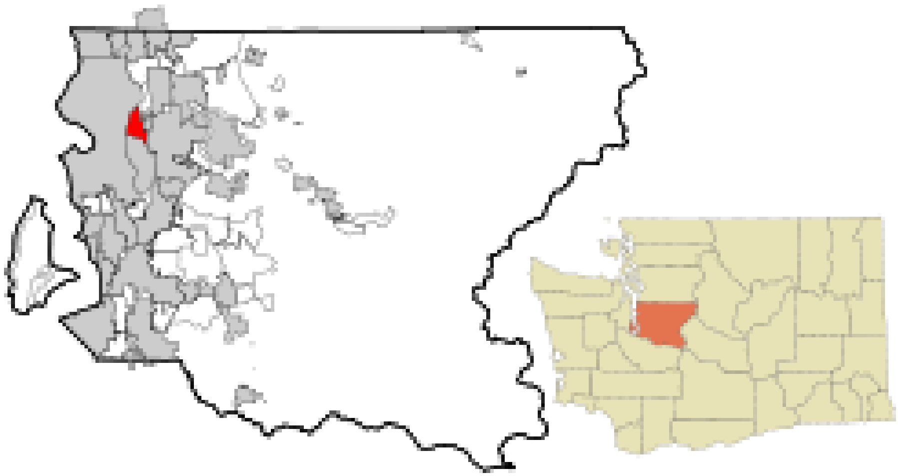
Medida de Posição 3: Média

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + \cdots + x_n).$$

```
> mean(cavendish)
[1] 5.447931
>
```



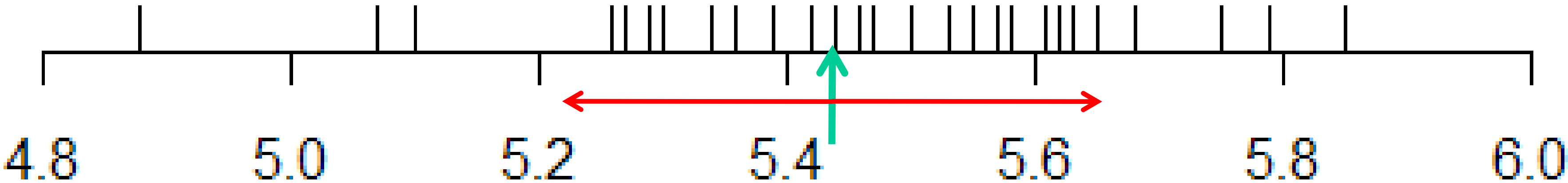
A média não é robusta



Medina, Washington, EUA
1079 residências (2000)

Patrimônio Líquido Médio = US$ 46 milhões
Patrimônio Líquido Médio sem Bill Gates = US$ 1,15 milhões

Medida de Dispersão 3:Desvio Padrão




```
> mean(cavendish)
[1] 5.447931
> var(cavendish)
[1] 0.04881700
>  sd(cavendish)
[1] 0.2209457
```

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

O **resíduo** é a diferença entre o valor observado e a média dos valores $(x_i - \overline{x})$

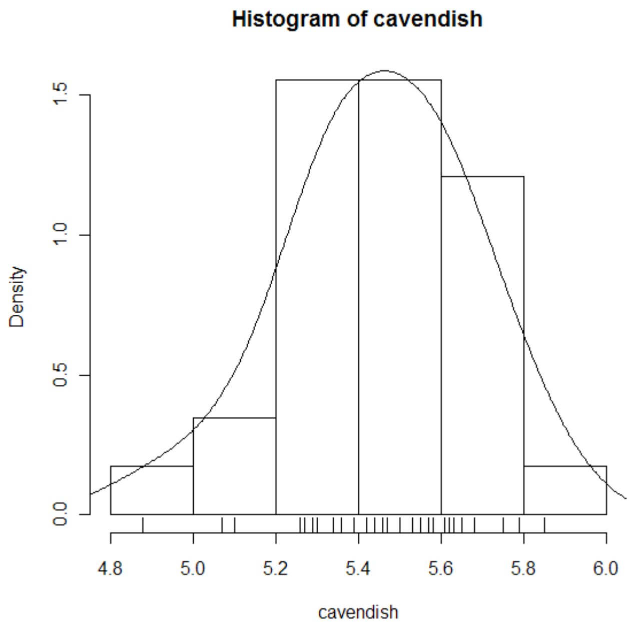
O **variância** é a média dos resíduos $\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2$

O **desvio padrão** é a raiz quadrada da variância. Se o resíduo tem unidades de comprimento (por ex. cm) então o desvio terá as mesmas unidades

Mas, e a densidade da Terra ?!

Erros de Medida

Cada medida individual da densidade da Terra tem uma precisão intrínseca revelada pela distribuição de valores. Esta precisão pode ser representada pelo desvio padrão σ=0,22.



Consideramos cada medida como sendo independente. Assim a primeira medida será 5,50 ± 0,22, a segunda seria 5,07 ± 0,22. Assim, não esperamos exatidão das medidas.

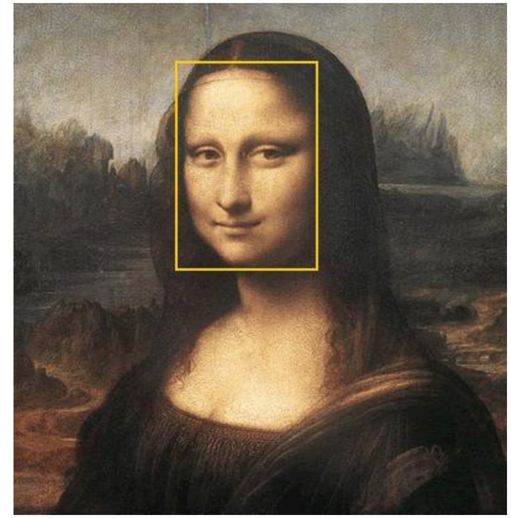
Uma forma de redução do efeito de erros de medida é através de repetições <u>independentes</u> do experimento. A precisão da média, neste caso será $\sigma/\sqrt{n}$.

A melhor estimativa para a densidade da Terra será, portanto,

$$\boxed{\bar{x} \pm \sigma/\sqrt{n} = 5{,}45 \pm 0{,}04}$$

Como comparar populações?

Razão Áurea e os Shoshoni

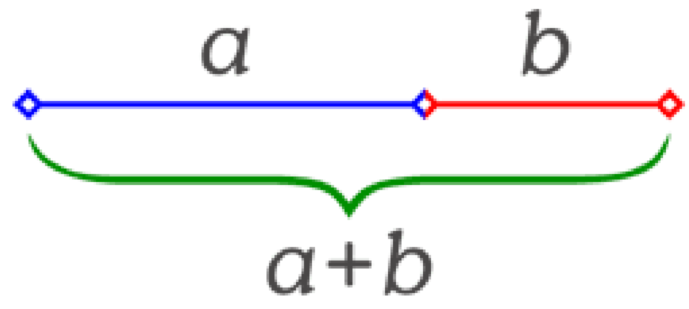


Os Shoshoni , tribos de nativos americanos do Wyoming, costumavam decorar seus objetos de couro com retângulos. Estariam os Shoshoni empregando a razão áurea em suas decorações.

Lowie's Selected Papers in Antrhopology (1970) Dubois,C. (ed)

Razão Áurea e os Shoshoni

Razão Áurea



a+b está para ***a*** assim como ***a*** está para ***b***

$$\frac{a+b}{a} = \frac{a}{b} = \varphi\,.$$

$$\varphi = \frac{1+\sqrt{5}}{2} \approx 1.6180339887\ldots$$

Um retângulo construído na razão áurea terá razão altura/largura = 1/1.618.... = 0.618034

Razão Áurea e os Shoshoni: Dados

Os dados abaixo representam 20 medidas da razão altura-largura de decorações Shoshoni. Os Shoshoni estão usando a razão áurea?

0.693	0.662	0.690	0.606	0.570
0.749	0.672	0.628	0.609	0.844
0.654	0.615	0.668	0.601	0.576
0.670	0.606	0.611	0.553	0.933

1. Construa o diagrama de ramo e folhas. 2. Calcule o sumário de 5 números. 3. Construa a tabela de freqüências e o histograma. 4. Calcule medidas resumo. 5. Tente responder a questão.

Comparando Populações

Para determinar se a população italiana moderna e a população etrusca antiga têm uma origem comum, Barnicot e Brothwell realizaram medidas da largura máxima de 84 crânios etruscos e de 70 crânios italianos modernos (em *Medical Biology and Etruscan Origins, 1959).*
As medidas (em mm) estão a

Etruscos

141	148	132	138	154	142	150	146	155	158	150	140
147	148	144	150	149	145	149	158	143	141	144	144
126	140	144	142	141	140	145	135	147	146	141	136
140	146	142	137	148	154	137	139	143	140	131	143
141	149	148	135	148	152	143	144	141	143	147	146
150	132	142	142	143	153	149	146	149	138	142	149
142	137	134	144	146	147	140	142	140	137	152	145

Italianos Modernos

133	138	130	138	134	127	128	138	136	131	126	120
124	132	132	125	139	127	133	136	121	131	125	130
129	125	136	131	132	127	129	132	116	134	125	128
139	132	130	132	128	139	135	133	128	130	137	133
140	143	144	137	140	136	135	126	139	131	133	138
141	140	130	137	134	130	148	135	138	135	138	

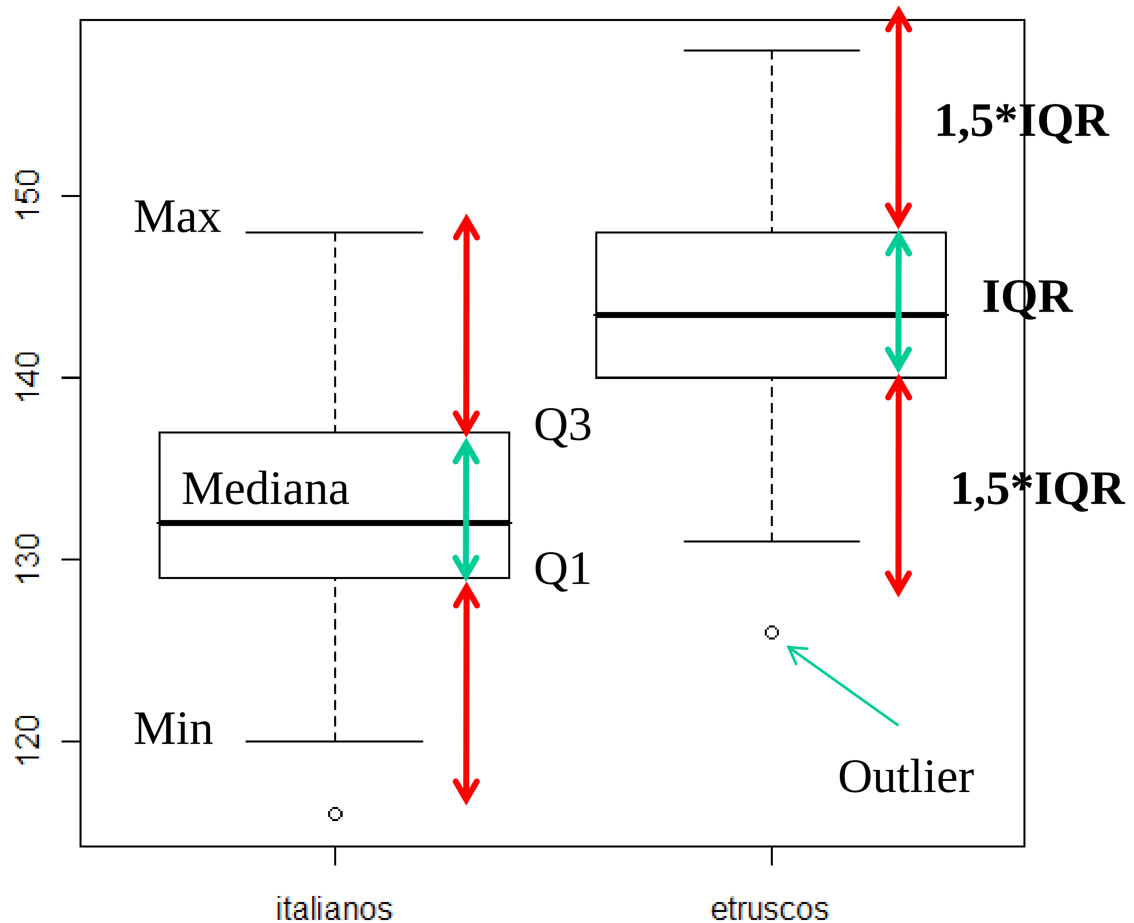
Comparando Populações: Box Plot

Começamos por calcular os sumários de 5 números para as duas amostras populacionais.

```
> italianos <- scan("C:/Documents and Settings/Renato Vicente/Desktop/Alesp/data/italian.dat")
Read 70 items

> etruscos <- scan("C:/Documents and Settings/Renato Vicente/Desktop/Alesp/data/etruscan.dat")
Read 84 items

> summary (italianos)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  116.0   129.0   132.0   132.4   136.8   148.0

> summary (etruscos)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  126.0   140.0   143.5   143.8   148.0   158.0

> boxplot(list(italianos=italianos,etruscos=etruscos))
```

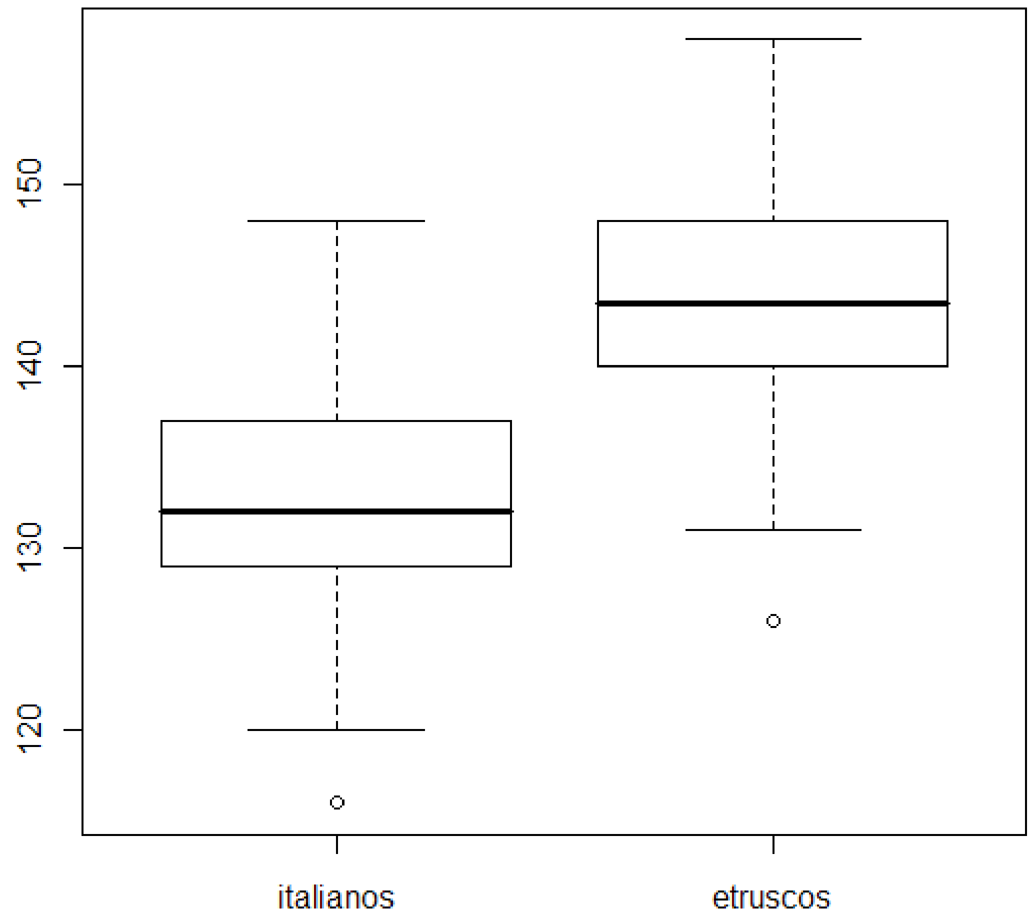
Comparando Populações: Box Plot

Dados abaixo de Q1− 1,5 IQR ou acima de Q3 + 1,5 IQR são considerados <u>outliers</u> e marcados de forma diferente.



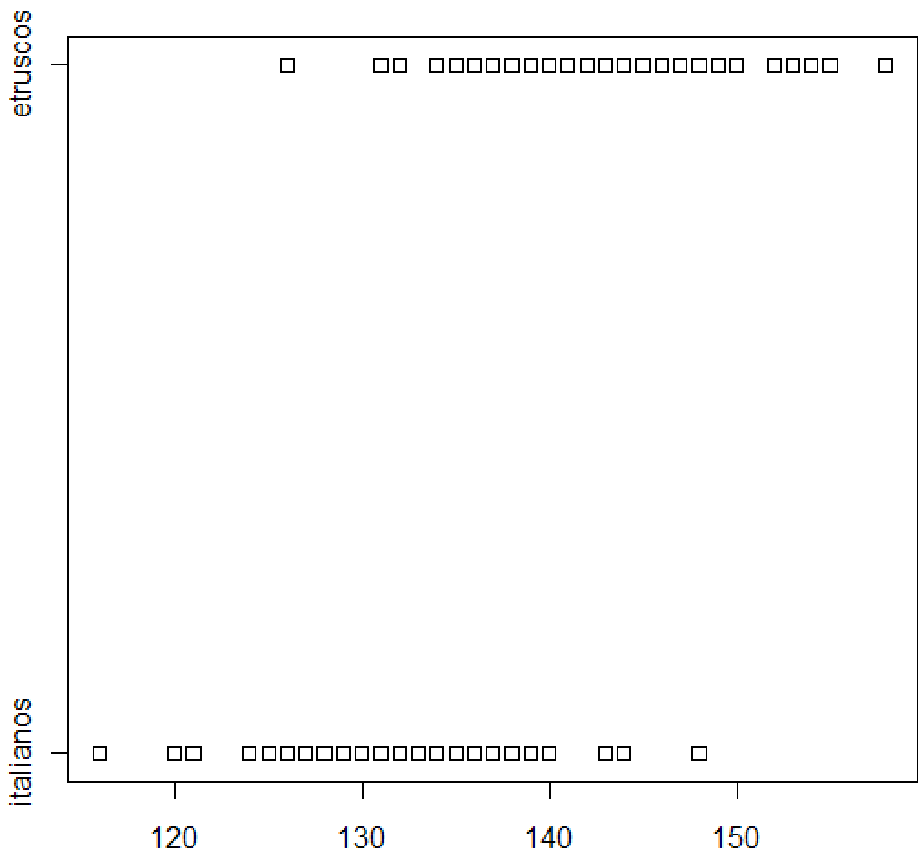
Comparando Populações: Box Plot

Os crânios etruscos são tipicamente mais largos do que os italianos.



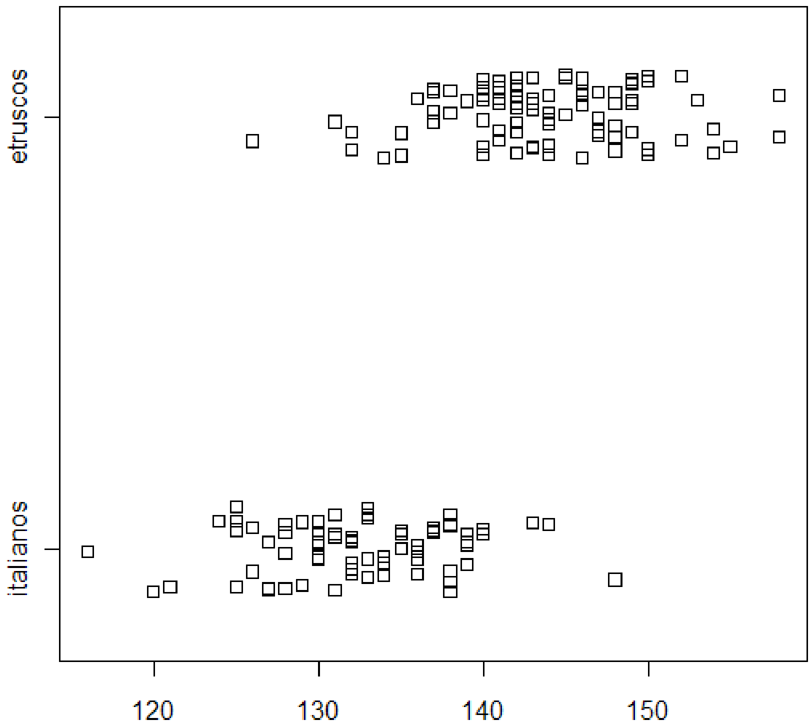
Comparando Populações: *Stripchart 2*

```
> stripchart(list(italianos=italianos,etruscos=etruscos))
```



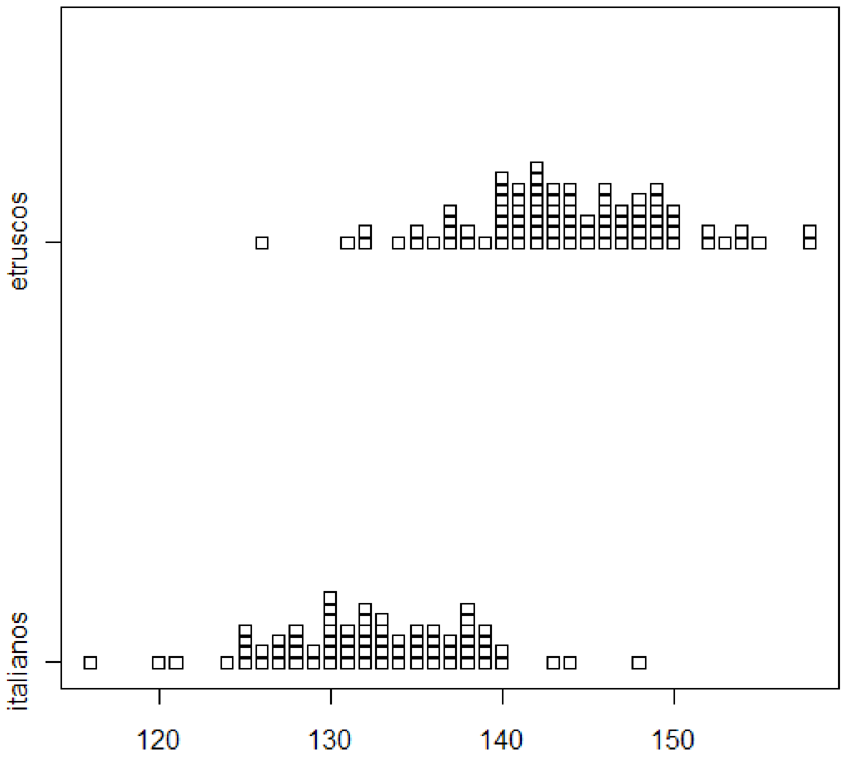
Comparando Populações: *Stripchart 2*

> stripchart(list(italianos=italianos,etruscos=etruscos),method="jitter")



Comparando Populações: *Stripchart 1*

```
> stripchart(list(italianos=italianos,etruscos=etruscos),method="stack")
```

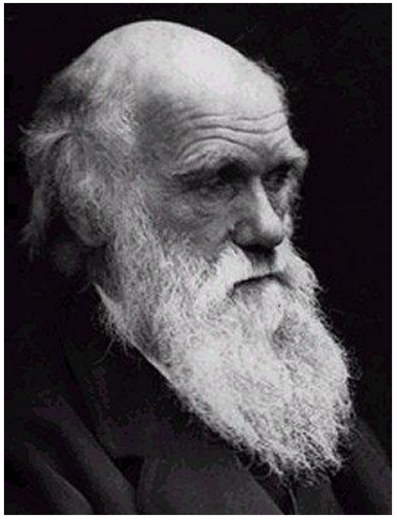


Testando a hipótese de diferença

```
> t.test(italianos,etruscos)

        Welch Two Sample t-test

data:  italianos and etruscos
t = -11.9659, df = 148.819, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -13.202123  -9.459782
sample estimates:
mean of x mean of y
 132.4429  143.7738
```

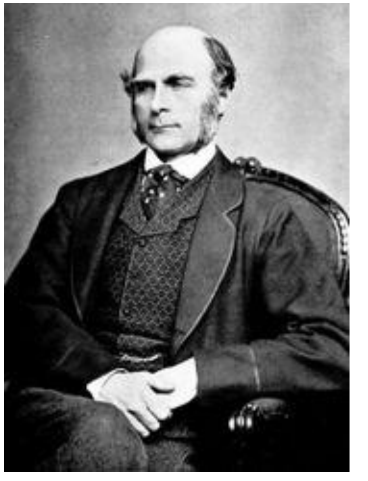
As populações são diferentes com p-valor praticamente nulo (quase certamente).

Galton e Darwin



Charles Darwin
(1809-1882)

Em seu livro 1876, *The Effect of Cross- and Self-fertilization in the Vegetable Kingdom,* Darwin descreveu uma série de experimentos projetados para demonstrar que a fertilização cruzada contribuiria para produzir plantas com crescimento mais vigoroso do que aquele observado em plantas auto-fertilizadas. Para a análise estatística dos dados Darwin consultou seu primo Francis Galton um dos pioneiros da bioestatística (e da eugenia).



Francis Galton
(1822-1911)

Darwin e Galton

Os dados ao lado representam as alturas finais em polegadas plantas provenientes de pares de sementes de mesma idade. Em um tratamento (Cross) a fertilização foi cruzada, no outro (Self) houve auto-fertilização.

Há evidência de diferença entre os tratamentos de fertilização?

1. Calcule sumários de 5 números.
2. Construa um box plot para efetuar a comparação.
3. Calcule média e desvio padrão.
4. Tente responder a questão.

Par	Cross	Self
1	23.5	17.4
2	12.0	20.4
3	21.0	20.0
4	22.0	20.0
5	19.1	18.4
6	21.5	18.6
7	22.1	18.6
8	20.4	15.3
9	18.3	16.5
10	21.6	18.0
11	23.3	16.3
12	21.0	18.0
13	22.1	12.8
14	23.0	15.5
15	12.0	18.0

Resumindo

1. Como surgiu a idéia?
2. O que é uma distribuição estatística?
3. Como utilizo as distribuições estatísticas?
4. Como comparo populações?
5. Resumindo