TP FINAL

Integrantes: Luis Paredes, Franco Gazzola

Collab para probar el codigo rapidamente https://colab.research.google.com/drive/1V3YPi8S0r9cAOL3UgJiguCJ5Kz94keli?usp=sharing

Introducción

Generar una muestra aleatoria de una variable con distribucion Ber(p) de tamaño n (esto se realiza usando la funcion rbinom(n,1,p)). Para esta muestra, calcular las dos estimaciones por intervalos para p de nivel asintotico 0.95, correspondientes a los dos metodos que se mencionan a continuacion. Repetir todo este procedimiento k veces, obteniendo asılk muestras independientes de tamaño n y k intervalos para cada uno de los dos métodos.

Metodo 1

Metodo de nivel asintotico donde se sustituye el valor de p en la varianza por X¯

$$\sum_{i=1}^{n} X_i \sim \text{Binomial}(n, p)$$

$$X_i \sim \text{Bernoulli}(p)$$

$$\mathbb{E}[X_i] = p$$

$$\text{Var}[X_i] = p \cdot (1 - p)$$

$$\text{Var}[\overline{X}] = \overline{X} \cdot (1 - \overline{X})$$

Dado que la esperanza es finita y la varianza también y considerando los xi independientes e identicamente distribuidas por ser muestras aleatorias entonces podemos usar el teorema central de limite.

$$\frac{\sum_{i=1}^{n} X_i - n \cdot \mathbb{E}[X_i]}{\sqrt{n \cdot \text{Var}[X_i]}} = \frac{\sum_{i=1}^{n} X_i - n \cdot p}{\sqrt{n \cdot \overline{X} \cdot (1 - \overline{X})}} \sim Normal(0, 1)$$

Como conozco la distribución y relaciona al parámetro desconocido, la uso como pivote.

$$U = \frac{\sum_{i=1}^{n} X_i - n \cdot p}{\sqrt{n \cdot \overline{X} \cdot (1 - \overline{X})}}$$

Entonces

$$P(a < U < b) = 1 - \alpha$$

Siendo U normal estándar busco el intervalo de confianza mas corto. Aprovechando la simetría de la normal nos quedamos con

$$\mathbb{P}(-z_{1-\alpha/2} \leq U \leq z_{1-\alpha/2}) = 1 - \alpha$$

reemplazando y despejo el parametro desconocido

$$\mathbb{P}(-z_{1-\alpha/2} \leq \frac{\sum_{i=1}^{n} X_i - n \cdot p}{\sqrt{n \cdot \overline{X} \cdot (1 - \overline{X})}} \leq z_{1-\alpha/2}) = 1 - \alpha$$

$$\mathbb{P}(-z_{1-\alpha/2} \cdot \sqrt{n \cdot \overline{X} \cdot (1 - \overline{X})} \leq \sum_{i=1}^{n} X_i - n \cdot p \leq z_{1-\alpha/2} \cdot \sqrt{n \cdot \overline{X} \cdot (1 - \overline{X})}) = 1 - \alpha$$

$$\mathbb{P}(\frac{\sum_{i=1}^{n} X_i - z_{1-\alpha/2} \cdot \sqrt{n \cdot \overline{X} \cdot (1 - \overline{X})}}{n} \leq p \leq \frac{\sum_{i=1}^{n} X_i + z_{1-\alpha/2} \cdot \sqrt{n \cdot \overline{X} \cdot (1 - \overline{X})}}{n}) = 1 - \alpha$$

Entonces el intervalo de confianza nos queda

$$IC(\underset{-}{X}) = (\frac{\sum_{i=1}^{n} X_i - z_{1-\alpha/2} \cdot \sqrt{n\,\bar{x}\,(1-\bar{x})}}{n}, \frac{\sum_{i=1}^{n} X_i + z_{1-\alpha/2} \cdot \sqrt{n\,\bar{x}\,(1-\bar{x})}}{n})$$

Metodo 2

Método de nivel asintótico cuando no se sustituye a p y se calcula los extremos del intervalo.

Presentemos el teorema del capitulo 9 visto en clase

$$\text{Bajo ciertas condiciones muy generales, sea } \Theta(\underline{X}) \text{ un EMV de } \Theta \text{ consistente y sea } q(\theta) \text{ derivable con } \frac{d}{d\theta}q(\theta) \neq 0 \text{ pa}$$
$$\text{Entonces vale que } \sqrt{n} \cdot \sqrt{\mathbb{I}(\theta)} \cdot (\hat{\Theta}_n - \theta) \sim N(0,1)$$

Usando de referencia el **ejemplo 8** del capitulo **PARAMETRIC INTERVAL ESTIMATION** del libro Alexander Mood ([3]) y tomando como pivote el resultado del teorema anterior desarrollamos:

Sabiendo que para una muestra aleatorio $X_1, ..., X_n \overset{iid}{\sim} Ber(p),\ p \in (0,1)$

$$\hat{p}_{EMV} = \bar{X} = \frac{\sum_{i=0}^{n} X_i}{n}$$

$$I(p) = -E[\frac{d^2}{dp^2} Ln(f_p(x))] = \frac{1}{p \cdot (1-p)}$$

entonces el pivote U es

$$U = \sqrt{n} \cdot \sqrt{\frac{1}{p \cdot (1-p)}} \cdot (\bar{X} - p) \sim N(0,1)$$

De nuevo, como U es normal estándar busco el intervalo de confianza mas corto aprovechando la simetría de la normal y reemplazo el pivote obtenemos

$$\mathbb{P}(-z_{1-\alpha/2} \leq \frac{\bar{X} - p}{\sqrt{\frac{p \cdot (1-p)}{n}}} \leq z_{1-\alpha/2}) = 1 - \alpha$$

Al despejar la variable desconocida (p) el intervalo de confianza nos queda

$$IC(\underline{X}) = (\frac{2n\hat{p} + z^2 - z\sqrt{4n\hat{p} + z^2 - 4n\hat{p}^2}}{2(n+z^2)}, \frac{2n\hat{p} + z^2 + z\sqrt{4n\hat{p} + z^2 - 4n\hat{p}^2}}{2(n+z^2)})$$

Longitud esperada y Cubrimiento Empirico

Para el método 1 y el método 2 realizando una única vez el experimento obtenemos los siguientes resultados:

n	p	metodo	ext_izq	ext_der	cubrimiento_emp	long_estimada
20	0.1	1	-0.03147838	0.2314784	1	0.2629568
20	0.1	2	0.02786648	0.3010336	1	0.2731672
20	0.5	1	0.33196777	0.7680322	1	0.4360645
20	0.5	2	0.34208534	0.7418021	1	0.3997168
50	0.1	1	0.02992691	0.2100731	1	0.1801462
50	0.1	2	0.05617600	0.2380482	1	0.1818722
50	0.5	1	0.44319508	0.7168049	1	0.2736098
50	0.5	2	0.44233442	0.7062500	1	0.2639155
100	0.1	1	0.05630871	0.1836913	1	0.1273826
100	0.1	2	0.06999406	0.1981210	1	0.1281269
100	0.5	1	0.47296694	0.6670331	1	0.1940661
100	0.5	2	0.47215390	0.6626670	1	0.1905131

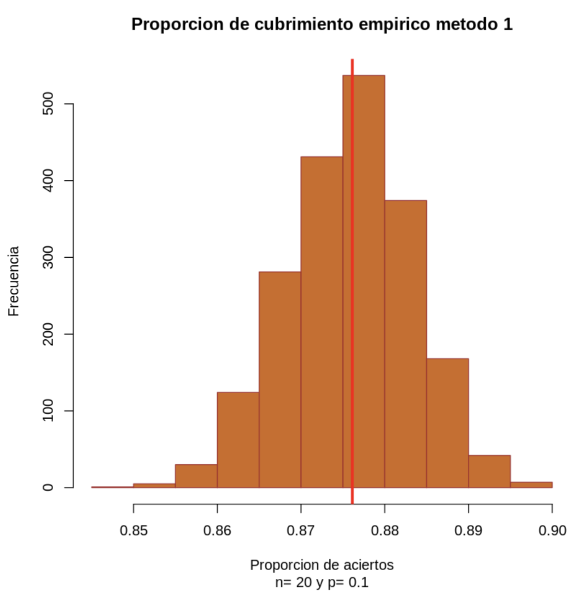
Para cada experimento se obtienen k muestras aleatorias y para cada una se calcula si el intervalo de confianza contiene el valor real de p.
De todas las k muestras aleatorias calculamos la proporcion de veces que el intervalo contenia al verdadero parametro.

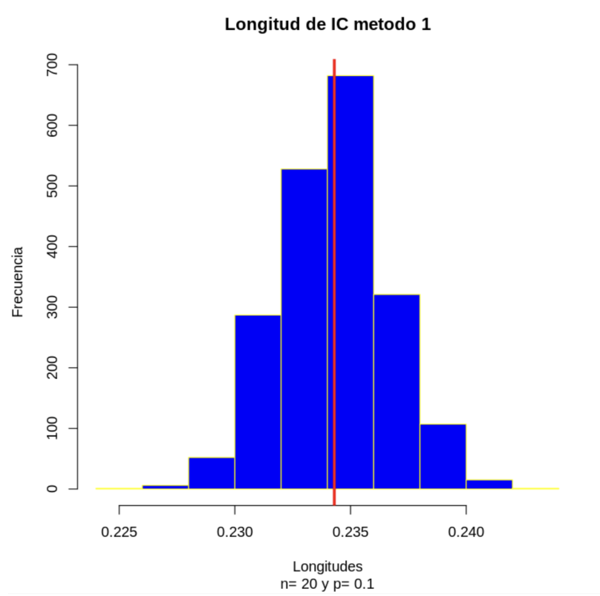
Observamos que cada vez que se volvia a correr el codigo la proporcion cambiaba ligeramente por ser los intervalos de confianza dependientes de la muestra aleatoria.
Para entender bien la proporcion real se corrio el experimento k veces para cada combinacion de n y p para observar hacia que valores tiende la proporcion real y se tomo la media como valor representativo.

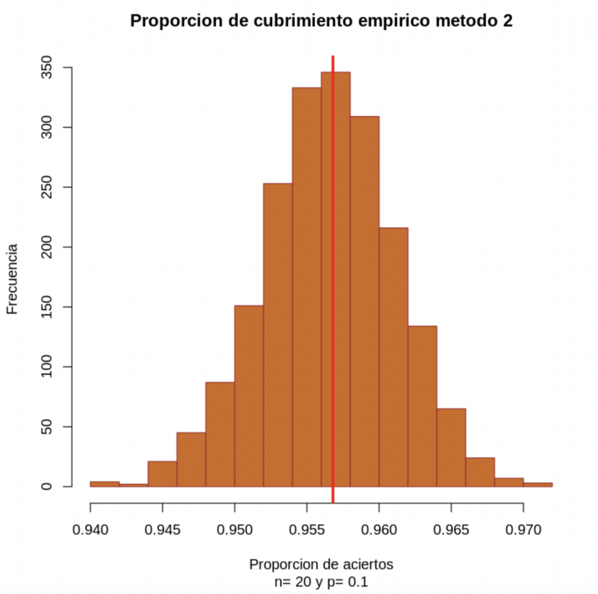
n	p	metodo	cubrimiento_emp	long_estimada
20	0.1	1	0.8761262	0.2342990
20	0.1	2	0.9568030	0.2632072
20	0.5	1	0.9588445	0.4268606
20	0.5	2	0.9588445	0.3927088
50	0.1	1	0.8786607	0.1610425
50	0.1	2	0.9703173	0.1664771
50	0.5	1	0.9351765	0.2743654
50	0.5	2	0.9351765	0.2645931
100	0.1	1	0.9325300	0.1158818
100	0.1	2	0.9365942	0.1176802
100	0.5	1	0.9434352	0.1950103
100	0.5	2	0.9434352	0.1914054

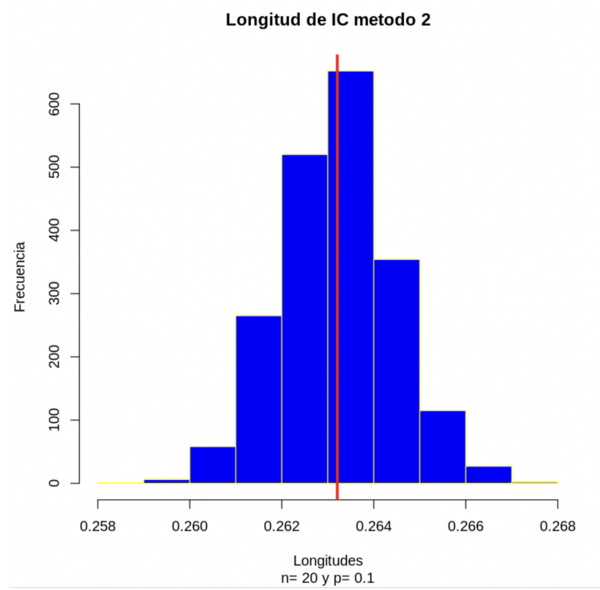
En los siguientes gráficos la linea roja representa el promedio luego de haber realizado k veces cada experimento.

Cubrimiento empírico con ambos metodos siendo n = 20, p = 0.1

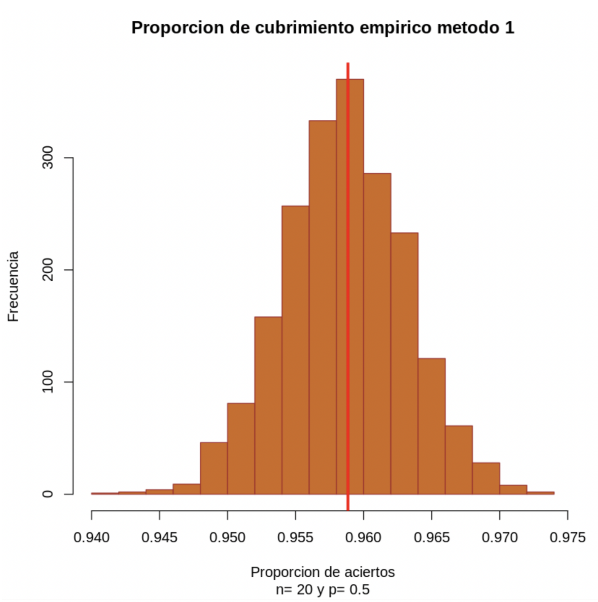


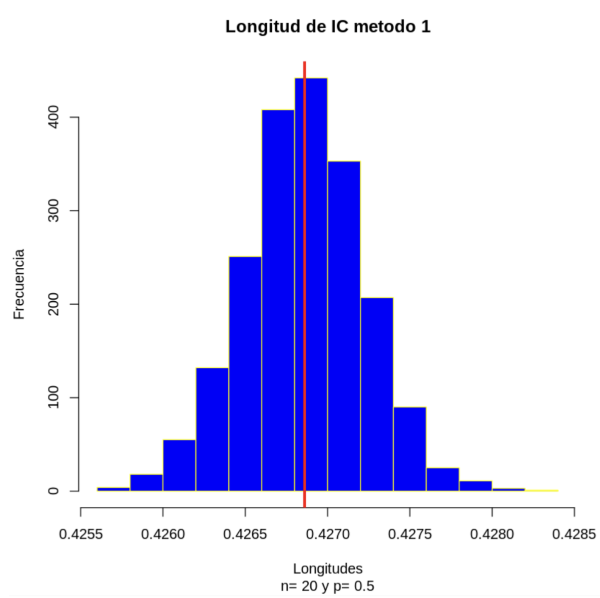


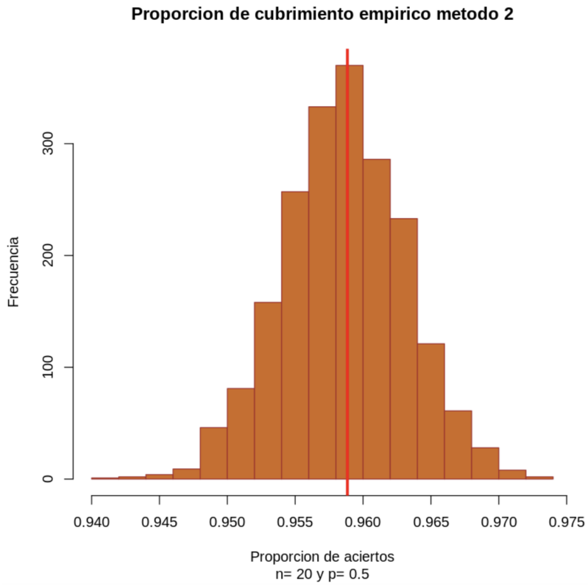


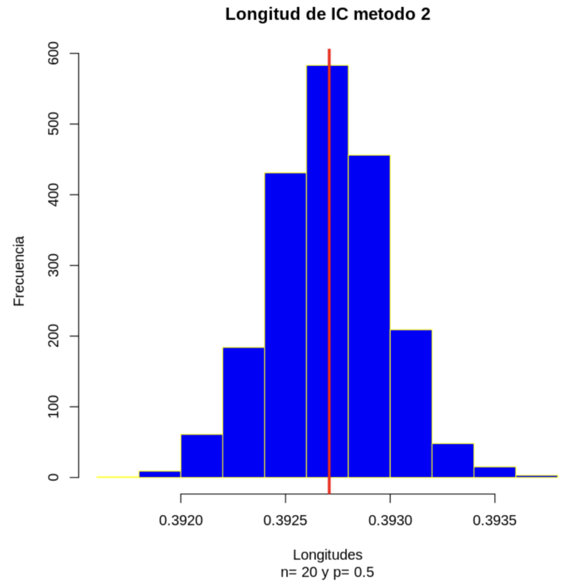


Cubrimiento empirico con ambos metodos siendo n = 20, p = 0.5

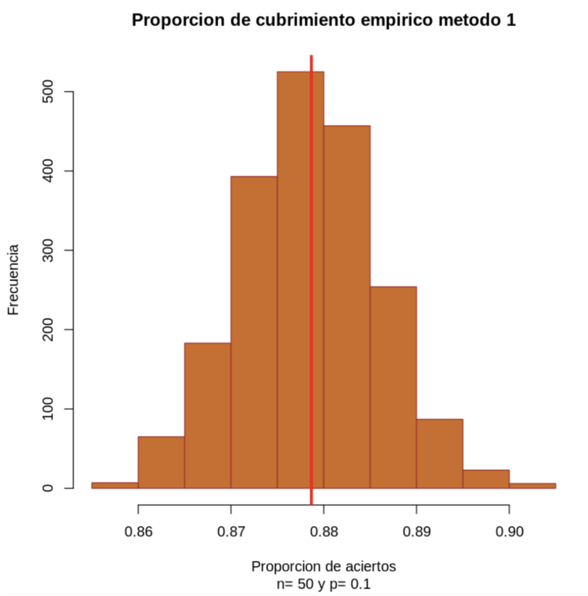


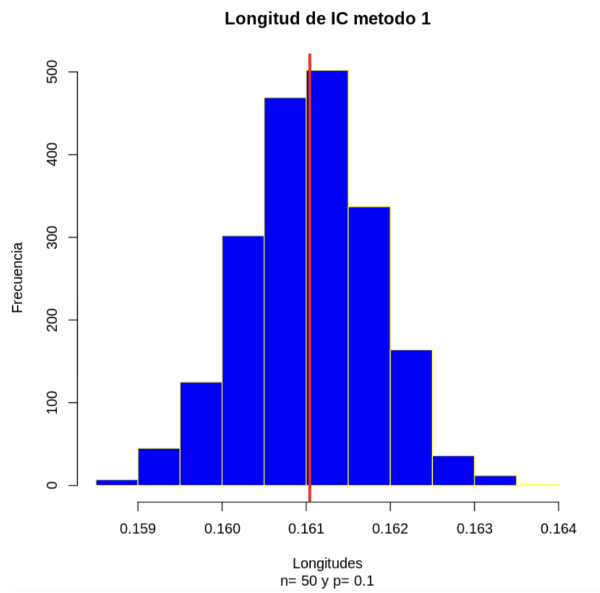


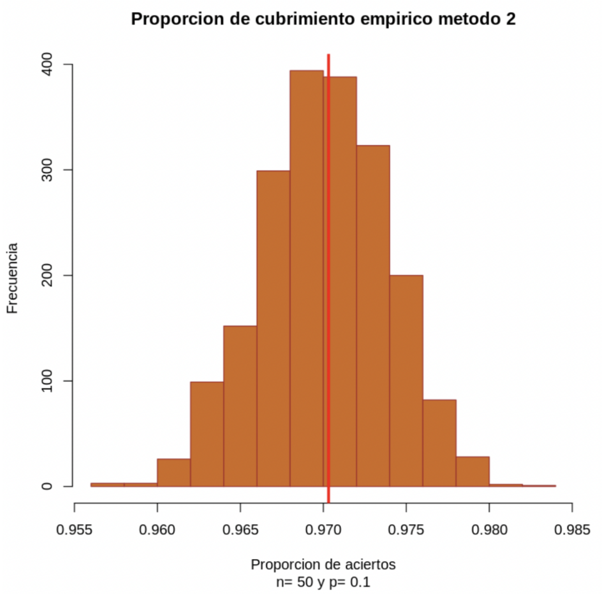


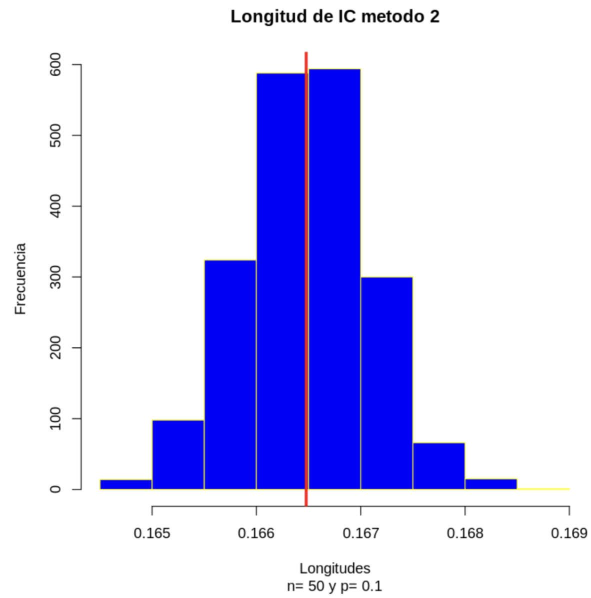


Cubrimiento empirico con ambos metodos siendo n = 50, p = 0.1

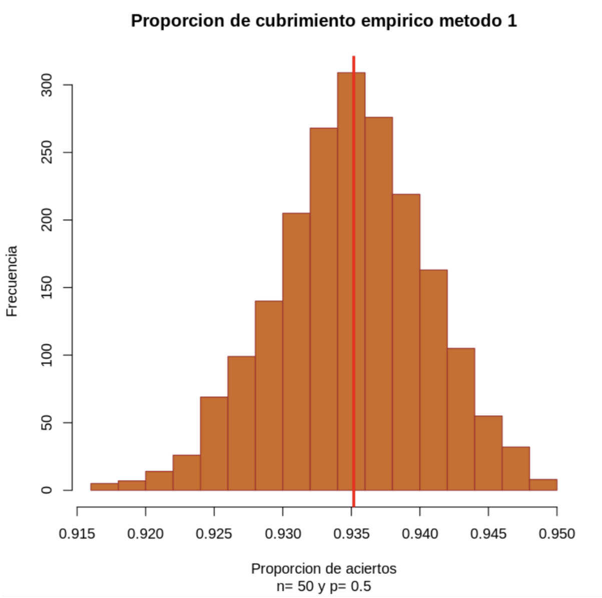


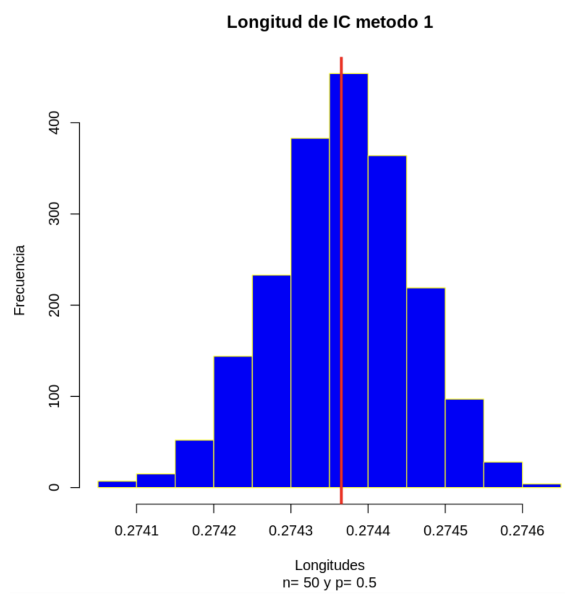


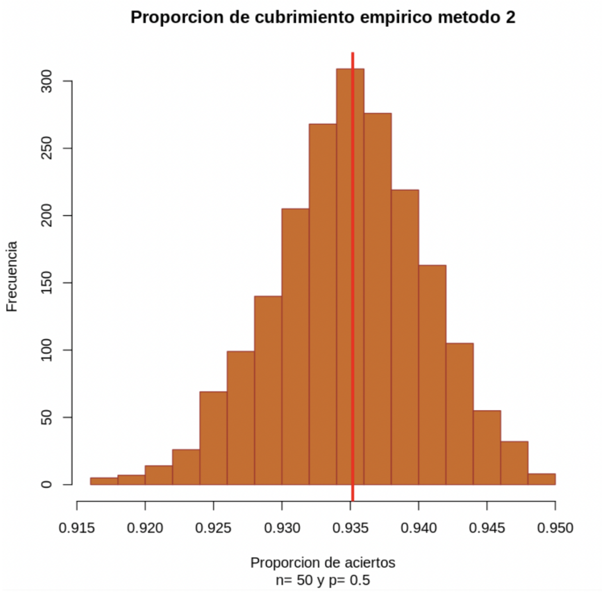


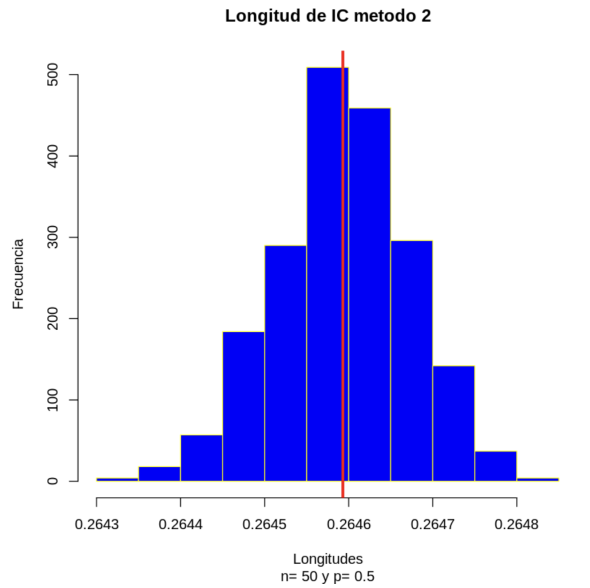


Cubrimiento empirico con ambos metodos siendo n = 50, p = 0.5

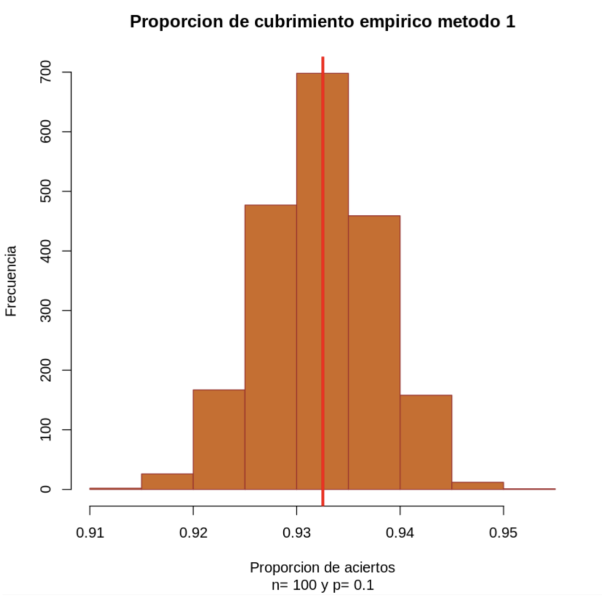


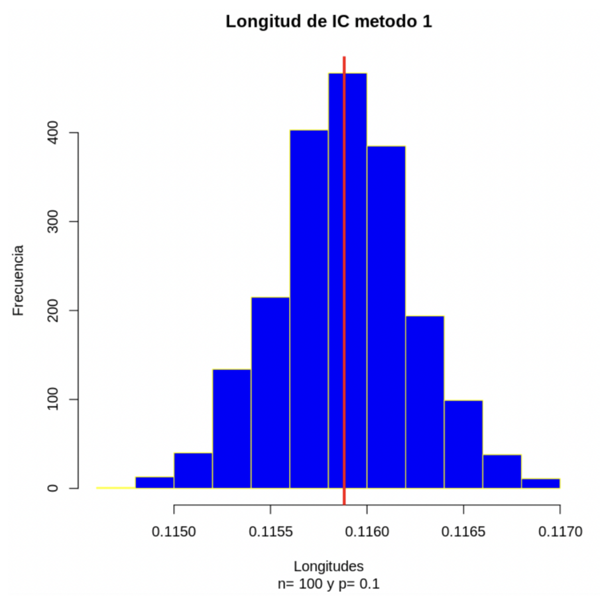


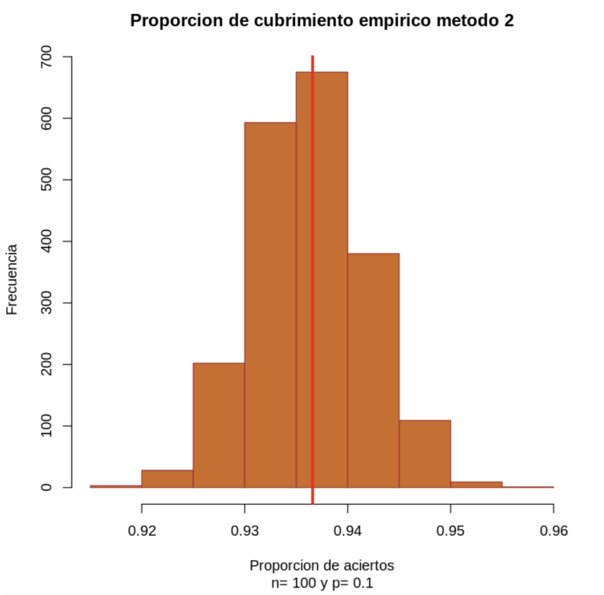


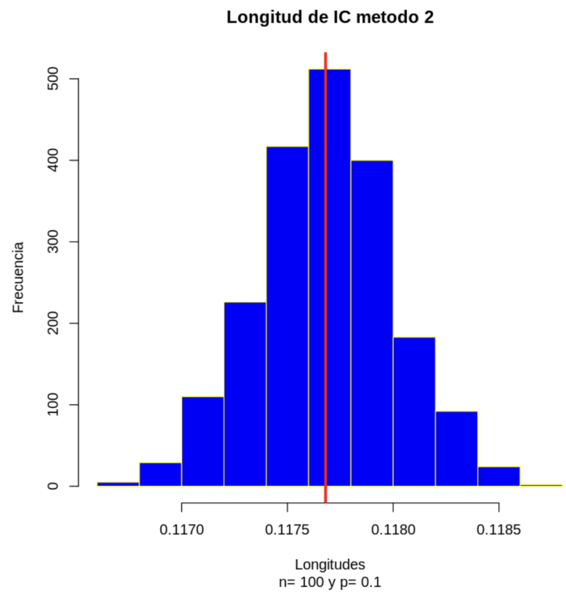


Cubrimiento empirico con ambos metodos siendo n = 100, p = 0.1

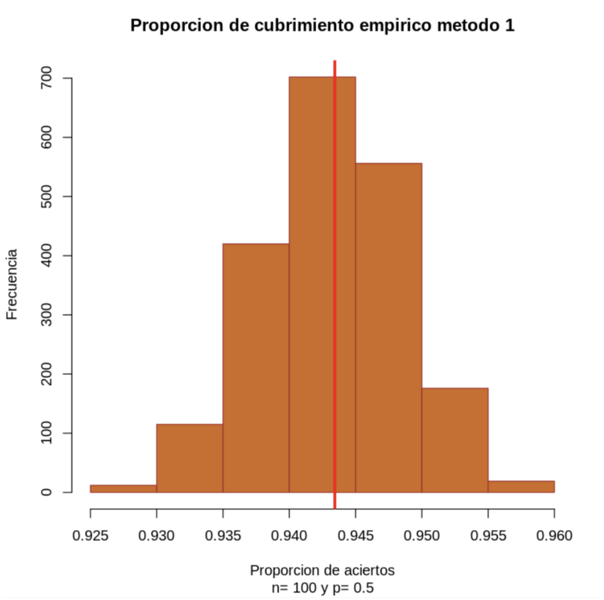


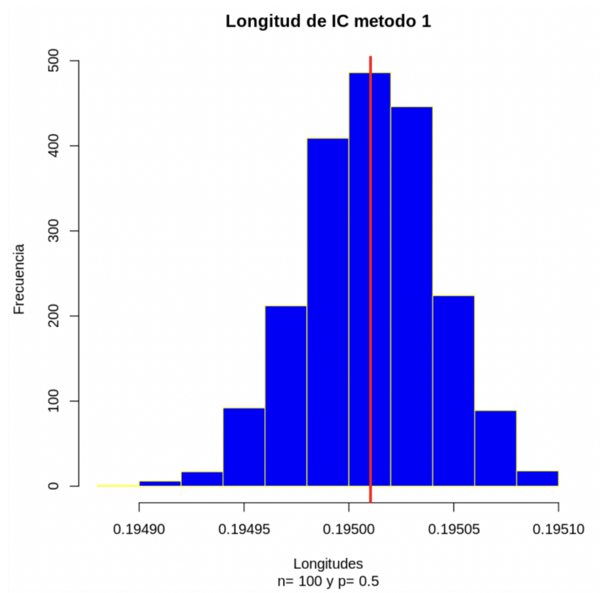


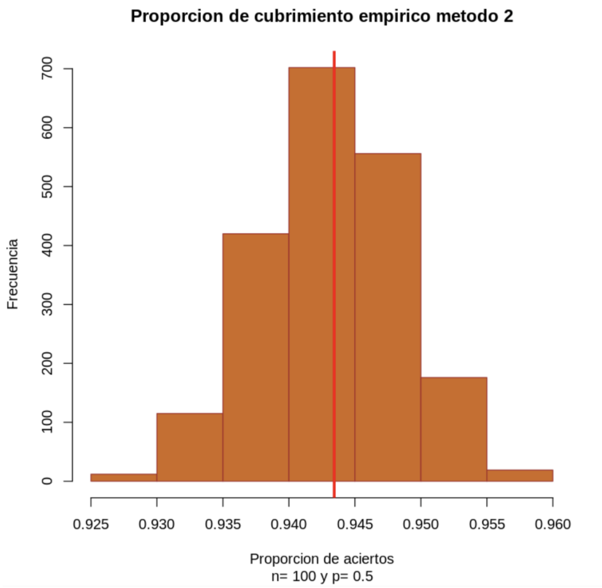


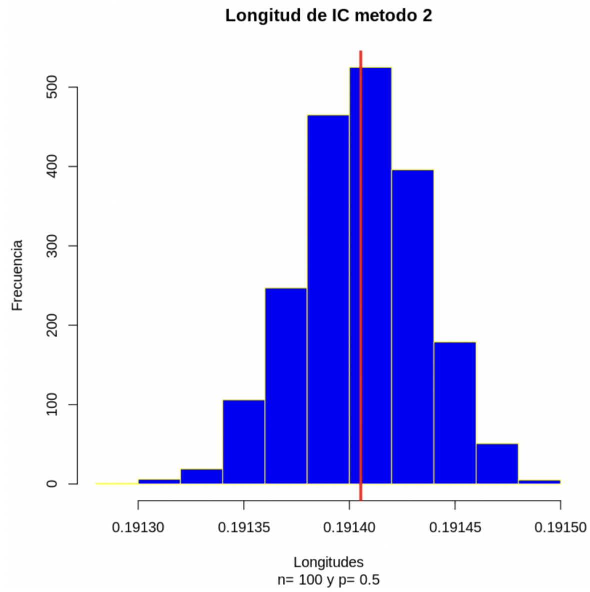


Cubrimiento empirico con ambos metodos siendo n = 100, p = 0.5









Conclusion

Observamos que al ir aumentando la muestra vemos que ambos metodos van mejorando sistematicamente su proporcion de cubrimiento empirico y su intervalo de confianza se va achicando.
Notamos que ambos metodos tienen la misma media para el cubrimiento empirico cuando p=0.5 aunque la longitud sea distinta (por lo general el metodo 2 tiene una longitud mas corta) y cuando p=0.1 la media para el cubrimiento empirico varia entre metodos.

Vemos que con una muestra pequeña (n=20) el metodo 2 tiene un cubrimiento empirico mejor que el metodo 1.
Esto se debe a que para p=0.1 la probabilidad de que salga 1 es baja entonces para una muestra pequeña es posible que salgan pocos 1 y el modelo va a pensar que p es cero y como tambien arrastra error de aproximacion (varianza = media de la muestra) entonces el error de cubrimiento empirico va aumentando.

Referencias

1. Apuntes de Estadistica - Ing Jemina M. Garcia - 2020
2. Probabilidad y Estadistica Elementales para Estudiantes de Ciencias
3. Introduction to the Theory of Statistics (Pag 394-396) - Alexander Mood