ME111 - Laboratório de Estatística

Aula 4 - Análise Bivariada

Profa. Larissa Avila Matos

Associação entre duas variáveis

- Sua opinião sobre o comportamento de uma variável muda na presença de informação de uma segunda variável?
- A **distribuição conjunta** das duas variáveis descreve a associação existente entre elas.
- Grau de dependência: como uma variável "explica" ou se "associa" a outra.
- Assim como na análise univariada estas relações podem ser resumidas por gráficos, tabelas e/ou medidas estatística. O tipo de resumo vai depender dos tipos das variáveis envolvidas. Vamos considerar três possibilidades:
  - as duas variáveis são qualitativas;
  - as duas variáveis são quantitativas; e
  - uma variável é quantitativa e a outra qualitativa.

- Salienta-se ainda que:
  - As análise mostradas a seguir não esgotam as possibilidades de análises envolvendo duas variáveis e devem ser vistas apenas como uma sugestão inicial;
  - Relações entre duas variáveis devem ser examinadas com cautela pois podem ser mascaradas por uma ou mais variáveis adicionais não considerada na análise. Estas são chamadas variáveis de confundimento. Análises com variáveis de confundimento não serão discutidas neste ponto;
  - Vamos continuar analisando o conjunto de dados `SleepStudy`.

- Reelembrando: Lendo e transformando os dados.

```r
# Carregando os dados
library(Lock5Data)
data(SleepStudy)
# Transformando a variável Gender em fator
dados<-SleepStudy
dados$Gender[dados$Gender==1]<-"Male"
dados$Gender[dados$Gender==0]<-"Female"
dados$Gender<-as.factor(dados$Gender)
# Transformando a variável AnxietyStatus em fator ordenado,
# ou seja, em variável qualitativa ordinal
dados$AnxietyStatus<-ordered(dados$AnxietyStatus)
```

Qualitativa vs Qualitativa

- Vamos considerar as variáveis `Gender` (Sexo) e `AnxietyStatus` (Ansiedade).
- A tabela envolvendo duas variáveis é chamada tabela de cruzamento ou tabela de contingência e pode ser apresentada de várias formas, conforme será discutido a seguir. A forma mais adequada de apresentação vai depender dos objetivos da análise e da interpretação desejada para os dados. Iniciamente obtemos com o comando `table()` a tabela de frequências absolutas. A tabela extendida incluindo os totais marginais pode ser obtida com `addmargins()`.

```r
GenderAnxiety.tb <- table(dados$Gender, dados$AnxietyStatus)
addmargins(GenderAnxiety.tb)
```

```
        moderate normal severe Sum
Female        37    102     12 151
Male          19     79      4 102
Sum           56    181     16 253
```

- 37 mulheres tem ansiedade no nível moderado.
- Na última coluna: frequência de cada nível da variável `Gender`.
- Na última linha: frequência de cada nível da variável `AnxietyStatus`.
- Parte interna da tabela: frequências conjuntas entre `Gender` e `AnxietyStatus`.

Frequências Relativas

- Podemos considerar também proporções condicionais (frequências relativas):
  - em relação ao total de elementos;
  - em relação ao total de cada linha;
  - em relação ao total de cada coluna.
- A proporção condicional escolhida depende do estudo que pretendemos fazer.

Frequências Relativas

- Distribuição das frequências relativas ao total da amostra.
- Total da amostra é 253.

```
round(addmargins(prop.table(GenderAnxiety.tb)),3)
```

```
        moderate normal severe   Sum
Female     0.146  0.403  0.047 0.597
Male       0.075  0.312  0.016 0.403
Sum        0.221  0.715  0.063 1.000
```

- 14,6% dos alunos são mulheres e sofrem de ansiedade no nível moderado.

Frequências relativas ao total das colunas

- Distribuição das frequências relativas ao total de cada coluna.

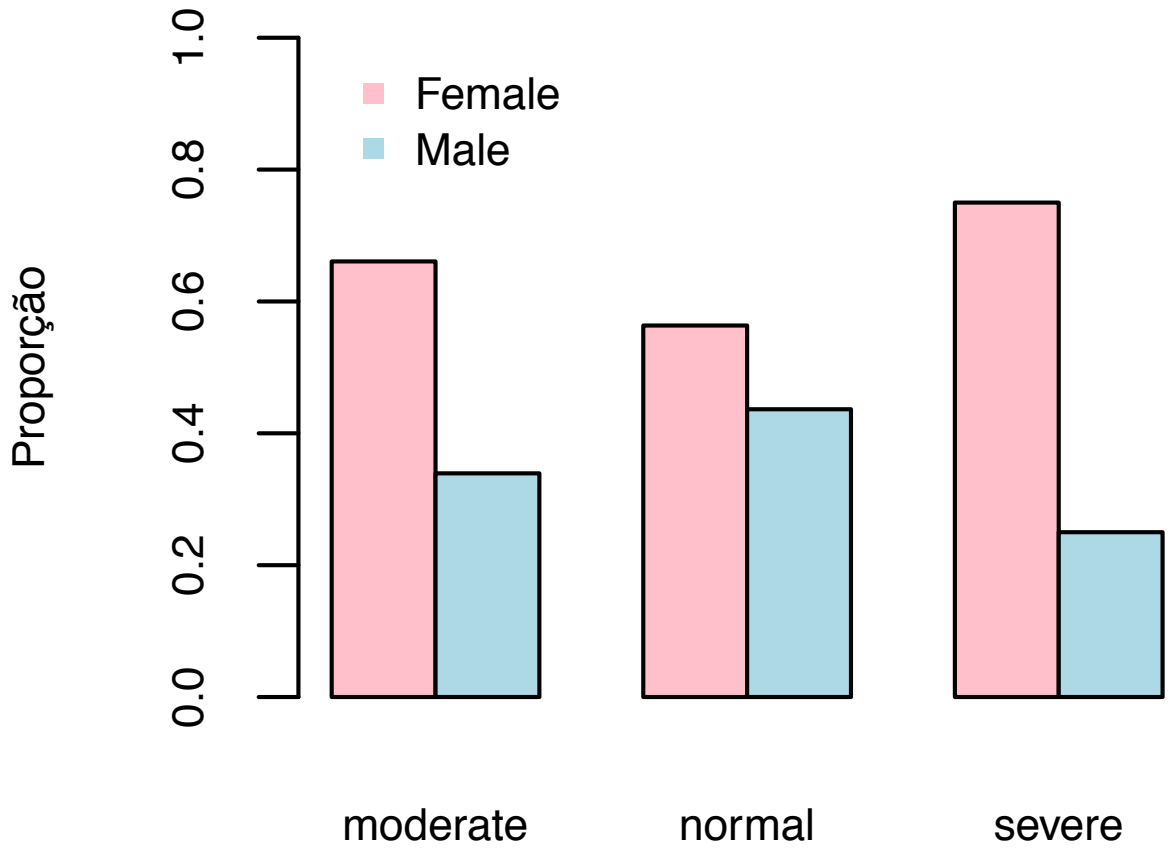
```
round(addmargins(prop.table(GenderAnxiety.tb, 2), 1),2)
```

```
        moderate normal severe
Female      0.66   0.56   0.75
Male        0.34   0.44   0.25
Sum         1.00   1.00   1.00
```

- Entre os alunos com ansiedade no nível normal:
  - 56% são mulheres.
  - 44% são homens.
- Permite comparar a distribuição do sexo (`Gender`) conforme o nível de ansiedade (`AnxietyStatus`).

Sexo conforme o nível de ansiedade

```
barplot(prop.table(GenderAnxiety.tb, 2), xlab=" ", ylab="Proporção",
        beside=TRUE, legend.text=TRUE, ylim=c(0,1), main=" ",
        col=c("pink","lightblue"), cex.axis=0.7,
        cex.lab=0.7, cex=0.7, las=0.8)
```



Frequências relativas ao total das linhas

- Distribuição das frequências relativas ao total de cada linha.

```
round(addmargins(prop.table(GenderAnxiety.tb, 1), 2),2)
```

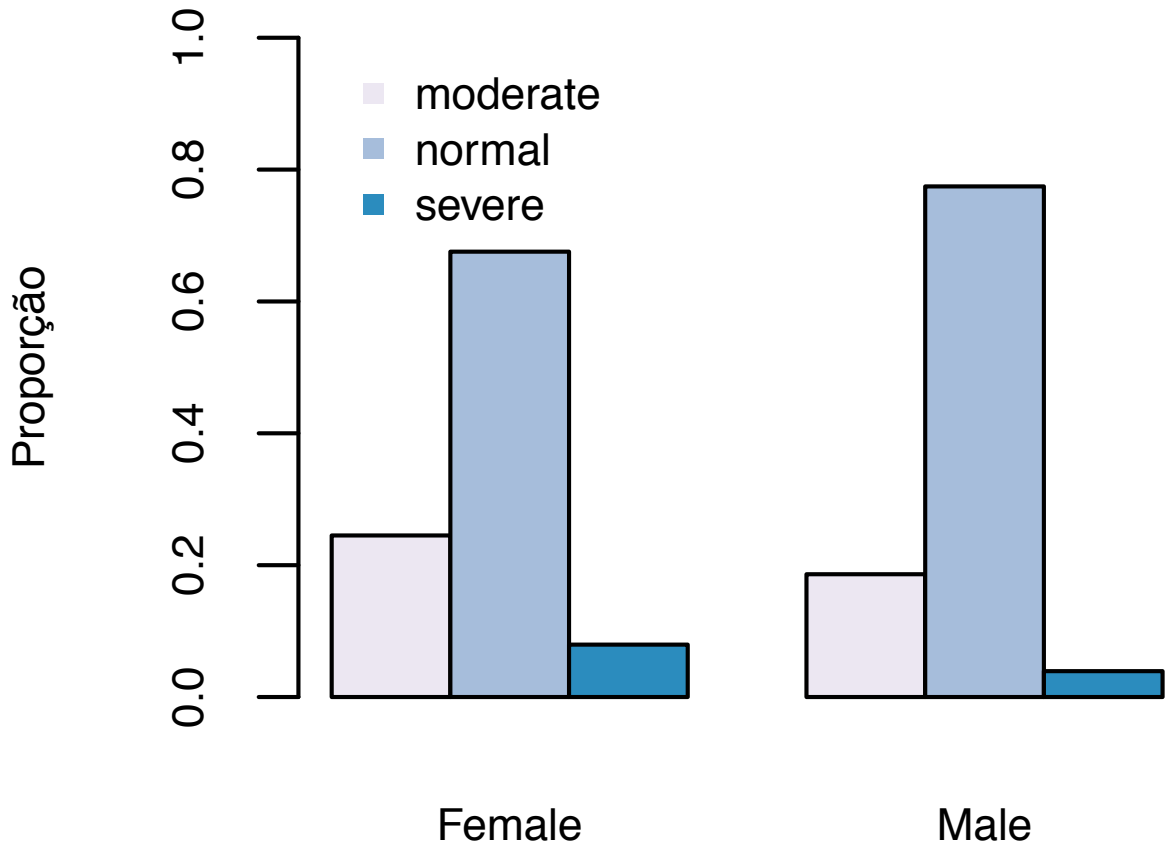
```
        moderate normal severe  Sum
Female      0.25   0.68   0.08 1.00
Male        0.19   0.77   0.04 1.00
```

- Entre os alunos do sexo masculino:
  - 19% sofrem de ansiedade no nível moderado.
  - 77% sofrem de ansiedade no nível normal.
  - 4% sofrem de ansiedade no nível severo.

Permite comparar a distribuição do nível de ansiedade (`AnxietyStatus`) conforme o sexo (`Gender`).

Nível de ansiedade conforme o sexo

```
barplot(t(prop.table(GenderAnxiety.tb, 1)), xlab=" ", ylab="Proporção",
        beside=TRUE, legend.text=TRUE, ylim=c(0,1), main=" ",
        col=c("#ece7f2","#a6bddb","#2b8cbe"), cex.axis=0.7,
        cex.lab=0.7, cex=0.7, las=0.7)
```

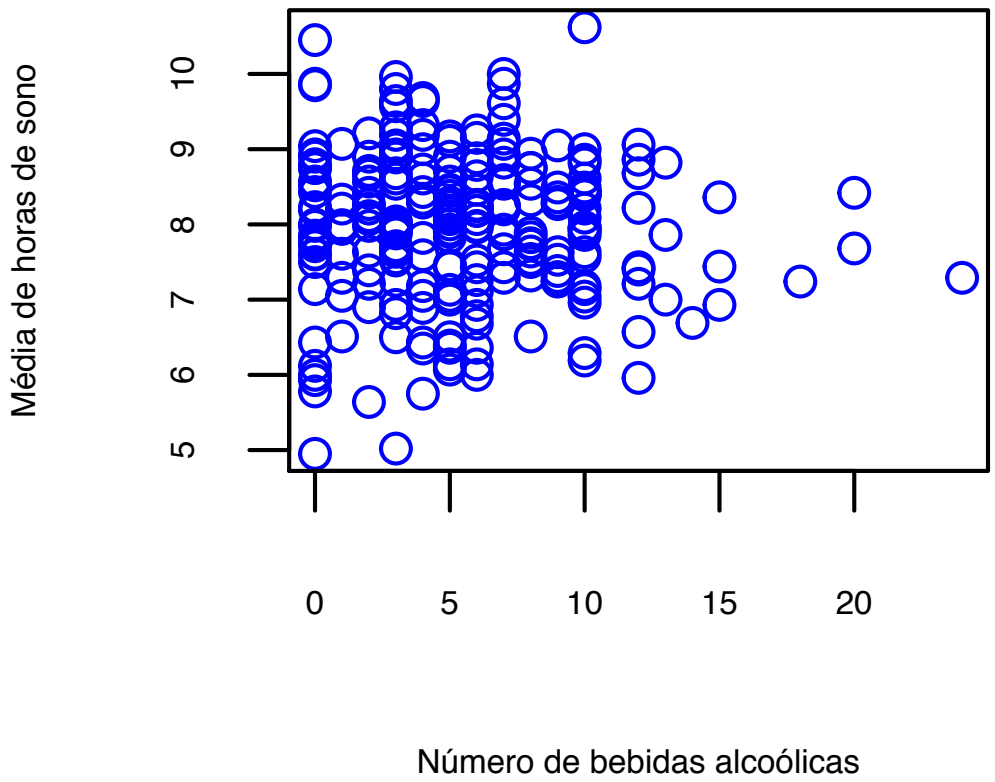


Associação entre duas variáveis quantitativas

- Associação entre duas variáveis **qualitativas**: comparar proporções condicionais.
- Associação entre duas variáveis **quantitativas**: comparamos como a mudança de uma variável afeta a outra variável.
- Nesse caso, vamos considerar o número de bebidas alcoólicas por semana e média de horas de sono

Diagrama de dispersão

```
plot(x=dados$Drinks,y=dados$AverageSleep, xlab="Número de bebidas alcoólicas",
     ylab="Média de horas de sono",main="",pch=21,cex.axis=0.5,cex.lab=0.5,col="blue")
```



- Nesse caso, parece não existir uma dependência entre o número de bebidas alcoólicas por semana e média de horas de sono.

Coeficiente de Correlação

- **Objetivo:** obter uma medida que permita quantificar a dependência que pode existir entre duas variáveis (positiva, negativa, muita ou pouca).
- Dado n pares de observações $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$:

$$Corr(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

  onde s_x é o desvio padrão de X e s_y é o desvio padrão de Y.
- Essa medida leva em consideração todos os desvios $(x_i - \bar{x})$ e $(y_i - \bar{y})$ padronizados da forma $z_{x_i} = \frac{x_i - \bar{x}}{s_x}$ e $z_{y_i} = \frac{y_i - \bar{y}}{s_y}$.
- Interpretação: z_{x_i} indica o número de desvios-padrão que a observação x_i está afastada da média de X.

Propriedades

- $-1 \leq Corr(X, Y) \leq 1$
- $Corr(X, Y)$ próxima de 1: X e Y estão positivamente associadas e o tipo de associação entre as variáveis é linear.
- $Corr(X, Y)$ próxima de -1: X e Y estão negativamente associadas e o tipo de associação entre as variáveis é linear.
- Se z_x e z_y têm o mesmo sinal, estamos somando um termo positivo na expressão da correlação.
- Se z_x e z_y têm sinais opostos, estamos somando um termo negativo na expressão da correlação.

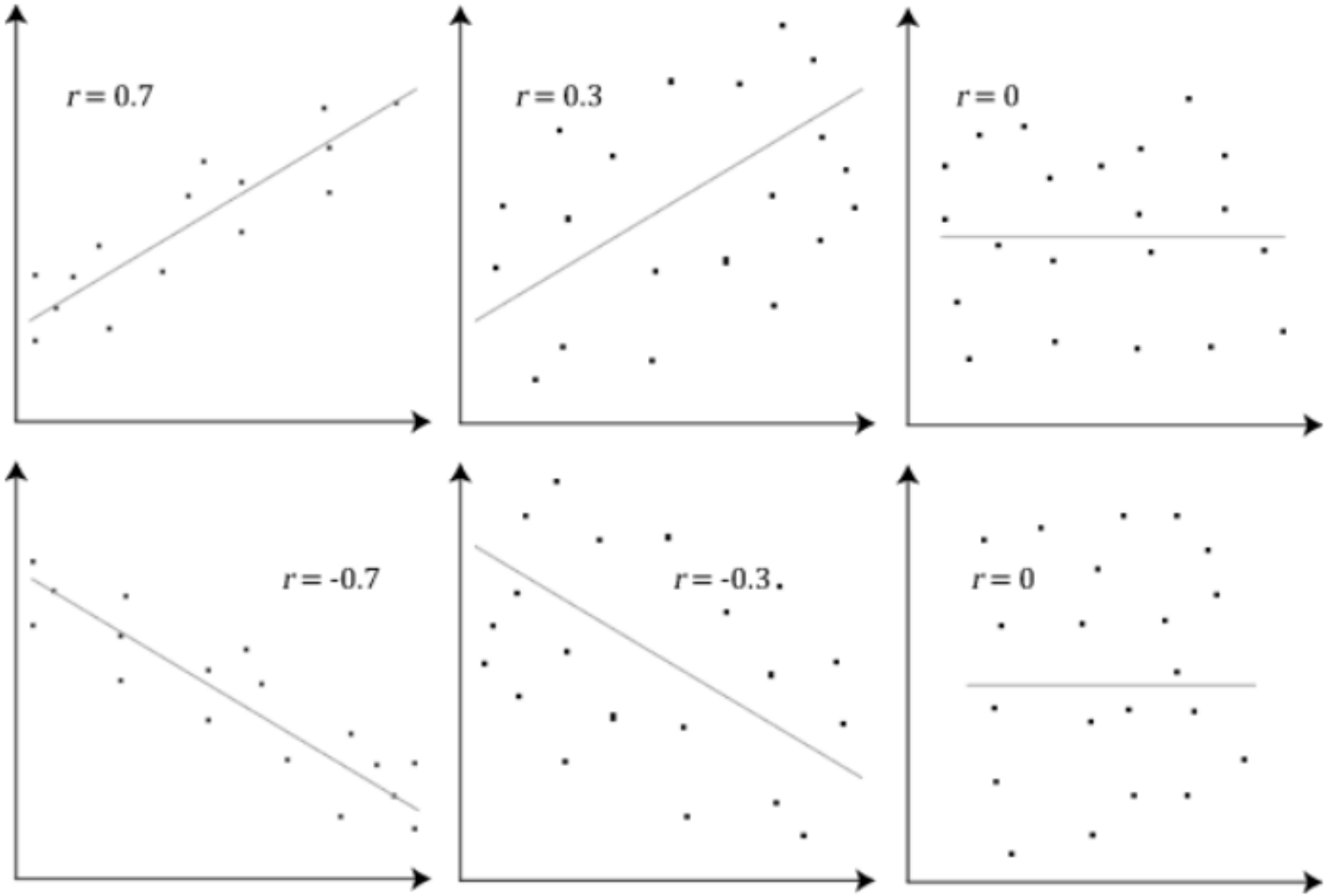


Figure 1: Correlação

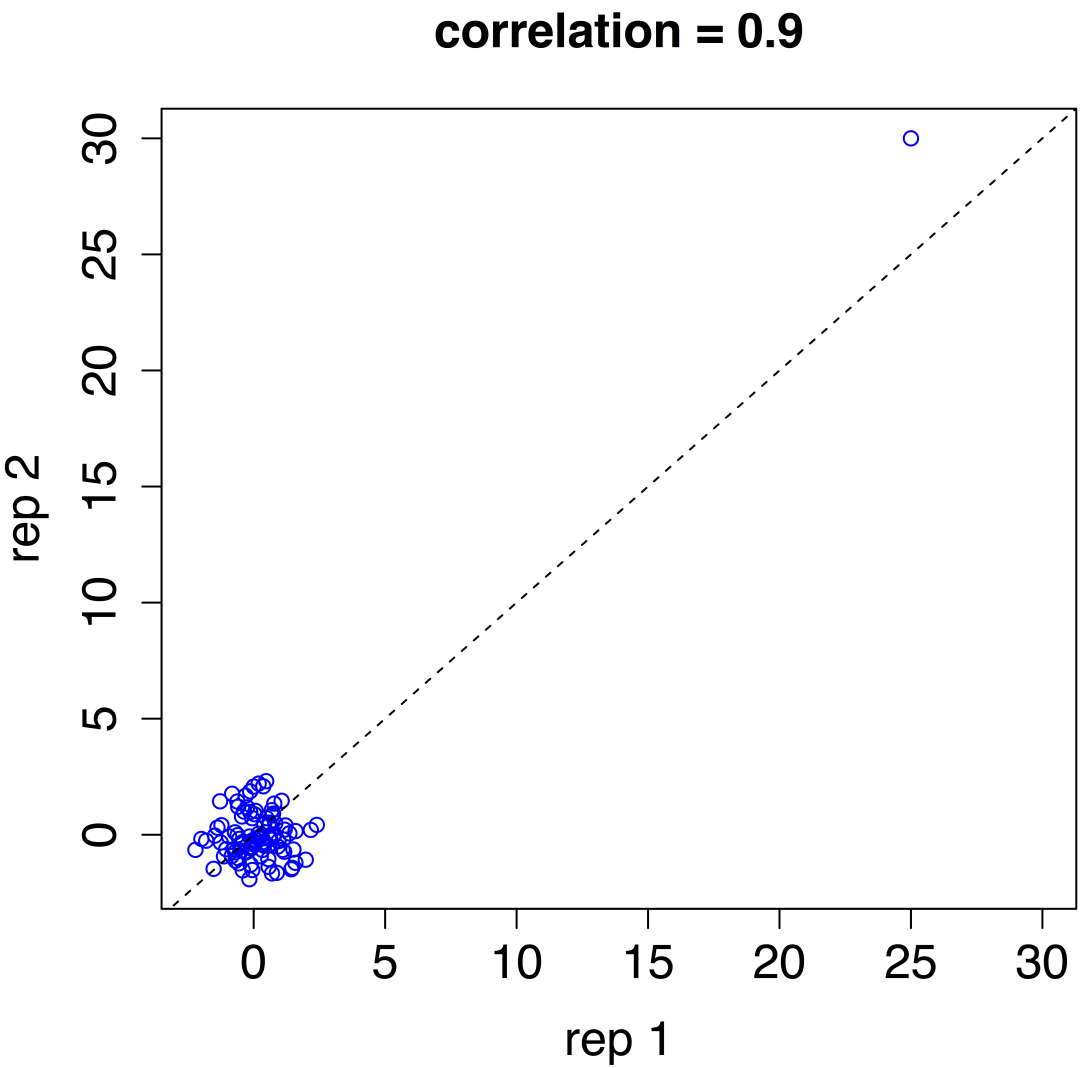
- Para as variáveis `Drinks` e `AverageSleep` nós temos que a correlação é

```
round(cor(dados$Drinks,dados$AverageSleep),3)
```

```
[1] -0.037
```

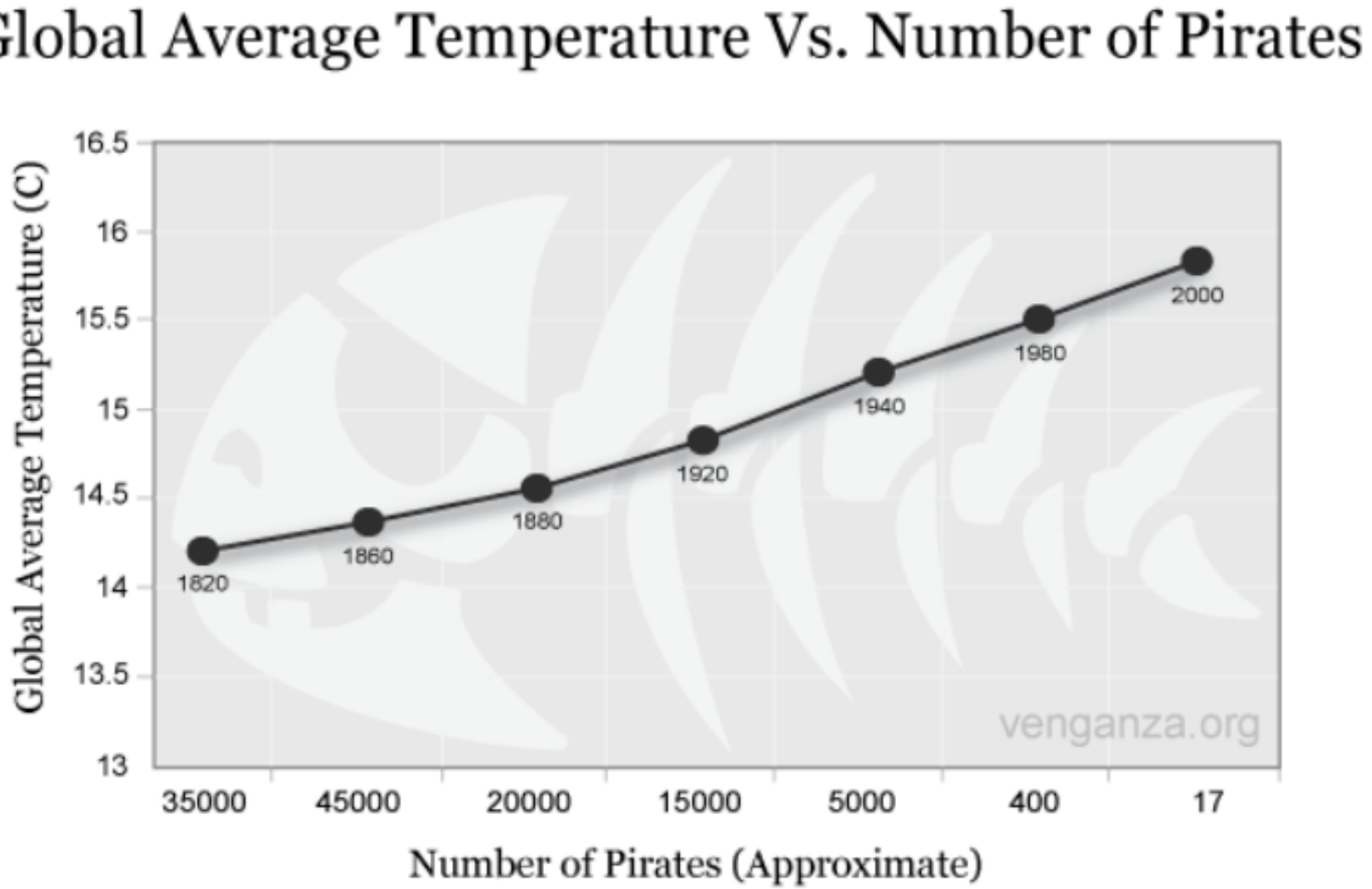
- Como a correlação é próxima de zero, significa que não existe correlação linear. Mas como é diferene de zero, podemos dizer que é uma correlação negativa bem pequena

Cuidado: correlação e *outliers*



Fonte:
<http://simplystatistics.org/2015/08/12/correlation-is-not-a-measure-of-reproducibility/>

Cuidado: correlação não implica causa!

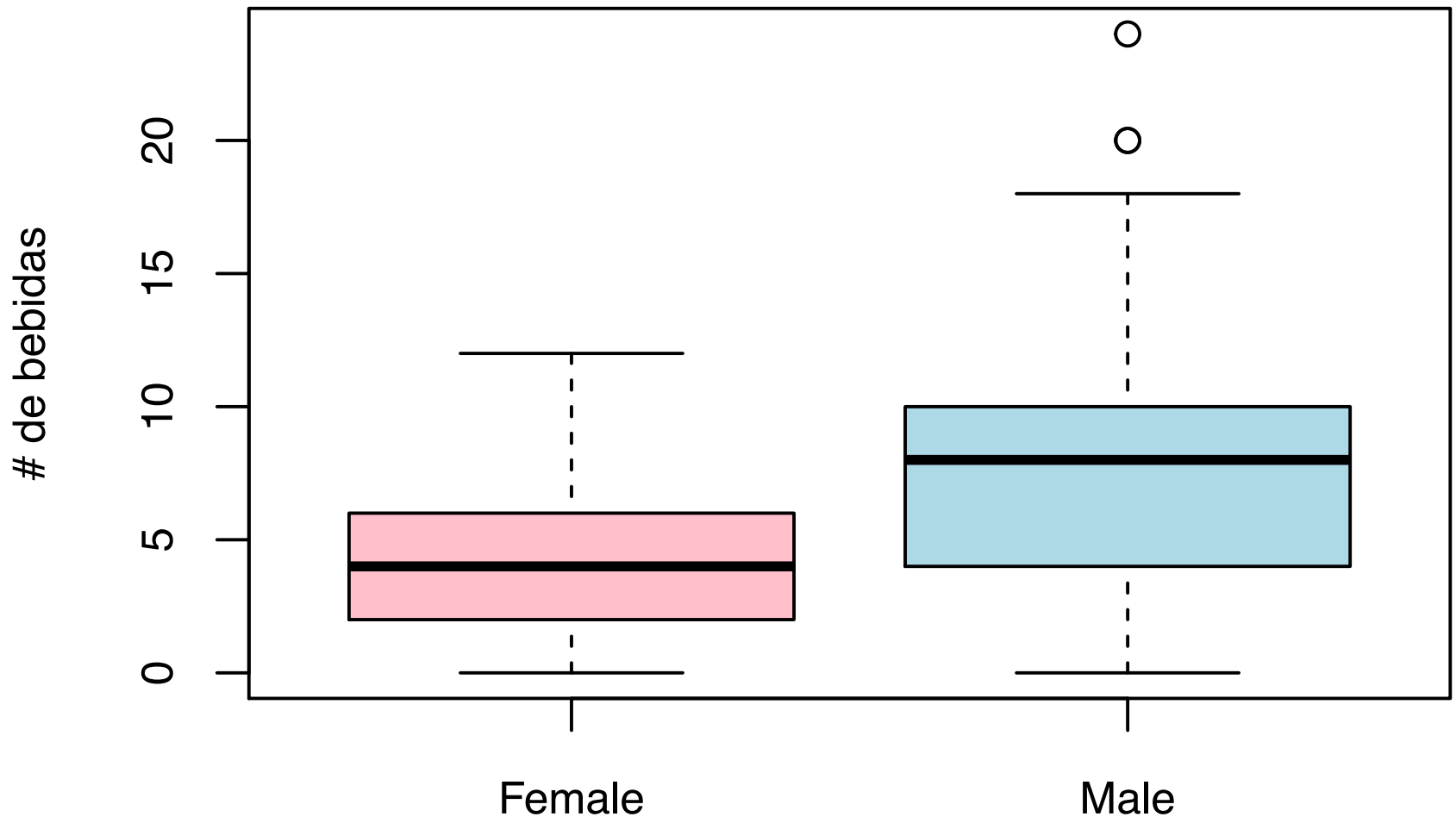


Qualitativa vs Quantitativa

- Associação entre uma variável **quantitativa** e uma variável **qualitativa**: É comum nessas situações analisar o que acontece com a variável quantitativa dentro de cada categoria da variável qualitativa. Essa análise pode ser conduzida por meio de medidas-resumo, histogramas, box plots ou ramo-e-folhas.

- Nesse caso, vamos considerar as variáveis `Gender` (Sexo) e `Drinks` (número de bebidas alcoólicas por semana).

```
boxplot(dados$Drinks~dados$Gender,col=c("pink","lightblue"),
        ylab="# de bebidas",cex.lab=0.8,cex.axis=0.8)
```



```
tapply(dados$Drinks,dados$Gender,summary)
```

```
$Female
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   2.000   4.000   4.238   6.000  12.000

$Male
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   4.250   8.000   7.539  10.000  24.000
```

```r
quantile(dados$Drinks)
```

```
  0%  25%  50%  75% 100%
   0    3    5    8   24
```

```r
Drinks.cl <- cut(dados$Drinks, quantile(dados$Drinks),
                 include.lowest = T)
DG.tb <- table(dados$Gender, Drinks.cl)
DG.tb
```

```
        Drinks.cl
         [0,3] (3,5] (5,8] (8,24]
  Female    64    39    38     10
  Male      24    10    21     47
```

```
round(addmargins(prop.table(DG.tb, margin = 1),2),2)
```

```
        Drinks.cl
         [0,3] (3,5] (5,8] (8,24]  Sum
  Female  0.42  0.26  0.25   0.07 1.00
  Male    0.24  0.10  0.21   0.46 1.00
```

Referências

- Bussab, W. O. & Morettin, P. A. (1987). *Estatística Básica. Atual Editora Ltda., São Paulo.*
- *http://www.leg.ufpr.br/~paulojus/embrapa/Rembrapa/Rembrapase8.html*
- *Magalhães, Marcos N.; Lima, Antonio Carlos P. (2010). Noções de probabilidade e estatística. São Paulo: Edusp, 2010.*
- *Wardrop, R. L. (1995). Statistics: Learning in the presence of variation.*