CE073 - Análise de Dados Categóricos

Trabalho No.1

Luiz Henrique Barretta Francisco - 20213026

agosto/2023

1-a) Descreva as suposições que seriam necessárias para usar uma distribuição binomial para a amostra.

Para que o uso da distribuição binomial seja justificado, as seguintes suposições sobre o cenário real devem ser atendidas:

1. Independência: Cada amostra deve ser independente das outras, ou seja, o resultado do teste de uma gestante não deve influenciar no resultado de outra. Essa suposição pode ser justificada pelo fato de essa doença ser sexualmente transmissível, resultando em baixa transmissibilidade. Essa suposição não seria possível para doenças transmissíveis pelo ar, como o Covid-19, por exemplo.
2. Dois resultados possíveis: Cada gestante deve ter apenas dois resultados possíveis, tratados como sucesso (positivo para clamídia) ou fracasso (negativo para clamídia).
3. Probabilidade constante de sucesso: A probabilidade de sucesso (positivo para clamídia) deve ser constante em todas as tentativas. Neste caso, estamos supondo que a prevalência de clamídia não varia ao longo das 750 gestantes. Essa suposição pode ser justificada pelo mesmo argumento, da baixa transmissibilidade, utilizada no item 1.
4. Espaço finito de valores possíveis para a variável resposta: A quantidade de sucessos (positivo para clamídia) é um intervalo fechado de valores, de 0 até 750 (número total de gestantes).

1-b) Presumindo que as suposições sejam satisfeitas, encontre um intervalo de confiança para estimar a prevalência de clamídia. Use o procedimento de intervalo de confiança mais apropriado para este problema e interprete os resultados.

Para estimar o intervalo de confiança de prevalência de clamídia com base nos dados fornecidos, utiliza-se o intervalo de confiança para a proporção populacional. O intervalo de confiança de Agresti-Coull é calculado utilizando a distribuição normal, que para um número grande amostras (750 gestantes) serve bem como aproximação à distribuição binomial, dado por:

$$IC = \hat{\pi} \pm Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n + Z^2_{1-\frac{\alpha}{2}}}}$$

com

$$\hat{\pi} = \frac{w + Z^2_{1-\frac{\alpha}{2}}/2}{n + Z^2_{1-\frac{\alpha}{2}}}$$

Substituindo pelos valores:

$$w = 48; n = 750; \alpha = 5\%$$

```
n <- 750
w <- 48
alpha <- 0.05
library( package = binom )
binom.confint(x = w, n = n, conf.level = 1-alpha, methods = "agresti-coull")
```

```
##          method  x   n  mean      lower     upper
## 1 agresti-coull 48 750 0.064 0.04847048 0.0839731
```

Obtemos o intervalo de confiança $(0.048; 0.084)$. Traduzindo para um contexto mais acessível, podemos dizer, com 95% de confiança, que a prevalência de clamídia na população estudada encontra-se entre 4,8% e 8,4%.

2-a) Calcule os intervalos de Clopper-Pearson para S_e e S_p entre os homens sintomáticos que forneceram amostras de swab para teste de clamídia. Interprete os intervalos.

```
aptima <- read.table( file="http://leg.ufpr.br/~lucambio/CE073/20222S/Aptima_combo.csv",
                      header=TRUE , sep=",")
c.table <- array (data = c(aptima [1,6], aptima [1,7],
                           aptima [1,9], aptima [1,8]),
                  dim = c(2,2),
                  dimnames = list(True = c("+", "-"),
                                  Assay = c("+", "-")))
w1 <- c.table [1,1]
n1 <- sum(c.table [1,])
Se.hat <- w1 / n1
w2 <- c.table [2,2]
n2 <- sum(c.table [2,])
Sp.hat <- w2 / n2

alpha <- 0.05
print("Intervalo de Confiança de Clopper-Pearson para Sensibilidade:")
round(qbeta(p=c(alpha/2, 1 - alpha/2), shape1=c(w1, w1+1), shape2=c(n1-w1+1, n1-w1)),4)
binom.confint(x = w1, n = n1, conf.level = 1-alpha, methods = "exact")

print("Intervalo de Confiança de Clopper-Pearson para Especificidade:")
round(qbeta(p=c(alpha/2, 1 - alpha/2), shape1=c(w2, w2+1), shape2=c(n2-w2+1, n2-w2)),4)
binom.confint(x = w2, n = n2, conf.level = 1-alpha, methods = "exact")
```

```
## [1] "Intervalo de Confiança de Clopper-Pearson para Sensibilidade:"
```

```
## [1] 0.9282 0.9856
```

```
##   method   x   n     mean     lower     upper
## 1  exact 190 197 0.964467 0.9281616 0.9855967
```

```
## [1] "Intervalo de Confiança de Clopper-Pearson para Especificidade:"
```

```
## [1] 0.9489 0.9824
```

```
##   method   x   n      mean     lower     upper
## 1  exact 464 479 0.9686848 0.9488757 0.9823693
```

Sensibilidade (S_e): A sensibilidade de um teste é a capacidade do teste de identificar corretamente os verdadeiros positivos, ou seja, as pessoas que realmente têm a condição (no caso, clamídia) e são diagnosticadas corretamente pelo teste como positivas. O intervalo de Clopper-Pearson para S_e está entre 0.9282 e 0.9856. Isso significa que, com um nível de confiança de 95%, a sensibilidade real do teste na população de homens sintomáticos que forneceram amostras de swab provavelmente está dentro dessa faixa.

Especificidade (S_p): A especificidade de um teste é a capacidade do teste de identificar corretamente os verdadeiros negativos, ou seja, as pessoas que realmente não têm a condição e são diagnosticadas corretamente pelo teste como negativas. O intervalo de Clopper-Pearson para S_p está entre 0.9489 e 0.9824. Isso significa

que, com um nível de confiança de 95%, a especificidade real do teste na população de homens sintomáticos que forneceram amostras de swab provavelmente está dentro dessa faixa.

2-b) Calcule os intervalos de Clopper-Pearson para as outras combinações de doença, gênero, espécime e sintoma e exiba-os de maneira organizada, por exemplo, eles podem ser colocados em um quadro de dados com a rotulagem apropriada.

```
aptima$Se_hat <- aptima$True_positive /
        (aptima$True_positive + aptima$False_negative)

aptima$Sp_hat <- aptima$True_negative/
        (aptima$False_positive + aptima$True_negative)

sensibilidade <- binom.confint(x = aptima$True_positive, n = aptima$True_positive +
                 aptima$False_negative, conf.level = 1-alpha, methods = "exact")
especificidade <- binom.confint(x = aptima$True_negative, n = aptima$False_positive +
                  aptima$True_negative, conf.level = 1-alpha, methods = "exact")

sensibilidade_rounded <- round(sensibilidade[2:6], 4)
especificidade_rounded <- round(especificidade[2:6], 4)

print("Tabela de Sensibilidade:")
df1 <- data.frame(aptima[0:4],sensibilidade_rounded)
print(df1)

print("Tabela de Especificidade:")
df2 <- data.frame(aptima[0:4],especificidade_rounded)
print(df2)
```

```
## [1] "Tabela de Sensibilidade:"
```

```
##      Disease Gender Specimen Symptoms_Status   x   n   mean  lower  upper
## 1  Chlamydia   Male     Swab     Symptomatic 190 197 0.9645 0.9282 0.9856
## 2  Chlamydia   Male     Swab    Asymptomatic  70  74 0.9459 0.8673 0.9851
## 3  Chlamydia   Male    Urine     Symptomatic 199 202 0.9851 0.9572 0.9969
## 4  Chlamydia   Male    Urine    Asymptomatic  77  80 0.9625 0.8943 0.9922
## 5  Chlamydia Female     Swab     Symptomatic 133 144 0.9236 0.8674 0.9613
## 6  Chlamydia Female     Swab    Asymptomatic  61  62 0.9839 0.9134 0.9996
## 7  Chlamydia Female    Urine     Symptomatic 136 145 0.9379 0.8854 0.9712
## 8  Chlamydia Female    Urine    Asymptomatic  60  62 0.9677 0.8883 0.9961
## 9  Gonorrhea   Male     Swab     Symptomatic 304 307 0.9902 0.9717 0.9980
## 10 Gonorrhea   Male     Swab    Asymptomatic  15  15 1.0000 0.7820 1.0000
## 11 Gonorrhea   Male    Urine     Symptomatic 311 316 0.9842 0.9635 0.9948
## 12 Gonorrhea   Male    Urine    Asymptomatic  13  13 1.0000 0.7529 1.0000
## 13 Gonorrhea Female     Swab     Symptomatic  94  94 1.0000 0.9615 1.0000
## 14 Gonorrhea Female     Swab    Asymptomatic  31  32 0.9688 0.8378 0.9992
## 15 Gonorrhea Female    Urine     Symptomatic  87  94 0.9255 0.8526 0.9695
## 16 Gonorrhea Female    Urine    Asymptomatic  28  32 0.8750 0.7101 0.9649
```

```
## [1] "Tabela de Especificidade:"
```

```
##      Disease Gender Specimen Symptoms_Status   x   n   mean  lower  upper
## 1  Chlamydia   Male     Swab     Symptomatic 464 479 0.9687 0.9489 0.9824
```

```
## 2  Chlamydia   Male     Swab    Asymptomatic 309 314 0.9841 0.9632 0.9948
## 3  Chlamydia   Male    Urine     Symptomatic 484 492 0.9837 0.9682 0.9930
## 4  Chlamydia   Male    Urine    Asymptomatic 316 320 0.9875 0.9683 0.9966
## 5  Chlamydia Female     Swab     Symptomatic 653 675 0.9674 0.9511 0.9795
## 6  Chlamydia Female     Swab    Asymptomatic 501 507 0.9882 0.9744 0.9956
## 7  Chlamydia Female    Urine     Symptomatic 668 676 0.9882 0.9768 0.9949
## 8  Chlamydia Female    Urine    Asymptomatic 502 507 0.9901 0.9771 0.9968
## 9  Gonorrhea   Male     Swab     Symptomatic 412 417 0.9880 0.9722 0.9961
## 10 Gonorrhea   Male     Swab    Asymptomatic 351 363 0.9669 0.9430 0.9828
## 11 Gonorrhea   Male    Urine     Symptomatic 433 434 0.9977 0.9872 0.9999
## 12 Gonorrhea   Male    Urine    Asymptomatic 368 370 0.9946 0.9806 0.9993
## 13 Gonorrhea Female     Swab     Symptomatic 772 787 0.9809 0.9688 0.9893
## 14 Gonorrhea Female     Swab    Asymptomatic 562 564 0.9965 0.9872 0.9996
## 15 Gonorrhea Female    Urine     Symptomatic 782 789 0.9911 0.9818 0.9964
## 16 Gonorrhea Female    Urine    Asymptomatic 564 567 0.9947 0.9846 0.9989
```

Cada tabela apresenta os resultados de sensibilidade e especificidade para diferentes combinações de doença, gênero, espécime e status de sintoma. A coluna "mean" representa a estimativa média, enquanto "lower" e "upper" representam os limites inferior e superior do intervalo de confiança de Clopper-Pearson, respectivamente. Essas tabelas fornecem uma visão geral das estimativas de desempenho do teste diagnóstico para cada cenário específico.

3-a) Calcule os intervalos de confiança de Wald e Agresti-Caffo para a diferença nas probabilidades de ser HIV positivo com base no uso de preservativo. Interprete os intervalos.

```r
hiv_table <- matrix(c(135, 434, 569, 15, 9, 24, 150, 443, 539), ncol = 3, byrow = TRUE)
colnames(hiv_table) <- c("Positivo", "Negativo", "Total")
rownames(hiv_table) <- c("Nunca", "Sempre", "Total")

prop_nunca <- hiv_table[1, "Positivo"] / hiv_table[1, "Total"]
prop_sempre <- hiv_table[2, "Positivo"] / hiv_table[2, "Total"]

se_nunca <- sqrt(prop_nunca * (1 - prop_nunca) / hiv_table[1, "Total"])
se_sempre <- sqrt(prop_sempre * (1 - prop_sempre) / hiv_table[2, "Total"])

diff_prop <- prop_nunca - prop_sempre

z <- qnorm(0.975)
wald_lower <- diff_prop - z * sqrt(se_nunca^2 + se_sempre^2)
wald_upper <- diff_prop + z * sqrt(se_nunca^2 + se_sempre^2)

prop_nunca_ag <- (hiv_table[1, "Positivo"] + 1) / (hiv_table[1, "Total"] + 2)
prop_sempre_ag <- (hiv_table[2, "Positivo"] + 1) / (hiv_table[2, "Total"] + 2)

se_nunca_ag <- sqrt(prop_nunca_ag * (1 - prop_nunca_ag) / hiv_table[1, "Total"])
se_sempre_ag <- sqrt(prop_sempre_ag * (1 - prop_sempre_ag) / hiv_table[2, "Total"])
diff_prop_ag <- prop_nunca_ag - prop_sempre_ag


agresti_caffo_lower <- diff_prop_ag - z * sqrt(se_nunca_ag^2 + se_sempre_ag^2 +
                       (1/(2 * hiv_table[1, "Total"])) + (1/(2 * hiv_table[2, "Total"])))
agresti_caffo_upper <- diff_prop_ag + z * sqrt(se_nunca_ag^2/(hiv_table[1, "Total"]+2) +
                       se_sempre_ag^2/(hiv_table[2, "Total"] +2 ))
```

```r
print("Intervalo de Confiança de Wald:")
cat(wald_lower, ";", wald_upper)

print("Intervalo de Confiança de Agresti-Caffo:")
cat(agresti_caffo_lower, ";", agresti_caffo_upper)
```

```
## [1] "Intervalo de Confiança de Wald:"

## -0.5845563 ; -0.190927

## [1] "Intervalo de Confiança de Agresti-Caffo:"

## -0.7272277 ; -0.3390061
```

Os intervalos de confiança de Wald e Agresti-Caffo indicam que a diferença nas probabilidades de ser HIV positivo entre aqueles que nunca usam preservativo e aqueles que sempre usam preservativo é estatisticamente significativa e maior para aqueles que nunca usam preservativo. Esses intervalos não incluem o valor zero, o que sugere que a diferença não é aleatória e há uma associação entre o uso de preservativo e o status do HIV.

3-b) Realize um teste escore, teste qui-quadrado de Pearson e LRT para testar a igualdade das probabilidades de sucesso.

```r
alpha <- 0.05
hiv_table <- matrix(c(135, 434, 15, 9), ncol = 2, byrow = TRUE)
colnames(hiv_table) <- c("Positivo", "Negativo")
rownames(hiv_table) <- c("Nunca", "Sempre")

score_test <- prop.test(x = hiv_table, conf.level = 1-alpha, correct = FALSE)
score_test
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  hiv_table
## X-squared = 18.322, df = 1, p-value = 1.866e-05
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.5845563 -0.1909270
## sample estimates:
##    prop 1    prop 2
## 0.2372583 0.6250000
```

```r
chisq <- chisq.test(x = hiv_table, correct = FALSE)
chisq
```

```
##
##  Pearson's Chi-squared test
##
## data:  hiv_table
## X-squared = 18.322, df = 1, p-value = 1.866e-05
```

```r
pi.bar <- colSums (hiv_table) [1]/ sum(hiv_table)
pi.hat.table <- hiv_table/rowSums(hiv_table)
log.Lambda <- hiv_table [1 ,1] * log(pi.bar / pi.hat.table [1 ,1]) +
  hiv_table [1 ,2] * log ((1 - pi.bar) / (1- pi.hat.table [1 ,1])) +
  hiv_table [2 ,1] * log(pi.bar / pi.hat.table [2 ,1]) +
  hiv_table [2 ,2] * log ((1 - pi.bar) /(1 - pi.hat.table [2 ,1]))
test.stat <- -2* log.Lambda
crit.val <- qchisq (p = 0.95 , df = 1)
p.val <- 1- pchisq (q = test.stat , df = 1)
round (data.frame (pi.bar , test.stat , crit.val , p.val , row.names = NULL ), 4)
```

```
##   pi.bar test.stat crit.val p.val
## 1  0.253    15.483   3.8415 1e-04
```

3-c) Estime a razão de chances e calcule o intervalo de confiança correspondente para ela. Interprete a estimativa e o intervalo.

```r
or.hat <- hiv_table[1,1]*hiv_table[2,2] / (hiv_table[2,1]*hiv_table[1,2])

var.log.or <- 1/hiv_table[1,1] + 1/hiv_table[1,2] +
              1/hiv_table[2,1] + 1/hiv_table[2,2]

or.ci <- exp(log
             (or.hat) + qnorm(p = c(alpha/2, .5, 1-alpha/2)) * sqrt(var.log.or))

print("Intervalo de Confiança para a razão de chances:")
round(or.ci,3)
print("Intervalo de Confiança para a razão de chances:")
round(rev(1/or.ci),3)
```

```
## [1] "Intervalo de Confiança para a razão de chances:"
```

```
## [1] 0.080 0.187 0.436
```

```
## [1] "Intervalo de Confiança para a razão de chances:"
```

```
## [1]  2.293  5.358 12.519
```

Para o primeiro intervalo da razão de chances $\widehat{\frac{odds_1}{odds_2}}$ concluímos que as chances de contrair HIV são 0.187 vezes, ou 81%, menores quando não se usa preservativo do que quando se usa.

Já para o segundo intervalo, sobre $\widehat{\frac{odds_2}{odds_1}}$, concluímos que ao se usar preservativo tem-se uma chance de 5.358 vezes maior de contrair HIV comparado a quando o preservativo não é usado.

3-d) Geralmente, pensa-se que o uso do preservativo ajuda a prevenir a transmissão do HIV. Os resultados aqui concordam com isso? Se não, quais fatores podem ter levado a esses resultados? Observe que Aseffa et al. (1998) fornecem a razão de chances estimada e um intervalo de Wald correspondente, mas falham em interpretá-los.

Variáveis de Confusão: Pode haver outros fatores não considerados na análise que poderiam estar influenciando os resultados. Por exemplo, indivíduos que optam por usar preservativos podem se envolver em

comportamentos mais arriscados ou ter características demográficas e socioeconômicas diferentes que podem impactar os resultados.

Viés de Amostragem: A amostra usada para a análise pode não ser representativa da população em geral, levando a resultados enviesados.

Erros de Medição: Erros na coleta de dados ou classificação incorreta do uso de preservativos e dos resultados de transmissão de HIV podem afetar os resultados.

Causalidade e Correlação: A análise pode mostrar uma correlação entre o uso de preservativos e a transmissão de HIV, mas é importante observar que correlação não implica causalidade. Outros fatores podem estar influenciando a relação observada.

Tamanho da Amostra Pequeno: Se o tamanho da amostra for pequeno, os resultados podem não ser estatisticamente robustos, e flutuações aleatórias podem afetar as descobertas.

4-) Calcule este intervalo score para os dados de Larry Bird. Compare o intervalo com os intervalos Wald e Agresti-Caffo para ele. Considere $d = 0$

```r
#install.packages("PropCIs")
library(PropCIs)
c.table <- array(data = c(251 , 48, 34, 5), dim = c(2 ,2),
            dimnames = list(First=c("made", "missed"),
                            Second = c("made", "missed")))

alpha <- 0.05
pi.hat.table <- c.table/rowSums(c.table)
pi.hat1 <- pi.hat.table[1 ,1]
pi.hat2 <- pi.hat.table[2 ,1]
var.wald <- pi.hat1*(1 - pi.hat1) / sum(c.table[1,]) +
            pi.hat2*(1 - pi.hat2) / sum(c.table[2 ,])
print("Intervalo de Confiança de Wald:")
pi.hat1 - pi.hat2 + qnorm (p = c( alpha/2, 1- alpha/2))*sqrt(var.wald )


pi.tilde1 <- (c.table[1,1] + 1) / (sum(c.table[1 ,]) + 2)
pi.tilde2 <- (c.table [2,1] + 1) / (sum(c.table [2 ,]) + 2)
var.AC <- pi.tilde1*(1 - pi.tilde1 ) / (sum(c.table [1 ,]) + 2) +
          pi.tilde2*(1 - pi.tilde2 ) / (sum(c.table [2 ,]) + 2)
print("Intervalo de Confiança de Agresti-Caffo:")
pi.tilde1 - pi.tilde2 + qnorm (p = c( alpha /2, 1- alpha /2)) * sqrt (var.AC)

intervalo_score = diffscoreci(c.table[1,1], sum(c.table[1,]),
                              c.table[2,1], sum(c.table[2,]),
                              conf.level = 1 - alpha)

print("Intervalo de Confiança Score:")
c(intervalo_score$conf.int[1],intervalo_score$conf.int[2])
```

```
## [1] "Intervalo de Confiança de Wald:"

## [1] -0.11218742  0.06227017

## [1] "Intervalo de Confiança de Agresti-Caffo:"

## [1] -0.10353254  0.07781192
```

```
## [1] "Intervalo de Confiança Score:"
```

```
## [1] -0.09529146  0.08792491
```

Podemos observar que o intervalo de confiança de Wald é o mais curto, provavelmente devido a suposições feitas sobre a distribuição normal da diferença de probabilidades e seu valor ser extremo, gerando assim um nível de confiança real menor que o nominal.

Já o intervalo de confiança de Agresti-Caffo é um intervalo proposto com uma correção para esses casos extremos. Observa-se que ele é maior e compreende todo o intervalo de Wald anterior, nos levando a acreditar que ele está tendo um nível real muito próximo ao nominal.

O intervalo de confiança de Score, também conhecido como intervalo de confiança de Razão de Verossimilhança, utiliza a distribuição da razão de verossimilhança para estimar a incerteza. Essa abordagem leva em consideração o valor da função de verossimilhança nos extremos do intervalo. Sabe-se que o intervalo de confiança de Score é mais robusto para tamanhos de amostra menores ou proporções extremas, esse último sendo o caso analisado para o problema das diferenças de probabilidade $\pi_1 - \pi_2$.

5-) Mostre que $OR = 1$ quando $\pi_1 = \pi_2$.

$$OR = \frac{odds_1}{odds_2} = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}$$

Tomando $\pi_1 = \pi_2$, com $\pi_1 \neq 0$ temos:

$$OR = \frac{\pi_1(1-\pi_1)}{\pi_1(1-\pi_1)} = 1, \pi_1 \neq 0$$

Portanto, se $\pi_1 = \pi_2$, então $OR = 1$.