

Curso de Pós-Graduação em Ciências Veterinárias - UFRRJ

Métodos Estatísticos

Prof: Wagner Tassinari

wagner.tassinari@ini.fiocruz.br

Medidas de Associação

Medidas de Associação

- Um problema no qual nos deparamos frequentemente é verificar se determinada característica de uma população está ou não relacionada com a outra(s) e em que grau;
- Até a década de 1970, boa parte da teoria estatística usada partia do princípio de que as variáveis eram contínuas e, se possível, tinham distribuição normal. O motivo para isto é que podia-se lançar mão de uma série de métodos referentes a dist. normal;
- Metodologias para lidar com variáveis não-normais ou não existiam ou se referiam a problemas bastante específicos.

Medidas de Associação - Variáveis Qualitativas

Medidas de Associação - Variáveis Qualitativas

- Se as duas variáveis em estudo são independentes, espera-se que a distribuição marginal de uma delas (sem discriminar por valores da outra) seja igual às distribuições condicionadas por valores da outra;
- A partir dessa idéia, podemos construir uma medida de associação entre duas variáveis qualitativas, conhecida como Qui-quadrado.

Teste Qui-Quadrado

- É um teste de associação entre variáveis independentes e cujas variáveis são qualitativas (nominal e/ou ordinal);
- Seu objetivo é verificar se a distribuição das frequências observadas se desvia significativamente das frequências esperadas;
- O qui-quadrado testa a associação entre variáveis, mas não permite obter qualquer evidência quanto a força ou sentido dessa inter-relação;
- Alguns autores trabalham como sendo o Coeficiente de Contingência (C), sendo $0 \leq C \leq 1$.

- A estatística de qui-quadrado é dado por:

$$\chi^2 = \frac{\sum_{i=1}^k (f_{\text{observado}} - f_{\text{esperado}})^2}{f_{\text{esperado}}}$$

- H_0 : As duas variáveis não são associadas ($C = 0$)
- H_1 : As duas variáveis são associadas ($C \neq 0$)

Exemplo

O efeito de uma nova droga contra a febre aftosa está sendo testada em um rebanho através de um estudo prospectivo, ou seja, os animais são selecionados aleatoriamente para participar do estudo, e registra-se sua evolução. Para saber se a droga tem efeito ou não, alguns dos animais selecionados recebem a droga, outros recebem um placebo de maneira randomizada. Veja a tabela a seguir.

- H_0 : O tratamento não funciona, ou seja, o tratamento não está associado com a condição do animal. ($C = 0$)
- H_1 : O tratamento é eficaz, ou seja, o tratamento está associado com a condição do animal. ($C \neq 0$)

Observações:

- A correção de continuidade de *yates* é utilizada quando a frequência observada de uma das 'caselas' for < 5 ;
- O teste exato de *fisher* é utilizado quando o nosso n é muito pequeno.

Exemplo:

Exemplo: Suponha a seguinte tabela de contingência

Tratamento	Melhora	Não-melhora	total
Droga	48	8	56
Placebo	38	13	51
total	86	21	107



Solução utilizando o Rcommander

- Não precisa importar banco algum
 - Rcommander → Estatísticas → Tabelas de Contingência → Digite e analise tabela de dupla entrada

Digite tabela de dupla-entrada (two-way)

Tabela Estatísticas

Name for Row Variable (optional):


Name for Column Variable (optional):

Número de linhas: 2

Número de Colunas: 2

Entrar número:

	1	2
1	48	8
2	38	13

 Ajuda  Resetar  Aplicar  Cancelar  OK

Solução utilizando o Rcommander

Digite tabela de dupla-entrada (two-way)






Tabela Estatísticas

Computar Percentagens

☐ Percentual nas linhas ☐ Percentual nas colunas
☒ Percentagens do total ☐ Sem percentual

Teste de Hipóteses

☒ Teste de independência de Qui-Quadrado ☒ Componentes da estatística do Qui-quadrado
☐ Apresente frequências esperadas ☒ Teste exato de Fisher

 Ajuda  Resetar  Aplicar  Cancelar  OK

Output

```
> library(abind, pos=16)
> .Table <- matrix(c(48,8,38,13), 2, 2, byrow=TRUE)
> dimnames(.Table) <- list("Tratamento"=c("1", "2"), "Situação do animal"=c("1", "2"))
> .Table   # Counts
           Situação do animal
Tratamento 1 2
1      48  8
2      38 13
> totPercents(.Table) # Percentage of Total
           1 2 Total
1      44.9 7.5 52.3
2      35.5 12.1 47.7
Total 80.4 19.6 100.0
> .Test <- chisq.test(.Table, correct=FALSE)
> .Test
Pearson's Chi-squared test

data:  .Table
X-squared = 2.1243, df = 1, p-value = 0.145
```

Solução utilizando o Rcommander e no plugin EZT

```
> round(.Test$residuals^2, 2) # Chi-square Components
      Situação do animal
Tratamento   1   2
1 0.20 0.81
2 0.22 0.89

> remove(.Test)

> fisher.test(.Table)

      Fisher's Exact Test for Count Data

data: .Table
p-value = 0.2226
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.6990424 6.3044903
sample estimates:
odds ratio
 2.038804

> remove(.Table)
```

- Como o $p\text{-valor} = 0,145$, não rejeita-se H_0 , verificamos que as variáveis não apresentam associação.
- Solução utilizando o plugin Rcommander.EZT
 - Rcommander → Statistical analysis → Discrete variables → Enter and analyze two-way table

Exemplo com um banco de dados

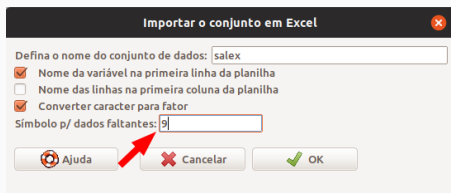
Banco salex.csv

Estes dados são provenientes de um surto de algum tipo de intoxicação alimentar, provavelmente originada pela ingestão de comidas estragadas, no dia 17 de outubro de 1992. Todas as variáveis são codificadas com 1= sim e 2=não (variáveis dicotômicas ou binárias).

- **ILL** : Doente ou não doentes
- **HAM** : Ingeriu ou não ingeriu presunto assado
- **BEEF** : Ingeriu ou não ingeriu bife de carne de boi
- **EGGS** : Ingeriu ou não ingeriu ovos
- **MUSHROOM** : Ingeriu ou não ingeriu cogumelo
- **PEPPER** : Ingeriu ou não ingeriu pimenta
- **PORKPIE** : Ingeriu ou não ingeriu "torta" de carne de porco
- **PASTA** : Ingeriu ou não ingeriu macarrão
- **RICE** : Ingeriu ou não ingeriu arroz
- **LETTUCE** : Ingeriu ou não ingeriu alface
- **TOMATO** : Ingeriu ou não ingeriu salada de tomates
- **COLESLAW** : Ingeriu ou não ingeriu repolho
- **CRISPS** : Ingeriu ou não ingeriu batatas fritas
- **PEACHCAKE** : Ingeriu ou não ingeriu bolo de pêssego
- **CHOCOLATE** : Ingeriu ou não ingeriu bolo de chocolate
- **FRUIT** : Ingeriu ou não ingeriu salada de fruta tropical
- **TRIFLE** : Ingeriu ou não ingeriu "uma espécie de *Waffles*"
- **ALMONDS** : Ingeriu ou não ingeriu amêndoas

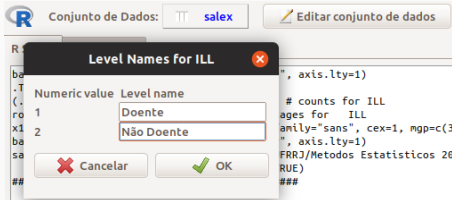
Solução utilizando o plugin Rcommander.EZT

- Importar o arquivo “salex.xls”
 - Rcommander → Arquivo → Importar arquivos de dados → from Excel data set

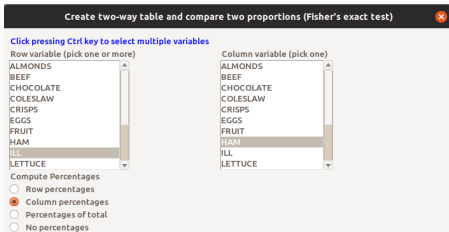


- Colocar os *labels* em todas as variáveis categóricas
- Rcommander → Conjunto de dados ativo → Variables → Convert numeric variables to factors

Solução utilizando o plugin Rcommander.EZT



- Verificando a possível associação via teste de Qui-quadrado
 - Rcommander → Statistical analysis → Discrete variables → Create two-way and compare two proportions



Solução utilizando o plugin Rcommander.EZT

```
> .Table <- NULL
> .Table <- xtabs(~ILL+HAM, data=salex)
> .Table
      HAM
ILL    Sin  Não
Doente  46    5
Não Doente 17    9
> colPercents(.Table) # Column Percentages
      HAM
ILL    Sin  Não
Doente  73  35.7
Não Doente 27  64.3
Total    100 100.0
Count     63  14.0
> .Test <- chisq.test(.Table, correct=TRUE)
> .Test
      Pearson's Chi-squared test with Yates' continuity correction
data:  .Table
X-squared = 5.5561, df = 1, p-value = 0.01842
```

```
> fisher.test(.Table)
      Fisher's Exact Test for Count Data
data:  .Table
p-value = 0.01206
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.22777 20.82921
sample estimates:
odds ratio
 4.75649
> res <- NULL
> res <- fisher.test(.Table)
> Fisher.summary.table <- rbind(Fisher.summary.table, summary.table.twoway(table=.Table, res=res))
> colnames(Fisher.summary.table)[length(Fisher.summary.table)] <- gettext(domain="R-RcmdrPlugin.EZR",
+   colnames(Fisher.summary.table)[length(Fisher.summary.table)])
> Fisher.summary.table
      HAM=Sin HAM=Não Fisher.p.value
ILL=Doente    46      5      0.0121
ILL=Não Doente 17      9
```

- Como o $p\text{-valor} = 0,01842$, rejeita-se H_0 , verificamos que as variáveis apresentam associação, ou seja, o presunto pode ter levado a intoxicação alimentar.
- Verifique se os outros alimentos apresentam alguma associação também.

Medidas de Associação - Variáveis Quantitativas

- A associação entre duas variáveis pode também ser expressa como um único valor, chamado de coeficiente de correlação linear;
- Coeficientes de correlação podem ou não ser baseados na distribuição das variáveis em estudo, dando origem a coeficientes paramétricos e não-paramétricos.

Coeficiente de Correlação de Pearson

- O Coeficiente de Correlação de Pearson é dada por:

$$\rho = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}}} = \frac{\text{cov}(x_i, y_i)}{\sigma_x \sigma_y}$$

- A estatística do teste t em função do Coeficiente de Correlação de Pearson é dada por:

$$t_p = \rho_p \sqrt{\frac{n-2}{1-\rho_p^2}}$$

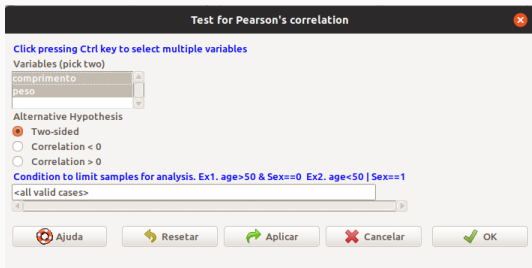
- Hipóteses:
 - $H_0 : \rho_p = 0$ (não existe correlação entre as variáveis)
 - $H_1 : \rho_p \neq 0$ (existe correlação entre as variáveis)

Exemplo:

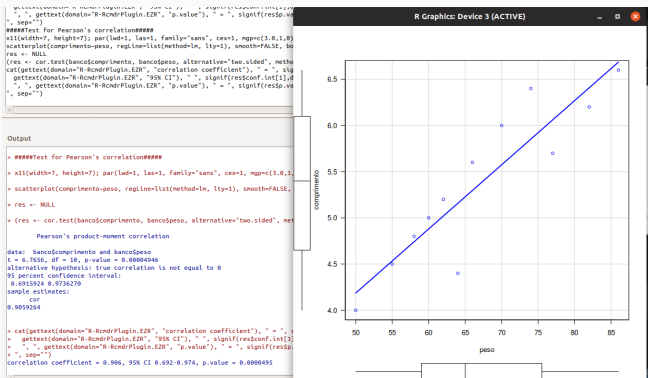
- Suponha que um pesquisador deseje saber se existe associação entre o peso e o comprimento de ratinhos. Foram coletados 12 ratinhos com pesos(g) iguais a: 50; 62; 70; 86; 60; 64; 66; 77; 58; 55; 82 e 74. E os comprimentos(cm) iguais a: 4,0; 5,2; 6,0; 6,6; 5,0; 4,4; 5,6; 5,7; 4,8; 4,5; 6,2 e 6,4.

Solução utilizando o plugin Rcommander.EZT

- Importar o arquivo “pesocompratos.xlsx”
 - Rcommander → Arquivo → Importar arquivos de dados → from Excel data set
- Testando a correlação
 - Rcommander → Statistical analysis → Continuous variables → Test for Pearson's correlation



Solução utilizando o plugin Rcommander.EZT



- Como $r = 0,9059$ e o p – valor $\leq 0,001$, rejeita-se H_0 .
Verificamos que as variáveis apresentam correlação muito alta.

Coeficiente de Correlação de Spearman

- O Coeficiente de Correlação de Spearman, que não supõe que que as variáveis envolvidas tenham uma distribuição em particular, sendo portanto um coeficiente de correlação não-paramétrico;
- Este coeficiente é particularmente útil quando uma (ou ambas) variável(eis) é(são) qualitativas de contagem ou ordinais;
- Este coeficiente é calculado sobre a ordenação (rank ou postos) dos dados obtidos, dentro de cada variável, daí calcula-se o Coeficiente de Correlação de Pearson entre posto-x e posto-y.

Coeficiente de Correlação de Spearman

- O Coeficiente de Correlação de Spearman ou “Rank Correlation” é dada por:

$$\rho_s = 1 - \frac{6 \sum d_i^2}{n^3 - n}$$

- Sendo $d_i = n$ de ordem de x_i de y_i
- A estatística do teste z em função do Coeficiente de Correlação de Spearman é dada por:

$$z_s = \frac{\sqrt{n-3}}{2} \ln\left(\frac{1+\rho_s}{1-\rho_s}\right) \sim N(0, 1)$$

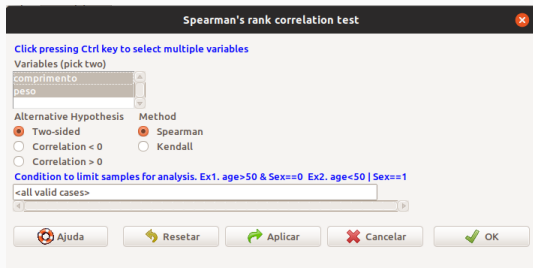
- Hipóteses:
 - $H_0 : \rho_s = 0$ (não existe correlação entre as variáveis)
 - $H_1 : \rho_s \neq 0$ (existe correlação entre as variáveis)

Exemplo:

- Voltando ao exemplo dos ratinhos, e supondo a não normalidade em pelo menos uma das variáveis, o pesquisador deseja saber se existe associação entre o peso e o comprimento de ratinhos utilizando o coeficiente de correlação de Spearman.

Solução utilizando o plugin Rcommander.EZT

- Importar o arquivo “pesocompratos.xlsx”
 - Rcommander → Arquivo → Importar arquivos de dados → from Excel data set
- Testando a correlação
 - Rcommander → Statistical analysis → Testes Não-Paramétricos → Sperman's rank correlation test



Solução utilizando o plugin Rcommander.EZT

```

", , gettext(domain="R-RcmdrPlugin.EZT", "p.value"), " = ", signif(res$p.value,
", sep=")
#####Spearman's rank correlation test#####
x11(width=7, height=7); par(lwd=1, las=1, family="sans", cex=1, mgp=c(3.0,1,0))
scatterplot(comprimento-peso, regLine=list(method=lm, lty=1), smooth=FALSE, boxplot
res <- NULL
(res <- cor.test(banco$comprimento, banco$peso, alternative="two.sided", method="
cat(gettext(domain="R-RcmdrPlugin.EZT", "Spearman's rank correlation coefficient"
gettext(domain="R-RcmdrPlugin.EZT", "p.value"), " = ", signif(res$p.value, digit
")

```

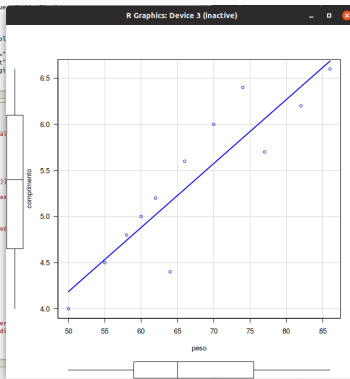
Output

```

", , gettext(domain="R-RcmdrPlugin.EZT", "p.value"), " = ", signif(res$p.va
", sep=")
correlation coefficient = 0.986, 95% CI 0.692-0.974, p.value = 0.0000495
> #####Spearman's rank correlation test#####
> x11(width=7, height=7); par(lwd=1, las=1, family="sans", cex=1, mgp=c(3.0,1,0))
> scatterplot(comprimento-peso, regLine=list(method=lm, lty=1), smooth=FALSE, box
> res <- NULL
> (res <- cor.test(banco$comprimento, banco$peso, alternative="two.sided", method
Spearman's rank correlation rho
data: banco$comprimento and banco$peso
S = 30, p-value = 0.000005042
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
0.8951049
> cat(gettext(domain="R-RcmdrPlugin.EZT", "Spearman's rank correlation coefficient
", gettext(domain="R-RcmdrPlugin.EZT", "p.value"), " = ", signif(res$p.value, digit
")
Spearman's rank correlation coefficient 0.895 p.value = 0.000005042

```

Mensagens



- Como $r_s = 0,8951$ e o p – valor $\leq 0,001$, rejeita-se H_0 .
Verificamos que as variáveis apresentam correlação muito alta.

O coeficiente de concordância de Kappa

- É utilizado para descrever a concordância entre dois ou mais avaliadores ou juizes quando realizam uma avaliação nominal ou ordinal de uma mesma amostra.
- Existem diversas doenças cujos diagnósticos dependem da avaliação do médico dos resultados de exames de imagem. Como exemplo ilustrativo da utilização do coeficiente de Kappa, apresentaremos uma situação onde dois médicos avaliam de forma independente o resultado de 180 exames de diagnóstico por imagem e o classificam como “normal”, “alterado” e “inconclusivo”.

O coeficiente de concordância de Kappa

$$\hat{K} = \frac{\hat{p}_{observada} - \hat{p}_{esperada}}{1 - \hat{p}_{esperada}}$$

- Interpretando o Coeficiente de Concordância de Kappa

Valor de Kappa	Interpretação
Menor que zero	insignificante (poor)
Entre 0 e 0,2	fraca (slight)
Entre 0,21 e 0,4	razoável (fair)
Entre 0,41 e 0,6	moderada (moderate)
Entre 0,61 e 0,8	forte (substantial)
Entre 0,81 e 1	quase perfeita (almost perfect)

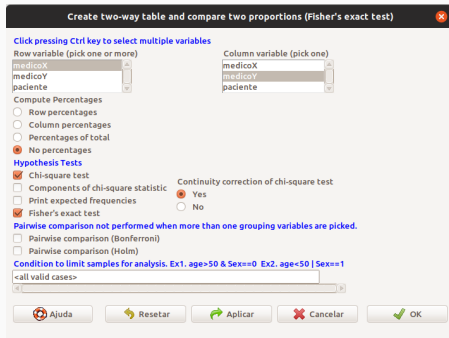
Exemplo de aplicação do Coeficiente de Concordância de Kappa

Em nosso exemplo, poderíamos formar uma base de dados com a seguinte estrutura:

Paciente	Médico X	Médico Y
1	Positivo	Positivo
2	Positivo	Negativo
3	Negativo	Positivo
4	Positivo	Positivo
...

Solução utilizando o plugin Rcommander.EZT

- Importar o arquivo “concordancia.xlsx”
 - Rcommander → Arquivo → Importar arquivos de dados → from Excel data set
- Primeiro vamos fazer a tabela de contingência
 - Rcommander → Statistical analysis → Discrete variables → Create two-way and compare two proportions



Solução utilizando o plugin Rcommander.EZR

```
> Fisher.summary.table <- rbind(Fisher.summary.table, summary.table.tway(table=Table, res=res))  
> colnames(Fisher.summary.table)[length(Fisher.summary.table)] <- gettext(domain="R-RcndrPlugin.EZR",  
+ colnames(Fisher.summary.table)[length(Fisher.summary.table)])  
  
> Fisher.summary.table  
      medicoV=1 medicoV=2 Fisher.p.value  
medicoX=1      7        2      0.0567  
medicoX=2      2        6
```

- Testando a concordância
 - Rcommander → Statistical analysis → Accuracy of diagnostic test → Kappa statistics for agreement of two tests

Output

```
> fisher.test(.Table)  
  
Fisher's Exact Test for Count Data  
  
Kappa statistics for agreement of two tests
```

Number	Test2 (+)	(-)
Test1 (+)	7	2
Test1 (-)	2	6

Ajuda Cancelar OK

```
> res <- fisher.test(.Table)  
  
> Fisher.summary.table <- rbind(Fisher.summary.table, summary.table.tway(table=Table, res=res))  
> colnames(Fisher.summary.table)[length(Fisher.summary.table)] <- gettext(domain="R-RcndrPlugin.EZR",  
+ colnames(Fisher.summary.table)[length(Fisher.summary.table)])  
  
> Fisher.summary.table  
      medicoV=1 medicoV=2 Fisher.p.value  
medicoX=1      7        2      0.0567  
medicoX=2      2        6
```

Solução utilizando o plugin Rcommander.EZT

```
> #####Kappa statistics for agreement of two tests#####  
> .Table <- NULL  
> .Table <- matrix(c(7, 2, 2, 6), 2, 2, byrow=TRUE)  
> colnames(.Table) <- gettext(domain="R-RcmdrPlugin.EZR",c("Test2 (+)", "Test2 (-)"))  
> rownames(.Table) <- gettext(domain="R-RcmdrPlugin.EZR",c("Test1 (+)", "Test1 (-)"))  
  
> .Table  
      Test2 (+) Test2 (-)  
Test1 (+)      7      2  
Test1 (-)      2      6  
  
> res <- NULL  
> res <- epi.kappa(.Table, conf.level = 0.95)  
> colnames(res$kappa) <- gettext(domain="R-RcmdrPlugin.EZR", colnames(res$kappa))  
  
> res[1]  
$kappa  
      est      lower      upper  
1 0.5277778 0.1230978 0.9324578
```

- Como $kappa = 0,5277$, podemos dizer que a concordância entre os dois médicos foi moderada.