Trabajo práctico número 2 de Astrometría

R.C. Martin[1]

[1] *Observatorio Astronómico de Córdoba, UNC, Argentina*

Received: ... / Accepted: ...
©The Authors 2023

Resumen / En este trabajo se simularán experimentos con el fin de estudiar sus distribuciones estadísticas. Particularmente se estudiarán la distribución de Fisher-Tippet, la distribución de Poisson, la distribución binomial y el experimento de Buffon. Además realizaremos un muestreo Bootstrap para estimar el error en la varianza.

Keywords / Estadística, Python.

1. Introducción

Definiremos las distribuciones a estudiar en este trabajo.

- La distribución de Fisher-Tippett o Distribución del Valor Extremo Generalizada (GEV) viene dada por:
  $\frac{1}{\sigma}t(x)^{\xi+1}e^{-t(x)}$
  donde si $\xi = 0$ tenemos $t(x) = e^{-\frac{x-\mu}{\sigma}}$. Para este trabajo utilizaremos $\mu = 0$ y $\lambda = \frac{1}{\sigma} = 1$.
  Luego, su CDF es:
  $\int_{-\infty}^{x} t(x') \cdot \exp(-t(x'))\, dx' = \exp(-t(x))$
  cuya inversa resulta:
  $y = \exp(-t(x)) \leftrightarrow -\ln(y) = t(x) \leftrightarrow \ln(\frac{1}{y}) = \exp(-x) \leftrightarrow x = -\ln(\ln(\frac{1}{y}))$
  Finalmente, mencionar que su media viene dada por $x_1 = \int_0^1 xt(x) \cdot \exp(-t(x))\, dx = \frac{0.57721}{\lambda} = 0.57721$
- Distribución exponencial: La distribución exponencial viene dada por:
  $\lambda e^{-\lambda x}$
  y podemos calcular su CDF como:
  $\int_0^x \lambda e^{-\lambda x'} dx' = 1 - e^{-\lambda x}$,
  cuya inversa resulta en:
  $y = 1 - e^{-\lambda x} \leftrightarrow x = -\frac{\ln(1-y)}{\lambda}$
  Si repetimos muchas veces el experimento, el resultado en un intervalo de x dado debe aproximar una distribución de Poisson, la cual viene dada por:
  $p(k) = \frac{e^{-\lambda}\lambda^k}{k!}$
- La distribución binomial viene dada por:
  $P(k) = \binom{n}{k} p^k (1-p)^{n-k}$
  Esta describe un fenómeno con una probabilidad p de suceder y $q = 1 - p$ de no suceder. Por lo tanto resulta discreta, no obstante para un número grande de ensayos esta se aproxima a una distribución normal.

2. Ejercicio 3

En el script "Ejercicio3.py" simulamos una realización de la distribución de Fisher-Tippet. Para ello generamos una distribución pseudo-aleatoria con Numpy entre 0 y 1. Por otra parte definimos la función inversa de la CDF,

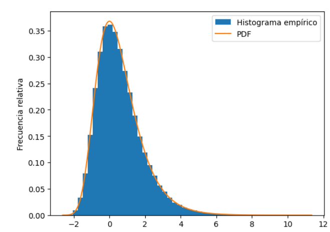

Figura 1: Distribución empírica de Fisher-Tippet como un histograma. La línea sólida representa la distribución teórica.

de esta forma podemos transformar los valores aleatorios entre 0 y 1 en una distribución de Fisher-Tippet que denominaremos *Empírica*. Finalmente, graficamos nuestro resultado junto a la distribución teórica para comparar los resultados (Figura 1).

Además calculamos la media de nuestra distribución simulada, obteniendo $\overline{x} = 0.5793510256369162$ similar a $\mu = 0.57721$

3. Ejercicio 4

Realizamos un proceso análogo al ejercicio 3 (Sección 2) pero con la distribución exponencial de $\lambda = 5/h$, obteniendo el histograma de la Figura 2.

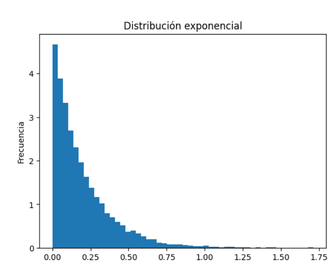

Figura 2: Histograma de la realización de la distribución exponencial con $\lambda = 5/h$

✉ *Contacto* / romanciromartin@mi.unc.edu.ar

Luego, realizamos esto mismo 100.000 veces y guardamos en una lista los eventos sucedidos hasta las 3 horas. Esto debería seguir una distribución de Poisson con $\lambda = 15$ pues son 3 horas de eventos a una tasa de $5/h$. Graficamos esta simulación junto con la Distribución teórica de Poisson, obteniendo lo visto en la Figura 3. Todo esto es realizado con el script "Ejercicio4.py".

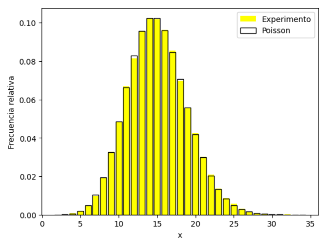

Figura 3: En amarillo tenemos la distribución obtenida con la simulación, mientras que los bordes negros representan la distribución teórica de Poisson.

4. Ejercicio 7

Realizaremos el Experimento de Buffon para estimar π. Para ello, supongamos que tenemos una mesa con dos líneas paralelas, separadas una distancia t. Sobre esta se lanzan agujas de manera aleatoria, siguiendo una distribución uniforme. Medimos la distancia x desde el centro de la aguja hasta la línea mas cercana y el ángulo θ que forma la aguja con la dirección de las rayas. Primero, deduzcamos las distribuciones de x y θ:

Todas las agujas se lanzan con una distribución uniforme, por lo que las distribuciones para (x, θ) deben ser constantes. Calculemos para x: Dado que se mide desde el centro de la aguja y la mesa mide t, la máxima distancia posible es $\frac{t}{2} \implies \int_0^{\frac{t}{2}} p(x)dx = 1 \leftrightarrow \frac{t}{2}\alpha = 1 \leftrightarrow \alpha = \frac{2}{t} \implies p(x) = \frac{2}{t}$

Análogamente, $\theta \in [0, \frac{\pi}{2}] \implies p(\theta) = \frac{2}{\pi}$

$\therefore p(x, \theta) = \frac{4}{t\pi}$ pues V.A. son independientes.

Luego, si $l < t$ la aguja tocará la raya cuando $cos(\frac{\pi}{2} - \theta) = \frac{x}{\frac{l}{2}}$ (trigonometría) por lo tanto, para que se produzca un contacto necesitamos que $\frac{l}{2}sen(\theta) \geq x \implies$ debemos integrar en esa región para x:

$\int_0^{\frac{\pi}{2}} \int_0^{\frac{l}{2}sen(\theta)} p(x, \theta)dxd\theta = \int_0^{\frac{\pi}{2}} \frac{l}{2}sen(\theta)\frac{2}{t} = \frac{2l}{t\pi}$ es la probabilidad de que toque una raya. Por lo tanto, tenemos que $p = \frac{2l}{t\pi} \leftrightarrow \pi = \frac{2l}{tp}$, entonces basta simular el experimento para poder estimar π. En el script "Ejercicio7.py" se simula el lanzamiento de 10.000.000 agujas, obteniendo $\pi = 3.143094401583114$.

5. Ejercicio 8

El Bootstrap Resampling consiste en tomar de una distribución aleatoria x con n datos un número n de valores, de forma aleatoria. Con estos se arma un nuevo conjunto de datos. Este proceso se repite muchas veces, por ejemplo lo haremos unas 100 veces. A cada uno de estos nuevos conjuntos obtenidos se le puede calcular la media y la varianza, y estas deberían seguir una distribución normal centrada en la media y varianza de la distribución original. Este método es útil para estimar el error de los momentos de orden superior a la media, como puede ser la varianza. En el script "Ejercicio8.py" generamos una distribución normal con parámetros $\mu = 0$ y $\sigma = 1$, a la cual le aplicamos 100 Bootrstraps y estudiamos lo que sucede con la varianza. Obtenemos la distribución de la Figura 4 con un $\overline{\sigma} = 0.9999746797143281$.

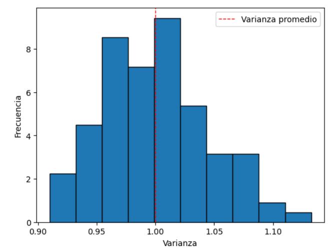

Figura 4: Distribución de la varianza obtenida mediante Bootstrap Resampling de una Gaussiana con 1000 datos.

Además, calculamos el intervalo de confianza de un 95 %, resultando en [0.9379948222637775 , 1.0619545371648786].

6. Ejercicio 9

En el script "Ejercicio9.py", simulamos 100 realizaciones de una distribución binomial con *n=10* y *p=0.4*. Luego, queremos realizar un test de χ^2, el cual viene dado por:

$\chi^2 = \sum_{i=1}^{n} \frac{(x_i - x_{teo})^2}{x_{teo}}$

por lo tanto generamos valores de una distribución teórica con los mismos parámetros para así estimar χ^2.

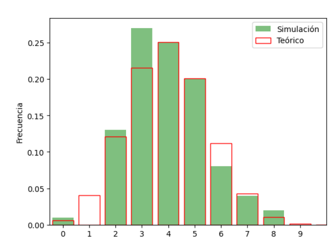

Figura 5: En verde observamos la distribución binomial simulada, mientras que en rojo la teórica.

Realizamos el histograma de ambas (Figura 5) y el cálculo nos devuelve un $\chi = 7.665712622068746$.

Ahora, realizamos una prueba de hipótesis nula para saber si mis datos están acorde al modelo:

H_0: La simulación sigue la distribución teórica

H_1: La simulación no sigue la distribución teórica.

Usamos un $\alpha = 0.01$ y tenemos 9 grados de libertad (1 menos que el número de ensayos). Por lo tanto $\chi^2_{1-\alpha,9} = 16.91898$

Recordemos que la zona de aceptación de H_0 es $\chi^2 < \chi^2_{1-\alpha,9}$ mientras que la región de rechazo es $\chi^2 > \chi^2_{1-\alpha,9}$ por lo que en este caso, la hipótesis nula es correcta, la simulación sigue la distribución teórica. Además realizamos el cálculo de *valor-p* dado por:

$p = \int_{\chi_{ref}}^{\infty} \chi^2(x)dx = 0.6614524813551338$

Ahora resulta interesante utilizar el *valor-p* para comparar la distribución binomial teórica con la distribución normal. Para ello generamos 100 gaussianas con $\mu \in [2, 7]$ y graficamos su *valor-p*.

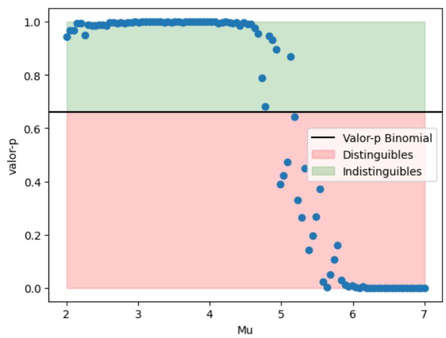

Figura 6: $Valor - p(\mu)$, en la región verde las gaussianas son indistinguibles de la binomial, mientras que en la roja son distinguibles.

Como se observa en la Figura 6, una distribución normal puede confundirse con una binomial para un rango de μ. Luego, en el script "Ejercicio9_f", simulamos 10000 realizaciones de una distribución binomial con *n=1000* y *p=0.4* para realizar el mismo test pero contra 100 gaussianas con $\mu \in [200, 700]$.

Podemos ver en las Figuras 7 y 8 como al aumentar el tamaño de la muestra, las distribuciones son cada vez más parecidas. El efecto es tal que para la distribución binomial simulada, $Valor - p = 1$ lo que provoca que las únicas gaussianas indistinguibles sean las que poseen un μ cercano a 400 (Ver Figura 9).

7. Conclusión

En este trabajo se profundizó en la simulación de distribuciones:

- En la sección 3 obtuvimos una distribución de Poisson a partir de 100.000 distribuciones exponenciales.
- En la sección 5 se mostró como estimar errores a los momentos de una distribución utilizando Boostrap Resampling (en particular, se aplicó para la varianza).
- En la sección 6 se mostró como analizar un análisis de hipótesis nula y, cualitativamente, como la distribución binomial tiende a una distribución normal para muestras grandes.

Además, en la sección 4 se logró estimar π a partir de la simulación de un experimento.

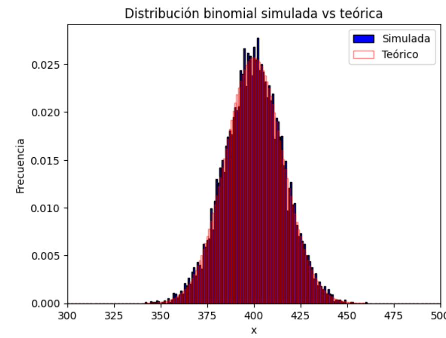

Figura 7: En azul observamos la distribución binomial simulada, mientras que en rojo la teórica.

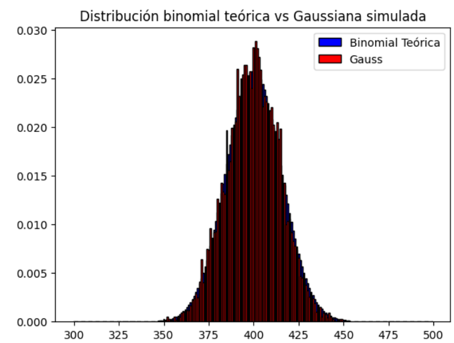

Figura 8: En azul observamos la distribución binomial teórica, mientras que en rojo una distribución gaussiana de $\mu = 400$ simulada.

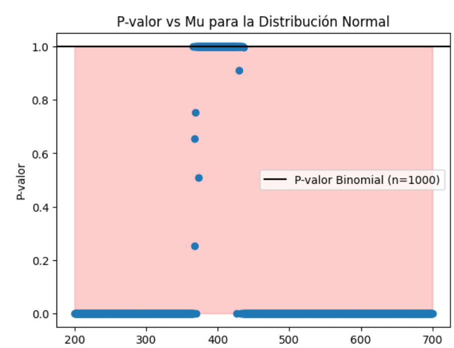

Figura 9: $Valor - p(\mu)$, la zona verde no es apreciable a diferencia de la Figura 6, puesto que Valor-p=1. Por lo tanto los puntos sobre la recta negra son los μ para los cuales son indistinguibles (μ cercano a 400)