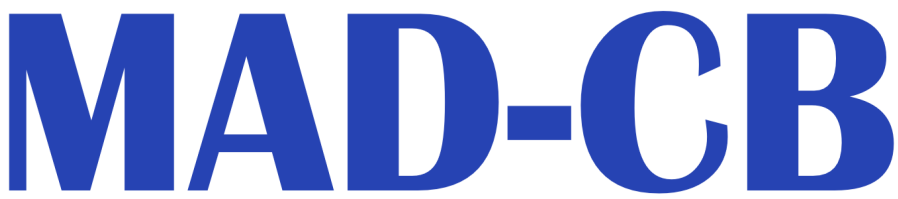


Figure 1:

Desvio Padrão e Variância

Medidas de Tendência Central e Dispersão Apropriadas

- As duas são *estatísticas* (para amostras) ou *paramêtros* (para populações)
- Devem ser apropriadas aos tipos de dados elas descrevem
- Ex: Média de uma variável não faz sentido

Exemplo: Número de Testes de Células CD4+ T em 2001 - 2009

```
## # A tibble: 9 × 2
##     ano     n
##   <dbl> <dbl>
## 1  2001  2196
## 2  2002 21217
## 3  2003 17017
## 4  2004 20366
## 5  2005 24753
## 6  2006 23480
## 7  2007 25354
## 8  2008 25347
## 9  2009 26507
```

Resumo (Summary) das Duas Variáveis - 1

```
summary(cd4n)
```

```
##       ano              n        
##  Min.   :2001   Min.   : 2196  
##  1st Qu.:2003   1st Qu.:20366  
##  Median :2005   Median :23480  
##  Mean   :2005   Mean   :20693  
##  3rd Qu.:2007   3rd Qu.:25347  
##  Max.   :2009   Max.   :26507
```

Resumo de n Usando Pacote DescTools

```
library(DescTools)
options(scipen = 1000)
Desc(cd4n$n, plotit = FALSE)
```

```
## ------------------------------------------------------------------------
## cd4n$n (numeric)
##
##      length           n         NAs      unique          0s        mean      meanSE
##           9           9           0         = n           0  20'693.00   2'520.67
##
##         .05         .10         .25      median         .75         .90         .95
##    8'124.40  14'052.80  20'366.00  23'480.00  25'347.00  25'584.60  26'045.80
##
##       range          sd       vcoef         mad         IQR        skew        kurt
##   24'311.00   7'562.01        0.37   3'355.12   4'981.00       -1.51        1.07
##
##
##    level  freq   perc  cumfreq  cumperc
## 1   2196     1  11.1%        1    11.1%
## 2  17017     1  11.1%        2    22.2%
## 3  20366     1  11.1%        3    33.3%
## 4  21217     1  11.1%        4    44.4%
## 5  23480     1  11.1%        5    55.6%
## 6  24753     1  11.1%        6    66.7%
## 7  25347     1  11.1%        7    77.8%
## 8  25354     1  11.1%        8    88.9%
## 9  26507     1  11.1%        9   100.0%
```

O Que Quer Dizer uma Média

- Média suficiente para descrever uma distribuição??
- Em média, esses duas pessoas parecem iguais?????

Uma Definição de Uma Média

Se você tem sua cabeça no congelador e seus pés no forno, em média, você sente confortável. Né?

- Precisa cuidar de que quer dizer com uma média!

Medindo Dispersão com Desvio da Média

```
suppressMessages(library(tidyverse))
dados1 <- tibble(x = c(1, 2, 3, 4, 5))
dados1
```

```
## # A tibble: 5 × 1
##       x
##   <dbl>
## 1     1
## 2     2
## 3     3
## 4     4
## 5     5
```

```
mean(dados1$x)
```

```
## [1] 3
```

Distância de Cada Ponto da Média

- $(x_i - \bar{x})$: diferença entre cada valor $i = 1 : 5$ e a média $(\bar{x})$

```
dev = dados1$x - mean(dados1$x)
(dados1 <- bind_cols(dados1, tibble(dev)))
```

```
## # A tibble: 5 × 2
##       x   dev
##   <dbl> <dbl>
## 1     1    -2
## 2     2    -1
## 3     3     0
## 4     4     1
## 5     5     2
```

```
sum(dados1$dev)
```

```
## [1] 0
```

Podemos Fazer Algo Útil dessa Tabela de Distância

- Truque: Fazer o quadrado: $(x_i - \bar{x})^2$

```
(dados1 <- bind_cols(dados1, tibble(devsq = dados1$dev^2)))
```

```
## # A tibble: 5 × 3
##       x   dev devsq
##   <dbl> <dbl> <dbl>
## 1     1    -2     4
## 2     2    -1     1
## 3     3     0     0
## 4     4     1     1
## 5     5     2     4
```

```
sum(dados1$devsq)
```

```
## [1] 10
```

Quadrado de Desvio Tem 2 Efeitos

1. Elimina os negativos
   - Negativos e positivos não podem cancelar um do outro
2. Aumenta desvios grandes mais que os pequenos
   - Dar para eles um peso maior

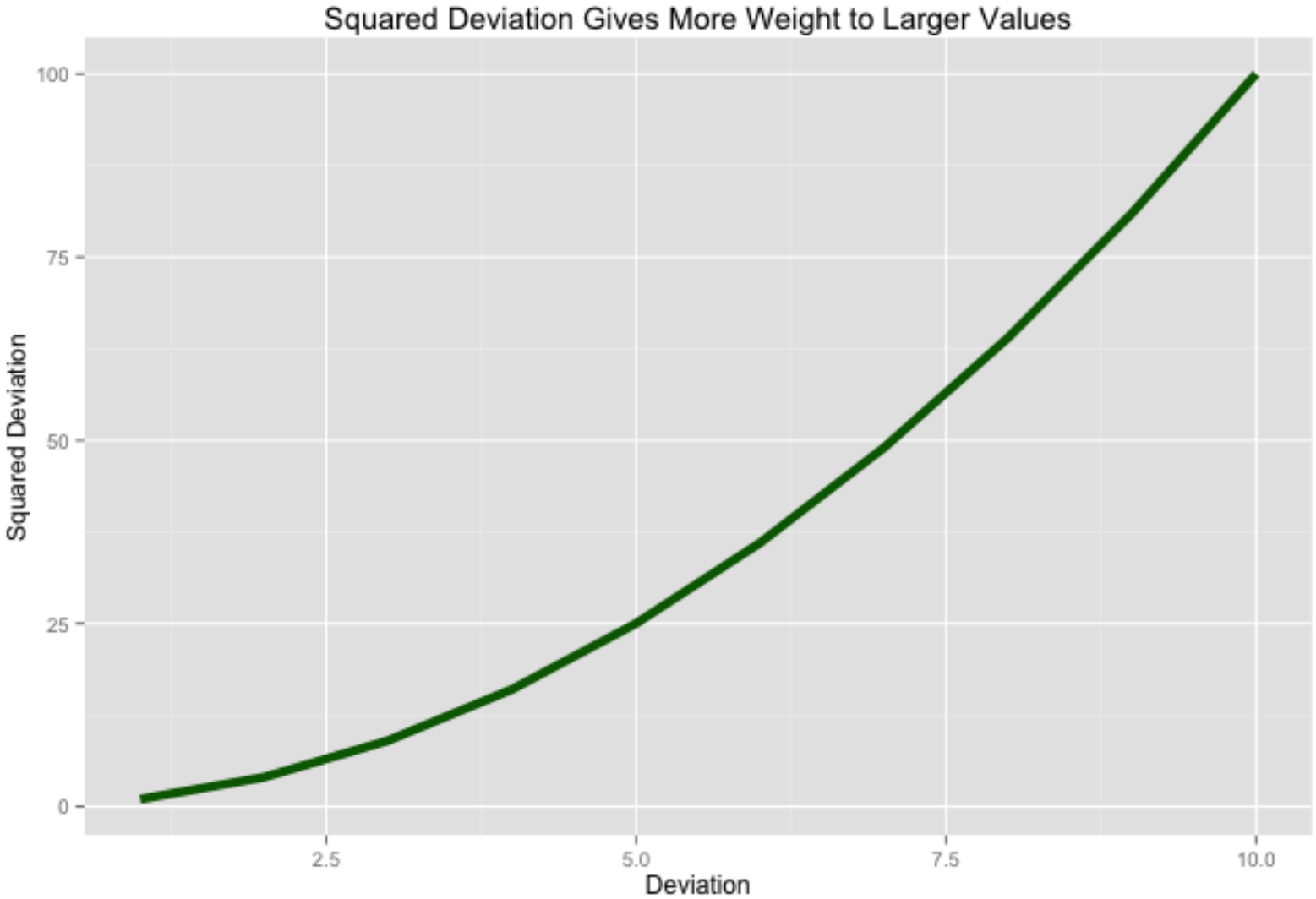

Squared Deviation Gives More Weight to Larger Values
Squared Deviation
100
75
50
25
0
2.5
5.0
7.5
10.0
Deviation

Figure 2:

Desvio Padrão/Standard Deviation

- Formula – População

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{n}}$$

- Formula – Amostra

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{(n-1)}}$$

- O Que é diferente entre as duas??????

Variância - Desvio Padrão ao Quadrado

- Duas formulas paralelas a Desvio Padrão
- Aqui – população

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{n}$$

- Esse formula parece parecido com uma outra que conhece
- Pode ver ele como a **média** dos desvios ao quadrado

O Que É Aquele (n-1)?

- Criou disputas desde o seculo 18
- Porque devemos dividir os desvios por um número menor?
- O que é o efeito no dp deste divisão por (n - 1)?
- Se n = 10, (n - 1) = 9, ou 10% menor
- Se n = 1000, (n - 1) = 999, ou 0.1% menor
- Pergunta prática:
  - Vale a pena reduzir n para (n - 1)?

Graus de Liberdade (Degrees of Freedom)

- A redução tem ligação ao conceito de **"graus de liberdade"**
  - Conceito que também usamos em relação a distribuições formais
  - Ex: t, F, χ^2, Normal
- Em nosso `dados`, podemos escolher o valor de todos exceto 1 livremente (n - 1)
  - Sem mudar a média e o desvio padrão
- MAS, se escolhemos o último valor e mudá-lo
  - Média e desvio padrão mudam

Tratamento de Graus de Liberdade nas Amostras

- Assim, não podemos escolher o valor de último número (n^{o}) livremente
  - O valor de já determinada média e do desvio padrão mudariam também
- $\therefore$ precisamos tirar 1 de denominador para as amostras
- Eu concordo, mas outros não
  - Eles acham n é suficiente

Médias, Medianas e Skewness

- Um conjunto de dados um pouco diferente

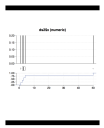
```
(dados2 <- tibble(x = c(1, 2, 3, 4, 50)))
```

```
## # A tibble: 5 × 1
##       x
##   <dbl>
## 1     1
## 2     2
## 3     3
## 4     4
## 5    50
```

- 50 - outlier

```r
Desc(dados2$x, plotit = FALSE)
```

```
## ------------------------------------------------------------------------
## dados2$x (numeric)
##
##   length       n     NAs  unique      0s    mean  meanSE
##        5       5       0     = n       0   12.00    9.51
##
##      .05     .10     .25  median     .75     .90     .95
##     1.20    1.40    2.00    3.00    4.00   31.60   40.80
##
##    range      sd   vcoef     mad     IQR    skew    kurt
##    49.00   21.27    1.77    1.48    2.00    1.07   -0.93
##
##
##    level  freq   perc  cumfreq  cumperc
## 1      1     1  20.0%        1    20.0%
## 2      2     1  20.0%        2    40.0%
## 3      3     1  20.0%        3    60.0%
## 4      4     1  20.0%        4    80.0%
## 5     50     1  20.0%        5   100.0%
```



Stephen Jay Gould e Estatística

- Ensaio: “The Median Isn’t the Message”
- Sofreu de um câncer cujo tempo de sobrevivência mediana foi 8 meses
- Ele sobreviveu mais 20 anos
  - Ele morreu de um outro câncer não relacionado

Medidas Robustas

- Robusto: insensível a outliers; pode lidar bem com skewness
- Não-robusto: sensível a outliers; utilidade cai com skewness

	Robust	Non-Robust
center	median	mean
spread	IQR	sd, range

Última Demonstração Disso - O Quarteto de Anscombe

- Conjunto de 4 distribuições
  - Todos os primeiros 3 têm os mesmos valores `x`
  - Todos têm a mesma média e desvio padrão
  - São totalmente diferentes

As Distribuições

```
anscombe
```

```
##    x1 x2 x3 x4    y1   y2    y3    y4
## 1  10 10 10  8  8.04 9.14  7.46  6.58
## 2   8  8  8  8  6.95 8.14  6.77  5.76
## 3  13 13 13  8  7.58 8.74 12.74  7.71
## 4   9  9  9  8  8.81 8.77  7.11  8.84
## 5  11 11 11  8  8.33 9.26  7.81  8.47
## 6  14 14 14  8  9.96 8.10  8.84  7.04
## 7   6  6  6  8  7.24 6.13  6.08  5.25
## 8   4  4  4 19  4.26 3.10  5.39 12.50
## 9  12 12 12  8 10.84 9.13  8.15  5.56
## 10  7  7  7  8  4.82 7.26  6.42  7.91
## 11  5  5  5  8  5.68 4.74  5.73  6.89
```

As Médias e os Desvios Padrões

```
ansres <- anscombe %>% summarize_all(funs(mean, sd))
(round(ansres, 2))
```

```
##   x1_mean x2_mean x3_mean x4_mean y1_mean y2_mean y3_mean y4_mean x1_sd
## 1       9       9       9       9     7.5     7.5     7.5     7.5  3.32
##   x2_sd x3_sd x4_sd y1_sd y2_sd y3_sd y4_sd
## 1  3.32  3.32  3.32  2.03  2.03  2.03  2.03
```

Os Gráficos de Distribuição

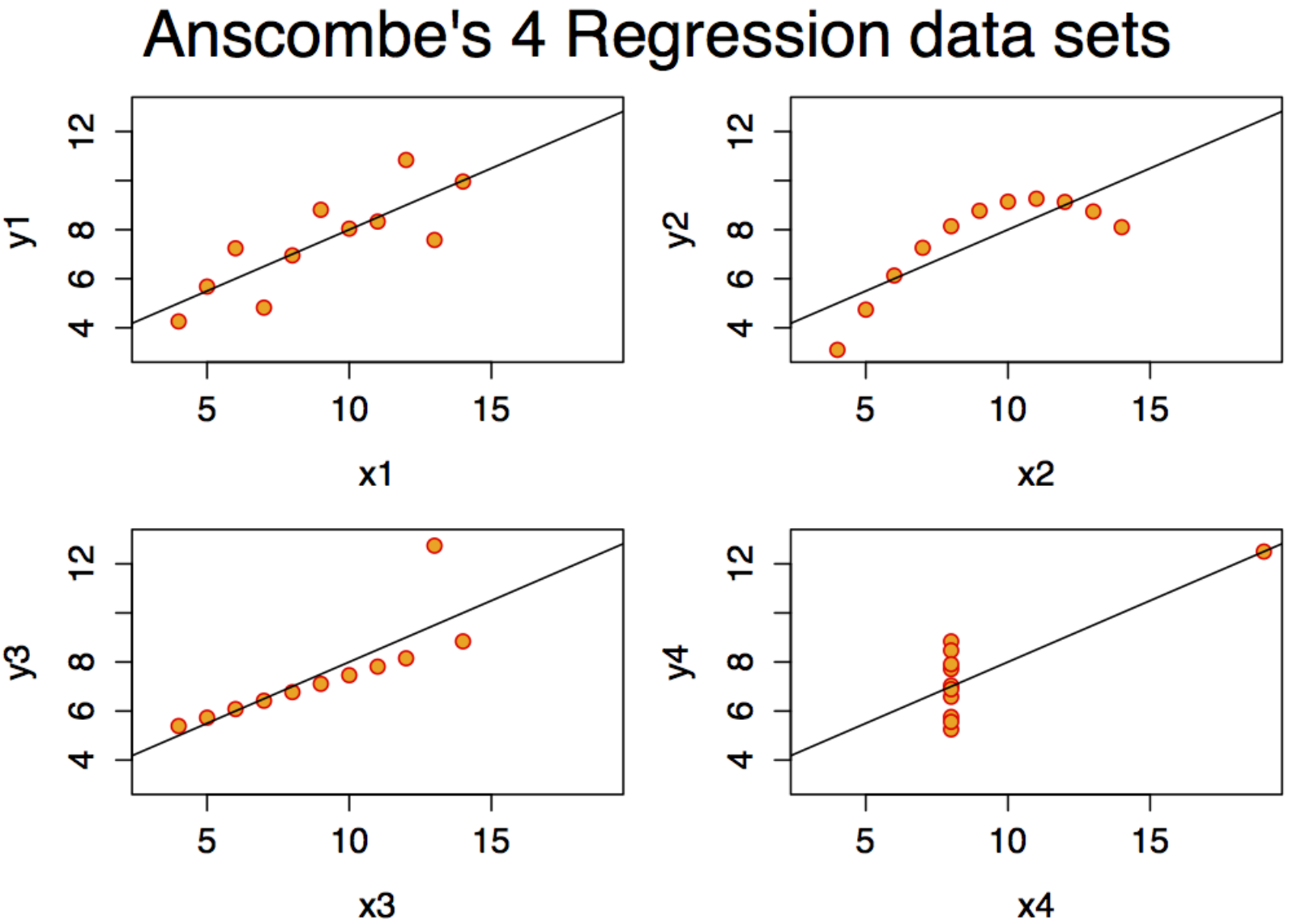


Figure 3:

O Que Anscombe Quer Dizer

- Não aceita as estatísticas que os programas produzem
- Visualizar os dados com gráficos
- Pense!