Análise de Dados Categóricos

Fernando Lucambio

Departamento de Estatística
Universidade Federal do Paraná

Outubro, 2022

Quase toda análise estatística começa com algum tipo de modelo estatístico. Um modelo estatístico geralmente assume a forma de uma distribuição de probabilidade que tenta quantificar a incerteza que vem com a observação de uma nova resposta.

O modelo pretende representar o fenômeno desconhecido que rege o processo de observação. Ao mesmo tempo, o modelo precisa ser conveniente para trabalhar matematicamente, para que procedimentos de inferência como intervalos de confiança e testes de hipóteses possam ser desenvolvidos.

A seleção de um modelo geralmente é um compromisso entre dois objetivos concorrentes: fornecer uma aproximação mais detalhada do processo que gera os dados e fornecer procedimentos de inferência fáceis de usar. No caso de respostas binárias, o modelo natural é a distribuição de Bernoulli. Seja Y uma variável aleatória de Bernoulli com resultados de 0 e 1.

Normalmente, diremos que $Y = 1$ é um sucesso e $Y = 0$ é um fracasso. Por exemplo, um sucesso seria uma tentativa de lance livre de basquete que é boa ou um indivíduo que é curado de uma doença por uma nova droga; uma falha seria uma tentativa de lance livre que é perdida ou um indivíduo que não é curado.

Denotamos a probabilidade de sucesso como $P(Y = 1) = \pi$ e a probabilidade de falha correspondente como $P(Y = 0) = 1 - \pi$.

A função de probabilidade Bernoulli para Y combina essas duas expressões em uma fórmula:

$$P(Y = y) = \pi^y (1 - \pi)^{1-y}$$

para $y = 0$ ou 1, onde usamos a convenção padrão de que uma letra maiúscula Y denota a variável aleatória e a letra minúscula y denota um possível valor de Y.

O valor esperado de Y é $\mathrm{E}(Y) = \pi$ e a variância de Y é $\mathrm{Var}(Y) = (1 - \pi)$.

Muitas vezes, observam-se múltiplas respostas de Bernoulli por meio de amostragem repetida ou tentativas em ambientes idênticos. Isso leva à definição de variáveis aleatórias separadas para cada tentativa, $Y_1, \cdots, Y_n$, onde n é o número de tentativas.

Se todas as tentativas forem idênticas e independentes, podemos tratar $W = \sum_{i=1}^{n} Y_i$ como uma variável aleatória binomial com função de probabilidade

$$P(W = w) = \binom{n}{w} \pi^w (1-\pi)^{1-w}$$

para $w = 0, \cdots, n$.

O valor esperado de W é $\mathrm{E}(W) = n$ e a variância de W é $\mathrm{Var}(W) = n(1-\pi)$. Observe que a distribuição de Bernoulli é um caso especial da distribuição binomial quando $n = 1$.

A distribuição binomial é um modelo razoável para a distribuição de sucessos em um determinado número de tentativas, desde que o processo de observação de tentativas repetidas satisfaça certas suposições. Essas suposições são:

- **Existem *n* tentativas idênticas**. Refere-se ao processo pelo qual os ensaios são conduzidos. A ação que resulta na tentativa e na medição realizada deve ser a mesma em cada tentativa. Os ensaios não podem ser uma mistura de diferentes tipos de ações ou medições.
- **Existem dois resultados possíveis para cada tentativa**. Isso geralmente é apenas uma questão de saber o que é medido. No entanto, há casos em que uma medida de resposta tem mais de dois níveis, mas o interesse reside apenas em saber se um determinado nível ocorre ou não. Neste caso, o nível especial pode ser considerado “sucesso” e todos os níveis restantes “fracasso”.

- **As tentativas são independentes umas das outras**. Em particular, não há nada na condução dos ensaios que faria com que qualquer subconjunto de ensaios se comportasse de forma mais semelhante entre si.
- **A probabilidade de sucesso permanece constante para cada tentativa**. Isso significa que todas as variáveis que podem afetar a probabilidade de sucesso precisam ser mantidas constantes de tentativa a tentativa. Como essas variáveis nem sempre são conhecidas antecipadamente, essa pode ser uma condição muito difícil de confirmar. Muitas vezes, podemos apenas confirmar que as variáveis "óbvias" não estão mudando e, em seguida, simplesmente presumimos que outras também não estão.
- **A variável aleatória de interesse W é o número de sucessos**. Especificamente, isso implica que não estamos interessados na ordem em que os sucessos e fracassos ocorrem, mas apenas em sua contagem total.

O objetivo desta seção é estimar e fazer inferências sobre o parâmetro de probabilidade de sucesso da distribuição de Bernoulli. Começamos estimando o parâmetro usando sua estimativa de máxima verossimilhança, porque é relativamente fácil de calcular e possui propriedades que o tornam atraente em grandes amostras.

A seguir, os intervalos de confiança para a verdadeira probabilidade de sucesso são apresentados e comparados. Muitos intervalos de confiança diferentes têm sido propostos na literatura estatística.

Apresentaremos primeiro o procedimento mais simples, e depois apresentaremos várias alternativas melhores. Concluímos esta seção com testes de hipóteses para π.

A função de verossimilhança é uma função de um ou mais parâmetros condicionais aos dados observados. A função de verossimilhança para π quando $y_1, \cdots, y_n$ são observações de uma distribuição de Bernoulli é

$$L(\pi|y_1, \cdots, y_n) = P(Y_1 = y_1) \times \cdots \times P(Y_n = y_n) = \pi^w (1-\pi)^{n-w}.$$

Alternativamente, quando registramos apenas o número de sucessos de um número de tentativas, a função de verossimilhança para π é simplesmente

$$L(\pi|w) = P(W = w) = \binom{n}{w} \pi^w (1-\pi)^{n-w}.$$

O valor de π que maximiza a função de verossimilhança é considerado o valor mais plausível para o parâmetro e é chamado de estimativa de máxima verossimilhança (MLE).

Pode-se mostrar que o MLE de π é $\widehat{\pi} = w/n$, que é simplesmente a proporção observada de sucessos. Isso é verdade tanto para $L(\pi|y_1, \cdots, y_n)$ quanto para $L(\pi|w)$ porque o termo $\binom{n}{w}$ não contém informações sobre π.

Como $\widehat{\pi}$ varia de amostra para amostra é uma estatística e tem uma distribuição de probabilidade correspondente. Como em todos os MLEs, $\widehat{\pi}$ tem uma distribuição normal aproximada em amostras grandes. A média da distribuição normal é π, e a variância é encontrada a partir de

$$\begin{aligned}\widehat{\text{Var}}(\widehat{\pi}) &= -\,\text{E}\left(\frac{\partial^2 \log L(\pi|W)}{\partial \pi^2}\right)^{-1}\Bigg|_{\pi=\widehat{\pi}} \\ &= \text{E}\left(\frac{n}{\pi} - \frac{n}{1-\pi}\right)^{-1}\Bigg|_{\pi=\widehat{\pi}} = \frac{\widehat{\pi}(1-\widehat{\pi})}{n},\end{aligned}$$

onde $\log(\cdot)$ é a função log natural.

Podemos escrever a distribuição como

$$\widehat{\pi} \sim N(\pi, \widehat{\text{Var}}(\widehat{\pi})).$$

A aproximação tende a ser melhor à medida que o tamanho da amostra aumenta.

Intervalos confidencias de Wald

Utilizando a distribuição normal, consideramos $(\widehat{\pi}-\pi)/\widehat{\text{Var}}(\widehat{\pi})^{1/2}$ como uma variável normal padrão. Então, para $0 < \alpha < 1$, temos

$$P\left(Z_{\alpha/2} < \frac{\widehat{\pi} - \pi}{\sqrt{\widehat{\text{Var}}(\widehat{\pi})}} < Z_{1-\alpha/2}\right) \approx 1 - \alpha,$$

onde Z_α é o α-ésimo quantil da distribuição normal padrão, ou seja, $Z_{0.975} = 1.96$. Depois de reorganizar os termos e reconhecer que $-Z_{\alpha/2} = Z_{1-\alpha/2}$, obtemos

$$P\left(\widehat{\pi} - Z_{1-\alpha/2}\sqrt{\widehat{\text{Var}}(\widehat{\pi})} < \pi < \widehat{\pi} + Z_{1-\alpha/2}\sqrt{\widehat{\text{Var}}(\widehat{\pi})}\right) \approx 1 - \alpha.$$

Agora, temos uma probabilidade aproximada que tem o parâmetro π centrado entre duas estatísticas. Quando substituímos $\widehat{\pi}$ e $\widehat{\text{Var}}(\widehat{\pi})$ pelos valores observados da amostra, obtemos o $(1 - \alpha)100$

$$\widehat{\pi} - Z_{1-\alpha/2}\sqrt{\widehat{\pi}(1-\widehat{\pi})/n} < \pi < \widehat{\pi} + Z_{1-\alpha/2}\sqrt{\widehat{\pi}(1-\widehat{\pi})/n}.$$

Este é o intervalo usual para uma probabilidade de sucesso que é dado na maioria dos livros de estatística introdutórios. Os intervalos de confiança baseados na normalidade aproximada dos MLEs são chamados de "intervalos de confiança de Wald" porque Wald (1943) foi o primeiro a mostrar essa propriedade dos MLEs em grandes amostras.

Quando w está próximo de 0 ou n, ocorrem dois problemas com este intervalo:

- Os limites calculados podem ser menores que 0 ou maiores que 1, o que está fora dos limites para uma probabilidade.
- Quando $w = 0$ ou 1, $\sqrt{\widehat{\pi}(1-\widehat{\pi})/n} = 0$ para $n > 0$. Isso faz com que os limites inferior e superior sejam exatamente os mesmos, 0 para $w = 0$ ou 1 para $w = 1$.

Intervalos de Wald.

Suponha que $w = 4$ sucessos sejam observados em $n = 10$ tentativas. O intervalo de confiança de Wald de 95

$$0.4 \pm 1.96\sqrt{0.4(1-0.4)/10} = (0.0964; 0.7036),$$

onde usamos a notação abreviada entre parênteses para significar $0.0964 < \pi < 0.7036$. O código R abaixo mostra como esses cálculos são realizados:

```
w <- 4
n <- 10
alpha <- 0.05
pi.hat <- w/n
var.wald <- pi.hat*(1 - pi.hat)/n
lower <- pi.hat - qnorm (p = 1- alpha /2) * sqrt ( var.wald )
upper <- pi.hat + qnorm (p = 1- alpha /2) * sqrt ( var.wald )
round ( data.frame (lower , upper ), 4)
```

```
##    lower  upper
## 1 0.0964 0.7036
```

Intervalos de Wald.

No código, usamos a função qnorm() para encontrar o quantil $1 - \alpha/2$ de uma distribuição normal padrão. Podemos calcular o intervalo mais rapidamente aproveitando como R realiza cálculos vetoriais:

```
round(pi.hat + qnorm (p = c( alpha /2, 1- alpha /2) )*sqrt(var.wald ), 4)
## [1] 0.0964 0.7036
```

O intervalo de confiança é bastante amplo e pode não ser significativo para algumas aplicações. No entanto, ele fornece informações sobre um intervalo que pode ser útil em situações de teste de hipóteses. Por exemplo, um teste de $H_0 : \pi = 0.5$ vs. $H_1 : \pi \neq 0.5$ não rejeitaria H_0 porque 0.5 está dentro desse intervalo. Se, em vez disso, o teste foi $H_0 : \pi = 0.8$ vs. $H_1 : \pi \neq 0.8$, há evidências para rejeitar a hipótese nula.

As inferências para π do intervalo de confiança de Wald dependem da aproximação da distribuição normal subjacente para o estimador de máxima verossimilhança. Para que essa aproximação funcione bem, precisamos de uma amostra grande e, infelizmente, o tamanho da amostra no último exemplo era bem pequeno.

Além disso, observe que $\widehat{\pi}$ pode assumir apenas 11 valores possíveis diferentes no último exemplo: 0/10, 1/10, $\cdots$, 10/10, mas uma distribuição normal é uma função contínua.

Esses problemas levam o intervalo de conferência de Wald a ser aproximado, no sentido de que a probabilidade de o intervalo cobrir o parâmetro, sua cobertura ou nível de confiança verdadeiro não é necessariamente igual ao nível declarado $1 - \alpha$. A qualidade da aproximação varia com *n* e π, e como veremos mais adiante, o intervalo de Wald geralmente tem cobertura < $1 - \alpha$.

Tal intervalo é chamado de intervalo liberal. Por outro lado, um intervalo com cobertura superior ao nível declarado é chamado de conservador. Embora esta última propriedade possa parecer de boa qualidade, ela pode levar a intervalos bastante amplos em comparação com outras.

Queremos intervalos de confiança que coloquem o parâmetro dentro de um intervalo tão estreito quanto possível, mantendo pelo menos o nível de confiança declarado. Se quiséssemos intervalos que tivessem maior cobertura, teríamos declarado um nível de confiança maior.

Tem havido muitas pesquisas para encontrar um intervalo como esse para π, incluindo Agresti and Caffo (2000), Agresti and Min (2001), dentre outros. Brown et ai. (2001) apresentam uma revisão completa da maioria dos intervalos concorrentes.

Intervalos confidencias de Wilson

Quando $n < 40$, Brown et al. (2001) recomendam o intervalo de Wilson ou o intervalo de Jeffreys porque mantêm níveis de confiança verdadeiros mais próximos do nível declarado do que outros intervalos. A fórmula do intervalo de Wilson é encontrada examinando a estatística de teste

$$Z_0 = \frac{\widehat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}},$$

que é a estatística de teste escore frequentemente usada para um teste de $H_0 : \pi = \pi_0$ vs. $H_1 : \pi \neq \pi_0$, onde $0 < \pi_0 < 1$.

A variância no denominador de Z_0 é calculada assumindo que a hipótese nula é verdadeira, em vez de usar a estimativa irrestrita baseada nos dados. Isso leva à vantagem de que o denominador não é 0 sempre que $w = 0$ ou n.

Podemos aproximar a distribuição de Z_0 com uma normal padrão para obter

$$P(-Z_{1-\alpha/2} < Z_0 < Z_{1-\alpha/2}) \approx 1 - \alpha.$$

Tratando a aproximação como uma igualdade, o intervalo de Wilson contém o conjunto de todos os valores possíveis de π_0 que satisfazem a equação. Por outro lado, o conjunto de todos os valores possíveis para pi_0 que levam à rejeição da hipótese nula estão fora do intervalo de confiança. O processo de formação de um intervalo a partir de um procedimento de teste de hipótese como esse é frequentemente chamado de "inversão do teste". Como o intervalo de Wilson é baseado em um teste escore, também é frequentemente chamado de intervalo escore.

Os pontos finais do intervalo são encontrados definindo Z_0 igual a $\pm Z_{1-\alpha/2}$ e aplicando a fórmula quadrática para resolver π_0.

Assim, o intervalo $(1-\alpha)100\%$ de Wilson é

$$\widehat{\pi} \pm \frac{Z_{1-\alpha/2}\sqrt{n}}{n+Z^2_{1-\alpha/2}}\sqrt{\widehat{\pi}(1-\widehat{\pi})+\frac{Z^2_{1-\alpha/2}}{4n}},$$

onde

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

pode ser pensado como uma estimativa ajustada de π.
Este intervalo recebeu o nome de Wilson (1927), que primeiro propôs encontrar um intervalo para π dessa maneira. Observe que o intervalo de Wilson sempre tem limites entre 0 e 1. Os intervalos de confiança de Wald e Wilson discutidos até agora são procedimentos de inferência frequentista. A "confiança" associada a esses tipos de procedimentos de inferência ocorre pela repetição do processo de coleta de uma amostra e cálculo de um intervalo de confiança a cada vez.

Intervalos confidencias de Agresti-Coull

Brown et al. (2001) recomendam o intervalo de Agresti-Coull (Agresti and Coull 1998) para $n \geq 40$, principalmente porque é um pouco mais fácil de calcular à mão e se assemelha mais ao popular intervalo de Wald.
O intervalo $(1-\alpha)100\%$ Agresti-Coull é

$$\widehat{\pi} - Z_{1-\alpha/2}\sqrt{\frac{\widehat{\pi}(1-\widehat{\pi})}{n+Z^2_{1-\alpha/2}}} < \pi < \widehat{\pi} + Z_{1-\alpha/2}\sqrt{\frac{\widehat{\pi}(1-\widehat{\pi})}{n+Z^2_{1-\alpha/2}}}.$$

O intervalo é essencialmente o intervalo Wald onde $Z^2_{1-\alpha/2}/2$ sucessos e $Z^2_{1-\alpha/2}/2$ falhas são adicionados aos dados observados. Especificamente, para $\alpha = 0.05$, isso significa que cerca de dois sucessos e duas falhas são adicionados porque $Z_{1-0.05/2} = 1.96 \approx 2$. Semelhante ao intervalo Wald, esse intervalo tem a propriedade indesejável que pode ter limites menores que 0 ou maiores que 1.

Intervalos de Wilson e Agresti-Coull.

Suponha novamente que $w = 4$ sucessos sejam observados em $n = 10$ tentativas. Para um intervalo de confiança de 95%, a estimativa ajustada de π é

$$\widehat{\pi} = \frac{w + Z^2_{1-\alpha/2}/2}{n + Z^2_{1-\alpha/2}} = \frac{4 + 1.96^2/2}{10 + 1.96^2} = 0.4278.$$

Os limites do intervalo de 95

$$\widehat{\pi} \pm Z_{1-\alpha/2}\sqrt{\frac{\widehat{\pi}(1-\widehat{\pi})}{n + Z^2_{1-\alpha/2}}} = 0.4278 \pm \sqrt{\frac{0.4278(1-0.4278)}{10 + 1.96^2}}$$

levando a um intervalo de $0.1671 < \pi < 0.6884$. Ambos os intervalos de confiança têm limites bastante semelhantes neste caso, mas são bastante diferentes dos limites do intervalo de Wald de (0.0964; 0.7036) que calculamos mais cedo.

Intervalos de Wilson e Agresti-Coull.

```
p.tilde <- (w + qnorm(p = 1- alpha/2)^2/2) / (n + qnorm(p = 1- alpha/2)^2)
p.tilde
```

```
## [1] 0.4277533
```

I.C. Wilson

```
round(p.tilde + qnorm(p = c( alpha/2, 1- alpha/2) ) * sqrt(n)/(n + qnorm(p = 1- alpha/2)^2)
      * sqrt(pi.hat*(1 - pi.hat) + qnorm (p = 1- alpha/2)^2/(4*n)), 4)
```

```
## [1] 0.1682 0.6873
```

I.C. Agresti - Coull

```
var.ac <- p.tilde*(1 -p.tilde ) / (n + qnorm(p = 1- alpha/2)^2)
round(p.tilde + qnorm(p = c( alpha/2, 1- alpha/2) ) * sqrt(var.ac), 4)
```

```
## [1] 0.1671 0.6884
```

Após calcular $\widehat{\pi}$, calculamos os intervalos de Wilson e Agresti-Coull através de uma linha de código para cada um. Observe que a execução de parte de uma linha de código pode ajudar a destacar como ela funciona. A função binom.confint() do pacote binom pode ser usada para simplificar os cálculos. Observe que este pacote não está na instalação padrão do R, portanto, ele precisa ser instalado antes de seu uso.

```
library( package = binom )
binom.confint(x = w, n = n, conf.level = 1-alpha , methods = "all")
```

```
##           method x  n      mean      lower     upper
## 1  agresti-coull 4 10 0.4000000 0.16711063 0.6883959
## 2     asymptotic 4 10 0.4000000 0.09636369 0.7036363
## 3          bayes 4 10 0.4090909 0.14256735 0.6838697
## 4        cloglog 4 10 0.4000000 0.12269317 0.6702046
## 5          exact 4 10 0.4000000 0.12155226 0.7376219
## 6          logit 4 10 0.4000000 0.15834201 0.7025951
## 7         probit 4 10 0.4000000 0.14933907 0.7028372
## 8        profile 4 10 0.4000000 0.14570633 0.6999845
## 9            lrt 4 10 0.4000000 0.14564246 0.7000216
## 10     prop.test 4 10 0.4000000 0.13693056 0.7263303
## 11        wilson 4 10 0.4000000 0.16818033 0.6873262
```

A função calcula 11 intervalos diferentes para quando o argumento métodos = "todos" é usado. O primeiro, segundo e décimo primeiro intervalos são os intervalos de Agresti-Coull, Wald e Wilson, respectivamente. Consulte a ajuda da função para obter mais informações sobre os outros intervalos.

Testes

Quando apenas um parâmetro simples é de interesse, como aqui, geralmente preferimos intervalos de confiança em vez de testes de hipóteses, porque o intervalo fornece uma faixa de valores de parâmetros possíveis. Normalmente, podemos inferir que um valor hipotético para um parâmetro pode ser rejeitado se não estiver dentro do intervalo de confiança para o parâmetro. No entanto, existem situações em que um valor fixo conhecido de π, digamos π_0, é de interesse especial, levando a uma hipótese formal de teste de $H_0 : \pi = \pi_0$ vs. $H_1 : \pi \neq \pi_0$.
Com relação ao intervalo de Wilson, notou-se que a estatística do teste escore

$$Z_0 = \frac{\widehat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}},$$

é frequentemente utilizado nestas situações.

Quando a hipótese nula é verdadeira, Z_0 deve ter aproximadamente uma distribuição normal padrão, onde a aproximação geralmente é melhor para amostras maiores.

A hipótese nula é rejeitada quando um valor incomum de Z_0 é observado em relação a esta distribuição, ou seja, algo menor que $-Z_{1-\alpha/2}$ ou maior que $Z_{1-\alpha/2}$. O *p*-valor é uma medida de quão extremo o valor da estatística de teste é em relação ao que é esperado quando H_0 é verdadeiro.

Este *p*-valor é calculado como $2P(Z > |Z_0|)$ onde Z tem uma distribuição normal padrão. Observe que este teste é equivalente a rejeitar a hipótese nula quando π_0 está fora do intervalo de Wilson. Se desejado, a função prop.test() pode ser usada para calcular Z_0, Z_0^2 é realmente dado e um *p*-valor correspondente.

Recomendamos usar o teste escore ao realizar um teste para π. No entanto, existem procedimentos de teste alternativos. Em particular, o teste de razão de verossimilhança (LRT) é uma maneira geral de realizar testes de hipóteses e pode ser usado aqui para testar. Informalmente, a estatística LRT é

$$\Lambda = \frac{\text{Máximo da função de verossimilhança sob } H_0}{\text{Máximo da função de verossimilhança sob } H_0 \text{ ou } H_1}.$$

Para o teste específico de $H_0 : \pi = \pi_0$ vs. $H_1 : \pi \neq \pi_0$, o denominador é $\widehat{\pi}^w(1-\widehat{\pi})^{n-w}$, porque o valor máximo possível da função de verossimilhança ocorre quando é avaliada no MLE. O numerador é $\pi_0^w(1-\pi_0)^{n-w}$ porque existe apenas um valor possível da função de verossimilhança se a hipótese nula for verdadeira.

A estatística transformada $-2\log(\Lambda)$ acaba por ter uma distribuição aproximada de χ^2_1 em grandes amostras se a hipótese nula for verdadeira. Para este teste, a estatística transformada pode ser reexpressa como

$$-2\log(\Lambda) = -2\left(w\log\left(\pi_0/\widehat{\pi}\right) + (n-w)\log\left((1-\pi_0)/(1-\widehat{\pi})\right)\right).$$

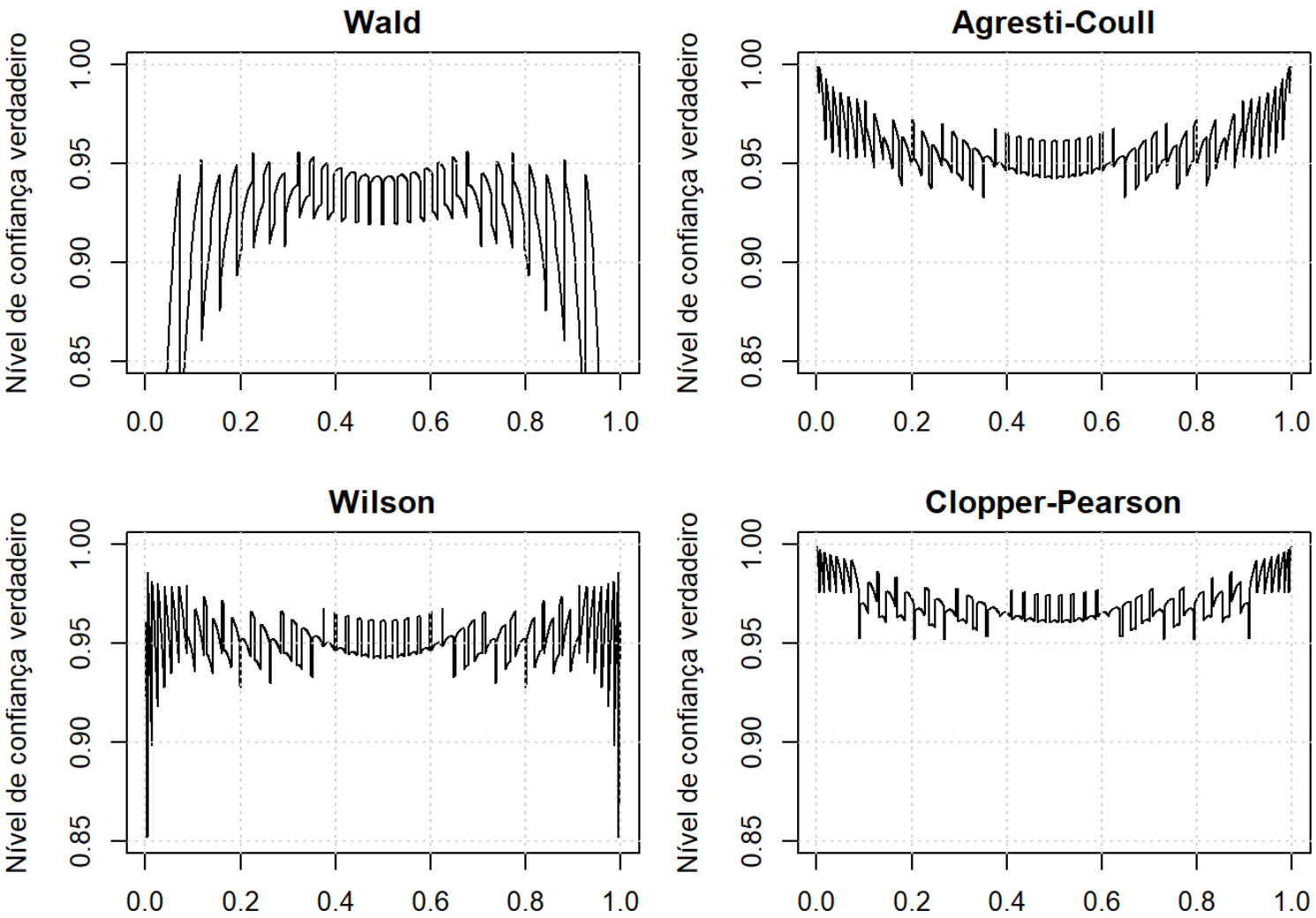
Rejeitamos a hipótese nula se $-2\log(\Lambda) > \chi^2_{1,1-\alpha/2}$, onde $\chi^2_{1,1-\alpha/2}$ é o quantil $1-\alpha/2$ de uma distribuição qui-quadrado com 1 grau de liberdade, por exemplo, $\chi^2_{1,0.095} = 3.84$ quando $\alpha = 0.05$.

O *p*-valor é $P(A > -2\log(\lambda))$ onde A tem uma distribuição χ^2_1.

Conforme discutido, um método de intervalo de confiança pode não atingir seu nível de confiança declarado. As razões para isso são explicadas brevemente. A figura abaixo fornece uma comparação dos níveis de confiança verdadeiros para os intervalos de Wald, Wilson, Agresti-Coull e Clopper-Pearson.

Para cada gráfico, *n* é 40 e o nível de confiança declarado é 0.95, α=0.05. O verdadeiro nível de confiança (cobertura) para cada método de intervalo é apresentado em função de π. Por exemplo, o nível de confiança verdadeiro em π=0.157 é 0.8760 para o Wald, 0.9507 para o Wilson, 0.9507 para o Agresti-Coull e 0.9740 para os intervalos de Clopper-Pearson, respectivamente.

Obviamente, nenhum desses intervalos atinge exatamente o nível de confiança declarado de forma consistente.


Wald
Agresti-Coull
Wilson
Clopper-Pearson
Nível de confiança verdadeiro
0.85
0.90
0.95
1.00
0.0
0.2
0.4
0.6
0.8
1.0

Abaixo estão algumas conclusões gerais do exame destes gráficos:

- O intervalo de Wald tende a ser o mais distante de 0.95 com mais frequência. Na verdade, o verdadeiro nível de confiança é muitas vezes muito baixo para estar no gráfico em valores extremos de π.
- O intervalo Agresti-Coull faz um trabalho muito melhor do que o Wald com seu verdadeiro nível de confiança geralmente entre 0.93 e 0.98. Para valores próximos de 0 ou 1, o intervalo pode ser muito conservador.
- O intervalo de Wilson tem um desempenho um pouco melhor do que o intervalo de Agresti-Coull com seu verdadeiro nível de confiança geralmente entre 0.93 e 0.97; no entanto, para π muito extremo, pode ser muito liberal.

- Esse desempenho para π extremo pode ser melhorado alterando o limite inferior do intervalo para $-\log(1-\alpha)/n$ se $w = 1$ e o limite superior do intervalo para $1 + \log(1-\alpha)/n$ se $w = n - 1$; veja pág. 112 de Brown et al. (2001) para justificação. Esta pequena modificação foi usada pela função binom.confint() no passado (versão 1.0-5 do pacote), mas agora não é mais implementada a partir da versão 1.1-1 do pacote.
- 4- O intervalo de Clopper-Pearson tem um nível de confiança verdadeiro igual ou superior ao nível declarado, onde geralmente oscila entre 0.95 e 0.98. Para valores próximos de 0 ou 1, o intervalo pode ser muito conservador.

Achados semelhantes podem ser mostrados para outros valores de n e α. Por que esses gráficos na figura acima têm padrões tão estranhos?

É tudo por causa da discrição de uma variável aleatória binomial. Para um dado n, existem apenas $n + 1$ intervalos possíveis que podem ser formados, um para cada valor de $w = 0, 1, \cdots, n$. Para um valor específico de π, alguns desses intervalos contêm π e outros não.

Assim, o verdadeiro nível de confiança em π, digamos $C(\pi)$, é a soma das probabilidades binomiais para todos os intervalos que contêm π:

$$C(\pi) = \sum_{w=0}^{n} I(w) \binom{n}{w} \pi^w (1-\pi)^{n-w}.$$

Cada uma dessas probabilidades binomiais muda lentamente à medida que muda π. Contanto que não ultrapassemos π em nenhum limite de intervalo, o verdadeiro nível de confiança também muda lentamente.

No entanto, assim que π ultrapassa um limite de intervalo, uma probabilidade é subitamente adicionada ou subtraída do nível de confiança verdadeiro, resultando nos picos que aparecem em todas as partes da figura acima.

Ilustramos encontrar o verdadeiro nível de confiança e quando esses picos ocorrem no próximo exemplo.

Nível de confiança verdadeiro para o intervalo de Wald.

Mostramos neste exemplo como calcular um nível de confiança verdadeiro para o intervalo de Wald com $n = 40$, $\pi = 0.157$ e $\alpha = 0.05$.

Segue abaixo a descrição do processo:

- Encontrar a probabilidade de obter cada valor possível de w usando a função dbinom() com n = 40 e $\pi = 0.157$,
- Calcular o intervalo de confiança de Wald de 95
- Somar as probabilidades correspondentes aos intervalos que contêm $\pi = 0.157$; este é o verdadeiro nível de confiança.

```
pi <- 0.157
alpha <- 0.05
n <- 40
w <- 0:n
pi.hat <- w/n
pmf <- dbinom (x = w, size = n, prob = pi)
var.wald <- pi.hat *(1 - pi.hat)/n
lower <- pi.hat - qnorm (p = 1- alpha /2) * sqrt(var.wald )
upper <- pi.hat + qnorm (p = 1- alpha /2) * sqrt(var.wald )
save <- ifelse ( test = pi >lower , yes = ifelse ( test = pi <upper, yes = 1, no = 0) , no = 0)
data.frame (w, pi.hat , round ( data.frame(pmf , lower , upper ) ,4), save ) [1:13 ,]
```

```
##     w pi.hat    pmf   lower  upper save
## 1   0  0.000 0.0011  0.0000 0.0000    0
## 2   1  0.025 0.0080 -0.0234 0.0734    0
## 3   2  0.050 0.0292 -0.0175 0.1175    0
## 4   3  0.075 0.0689 -0.0066 0.1566    0
## 5   4  0.100 0.1187  0.0070 0.1930    1
## 6   5  0.125 0.1591  0.0225 0.2275    1
## 7   6  0.150 0.1729  0.0393 0.2607    1
## 8   7  0.175 0.1564  0.0572 0.2928    1
## 9   8  0.200 0.1201  0.0760 0.3240    1
## 10  9  0.225 0.0795  0.0956 0.3544    1
## 11 10  0.250 0.0459  0.1158 0.3842    1
## 12 11  0.275 0.0233  0.1366 0.4134    1
## 13 12  0.300 0.0105  0.1580 0.4420    0
```

```
sum( save*pmf )
```

```
## [1] 0.875905
```

```
sum( dbinom (x = 4:11 , size = n, prob = pi))
```

```
## [1] 0.875905
```

Nível de confiança verdadeiro para o intervalo de Wald.

O código acima permite calcular um intervalo para cada valor possível de *w* em vez de um intervalo para apenas um.

Uma nova parte dentro do código é a função ifelse(). Esta função faz uma verificação lógica para saber se está ou não dentro de cada um dos 41 intervalos. Por exemplo, o segundo intervalo é (-0.0234; 0.0734), que não contém $\pi = 0.157$, portanto, o objeto de save tem um valor de 0 para seu segundo elemento.

Nível de confiança verdadeiro para o intervalo de Wald.

Observe que o limite superior do intervalo em $w = 3$ mal contém $\pi = 0.157$ e $P(W = 3) = 0.0689$. Usando o mesmo código com a mudança de pi <- 0.156, o limite superior em $w = 3$ agora contém $\pi = 0.156$, de modo que $P(W = 3) = 0.0706$ é incluído ao somar probabilidades para o verdadeiro nível de confiança.

No geral, $w = 3$ a 11 agora têm intervalos de confiança que contêm $\pi = 0.156$, levando a um nível de confiança verdadeiro de 0.9442! Isso demonstra o que mencionamos anteriormente como a causa dos picos na figura acima.

Em problemas simples como este, podemos determinar exatamente as probabilidades de cada intervalo que contém um dado π, de modo que os gráficos como na figura acima possam ser feitos exatamente. Em outros casos, podemos ter que confiar na simulação de Monte Carlo.

Exploramos a abordagem de simulação a seguir para nos permitir comparar um nível de confiança real exato com um estimado por simulação. Isso será útil mais adiante no texto, quando o método de simulação for o único método de avaliação disponível.

Nível de confiança verdadeiro estimado para o intervalo de Wald.

Suponha novamente que $n = 40$, $\pi = 0.157$ e $\alpha = 0.05$.

Abaixo está uma descrição do processo para estimar o verdadeiro nível de confiança por meio de simulação:

- Simule 1.000 amostras usando a função rbinom() com n = 40 e π= 0.157,
- Calcule o intervalo de confiança de Wald de 95
- Calcule a proporção de intervalos que contêm π= 0.157; este é o nível de confiança verdadeiro estimado.

```r
numb.bin.samples <- 1000 # Binomial samples of size n
set.seed (4516)
w <- rbinom (n = numb.bin.samples, size = n, prob = pi)
pi.hat <- w/n
var.wald <- pi.hat *(1 - pi.hat)/n
lower <- pi.hat - qnorm (p = 1- alpha /2) * sqrt (var.wald )
upper <- pi.hat + qnorm (p = 1- alpha /2) * sqrt (var.wald )
data.frame(lower , upper ) [1:10 ,]
```

```
##           lower     upper
## 1   0.039344453 0.2606555
## 2   0.039344453 0.2606555
## 3   0.057249138 0.2927509
## 4   0.076040994 0.3239590
## 5   0.076040994 0.3239590
## 6   0.039344453 0.2606555
## 7   0.076040994 0.3239590
## 8  -0.006624323 0.1566243
## 9   0.022511030 0.2274890
## 10  0.007030745 0.1929693
```

```r
save <- ifelse ( test = pi >lower, yes = ifelse ( test = pi < upper, yes = 1, no = 0) , no = 0)
save[1:10]
```

```
##  [1] 1 1 1 1 1 1 1 0 1 1
```

```r
mean( save )
```

```
## [1] 0.878
```

Nível de confiança verdadeiro estimado para o intervalo de Wald.

Novamente, estamos usando muito do mesmo código do passado. A função ifelse() é usada para verificar se π= 0.157 está dentro de cada um dos intervalos. Por exemplo, vemos que a amostra 8 resulta em $\widehat{\pi}$ = 0.075 e um intervalo de (-0.0066; 0.1566), que não contém 0.157, então o valor correspondente de save é 0.

A média de todos os 0's e 1 no save é 0.878. Este é o nosso nível de confiança verdadeiro estimado para o intervalo de Wald em n = 40 e π= 0.157. Neste problema de simulação relativamente simples, já sabemos que os intervalos para $w = 4, \cdots, 11$ contém π= 0.157 enquanto os outros não. Para ver que a simulação está, de fato, estimando $P(4 \leq W \leq 11)$, a função table() é usada para calcular o número de vezes que cada w ocorre:

```
counts <- table (w)
counts
```

```
## w
##   1   2   3   4   5   6   7   8   9  10  11  12  13
##   8  35  64 123 147 165 172 123  76  46  26  11   4
```

```
sum( counts[4:11]) / numb.bin.samples
```

```
## [1] 0.878
```

Nível de confiança verdadeiro estimado para o intervalo de Wald.

Por exemplo, havia 64 das 1.000 observações que resultaram em um $w = 3$. Isso é muito semelhante ao $P(W = 3) = 0.0689$ que obtivemos para o exemplo anterior. Somando as contagens para $w = 4, \cdots, 11$ e dividindo por 1000, obtemos a mesma estimativa de 0.878 para o nível de confiança verdadeiro.

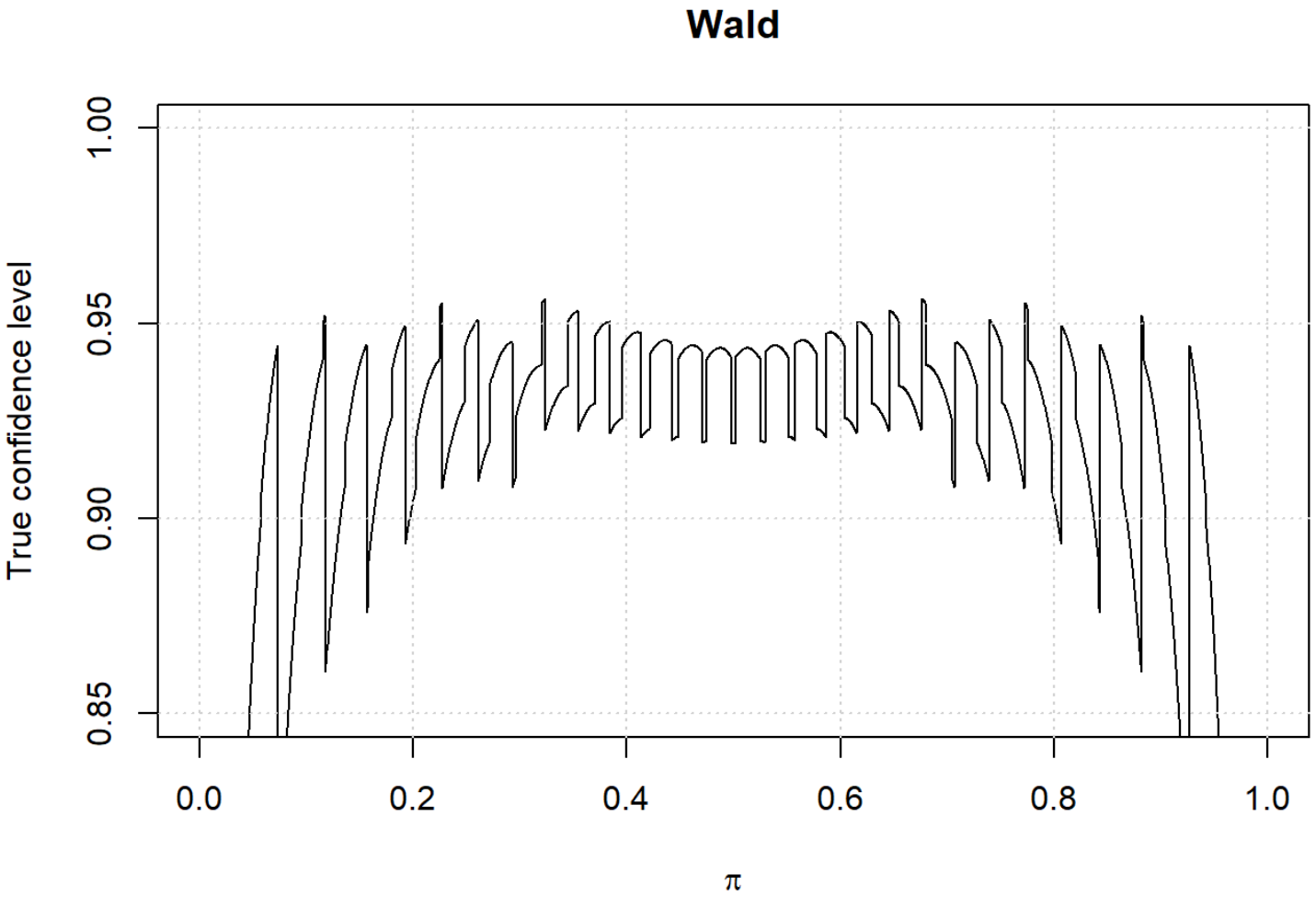
Nível de confiança verdadeiro estimado para o intervalo de Wald.

A estimativa do nível de confiança verdadeiro aqui é quase igual ao nível de confiança verdadeiro encontrado no exemplo anterior.

Devido ao uso de um grande número de amostras, a lei dos grandes números garante que isso aconteça. Poderíamos até chegar a encontrar um intervalo de confiança para o verdadeiro nível de confiança! Para este caso, temos 878 “sucessos” em 1.000 “ensaios”.

Um intervalo de Wilson de 95% para o próprio nível de confiança verdadeiro é (0.8563; 0.8969), que por acaso contém 0.8759, o nível de confiança verdadeiro conhecido.

```r
alpha <- 0.05
n <- 40
w <- 0:n
pi.hat <- w/n
pi.seq <- seq( from = 0.001 , to = 0.999 , by = 0.0005)
# Wald
var.wald <- pi.hat*(1 - pi.hat)/n
lower.wald <- pi.hat - qnorm (p = 1- alpha /2) * sqrt (var.wald )
upper.wald <- pi.hat + qnorm (p = 1- alpha /2) * sqrt (var.wald )
# Save true confidence levels in a matrix
save.true.conf <- matrix ( data = NA , nrow = length (pi.seq), ncol = 2)
# Create counter for the loop
counter <- 1
# Loop over each pi
for(pi in pi.seq) {
  pmf <- dbinom (x = w, size = n, prob = pi)
  save.wald <- ifelse ( test = pi > lower.wald, yes = ifelse ( test = pi < upper.wald, yes = 1, no = 0), no = 0)
  wald <- sum( save.wald * pmf)
  save.true.conf[ counter ,] <- c(pi , wald )
  # print ( save . true . conf [ counter ,])
  counter <- counter +1
}
plot (x = save.true.conf[,1] , y = save.true.conf[,2], main = "Wald", xlab = expression(pi),
      ylab = "True confidence level", type = "l", ylim = c (0.85 ,1) )
abline (h = 1-alpha , lty = "dotted")
grid()
```


Wald
True confidence level
1.00
0.95
0.90
0.85
0.0
0.2
0.4
0.6
0.8
1.0
π

Criamos um vetor pi.seq que é uma sequência de números de 0.001 a 0.999 por 0.0005. O código de função for(pi in pi.seq) (frequentemente chamado de “loop for”) instrui R a tirar um valor de pi.seq por vez. O código entre chaves, então, encontra o verdadeiro nível de confiança para π.

O objeto save.true.conf é uma matriz criada para ter 1.997 linhas e 2 colunas. A princípio, todos os seus valores são inicializados como "NA" dentro de R. Seus valores são atualizados uma linha por vez inserindo o valor de e o nível de confiança verdadeiro. Finalmente, o objeto counter nos permite alterar o número da linha de save.true.conf dentro do loop.

Após o loop for, usamos a função plot() para plotar o valor de π no eixo *x* e o nível de confiança verdadeiro no eixo *y* usando as colunas apropriadas de save.true.conf. O argumento type = "l" instrui R a construir um gráfico de linha onde cada par de nível de confiança verdadeiro é conectado por uma linha. A função abline() desenha uma linha pontilhada horizontal em 0.95, que é o nível de confiança declarado.

O pacote binom em R também pode ser usado para calcular os verdadeiros níveis de confiança. A função binom.coverage() calcula o nível de confiança verdadeiro para um de cada vez, e a função binom.plot() plota os níveis de confiança verdadeiros em um conjunto de valores diferentes de π.