

Análise de Dados Categóricos

Fernando Lucambio

Departamento de Estatística
Universidade Federal do Paraná

Outubro, 2022

Consideramos agora a situação em que o mesmo ensaio de Bernoulli é medido em unidades que podem ser classificadas em grupos.

O caso mais simples é quando uma população consiste em dois grupos, como fêmeas e machos, peixes de água doce e salgada, ou empresas americanas e estrangeiras.

Abaixo estão dois exemplos com uma resposta binária em ensaios que formam dois grupos.

Arremesso de lance livre de Larry Bird.

Um lance livre é um arremesso no basquete onde o arremessador pode arremessar livremente (sem oposição de outro jogador) de um local específico na quadra.

O arremesso é feito (um sucesso) ou perdido (uma falha). Na maioria das vezes, durante um jogo da National Basketball Association (NBA), os lances livres são feitos em pares.

Isso significa que um arremessador de lance livre tem uma tentativa e, posteriormente, tem uma segunda tentativa, não importa o que aconteça na primeira.

Arremesso de lance livre de Larry Bird.

O ex-jogador da NBA Larry Bird foi um dos mais bem sucedidos em fazer lances livres durante sua carreira com uma taxa de sucesso de 88,6%. Em comparação, a média da NBA durante esse período foi de cerca de 75% (<http://www.basketball-reference.com>).

A excelente taxa de sucesso de Bird está entre suas muitas conquistas, pelas quais ele foi reconhecido como um dos 50 maiores jogadores da história da NBA.

Durante as temporadas da NBA de 1980-81 e 1981-82, foram registrados os resultados das tentativas de lance livre de Bird arremessadas em duplas, e um resumo é mostrado na Tabela abaixo.

Arremesso de lance livre de Larry Bird.

Por exemplo, Bird fez sua primeira e segunda tentativa 251 vezes. Além disso, Bird fez a primeira tentativa, mas depois posteriormente perdeu a segunda tentativa 34 vezes. No geral, ele fez a primeira tentativa 285 vezes sem levar em conta o que aconteceu na segunda tentativa. No total, Bird acertou 338 pares de lances livres durante a temporada.

TABELA: os resultados dos lances livres de Larry Bird; fonte de dados Wardrop (1995).

		Segundo lance		Total
		Acerto	Erro	
Primeiro lance	Acerto	251	34	285
	Erro	48	5	53
Total		299	39	338

Arremesso de lance livre de Larry Bird.

Fãs e comentaristas de basquete muitas vezes especulam que os resultados de um segundo lance livre podem depender dos resultados do primeiro. Por exemplo, se um atirador erra a primeira tentativa, o desapontamento ou a determinação levarão a alterar sua abordagem para a segunda tentativa? Se sim, então devemos ver que a probabilidade de sucesso na segunda tentativa é diferente dependendo se a primeira tentativa foi feita ou perdida.

Assim, os dois grupos neste problema são formados pelos resultados da primeira tentativa, e os ensaios de Bernoulli que observamos são os resultados da segunda tentativa. Dados os dados da Tabela acima, investigaremos se o resultado da segunda tentativa depende do que acontece na primeira tentativa.

Ensaio clínico da vacina Salk.

Ensaio clínico são realizados para determinar a segurança e eficácia de novos medicamentos. Frequentemente, as respostas de segurança e eficácia são de natureza categórica; por exemplo, a resposta de eficácia pode ser simplesmente se um medicamento cura ou não um paciente de uma doença.

Para garantir que um novo medicamento é realmente melhor do que não fazer nada (os pacientes às vezes melhoram sem intervenção), é essencial ter um grupo de controle no estudo. Isso é alcançado em ensaios clínicos randomizando os pacientes em dois grupos: novo medicamento ou controle. O grupo de controle geralmente é um placebo, que é administrado exatamente como o novo medicamento, mas não contém medicação.

Ensaio clínico da vacina Salk.

Um dos mais famosos e maiores ensaios clínicos já realizados foi em 1954. Mais de 1,8 milhão de crianças participaram do ensaio clínico para determinar a eficácia da vacina contra a poliomielite desenvolvida por Jonas Salk (Francis et al., 1955).

Embora o projeto real do estudo tenha suscitado debate (Brownlee, 1955; Dawson, 2004), renunciamos a essa discussão e focamos nos dados obtidos da parte randomizada e controlada por placebo do estudo.

Os dados, apresentados na Tabela a seguir, mostram que 57 das 200.745 crianças do grupo vacinado desenvolveram poliomielite durante o período do estudo, em oposição a 142 das 201.229 crianças do grupo placebo.

Ensaio clínico da vacina Salk.

A questão de interesse para o ensaio clínico foi “A vacina ajuda a prevenir a poliomielite?” Desenvolveremos medidas de comparação nesta seção para responder a essa pergunta.

TABELA: Resultados do ensaio clínico da vacina Salk; fonte de dados Francis et al. (1955, pág. 25).

	Polio	Livre de Polio	Total
Vacina	57	200,688	200,745
Placeo	142	201,087	201,229
Total	199	401,775	401,974

O modelo e a notação seguem aqueles usados para uma única variável aleatória binomial. Começamos considerando duas variáveis aleatórias de Bernoulli separadas, Y_1 e Y_2 , uma para cada grupo. As probabilidades de sucesso para os dois grupos são denotadas por π_1 e π_2 , respectivamente. Observamos n_j tentativas de Y_j levando a w_j sucessos observados, $j = 1, 2$. Substituímos um subscrito por “+” para denotar uma soma nesse subscrito. Assim, $n_+ = n_1 + n_2$ é o número total de tentativas e $w_+ = w_1 + w_2$ é o número total de sucessos observados.

Essa notação está representada na Tabela abaixo. A tabela do lado direito é chamada de tabela de contingência bidirecional, porque fornece uma lista de todas as tabulações cruzadas possíveis de duas variáveis categóricas.

TABELA: Probabilidade e estruturas de contagem observadas para duas variáveis aleatórias binomiais independentes.

		Resposta				Resposta			
		1	2			1	2		
Grupo	1	π_1	$1 - \pi_1$	1	Grupo	1	w_1	$n_1 - w_1$	n_1
Grupo	2	π_2	$1 - \pi_2$	1	Grupo	2	w_2	$n_2 - w_2$	n_2
							w_+	$n_+ - w_+$	n_+

Denotamos a variável aleatória que representa o número de sucessos no grupo j por W_j e escrevendo sua função de probabilidade como

$$P(W_j = w_j) = \binom{n_j}{w_j} \pi_j^{w_j} (1 - \pi_j)^{n_j - w_j}, \quad w_j = 0, 1, \dots, n_j, \quad j = 1, 2.$$

Assumimos que as duas variáveis aleatórias Y_1 e Y_2 são independentes, de modo que o resultado de uma não pode afetar o resultado da outra.

No ensaio clínico da vacina Salk, por exemplo, isso significa que as crianças designadas para receber a vacina não podem transmitir imunidade ou doença às crianças do grupo placebo e vice-versa. Essa suposição é crítica no que se segue e, portanto, esse modelo é chamado de modelo binomial independente. Quando a independência não é satisfeita, outros modelos precisam ser usados para levar em conta a dependência entre as variáveis aleatórias, por exemplo, dados pareados.

Quando quisermos simular dados deste modelo, usaremos o código R como mostrado abaixo. Esse procedimento de amostragem será importante em breve quando usarmos essas contagens simuladas para avaliar procedimentos de inferência estatística, como intervalos de confiança, para determinar se eles funcionam conforme o esperado.

Simulando contagens em uma tabela de contingência.

Considere o caso de $\pi_1 = 0.2$, $\pi_2 = 0.4$, $n_1 = 10$ e $n_2 = 10$. O código R abaixo mostra como simular um conjunto de contagens para uma tabela de contingência.

```
pi1 <- 0.2
pi2 <- 0.4
n1 <- 10
n2 <- 10
set.seed(8191)
w1 <- rbinom(n = 1, size = n1, prob = pi1)
w2 <- rbinom(n = 1, size = n2, prob = pi2)
c.table <- array(data = c(w1, w2, n1 - w1, n2 - w2), dim = c(2, 2),
dimnames = list(Group = c(1, 2), Response = c(1, 2)))
c.table
##      Response
## Group 1 2
##      1 1 9
##      2 3 7
c.table[1,1] # w1
## [1] 1
c.table[1,2] # n1 - w1
## [1] 9
c.table[1,] # w1 and n1 - w1
## 1 2
## 1 9
sum(c.table[1,]) # n1
## [1] 10
```

Verossimilhança e estimativas

Os principais interesses neste problema são estimar as probabilidades de sucesso, π_1 e π_2 , para cada grupo e comparar essas probabilidades. A estimativa de máxima verossimilhança novamente fornece uma solução conveniente e poderosa. Como Y_1 e Y_2 são independentes, W_1 e W_2 também são. A função de verossimilhança formada por variáveis aleatórias independentes é apenas o produto de suas respectivas funções de verossimilhança. Portanto, a função de verossimilhança é

$$L(\pi_1, \pi_2 | w_1, w_2) = L(\pi_1 | w_1) \times L(\pi_2 | w_2).$$

Maximizar essa verossimilhança sobre π_1 e π_2 resulta nas estimativas “óbvias” $\hat{\pi}_1 = w_1/n_1$ e $\hat{\pi}_2 = w_2/n_2$, as proporções amostrais nos dois grupos. Em outras palavras, quando as variáveis aleatórias são independentes, cada probabilidade é estimada usando apenas os dados de seu próprio grupo.

Arremesso de lance livre de Larry Bird.

O objetivo deste exemplo é estimar a probabilidade de sucesso usando uma estrutura de tabela de contingência em R. Se os resultados do teste de Bernoulli já estiverem resumidos em contagens como na tabela do exemplo Larry Bird anterior, então uma tabela de contingência é criada em R usando a função `array()`:

```
c.table <- array ( data = c(251 , 48, 34, 5) , dim = c(2 ,2) ,  
                  dimnames = list ( First = c(" made " , " missed " ) , Second = c(" made " , " missed ")))  
c.table  
##           Second  
## First      made   missed  
## made       251     34  
## missed     48      5  
list ( First = c(" made " , " missed " ) , Second = c(" made " , " missed " ))  
## $First  
## [1] " made " " missed "  
##  
## $Second  
## [1] " made " " missed "
```

Arremesso de lance livre de Larry Bird.

Como os níveis das variáveis de linha e coluna têm nomes, usamos esses nomes no argumento `dimnames`. As estimativas da probabilidade de sucessos ou proporções amostrais, são encontradas aproveitando-se de como o R realiza os cálculos:

```
rowSums(c.table) # n1 and n2
##   made   missed
##   285     53
pi.hat.table <- c.table/rowSums(c.table)
pi.hat.table
##           Second
## First      made   missed
##   made  0.8807018 0.11929825
##   missed 0.9056604 0.09433962
```

Arremesso de lance livre de Larry Bird.

A função `rowSums()` encontra a soma das contagens em cada linha para obter n_1 e n_2 . Tomando a tabela de contingência de contagens divididas por essas somas de linhas, obtemos $\hat{\pi}_1$ e $\hat{\pi}_2$ na primeira coluna e $1 - \hat{\pi}_1$ e $1 - \hat{\pi}_2$ na segunda coluna.

Observe que R faz essa divisão corretamente tomando as contagens na primeira (segunda) linha de `c.table` dividida pelo primeiro (segundo) elemento de `rowSums(c.table)`.

Estamos interessados em encontrar intervalos de confiança para a diferença de duas probabilidades. Uma abordagem relativamente fácil para comparar π_1 e π_2 pode ser desenvolvida tomando sua diferença $\pi_1 - \pi_2$.

A estimativa correspondente de $\pi_1 - \pi_2$ é $\hat{\pi}_1 - \hat{\pi}_2$. Cada estimativa de probabilidade de sucesso tem uma distribuição de probabilidade que é aproximada por uma distribuição normal em grandes amostras

$$\hat{\pi}_j \sim N(\pi_j, \widehat{\text{Var}}(\hat{\pi}_j)),$$

onde $\widehat{\text{Var}}(\hat{\pi}_j) = \hat{\pi}_j(1 - \hat{\pi}_j)/n_j, j = 1, 2$.

Como as combinações lineares de variáveis aleatórias normais são elas próprias variáveis aleatórias normais (Casella e Berger, 2002), a distribuição de probabilidade para $\hat{\pi}_1 - \hat{\pi}_2$ é aproximada por

$$N(\hat{\pi}_1 - \hat{\pi}_2, \widehat{\text{Var}}(\hat{\pi}_1 - \hat{\pi}_2)),$$

sendo

$$\widehat{\text{Var}}(\hat{\pi}_1 - \hat{\pi}_2) = \hat{\pi}_1(1 - \hat{\pi}_1)/n_1 + \hat{\pi}_2(1 - \hat{\pi}_2)/n_2.$$

Essa distribuição forma a base para uma série de procedimentos de inferência.

O intervalo de confiança mais fácil de formar para $\pi_1 - \pi_2$ usa a aproximação normal para $\hat{\pi}_1 - \hat{\pi}_2$ diretamente para criar um intervalo Wald:

$$\hat{\pi}_1 - \hat{\pi}_2 \pm Z_{1-\alpha/2} \sqrt{\frac{\hat{\pi}_1(1 - \hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1 - \hat{\pi}_2)}{n_2}}.$$

Infelizmente, o intervalo de Wald para $\pi_1 - \pi_2$ tem problemas semelhantes em alcançar o nível de confiança declarado como o intervalo de Wald para π . Investigaremos isso em breve.

Devido a esses problemas, vários outros intervalos de confiança para $\pi_1 - \pi_2$ foram propostos. Inspirados pelo bom desempenho geral do intervalo de Agresti-Coull para uma única probabilidade, Agresti and Caffo (2000) investigaram vários intervalos construídos como intervalos do tipo Wald em dados com diferentes números de sucessos e falhas adicionados.

Eles descobriram que adicionar um sucesso e um fracasso para cada grupo resulta em um intervalo que faz um bom trabalho para atingir o nível de confiança declarado. Especificamente, sejam

$$\tilde{\pi}_1 = \frac{w_1 + 1}{n_1 + 2} \quad \text{e} \quad \tilde{\pi}_2 = \frac{w_2 + 1}{n_2 + 2}$$

as estimativas alteradas de π_1 e π_2 .

Observe que, diferentemente de $\tilde{\pi}$ para o intervalo de Agresti-Coull, as estimativas $\tilde{\pi}_1$ e $\tilde{\pi}_2$ não mudam quando o nível de confiança muda.

O intervalo de confiança $(1 - \alpha)100\%$ Agresti-Caffo para $\pi_1 - \pi_2$ é

$$\tilde{\pi}_1 - \tilde{\pi}_2 \pm Z_{1-\alpha/2} \sqrt{\frac{\tilde{\pi}_1(1 - \tilde{\pi}_1)}{n_1 + 2} + \frac{\tilde{\pi}_2(1 - \tilde{\pi}_2)}{n_2 + 2}}.$$

No geral, recomendamos o método Agresti-Caffo para uso geral.

Arremesso de lance livre de Larry Bird.

O objetivo deste exemplo é calcular um intervalo de confiança para a diferença nas probabilidades de sucesso do segundo lance livre dados os resultados da primeira tentativa.

Continuando o código anterior, obtemos o seguinte:

```
alpha <- 0.05
pi.hat1 <- pi.hat.table[1 ,1]
pi.hat2 <- pi.hat.table[2 ,1]
# Wald
var.wald <- pi.hat1*(1 - pi.hat1) / sum(c.table[1,]) + pi.hat2*(1 - pi.hat2) / sum(c.table[2 ,])
pi.hat1 - pi.hat2 + qnorm (p = c( alpha/2, 1- alpha/2))*sqrt(var.wald )
## [1] -0.11218742  0.06227017

# Agresti - Caffo
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)
pi.tilde1 - pi.tilde2 + qnorm (p = c( alpha /2, 1- alpha /2)) * sqrt (var.AC)
## [1] -0.10353254  0.07781192
```

O intervalo de Wald de 95% é

$$(-0.1122 < \pi_1 - \pi_2 < 0.0623)$$

e o intervalo de 95% Agresti-Caffo é

$$(-0.1035 < \pi_1 - \pi_2 < 0.0778).$$

Os intervalos são semelhantes com o intervalo Wald sendo deslocado para a esquerda do intervalo Agresti-Caffo.

O intervalo Agresti-Caffo, com 95% de confiança, a diferença nas probabilidades de sucesso do segundo lance livre dado o resultado do primeiro está entre -0.1035 e 0.0778.

Como esse intervalo contém 0, não podemos detectar uma mudança na probabilidade de Bird ter um segundo lance livre bem-sucedido após as primeiras tentativas feitas e perdidas. Isso significa que ou não há diferença, ou há uma diferença, mas não foi detectada nesta amostra.

A última situação pode ser causada por má sorte (uma amostra incomum) ou por um tamanho de amostra muito pequeno.

Os mesmos intervalos de confiança podem ser obtidos de outras maneiras. Primeiro, podemos usar a função `array()`:

```
w1 <- 251
n1 <- 285
w2 <- 48
n2 <- 53
alpha <- 0.05
pi.hat1 <- w1/n1
pi.hat2 <- w2/n2
var.wald <- pi.hat1 *(1 - pi.hat1 ) / n1 + pi.hat2 *(1 - pi.hat2 ) / n2
pi.hat1 - pi.hat2 + qnorm (p = c( alpha /2, 1- alpha /2)) * sqrt (var.wald )
## [1] -0.11218742  0.06227017
```

Em segundo, a função `prop.test()` mostrada posteriormente fornece o intervalo de confiança de Wald como parte de sua saída. Por fim, a função `wald2ci()` no pacote `PropCIs` também calcula os intervalos de confiança de Wald e Agresti-Caffo.

Para encontrar um nível de confiança verdadeiro para um desses intervalos de confiança, é necessária a distribuição de probabilidade conjunta para todas as combinações possíveis de (W_1, W_2) .

Essa distribuição é o produto de duas probabilidades binomiais porque essas variáveis aleatórias são independentes.

Para dados n_1 e n_2 , existem $(n_1 + 1)(n_2 + 1)$ possíveis combinações observadas de (w_1, w_2) , e um intervalo de confiança pode ser calculado para cada uma dessas combinações. Para valores definidos de π_1 e π_2 , alguns desses intervalos contêm $\pi_1 - \pi_2$ e outros não.

Assim, o verdadeiro nível de confiança em π_1 e π_2 , $C(\pi_1, \pi_2)$, é a soma das probabilidades conjuntas para todos os intervalos que contêm $\pi_1 - \pi_2$:

$$C(\cdot) = \sum_{w_2=0}^{n_2} \sum_{w_1=0}^{n_1} I(\cdot) \binom{n_1}{w_1} \pi_1^{w_1} (1 - \pi_1)^{n_1 - w_1} \binom{n_2}{w_2} \pi_2^{w_2} (1 - \pi_2)^{n_2 - w_2},$$

onde a função indicadora $I(w_1, w_2)$ é 1 se o intervalo correspondente contiver $\pi_1 - \pi_2$ e 0 caso contrário.

Os detalhes do cálculo são fornecidos no próximo exemplo.

Níveis de confiança verdadeiros para os intervalos Wald e Agresti-Caffo.

O verdadeiro nível de confiança para o intervalo de Wald pode ser encontrado de maneira semelhante à discutida anteriormente. Considere o caso de $\alpha = 0.05$, $\pi_1 = 0.2$, $\pi_2 = 0.4$, $n_1 = 10$ e $n_2 = 10$.

Para encontrar todas as combinações possíveis de (w_1, w_2) , usamos o `expand.grid()`, que encontra todas as combinações possíveis dos argumentos, separados por vírgulas, dentro de seus parênteses.

Repetimos esse mesmo processo para encontrar todos os valores possíveis de $(\hat{\pi}_1, \hat{\pi}_2)$ e $P(W_1 = w_1, W_2 = w_2)$. Abaixo segue o código R:

```
alpha <- 0.05
pi1 <- 0.2
pi2 <- 0.4
n1 <- 10
n2 <- 10
# All possible combinations of w1 and w2
w.all <- expand.grid(w1 = 0:n1 , w2 = 0: n2)
# All possible combinations of pi^_1 and pi^_2
pi.hat1 <- (0: n1)/n1
pi.hat2 <- (0: n2)/n2
pi.hat.all <- expand.grid (pi.hat1 = pi.hat1 , pi.hat2 = pi.hat2 )
# Find joint probability for w1 and w2
prob.w1 <- dbinom (x = 0:n1 , size = n1 , prob = pi1)
prob.w2 <- dbinom (x = 0:n2 , size = n2 , prob = pi2)
prob.all <- expand.grid ( prob.w1 = prob.w1 , prob.w2 = prob.w2)
pmf <- prob.all$prob.w1*prob.all$prob.w2
# P(W1 = w1 , W2 = w2)
head ( data.frame(w.all , pmf = round (pmf ,4)))
```

##	w1	w2	pmf
## 1	0	0	0.0006
## 2	1	0	0.0016
## 3	2	0	0.0018
## 4	3	0	0.0012
## 5	4	0	0.0005
## 6	5	0	0.0002

Por exemplo, a probabilidade de observar $P(W_1 = 1, W_2 = 0) = 0.0016$. Usando essas probabilidades, calculamos o nível de confiança verdadeiro para o intervalo:

```
var.wald <- pi.hat.all [ ,1] * (1- pi.hat.all [ ,1]) / n1 + pi.hat.all [ ,2] * (1- pi.hat.all [ ,2]) / n2
lower <- pi.hat.all [ ,1] - pi.hat.all [ ,2] - qnorm (p = 1- alpha /2) * sqrt (var.wald )
upper <- pi.hat.all [ ,1] - pi.hat.all [ ,2] + qnorm (p = 1- alpha /2) * sqrt (var.wald )
save <- ifelse ( test = pi1 -pi2 > lower , yes = ifelse ( test = pi1 - pi2 < upper , yes = 1, no = 0) , no = 0)
sum( save *pmf )
```

```
## [1] 0.9281274
```

```
data.frame (w.all , round ( data.frame (pmf , lower , upper ) ,4) , save ) [1:15 ,]
```

##	w1	w2	pmf	lower	upper	save
## 1	0	0	0.0006	0.0000	0.0000	0
## 2	1	0	0.0016	-0.0859	0.2859	0
## 3	2	0	0.0018	-0.0479	0.4479	0
## 4	3	0	0.0012	0.0160	0.5840	0
## 5	4	0	0.0005	0.0964	0.7036	0
## 6	5	0	0.0002	0.1901	0.8099	0
## 7	6	0	0.0000	0.2964	0.9036	0
## 8	7	0	0.0000	0.4160	0.9840	0
## 9	8	0	0.0000	0.5521	1.0479	0
## 10	9	0	0.0000	0.7141	1.0859	0
## 11	10	0	0.0000	1.0000	1.0000	0
## 12	0	1	0.0043	-0.2859	0.0859	1
## 13	1	1	0.0108	-0.2630	0.2630	1
## 14	2	1	0.0122	-0.2099	0.4099	1
## 15	3	1	0.0081	-0.1395	0.5395	0

Todos os intervalos Wald possíveis são calculados e a função `ifelse()` é usada para verificar se $\pi_1 - \pi_2 = 0.2 - 0.4 = -0.2$ está dentro de cada intervalo. O último quadro de dados mostra os primeiros 15 intervalos com os resultados desta verificação. As probabilidades correspondentes aos intervalos que contêm -0.2 são somadas para produzir o nível de confiança verdadeiro de 0.9281, que é menor que o nível declarado de 0.95.

Também podemos calcular o nível de confiança verdadeiro mantendo uma das probabilidades constante e permitindo que a outra varie. A figura mostra um gráfico em que $\pi_2 = 0.4$ e π_1 varia de 0.001 a 0.999 por 0.0005. As linhas Agresti-Caffo e Wald são desenhadas no gráfico simultaneamente usando a função `plot()` primeiro para os níveis de confiança verdadeiros de Wald e depois usando a função `lines()` para os níveis de confiança verdadeiros do Agresti-Caffo.

```

# Initial settings
alpha<-0.05
pi1<-0.2
pi2<-0.4
n1<-10
n2<-10
numb.bin.samples<-1000 # Number of binomial samples
#####
# Estimated true confidence level
# Simulate w1 and w2
set.seed(2349)
w1<-rbinom(n = numb.bin.samples, size = n1, prob = pi1)
w2<-rbinom(n = numb.bin.samples, size = n2, prob = pi2)
pi.hat1<-w1/n1
pi.hat2<-w2/n2
# Wald
var.wald<-pi.hat1*(1-pi.hat1) / n1 + pi.hat2*(1-pi.hat2) / n2
lower<-pi.hat1 - pi.hat2 - qnorm(p = 1-alpha/2) * sqrt(var.wald)
upper<-pi.hat1 - pi.hat2 + qnorm(p = 1-alpha/2) * sqrt(var.wald)
# Intervals 1-5
data.frame(w1, w2, lower, upper)[1:5,]
##   w1 w2   lower   upper
## 1  0  6 -0.9036363 -0.29636369
## 2  2  3 -0.4770066  0.27700660
## 3  1  4 -0.6560451  0.05604514
## 4  2  1 -0.2098975  0.40989752
## 5  3  4 -0.5157711  0.31577115

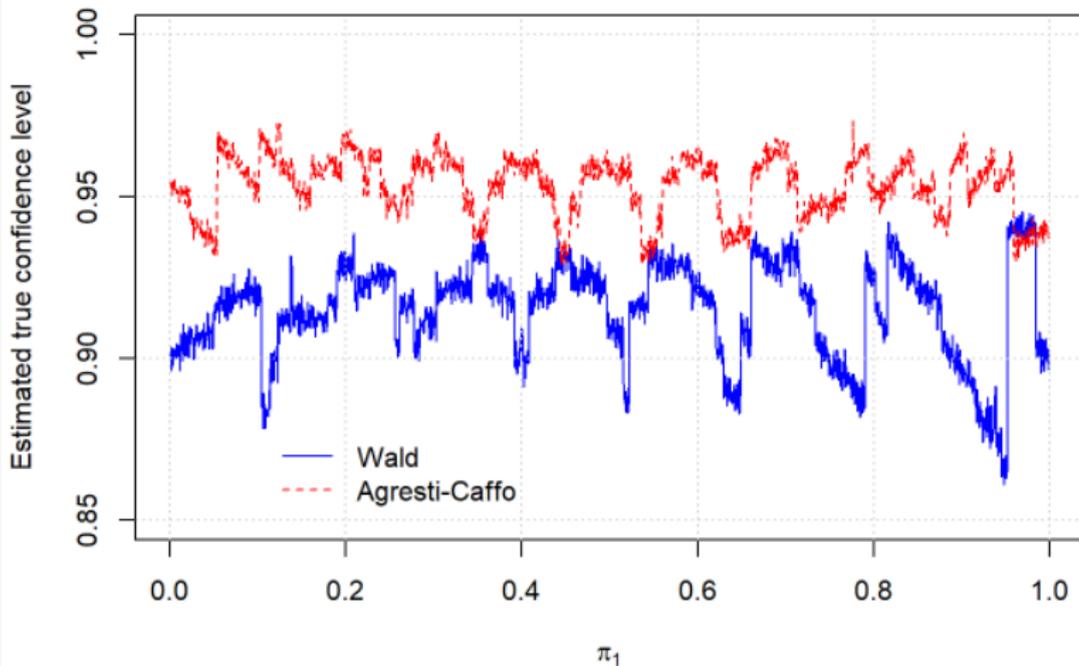
```

```

# Calculate estimated true confidence Level
save<-ifelse(test = pi1-pi2 > lower,
             yes = ifelse(test = pi1-pi2 < upper, yes = 1, no = 0), no = 0)
save[1:5]
## [1] 0 1 1 1 1
true.conf<-mean(save)
round(true.conf,4)
## [1] 0.92
# Agresti-Caffo
pi.tilde1<-(w1+1)/(n1+2)
pi.tilde2<-(w2+1)/(n2+2)
var.AC<-pi.tilde1*(1-pi.tilde1) / (n1+2) + pi.tilde2*(1-pi.tilde2) / (n2+2)
lower.AC<-pi.tilde1 - pi.tilde2 - qnorm(p = 1-alpha/2) * sqrt(var.AC)
upper.AC<-pi.tilde1 + pi.tilde2 + qnorm(p = 1-alpha/2) * sqrt(var.AC)
save.AC<-ifelse(test = pi1-pi2 > lower.AC,
                yes = ifelse(test = pi1-pi2 < upper.AC, yes = 1, no = 0), no = 0)
save.AC[1:10]
## [1] 1 1 1 1 1 1 1 1 1 1
true.conf.AC<-mean(save.AC)
round(true.conf.AC,4)

```

Verdadeiros níveis de confiança com $n_1=10$, $n_2=10$, $\pi_2=0.2$ e $\alpha=0.05$



A figura acima mostra que o intervalo de Wald nunca atinge o nível de confiança verdadeiro! O intervalo Agresti-Caffo tem um nível de confiança verdadeiro entre 0.93 e 0.97.

Incentivamos os leitores a alterar o valor predefinido para π_2 no programa para examinar o que acontece em outras situações.

Por exemplo, o intervalo Wald é extremamente liberal e o intervalo Agresti-Caffo é extremamente conservador para π_1 pequeno quando $\pi_2 = 0.1$ com $n_1 = 10$, $n_2 = 10$ e $\alpha = 0.05$.

```
#####

# Alternatively, these true confidence levels can be found using the binom.coverage()
# function in the binom package
library(binom)
binom.coverage(p = 0.16, n = n, conf.level = 0.95, method = "asymptotic")
##      method   p   n coverage
## 1 asymptotic 0.16 40 0.894039

# Plots of the true confidence levels can be found from the binom.plot()
# function in the binom package. Note that the method argument uses a function name
# that calculates the individual confidence interval. (binom.asymp is within the binom
# package. Also, the y-axis is the true confidence level and the x-axis is pi. The
# lattice package is used for plotting so the usual arguments (like xlab = ) for basic
# R plots do not always work.
binom.plot(n = 40, method = binom.asymp, np = 500, conf.level = 0.95)
```

