Entrega #3: Cobertura en función del parámetro de una distribución de Poisson

Autor: Tomás Notenson

1 de diciembre de 2021

1. Respuestas

Todos los $log()$ en este trabajo representan logaritmos naturales.

Principalmente se definieron las funciones `IsInside1`, `IsInside2`, `IsInside3` (ya estaba definida), y `IsInside4` que reciben como parámetros `nobs` y `mu`, calculaban el intervalo según el método que corresponda y devuelven una variable booleana con valor `True` si el valor real `mu` se encuentra en el intervalo y con valor `False` si el valor real no se encuentra en dicho intervalo. A continuación se detalla brevemente la implementación de la función y se comenta brevemente lo observado en las figuras resultante.

1) **Aproximación gaussiana + estimador de Neyman para la varianza**

   Para calcular el intervalo mediante este método se calculó $\hat{\sigma}_{neyman} = \sqrt{N_{obs}}$ (al no conocer el verdero valor de μ, se utilizó directamente N_{obs}) para obtener el intervalo definido por $N_{obs} \pm \sqrt{N_{obs}}$. Dentro de la función `IsInside1(nobs,mu)` se calculó este intervalo para luego retornar una variable booleana con valor `True` si el valor de `mu` se encuentra en el intervalo y con valor `False` si este no se encuentra en dicho intervalo.

   Para el valor de $\mu = 0$ el único valor observado es $N_{obs} = 0$ y este método arroja un intervalo asociado $[0; 0]$ con lo cual, la cobertura de este es del 100 %. Luego el valor de cobertura decae a 0 y empieza a crecer hasta llegar a un valor "cercano" al `CL` de 68.3 % requerido para el intervalo en un $\mu > 1$ para seguir oscilando entre una cobertura de ≈ 0.55 y ≈ 0.70. Esto se debe a que los intervalos para los distintos valores de N_{obs} están fijos pero van variando en probabilidad (teniendo mayor probabilidad los intervalos que contengan o estén más cercanos a μ) y al ir variando μ, este puede o no estar incluído en este intervalo con lo cual esa probabilidad puede aportar o no a la cobertura respectivamente. En las bajas de cobertura lo que ocurre es que al aumentar el valor de μ, este se sale del intervalo de valores más pequeños "perdiendo" la probabilidad asociada a ese intervalo y ahora está incluido en el intervalo siguiente al último de los intervalos que incluían μ (ordenando los intervalos en orden creciente de N_{obs} asociado) "ganando" la probabilidad asociada a este intervalo. En el caso de las subidas tenemos el efecto contrario (ver figura 1).

Cobertura de $N_{\rm obs} \pm \sqrt{N_{\rm obs}}$

		$\mu = 1.9$		$\mu = 2.1$		$\mu = 1.3$	
$N_{\rm obs}$	$\mu_1 - \mu_2$	$P(N_{\rm obs})$	in?	$P(N_{\rm obs})$	in?	$P(N_{\rm obs})$	in?
0	0.00 - 0.00	0.15	–	0.12	–	0.27	–
1	0.00 - 2.00	0.28	✓	0.26	–	0.35	✓
2	0.58 - 3.41	0.27	✓	0.27	✓	0.23	✓
3	1.27 - 4.73	0.17	✓	0.19	✓	0.10	✓
4	2.00 - 6.00	0.08	–	0.10	✓	0.03	–
5	2.76 - 7.23	0.03	–	0.04	–	0.01	–
		0.72		0.56		0.68	

Figura 1: Ejemplos de intervalos de confianza calculados como $N_{obs} \pm \sqrt{N_{obs}}$. En gris se pueden observar los valores cobertura correspondientes a cada μ.

Es importante destacar la cualidad más importante de este método: su simplicidad.

2) **Log Likelihood (Wilks)**

Para calcular el intervalo mediante este método se calculó $log(\mathcal{L}(N_{obs},\mu))$, así como su valor máximo para (justamente) el estimador de máxima verosimilitud $\hat{\mu}_{MLE} = N_{obs}$ y se hallaron los dos valores de μ asociados a un valor de $log(\mathcal{L}(N_{obs},\mu)) = log(\mathcal{L}(N_{obs},\hat{\mu}_{MLE})) - \frac{1}{2}$ para obtener el intervalo. Esto se implementó en la función `IsInside2(nobs,mu)` que cumple el mismo rol señalado anteriormente para las funciones `IsInside`. En esta función se separó el caso particular en el que $N_{obs} = 0$ donde $\mathcal{L}(0,\mu) = -\mu$ y entonces tenemos que buscar los valores de μ tales que:

$$\begin{aligned} \mathcal{L}(0,\mu) &= \mathcal{L}(0,\hat{\mu}_{MLE}) - \frac{1}{2} \\ -\mu &= 0 - \frac{1}{2} \\ \implies \mu_u &= \frac{1}{2} \end{aligned} \tag{1}$$

donde en el segundo paso se utilizó el hecho de que $\mu > 0$ (observar que $\mu_u \in [N_{obs};\infty]$. El límite inferior se obtiene teniendo en cuenta que el valor de $\mu_l \in [0;N_{obs}]$ y $N_{obs} = 0$ lo que implica que $\mu_l = 0$.

Es importante destacar la simplicidad de este método. Así como también notar que la zona en la cual la cobertura está por debajo del `CL` es menor que para el caso de la aproximación gaussiana. Se obtiene un intervalo más confiable que el correspondiente al método anterior para valores $\mu \ll 1$, es decir, muy cercanos a 0.

3) **Frecuentista**

La implementación de esta función estuvo a cargo del cuerpo docente.

Este método a pesar de su complejidad, tiene como ventaja que por constucción la cobertura siempre está por encima del `CL` o justo en su valor[1] como se puede apreciar en la figura 2.

4) **Bayesiano**

En la implementación de la función `IsInside4(nobs,mu)` también se resolvió analíticamente el caso particular en el que $N_{obs} = 0$ por separado de los casos $N_{obs} > 0$. Se computó el posterior a partir de la distribución de Poisson y un prior uniforme $p(\mu|N_{obs}) = \mathcal{N}e^{-\mu}\mu^{N_{obs}}$ donde $\mathcal{N}$ es una constante de normalización. Luego se buscaron los cuantiles tales que $\int_0^{\mu_l} p(\mu|N_{obs})d\mu = \int_{\mu_u}^{\infty} p(\mu|N_{obs})d\mu = \alpha = \frac{1-CL}{2}$.

En este caso vemos que la cobertura es nula para valores pequeños de $\mu < 0.1744$ para la figura 2. Esto se debe a que el intervalo de confianza para el caso particular en el que $N_{obs} = 0$ se calcula como $[\mu_l;\mu_u]$ con sus extremos definidos según:

$$\begin{aligned} \alpha = \frac{1-CL}{2} &= \int_0^{\mu_l} e^{-\mu}d\mu = -e^{-\mu}\Big|_0^{\mu_l} = 1 - e^{-\mu_l} \\ \implies \mu_l &= -log(1-\alpha) \end{aligned} \tag{2}$$

Por otro lado,

$$\begin{aligned} \alpha = \frac{1-CL}{2} &= \int_{\mu_u}^{\infty} e^{-\mu}d\mu = -e^{-\mu}\Big|_{\mu_u}^{\infty} = e^{-\mu_d} \\ \implies \mu_u &= -log(\alpha) \end{aligned} \tag{3}$$

Para un nivel de confianza de $CL = 0.683$ tenemos que $\alpha = 0.159$ con lo cual $\mu_l = 0.1733$ y $\mu_u = 1.833$. Además tenemos que para $N_{obs} > 0$ los intervalos van a tener un μ_l mayor a 0.1733. Como conclusión los valores de $\mu < 0.1733$ no pertenecen a ninguno de los intervalos de confianza generados para $N_{obs} \geq 0$.

Este método tiene una cobertura más cercana al `CL` que los métodos de aproximación gaussiana y de Wilks, pero oscila alrededor de su valor a diferencia del método frecuentista que siempre está (oscilando) por encima.

[1] Recordar que esto se debe a que una variable con distribución de Poisson es discreta.

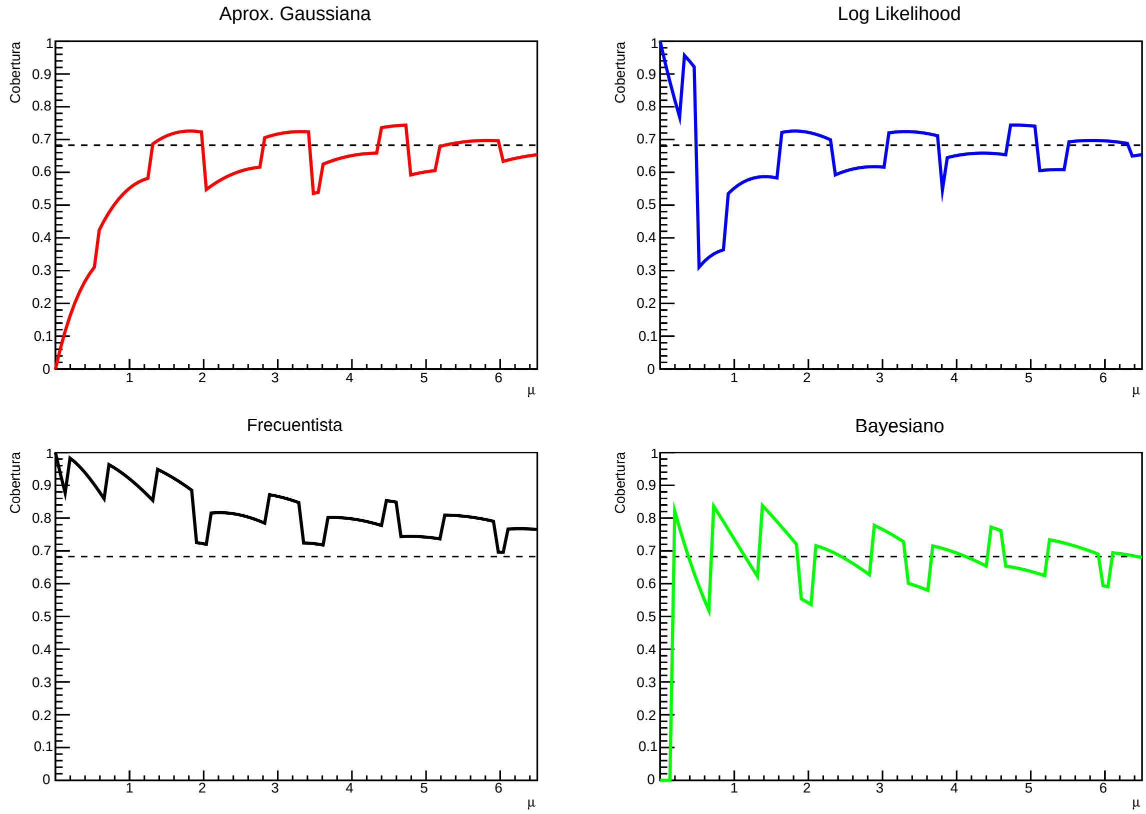

Figura 2: Cobertura en función del parámetro μ de la distribución de Poisson para cada uno de los métodos mencionados. El *confidence level* es de 68.3 %.

Para todos los métodos se obtienen intervalos de confianza que tienen decaimientos abruptos de cobertura para ciertos valores de μ asociados con los límites de los intervalos a N_{obs} fijo. Esta forma de estudiar la cobertura de manera "ordenada" estudiando los intervalos a $N_{obs} = cte$ permite sacar estas conclusiones que serían más difícil de deducir si se analizara desde otro punto de vista. En general las oscilaciones alrededor del `CL` van teniendo menor frecuencia porque los intervalos se hacen más grandes y se achican un poco en amplitud a medida que μ crece (respecto a los valores pequeños de μ) para todos los métodos (ver figuras 3 y 4).

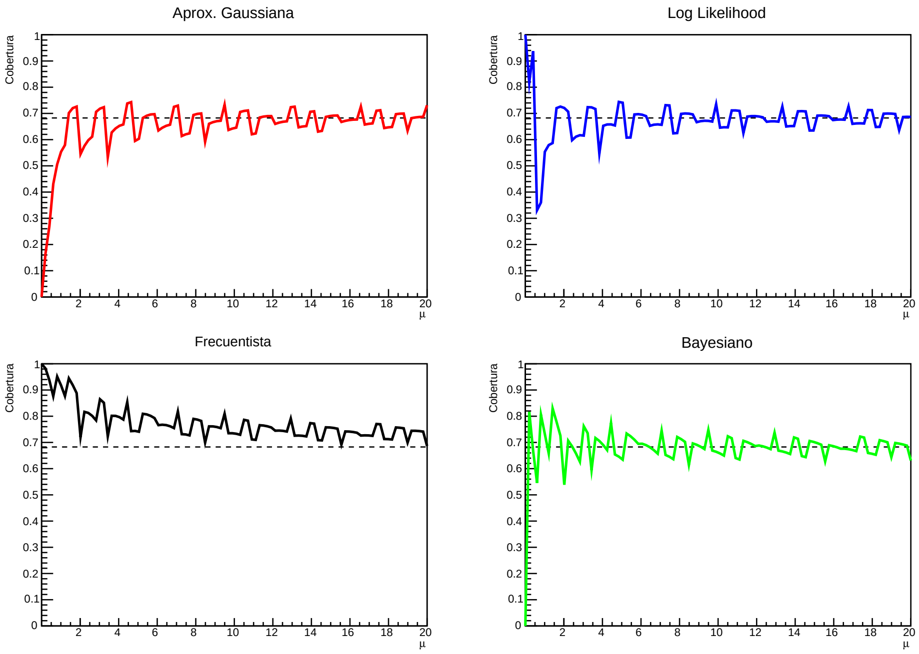

Figura 3: Cobertura de cada método en el límite asintótico. $\mu_{max} = 20$

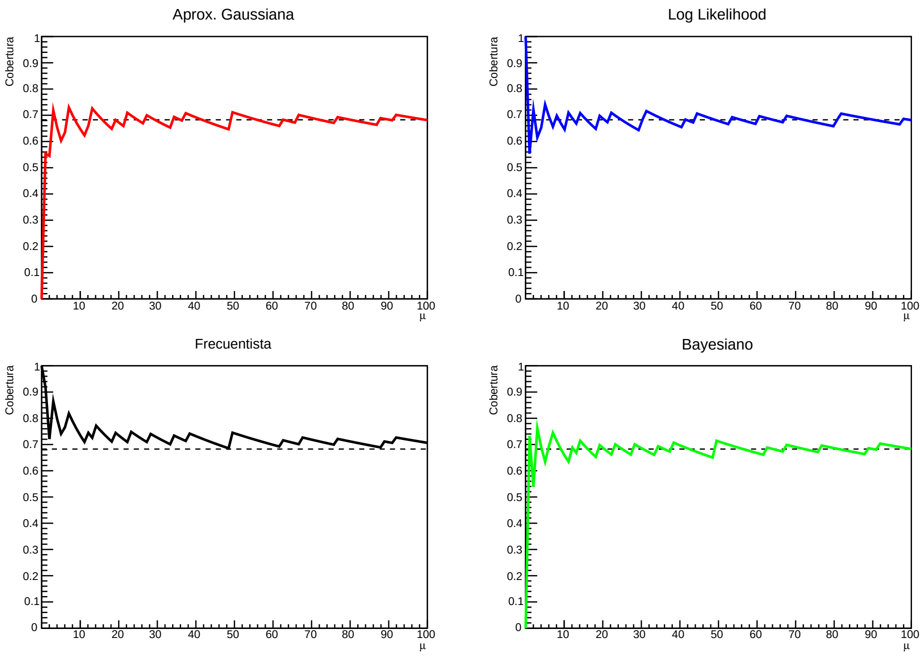

Figura 4: Cobertura de cada método en el límite asintótico. $\mu_{max} = 100$