Análisis de respuestas binarias

MCP

19 de febrero de 2018

Distribución Binomial

Consideremos una variable aleatoria binomial que cuenta el número de éxitos de un experimento repetido $n = 5$ veces, y supongamos que la probabilidad de éxito es $\pi = 0.6$. Se puede calcular la probabilidad de cada número de éxitos $w = 0, 1, 2, 3, 4, 5$. Por ejemplo, la probabilidad de 1 éxito en 5 intentos es

$$P(W = 1) = \binom{5}{1}(0.6)^1(1 - 0.6)^{5-1} = 0.0768$$

En R se usa la función "dbinom()"

```
dbinom( x = 1 , size = 5, prob = 0.6 )
```

```
## [1] 0.0768
```

Podemos encontrar las probabilidades $w = 0, \dots, 5$ cambiando el argumento x

```
dbinom(0:5, 5, 0.6)
```

```
## [1] 0.01024 0.07680 0.23040 0.34560 0.25920 0.07776
```

Para representar los datos de manera m'as descriptiva:

```
prob <- dbinom( x = 0:5 , size = 5 , prob = 0.6)
prob_df <- data.frame( w = 0:5 , prob = round( x = prob , digits = 4) )
prob_df
```

```
##   w   prob
## 1 0 0.0102
## 2 1 0.0768
## 3 2 0.2304
## 4 3 0.3456
## 5 4 0.2592
## 6 5 0.0778
```

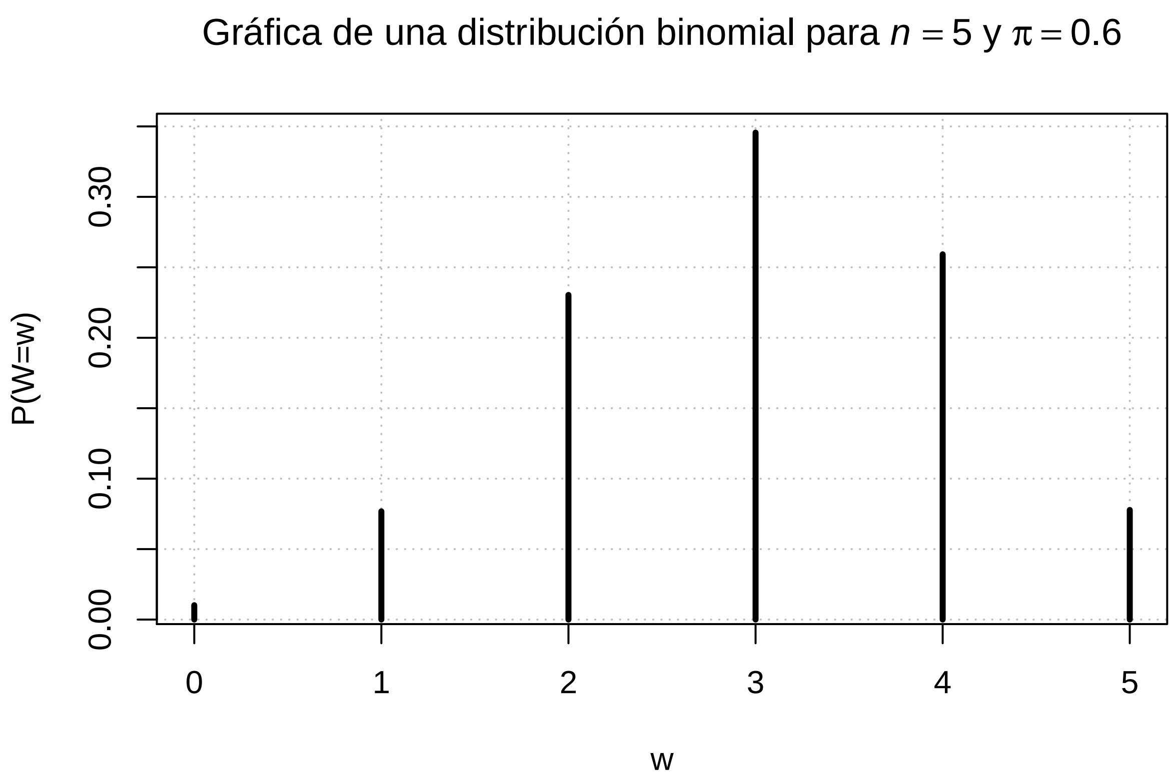
Graficamos:

```
pdf(file = "Figure1.1.pdf", width = 6, height = 6, colormodel = "cmyk")
plot( x = prob_df$w , y = prob_df$prob , type = "h" , xlab = "w" , ylab ="P( W = w )" , main =
        "Gráfica de una dist. binomial para n =5 , pi =0.6" , panel.first = grid( col = "gray"
  , lty = "dotted") , lwd = 3)
abline(h=0)
dev.off()
```

```
## pdf
##   2
```

De forma alterna:

```
plot(x = prob_df$w, y = prob_df$prob, type = "h", xlab = "w", ylab = "P(W=w)", main =
       expression(paste("Gráfica de una distribución binomial para ", italic(n) == 5, " y ",
      italic(pi) == 0.6)), panel.first = grid(col="gray", lty="dotted"), lwd = 3)
```



¿Cuáles son nuestras hipótesis?

La distribución binomial es un modelo rasonable para la distribución de éxitos en un número dado de ensayos siempre y cuando se satisfagan ciertas condiciones, a saber:

1. *Hay n ensayos idénticos.*
   La acción que resulta en el ensayo y la medida tomada deben ser las mismas en cada ensayo.
2. *Existen dos posibles resultados para cada ensayo.*
3. *Los ensayos son independientes unos de otros.* No existe factor alguno en la ejecución de los ensayos que pueda causar que un subconjunto de los ensayos se comporte de manera similar a otro.
4. *La probabilidad de éxito permanece constante para cada ensayo.*
5. *La variable aleatoria de interés W es el número de éxitos.*

Simulación de una muestra binomial

Simularemos 1000 osbervaciones aleatorias de W a partir de una distribución binomial con $\pi = 0.6$ y $n = 5$.

```
set.seed(4848)
bin5<-rbinom(n = 1000, size = 5, prob = 0.6)
bin5[1:10]
```

```
##  [1] 3 2 4 1 3 1 3 3 3 4
```

En la teoría

$$E(W) = n\pi = 5(0.6) = 3$$

$$Var(W) = n\pi(1-\pi) = 5(0.6)(0.4) = 1.2$$

Calculamos media y varianza muestrales:

```r
mean(bin5)
```

```
## [1] 2.991
```

```r
var(bin5)
```

```
## [1] 1.236155
```

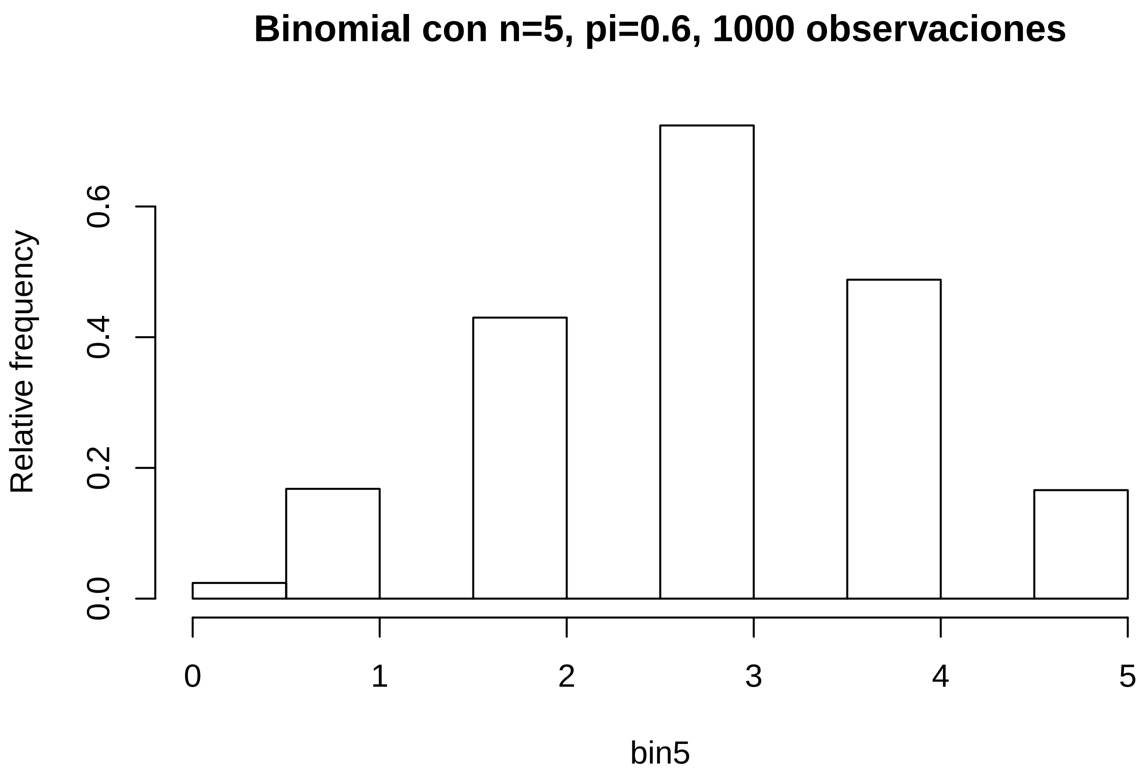
Por supuesto, se esperarían valores más cercanos a los poblacionales con un tamaño de muestra mayor.

Para tratar de ver qué tan bien la distribución observada sigue a la binomial, usamos "table()" para encontrar las frecuencias de cada posible respuesta y luego utilizamos "hist()" para graficar un histograma de frecuencias relativas.

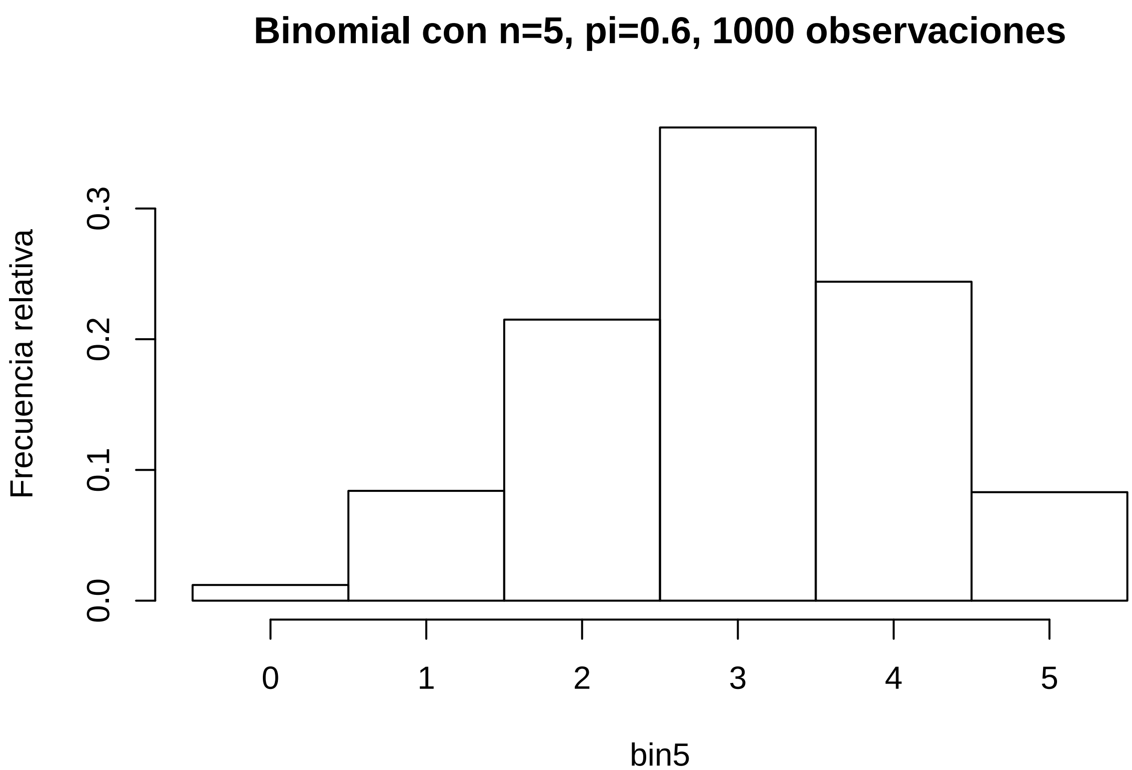
```r
table(x = bin5)
```

```
## x
##   0   1   2   3   4   5
##  12  84 215 362 244  83
```

```r
# Histograma de frecuencias relativas
hist(x = bin5, main = "Binomial con n=5, pi=0.6, 1000 observaciones", probability = TRUE,
     ylab = "Relative frequency")  # La columna de la izquierda no se despliega correctamente
```




```r
hist(x = bin5, main = "Binomial con n=5, pi=0.6, 1000 observaciones", probability = TRUE,
  breaks = c(-0.5:5.5), ylab = "Frecuencia relativa")
```

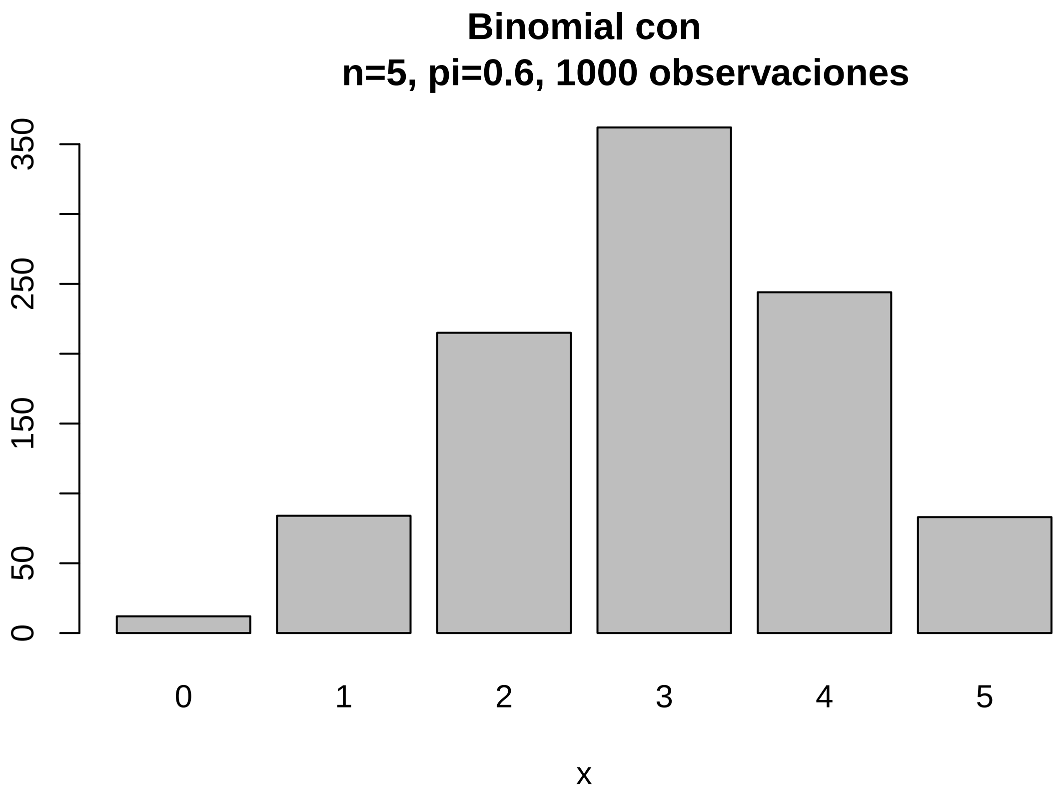


De otra manera

```
save.count<-table(bin5)
save.count
```

```
## bin5
##   0   1   2   3   4   5
##  12  84 215 362 244  83
```

```
barplot(height = save.count, names = c("0", "1", "2", "3", "4", "5"), main = "Binomial con
        n=5, pi=0.6, 1000 observaciones", xlab = "x")
```



Inferencia para la probabilidad de éxito

El objetivo es estimar y hacer inferencias acerca de la probabilidad del parámetro π de la distribución de Bernoulli.

Estimación e inferencia de máxima verosimilitud

La función de verosimilitud es una función de uno o más parámetros condicionados a los datos observados. La función de verosimilitud para π cuando $y_1, \cdots, y_n$ son observaciones de una distribución de Bernoulli es

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

Cuando se registra el número de éxitos en un determinado número de ensayos, la función de verosimilitud para π es simplemente $L(\pi|w) = P(W = w) = \binom{n}{w}\pi^w(1-\pi)^{n-w}$. El valor de π que maximiza la función de verosimilitud es considerado el valor más plausible para el parámetro y es llamado el estimador de máxima verosimilitud (MLE en inglés).

En este caso, el MLE de π es $\hat{\pi} = w/n$, la proporción observada de éxitos. Ya que $\bar{\pi}$ puede variar de muestra a muestra, es un estadístico y tiene su correspondiente distribución de probabilidad. Se puede mostrar que $\hat{\pi}$ tiene una distribución aproximadamente normal para muestras suficientemente grandes. La media es π y la varianza se calcula:

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

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

Intervalo de confianza de Wald

Utilizando esta distribución normal, podemos tratar a $\frac{\hat{\pi}-\pi}{\sqrt{\widehat{Var}(\hat{\pi})}}$ como aproximadamente normal. Por ello, para $0 < \alpha < 1$ se tiene

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

donde Z_α es el α-ésimo cuantil de una distribución normal estándar. Reorganizando términos:

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

Entonces, ahora tenemos una probabilidad aproximada que tiene el parámetro π centrado entre dos estadísticos. Cuando se reemplazan $\hat{\pi}$ y $\widehat{Var}(\hat{\pi})$ con los valores observados de la muestra, se obtiene un intervalo de confianza del $(1-\alpha)100\%$ para π

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

Los intervalos de confianza basados en la aproximación a la normal de los MLE's son llamados "Intervalos de confianza de Wald".

Cuando w está cerca de 0 o n, ocurren dos problemas:

1. Los límites calculados podrían ser menores a 0 o mayores a 1.
2. Cuando w es 0 o 1, $\sqrt{\hat{\pi}(1-\hat{\pi})} = 0$ para $n > 0$. Esto implica que los límites inferior y superior son iguales.

Supongamos que $w = 4$ éxitos son observados en $n = 10$ ensayos. El intervalo de Wald para π es $0.0964 < \pi < 0.7036$.

```r
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
```

O bien

```r
round(pi.hat + qnorm(p = c(alpha/2, 1-alpha/2)) * sqrt(var.wald),4)
```

```
## [1] 0.0964 0.7036
```

Este intervalo es algo extenso, sin embargo da información de un rango para π que posiblemente sea útil en prueba de hipótesis. Por ejemplo, una prueba de $H_0 : \pi = 0.5$ contra $H_a : \pi \neq 0.5$ no rechazaría H_0 puesto que 0.5 se encuentra en este rango. Pero si la prueba fuera $H_0 : \pi = 0.8$ contra $H_a : \pi \neq 0.8$, hay evidencia para rechazar la hipótesis nula.

Intervalo de confianza de Wilson:

Cuando $n < 40$ se suele recomendar usar el intervalo de Wilson, el cual se obtiene a partir del estadístico de prueba

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

el cual es llamado *estadístico de prueba score*; se utiliza frecuentemente en la prueba de $H_0 : \pi = \pi_0$ contra $H_a : \pi \neq \pi_0$, donde $0 < \pi_0 < 1$. Se puede aproximar la distribución de Z_0 a la distribución normal estándar para obtener $P(-Z_{1-\alpha/2} < Z_0 < Z_{1-\alpha/2}) \approx 1 - \alpha$. Ya que el intervalo de Wilson está basado en una prueba "score", comunmente es referido por *intervalo score*.

El intervalo de Wilson de $(1-\alpha)100\%$ es

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

donde

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

Obsérvese que el intervalo de Wilson siempre tiene límites entre 0 y 1.

Intervalo de confianza de Agresti-Coull:

Este intervalo es recomendado cuando $n \geq 40$, la fórmula está dada por:

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

Supongamos también que $w = 4$ éxitos son observados en $n = 10$ ensayos. En R:

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

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

```r
# Intervalo de 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
```

```r
# Intervalo de 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
```

```r
# qnorm(p=c(alpha/2,1-alpha/2))
```

Se puede simplificar todavía más el trabajo haciendo uso de la paquetería "binom":

```r
library(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
```

También se pueden obtener los intervalos uno a la vez y guardarlos en objetos

```
# Intervalo Agresti-Coull
save.ci<-binom.confint(x = w, n = n, conf.level = 1-alpha, methods = "ac")
save.ci
```

```
##          method x  n mean     lower     upper
## 1 agresti-coull 4 10  0.4 0.1671106 0.6883959
```