1. Dados pareados

```
## Dados sobre escolha de marcas de café em duas compras
## Tabela 9.5, p. 236 em Agresti (1996), Introduction to Categorical
##   Data Analysis

tab95 <- matrix(c(93, 17, 44, 7, 10,
                  9, 46, 11, 0, 9,
                  17, 11, 155, 9, 12,
                  6, 4, 9, 15, 2,
                  10, 4, 12, 2, 27), ncol = 5, byrow = TRUE)
rownames(tab95) <- colnames(tab95) <- c("High Point", "Taster's",
         "Sanka", "Nescafe", "Brim")

addmargins(tab95)

           High Point Taster's Sanka Nescafe Brim Sum
High Point         93       17    44       7   10 171
Taster's            9       46    11       0    9  75
Sanka              17       11   155       9   12 204
Nescafe             6        4     9      15    2  36
Brim               10        4    12       2   27  55
Sum               135       82   231      33   60 541

n <- sum(tab95)
cat("\n n =", n, "\n")

      n = 541

## Proporções amostrais (%)
print(100 * tab95 / n, digits = 2)

           High Point Taster's Sanka Nescafe Brim
High Point       17.2     3.14   8.1    1.29 1.85
Taster's          1.7     8.50   2.0    0.00 1.66
Sanka             3.1     2.03  28.7    1.66 2.22
Nescafe           1.1     0.74   1.7    2.77 0.37
Brim              1.8     0.74   2.2    0.37 4.99
```

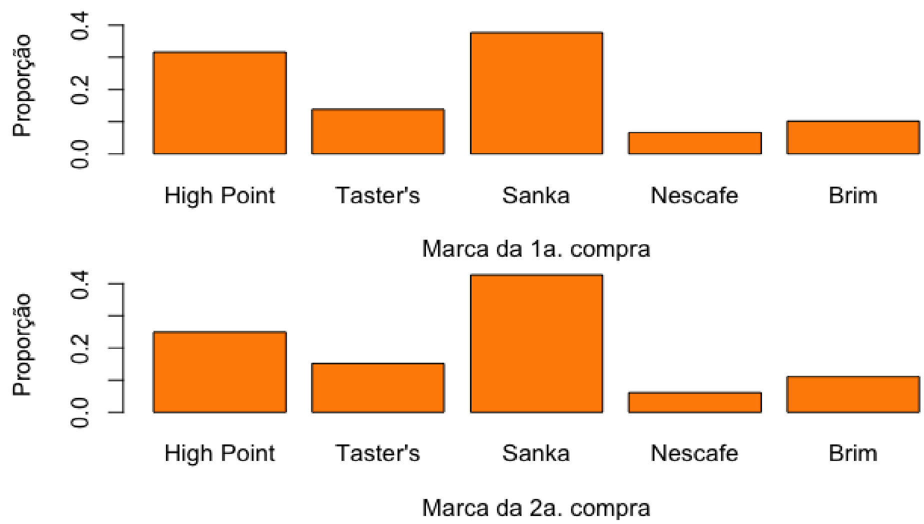
Nota 1. Comente sobre a simetria da distribuição conjunta.

```
## Distribuições marginais
compra1 <- margin.table(tab95, margin = 1) / n
compra2 <- margin.table(tab95, margin = 2) / n

round(rbind(compra1, compra2), digits = 2)

        High Point Taster's Sanka Nescafe Brim
compra1       0.32     0.14  0.38    0.07 0.10
compra2       0.25     0.15  0.43    0.06 0.11
```

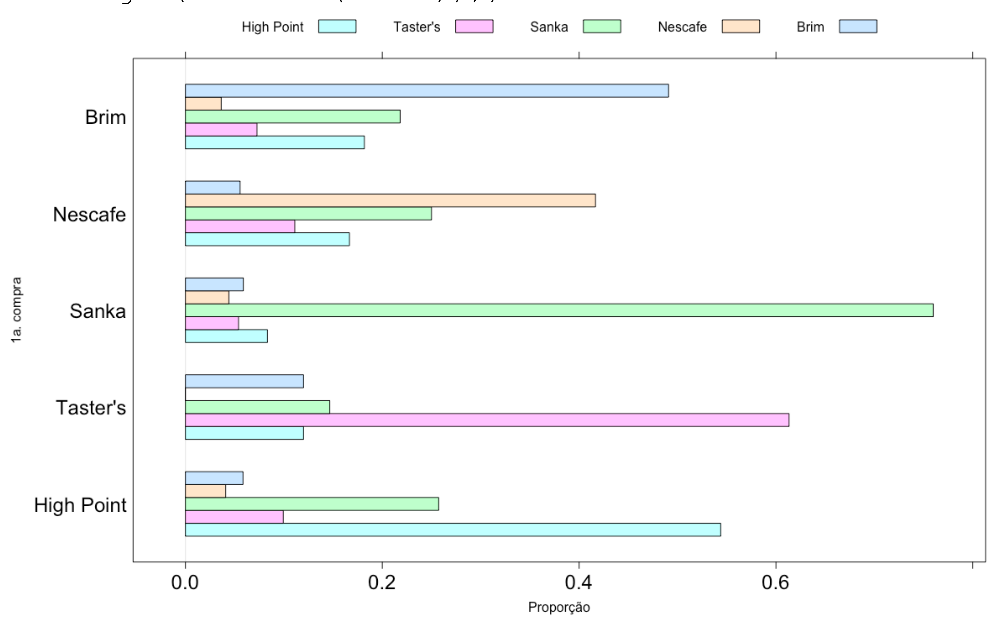
```
r12 <- c(0, max(compra1, compra2)) # Intervalo (eixo vertical)
par(mfrow = c(2, 1))
barplot(compra1, xlab = "Marca da 1a. compra", ylab = "Proporção",
        ylim = r12)
barplot(compra2, xlab = "Marca da 2a. compra", ylab = "Proporção",
        ylim = r12)
```



<u>Nota 2.</u> Comente sobre a homogeneidade marginal das distribuições.

```
## Distribuições condicionais (na 1a. compra)
library(lattice)

tab95c <- prop.table(tab95, margin = 1)
barchart(tab95c, xlab = "Proporção", ylab = "1a. compra", stack = FALSE,
   scale = list(cex = 1.5), auto.key = list(space = "top",
   columns = length(colnames(tab95))))
```



Nota 3. Interprete o gráfico acima e comente sobre a independência entre as variáveis.

1.1 Modelos

Em seguida os dados são organizados de uma forma conveniente para o ajuste de modelos log-lineares. A função `as.vector` empilha as colunas de uma matriz.

```
## Dados
compra1 <- factor(colnames(tab95), levels = colnames(tab95))
compra2 <- compra1
dados <- expand.grid(compra1 = compra1, compra2 = compra2)
dados$freq <- as.vector(tab95)
dados
```

```
      compra1    compra2 freq
1  High Point High Point   93
2    Taster's High Point    9
3       Sanka High Point   17
4     Nescafe High Point    6
5        Brim High Point   10
6  High Point   Taster's   17
7    Taster's   Taster's   46
8       Sanka   Taster's   11
9     Nescafe   Taster's    4
10       Brim   Taster's    4
11 High Point      Sanka   44
12   Taster's      Sanka   11
13      Sanka      Sanka  155
14    Nescafe      Sanka    9
15       Brim      Sanka   12
16 High Point    Nescafe    7
17   Taster's    Nescafe    0
18      Sanka    Nescafe    9
19    Nescafe    Nescafe   15
20       Brim    Nescafe    2
21 High Point       Brim   10
22   Taster's       Brim    9
23      Sanka       Brim   12
24    Nescafe       Brim    2
25       Brim       Brim   27
```

```
## Modelo de independência
mind <- glm(freq ~ compra1 + compra2, family = poisson, data = dados)

# G2 e X2
X2ind <- sum(resid(mind, type = "pearson")^2)
cat("\n Modelo de independência (g.l. = ", mind$df.residual,")")
cat("\n G2 = ", mind$deviance, "(p =", pchisq(mind$deviance,
      mind$df.residual, lower.tail = FALSE), ")")
cat("\n X2 = ", X2ind, "(p =", pchisq(X2ind, mind$df.residual,
      lower.tail = FALSE), ")")
```

```
Modelo de independência (g.l. =  16 )
G2 =  346.381 (p = 5.879941e-64 )
X2 =  463.3044 (p = 1.817251e-88 )
```

O ajuste do modelo de simetria requer uma variável auxiliar formada por pares de linhas e colunas (i, j) e tais que os pares (i, j) e (j, i) tenham o mesmo valor da variável sempre que $i \neq j$, para i e $j = 1,...,I$, significando que os coeficientes são iguais. Se todos os I^2 (= 25) valores da variável auxiliar forem diferentes, obtemos o modelo saturado.

As funções `pmax` (*parallel maxima*) e `pmin` (*parallel minima*) são usadas para criar a variável auxiliar. Aplicada aos vetores $(a_1, a_2, \ldots, a_N)$ e $(b_1, b_2, \ldots, b_N)$, a função `pmax` retorna o vetor com elementos $\max(a_1, b_1), \max(a_2, b_2), \ldots, \max(a_N, b_N)$.

```
## Modelo de simetria
# Variável auxiliar com um valor para cada diferente probabilidade
dados$aux <- factor(paste(pmax(as.numeric(dados$compra1),
  as.numeric(dados$compra2)), pmin(as.numeric(dados$compra1),
  as.numeric(dados$compra2)), sep = ","))

dados
      compra1    compra2 freq aux
1  High Point High Point   93 1,1
2    Taster's High Point    9 2,1
3       Sanka High Point   17 3,1
4     Nescafe High Point    6 4,1
5        Brim High Point   10 5,1
6  High Point   Taster's   17 2,1
7    Taster's   Taster's   46 2,2
8       Sanka   Taster's   11 3,2
9     Nescafe   Taster's    4 4,2
10       Brim   Taster's    4 5,2
11 High Point      Sanka   44 3,1
12   Taster's      Sanka   11 3,2
13      Sanka      Sanka  155 3,3
14    Nescafe      Sanka    9 4,3
15       Brim      Sanka   12 5,3
16 High Point    Nescafe    7 4,1
17   Taster's    Nescafe    0 4,2
18      Sanka    Nescafe    9 4,3
19    Nescafe    Nescafe   15 4,4
20       Brim    Nescafe    2 5,4
21 High Point       Brim   10 5,1
22   Taster's       Brim    9 5,2
23      Sanka       Brim   12 5,3
24    Nescafe       Brim    2 5,4
25       Brim       Brim   27 5,5
```

Nota 4. Crie a variável auxiliar (`aux`) sem utilizar as funções `pmax` e `pmin`.

O número de pares é I^2, que correspondem a $I + I(I\text{-}1)/2 = 15$ pares não redundantes (ou únicos).

```
length(unique(dados$aux))

     15

msim <- glm(freq ~ aux, family = poisson, data = dados)

# G2 e X2
X2s <- sum(resid(msim, type = "pearson")^2)
cat("\n Modelo de simetria (g.l. = ", msim$df.residual,")")
cat("\n G2 = ", msim$deviance, "(p =", round(pchisq(msim$deviance,
   msim$df.residual, lower.tail = FALSE), 4), ")")
cat("\n X2 = ", X2s, "(p =", round(pchisq(X2s, msim$df.residual,
   lower.tail = FALSE), 4), ")")

      Modelo de simetria (g.l. =  10 )
      G2 =  22.47293 (p = 0.0129 )
      X2 =  20.41236 (p = 0.0256 )
```

Os resultados acima indicam que o modelo de simetria não faz um bom ajuste aos dados ($p <$ 0,05). O modelo de quase simetria pode ser obtido a partir do modelo simetria pela adição dos efeitos individuais das compras ("quebram" a simetria).

```
## Modelo de quase simetria
mqsim <- update(msim, . ~ . + compra1 + compra2)

# G2 e X2
X2qs <- sum(resid(mqsim, type = "pearson")^2)
cat("\n Modelo de quase simetria (g.l. = ", mqsim$df.residual,")")
cat("\n G2 = ", mqsim$deviance, "(p =", round(pchisq(mqsim$deviance,
   mqsim$df.residual, lower.tail = FALSE), 4), ")")
cat("\n X2 = ", X2qs, "(p =", round(pchisq(X2qs, mqsim$df.residual,
   lower.tail = FALSE), 4), ")")

Modelo de quase simetria (g.l. =  6 )
G2 =  9.974047 (p = 0.1257 )
X2 =  8.530328 (p = 0.2018 )
```

Os resultados acima indicam que não se rejeita o ajuste do modelo de quase simetria ($p >$ 0,05).

```
# Frequências esperadas estimadas
festqsim <- matrix(mqsim$fitted.values, ncol = ncol(tab95), byrow = FALSE,
    dimnames = list(rownames(tab95), colnames(tab95)))
round(festqsim, 1)
```

```
           High Point Taster's Sanka Nescafe Brim
High Point       93.0     16.8  40.9     7.4 12.9
Taster's          9.2     46.0  11.6     1.7  6.5
Sanka            20.1     10.4 155.0     7.2 11.3
Nescafe           5.6      2.3  10.8    15.0  2.3
Brim              7.1      6.5  12.7     1.7 27.0
```

Com as estimativas acima, obtemos a estimativa de $(m_{22} / m_{23}) / (m_{12} / m_{13})$:

```
festqsim[2, 2] * festqsim[1, 3]  / (festqsim[1, 2] * festqsim[2, 3])

     9.65773
```

A estimativa acima pode ser escrita como

$$\frac{P(Y = \textit{Taster's} \mid X = \textit{Taster's})}{P(Y = \textit{Sanka} \mid X = \textit{Taster's})} = 9{,}7 \times \frac{P(Y = \textit{Taster's} \mid X = \textit{High Point})}{P(Y = \textit{Sanka} \mid X = \textit{High Point})}$$

São duas opções na primeira compra: *Taster's* ou *High Point*. A chance de escolher entre *Taster's* e *Sanka* na segunda compra é 9,7 vezes maior na primeira opção em relação à segunda opção.

Comparando as marcas *Taster's* e *High Point* como primeira compra, a chance entre *Taster's* e *Sanka* na segunda compra é 9,7 vezes maior quando a primeira compra é *Taster's.*

De acordo com o modelo de quase simetria, a estimativa de $(m_{31} / m_{32}) / (m_{21} / m_{22})$ é a mesma dada acima. De fato,

```
festqsim[3, 1] * festqsim[2, 2]  / (festqsim[3, 2] * festqsim[2, 1])

     9.65773
```

A estimativa acima pode ser escrita como

$$\frac{P(Y = \textit{High Point} \mid X = \textit{Sanka})}{P(Y = \textit{Taster's} \mid X = \textit{Sanka})} = 9{,}7 \times \frac{P(Y = \textit{High Point} \mid X = \textit{Taster's})}{P(Y = \textit{Taster's} \mid X = \textit{Taster's})}$$

São duas opções na primeira compra: *Sanka* ou *Taster's*. A chance de escolher entre *High Point* e *Taster's* na segunda compra é 9,7 vezes maior na primeira opção em opção à segunda opção.

Comparando as marcas *Sanka* e *Taster's* como primeira compra, a chance entre *High Point* e *Taster's* na segunda compra é 9,7 vezes maior quando a primeira compra é *Sanka.*

Para ajustar o modelo de quase independência criamos I (= 5) variáveis auxiliares assumindo em cada par (i, j) valor 0, se $i \neq j$, e valor 1, se $i = j$.

```
## Modelo de quase independência
# Variáveis auxiliares
dados$D1 <- as.numeric(dados$aux == "1,1")
dados$D2 <- as.numeric(dados$aux == "2,2")
dados$D3 <- as.numeric(dados$aux == "3,3")
dados$D4 <- as.numeric(dados$aux == "4,4")
dados$D5 <- as.numeric(dados$aux == "5,5")

mqind <- update(mind, . ~ . + D1 + D2 + D3 + D4 + D5)

# G2 e X2
X2qind <- sum(resid(mqind, type = "pearson")^2)
cat("\n Modelo de quase independência (g.l. = ", mqind$df.residual,")")
cat("\n G2 = ", mqind$deviance, "(p =", round(pchisq(mqind$deviance,
   mqind$df.residual, lower.tail = FALSE), 4), ")")
cat("\n X2 = ", X2qind, "(p =", round(pchisq(X2qind, mqind$df.residual,
  lower.tail = FALSE), 4), ")")
```

```
      Modelo de quase independência (g.l. =  11 )
      G2 =  13.78563 (p = 0.2451 )
      X2 =  12.24792 (p = 0.3453 )
```

Os resultados acima indicam que não se rejeita o ajuste do modelo de quase independência ($p > 0,05$).

Nota 5. Compare as frequências observadas e as frequências esperadas estimadas usando o modelo de quase independência.

Nota 6. Explique o significado do modelo de quase independência para as escolhas das marcas de café.

Como o modelo de quase independência está encaixado no modelo de quase simetria, os dois modelos podem ser comparados utilizando a estatística de teste baseada na razão de verossimilhanças (G^2), que é especificada com `test = "LRT"`.

```
# Quase independência x quase simetria
anova(mqind, mqsim, test = "LRT")
```

```
Model 1: freq ~ compra1 + compra2 + D1 + D2 + D3 + D4 + D5
Model 2: freq ~ aux + compra1 + compra2
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        11     13.786
2         6      9.974  5   3.8116   0.5769
```

Nota 7. Com base no teste acima, se você pretende apresentar resultados de apenas um modelo, qual modelo você escolheria?

```
## Teste de homogeneidade marginal
g2hm <- msim$deviance - mqsim$deviance
gl <- msim$df.residual - mqsim$df.residual
cat("\n Teste de homogeneidade marginal \n G2 = ", g2hm,
   "(p =", round(pchisq(g2hm, gl, lower.tail = FALSE), 4),
   ", g.l. = ", gl,")")
```

```
 Teste de homogeneidade marginal
 G2 =  12.49889 (p = 0.014 , g.l. =  4 )
```

1.2 *kappa* de Cohen

```
# kappa de Cohen
library(vcd)
(k95 <- Kappa(tab95))
confint(k95)
```

Nos resultados abaixo, `ASE` indica o erro padrão assintótico (*asymptotic standard error*). Os limites do intervalo de confiança assintótico de 95% são indicados por `lwr` (*lower*: inferior) e `upr` (*upper*: superior).

```
            value     ASE     z  Pr(>|z|)
Unweighted 0.4765 0.02805 16.99 1.060e-64
Weighted   0.4527 0.03297 13.73 6.473e-43
```

```
k95$Weights
```

```
     [,1] [,2] [,3] [,4] [,5]
[1,] 1.00 0.75 0.50 0.25 0.00
[2,] 0.75 1.00 0.75 0.50 0.25
[3,] 0.50 0.75 1.00 0.75 0.50
[4,] 0.25 0.50 0.75 1.00 0.75
[5,] 0.00 0.25 0.50 0.75 1.00
```

```
confint(k95)
```

```
Kappa              lwr       upr
  Unweighted 0.4214736 0.5314329
  Weighted   0.3880910 0.5173130
```

<u>Nota 8.</u> Para estes dados, a medida *kappa* ponderada (`Weighted`) é aplicável?

2. Inferência em uma tabela 2 × 2 com dados pareados

```
## Teste de McNemar
# Dados
# Tabela 10.1, p. 350 em Agresti (1990), Categorical Data Analysis
tab101 <- matrix(c(794, 150, 86, 570), byrow = TRUE, ncol = 2)
```

```
rownames(tab101) <- c("1. Aprovação", "1. Desaprovação")
colnames(tab101) <- c("2. Aprovação", "2. Desaprovação")
addmargins(tab101)

                2. Aprovação 2. Desaprovação  Sum
1. Aprovação             794             150  944
1. Desaprovação           86             570  656
Sum                      880             720 1600

n <- sum(tab101)
cat("\n n =", n, "\n")

   n = 1600

# Proporções amostrais e teste
prop101 <- prop.table(tab101)
round(prop101, 3)

                2. Aprovação 2. Desaprovação
1. Aprovação           0.496           0.094
1. Desaprovação        0.054           0.356

mcnemar.test(tab101, correct = FALSE)

     McNemar's chi-squared = 17.3559, df = 1, p-value = 3.099e-05

mcnemar.test(tab101, correct = TRUE)

     McNemar's chi-squared = 16.818, df = 1, p-value = 4.115e-05

# Teste exato (distribuição binomial)
# H1: pi1+> pi+1 (aprovação 1o. > aprovação 2o.)
ns <- tab101[1, 2] + tab101[2, 1] # n*
```

No cálculo do valor-*p*, como a distribuição é binomial (discreta), temos que $P(n_{12} \geq n_{12,obs}) = 1 - P(n_{12} < n_{12,obs}) = 1 - P(n_{12} \leq n_{12,obs} - 1)$.

```
valorp <- pbinom(tab101[1, 2] - 1, ns, 0.5, lower.tail = FALSE)
cat("\n Teste exato para H1: pi1+ > pi+1 \n n12 =", tab101[1, 2],
    ", n* =", ns, "(p =", valorp, ") \n")

     Teste exato para H1: pi1+ > pi+1
     n12 = 150 , n* = 236 (p = 1.857968e-05 )

# Diferença de aprovação (2o. - 1o.)
(estd <- prop101[2, 1] - prop101[1, 2])

     -0.04
```

```
# Erro padrão da estimativa
p1m <- sum(prop101[1, ])    # p1+
pm1 <- sum(prop101[, 1])    # p+1
epest <- sqrt((p1m * (1 - p1m) + pm1 * (1 - pm1) - 2 *
   (prod(diag(prop101)) - prop101[1, 2] * prop101[2, 1])) / n)
cat("\n e.p.(d) =", epest, "\n")

     e.p.(d) = 0.009549215

# Erro padrão da estimativa supondo independência
sqrt((p1m * (1 - p1m) + pm1 * (1 - pm1)) / n)

     0.01748928
```

Houve ganho de precisão (0,01748928 / 0,009549215 = 1,83).

```
# IC para a diferença de aprovação (2o. - 1o.)
conf <- 0.95
icdif <- estd + c(-1, 1) * qnorm((1 + conf) / 2) * epest
cat("\n IC de", conf * 100, "% para a diferença (2o. - 1o.): \n", icdif,
     "\n")

     IC de 95 % para a diferença (2o. - 1o.):
     -0.05871612 -0.02128388
```

Nota 9. Interprete o resultado acima. Como você escreveria em um relatório?