Trabajo Práctico Final de MEFE

Leila Prelat

4 de abril de 2019

Índice

1. Generalidades

Leila Rocío Prelat

LU: 400/15

Para los errores en los bines de los histogramas se utilizó el mismo código que en los trabajos prácticos computacionales anteriores (la función se llama Error_bines) y se utilizó el density=1 en los histogramas (correción del trabajo anterior).

2. Estadística en la calle

El objetivo de esta sección es asociar cada patente con un número natural, de manera tal que la patente AA000AA sea el número 1. Se usan 26 letras del abecedario para las patentes (la ñ no se utiliza) y los números van del 0 al 9 (se usan diez números en total). Por lo tanto, para asociar cada patente a un número natural, se utilizará una base mixta de 26 y 10. El primer paso es asociar la patente vista con una lista que indique la posición de cada letra en el abc (A se asocia con 1, B con 2, C con 3, ..., Z con 26). Sea una patente canónica (a,b,c,d,e,f,g), y sean i,j,k,l las posiciones de a,b,f,g en la lista abc, respectivamente. La patente que vimos, (a,b,c,d,e,f,g), la asociamos a los dígitos (i,j,c,d,e,k,l). Por ejemplo, la patente (A,B,0,2,5,F,J) primero la asociamos a (1,2,0,2,5,6,10). Una vez que tenemos los dígitos de la patente (i,j,c,d,e,k,l) le asociamos un número natural utilizando una base mixta de 26 y 10:

$$\text{Def(patentes)} = l\cdot 26^0+(k-1)\cdot 26^1+e\cdot 26^2+d\cdot 26^2\cdot 10^1+c\cdot 26^2\cdot 10^2+(j-1)\cdot 26^2\cdot 10^3+(i-1)\cdot 26^3\cdot 10^3 \quad (1)$$

De esta manera, se puede observar fácilmente que la patente (A,A,0,0,0,A,A) $\rightarrow$ (1,1,0,0,0,1,1) se asocia al natural 1. Y la última patente (que todavía no existe) (Z,Z,9,9,9,Z,Z) $\rightarrow$ (26,26,9,9,9,26,26) se asocia al natural $26 + 25 \cdot 26 + 9 \cdot 26^2 + 9 \cdot 26^2 \cdot 10 + 9 \cdot 26^2 \cdot 10^2 + 25 \cdot 26^2 \cdot 10^3 + 25 \cdot 26^3 \cdot 10^3$. Ese número resulta igual a la cantidad total de patentes diferentes que podemos construir con este sistema de 4 letras y 3 números: $26^4 \cdot 10^3$.

Se le aplicó la función patentes a cada patente observada en el barrio de Recoleta (incluidas en cada renglón (line) del txt patentesrecoleta.txt), y con cada resultado se formó la lista patentesrecoleta de números naturales asociados a las patentes observadas. Se eliminaron las patentes repetidas de dicha lista puesto que se consideró que no tenía sentido incluirlas.

También puede ser útil tener la inversa de la función patentes: a partir del número num que se le asoció a una patente recuperar cuál es la patente. Primero se divide al num por la base mixta utilizada en la formula 1:

$$\text{rta6} + (k-1)\cdot \text{rta5} + e\cdot \text{rta4} + d\cdot \text{rta3} + c\cdot \text{rta2} + (j-1)\cdot \text{rta1} + (i-1)\cdot \text{rta0}$$

Los corchetes significan parte entera:

$$\text{rta0} = \left[\frac{\text{num}}{26^3\cdot 10^3}\right]$$

$$\text{rta1} = \left[\frac{\text{num} - \text{rta0}\cdot 26^3\cdot 10^3}{26^2\cdot 10^3}\right] = \left[\frac{\text{num}}{26^2\cdot 10^3} - \text{rta0}\cdot 26\right]$$

$$\text{rta2} = \left[\frac{\text{num} - \text{rta0}\cdot 26^3\cdot 10^3 - \text{rta1}\cdot 26^2\cdot 10^3}{26^2\cdot 10^2}\right] = \left[\frac{\text{num}}{26^2\cdot 10^2} - \text{rta0}\cdot 26\cdot 10 - \text{rta1}\cdot 10^1\right]$$

$$\text{rta3} = \left[\frac{\text{num} - \text{rta0}\cdot 26^3\cdot 10^3 - \text{rta1}\cdot 26^2\cdot 10^3 - \text{rta2}\cdot 26^2\cdot 10^2}{26^2\cdot 10^1}\right]$$

$$\text{rta3} = \left[\frac{\text{num}}{26^2\cdot 10^1} - \text{rta0}\cdot 26\cdot 10^2 - \text{rta1}\cdot 10^2 - \text{rta2}\cdot 10\right]$$

$$\text{rta4} = \left[\frac{\text{num} - \text{rta0}\cdot 26^3\cdot 10^3 - \text{rta1}\cdot 26^2\cdot 10^3 - \text{rta2}\cdot 26^2\cdot 10^2 - \text{rta3}\cdot 26^2\cdot 10^1}{26^2}\right]$$

$$\text{rta4} = \left[\frac{\text{num}}{26^2} - \text{rta0}\cdot 26\cdot 10^3 - \text{rta1}\cdot 10^3 - \text{rta2}\cdot 10^2 - \text{rta3}\cdot 10^1\right]$$

$$\text{rta5} = \left[\frac{\text{num} - \text{rta0}\cdot 26^3\cdot 10^3 - \text{rta1}\cdot 26^2\cdot 10^3 - \text{rta2}\cdot 26^2\cdot 10^2 - \text{rta3}\cdot 26^2\cdot 10^1 - \text{rta4}\cdot 26^2}{26}\right]$$

$$\text{rta5} = \left[\frac{\text{num}}{26} - \text{rta0}\cdot 26^2\cdot 10^3 - \text{rta1}\cdot 10^3\cdot 26 - \text{rta2}\cdot 10^2\cdot 26 - \text{rta3}\cdot 10^1\cdot 26 - \text{rta4}\cdot 26\right]$$

$$\text{rta6} = [\text{num} - \text{rta0}\cdot 26^3\cdot 10^3 - \text{rta1}\cdot 26^2\cdot 10^3 - \text{rta2}\cdot 26^2\cdot 10^2 - \text{rta3}\cdot 26^2\cdot 10^1 - \text{rta4}\cdot 26^2 - \text{rta5}\cdot 26]$$

Recuperar la patente a partir de su número asociado puede resultar útil para saber cuál es la patente del estadístico observado en el Test de Kolmogorov-Smirnov, o para saber cual es la estimación bayesiana de la patente máxima existente.

3. Test de Kolmogorov-Smirnov

EL test de Kolmogorov-Smirnov está dentro de los tests de *goodness-of-fit* [1], los cuales sirven para cuantificar qué tan bueno es el ajuste de un modelo teórico respecto de los datos experimentales. El test de Kolmogorov-Smirnov propone construir una función de distribución acumulativa $S_n(x)$ (cdf empírica) utilizando las n mediciones realizadas $\{x_1,x_2,(...),x_n\}$:

$$S_n(x) = \begin{cases} 0 \;\; \text{si} \;\; x < x_1 \\ \frac{i}{n} \;\; \text{si} \;\; x_i \leq x < x_{i+1} \;\; \text{para cada} \;\; 1 \leq i \leq n-1 \\ 1 \;\; \text{si} \;\; x \geq x_n \end{cases}$$

En este test se compara la $F_o(x)$ (ajuste del modelo teórico, cdf teórica) con la $S_n(x)$. La hipótesis nula de este test es H_o: $S_n(x) = F_o(x)$, uno espera que las dos funciones sean parecidas entonces se espera que no difieran mucho entre sí. El estadístico del test de Kolmogorov es la norma infinito de la diferencia:

$$D_n = \|S_n - F_o\|_\infty = \sup |S_n(x) - F_o(x)| \tag{2}$$

Si no se utilizaron las mediciones para estimar parámetros de la $F_o(x)$ entonces la distribución del estadístico D_n es independiente de $F_o(x)$. En ese caso, la distribución de D_n sólo depende de n (cantidad de mediciones realizadas) y se pueden utilizar las tablas de Kolmogorov para hallar el valor crítico de D_n. En las tablas de Kolmogorov las únicas variables son el valor de n y la significancia α. La $F_o(x)$ no es variable porque dichas tablas se utilizan cuando D_n es independiente de $F_o(x)$. Para utilizar las tablas también se debe cumplir que las mediciones sean independientes entre sí, y que la función $F_o(x)$ sea continua [2].

3.1. Ventajas del test Kolmogorov-Smirnov

1) Cuando la $F_o(x)$ es continua y no se utilizaron mediciones para estimar sus parámetros, la D_n es *distribution free* (no depende de $F_o(x)$) y se pueden utilizar las tablas de Kolmogorov.

2) El estadístico del test de Kolmogorov tiene una distribución conocida:

$$\lim_{n\to\infty} P(D_n \leq \frac{z}{\sqrt{n}}) = 1 - 2\sum_{r=1}^{\infty}(-1)^{r-1}e^{-2r^2z^2} \tag{3}$$

El límite vale para un $n \geq 80$ (nuestro caso). No se puede hacer una suma infinita computacionalmente, sin embargo se puede aproximar ese límite por una sumatoria de r hasta 1500 dado que

la serie converge y decrece de forma gaussiana (se consideró que la contribución a la serie de los términos $e^{-2r^2z^2}$ con r mayor a 1500 y z menor a 10 es despreciable).

3) El test K-S es más sensible a los datos que el test χ^2. En muestras cortas, se utiliza el test K-S en lugar del test de χ^2.

P value:

El P value de un test de hipótesis es una función del estadístico t_{obs} observado del test dada por la probabilidad de obtener un estadístico igual o más extremo que el estadístico t observado, dado que la hipótesis nula H_o es cierta [4]:

$$\text{P value} = \Pr(\text{Observar un estadístico igual o más extremo que el observado } t_{obs}|H_o) \tag{4}$$

Según si el P value es menor o mayor al nivel de significancia α del test se rechaza o no se rechaza la hipótesis nula, respectivamente.

La interpretación matemática de si un estadístico es igual o más extremo que el observado, depende de si el test es de cola derecha, de cola izquierda o a dos colas.

En el test de Kolmogorov en particular, puesto que el estadístico D_n es más extremo respecto del observado $D_{n,obs}$ si $D_n \geq D_{n,obs}$ (ya que éste estadístico es la distancia entre la cdf teórica y la cdf empírica, y cuanto mayor sea la distancia, más lejos se encuentran dichas funciones), entonces se interpreta de esa forma que un estadístico sea más extremo que otro. Se dice que los tests en los que se realiza dicha interpretación son de cola derecha. En los mismos se calcula el P value como [4]:

$$\text{P value}|_{\text{cola derecha}} = \Pr(t \geq t_{obs}|H_o) = 1 - F_t(t_{obs}) \tag{5}$$

Donde F_t es la cdf del estadístico t evaluada en el estadístico observado t_{obs}. Dicha fórmula es válida para el test de Kolmogorov, con $t = D_n$.

Zona crítica:

Notemos que, para un test de cola derecha como el de Kolmogorov-Smirnov, se rechaza la hipótesis nula si y sólo si P value$= 1 - F_D(D_{n,obs}) < \alpha$, lo cual es equivalente a $1 - \alpha < F_D(D_{n,obs})$, y como F_t es creciente (estrictamente, en el caso del estadístico de Kolmogorov), e invertible (la inversa es la función cuantil $Q_D = F_D^{-1}$) entonces se rechaza la hipótesis sólo si $D_{n,obs} > Q_D(1-\alpha) \coloneqq D_{n,crit}$. Los valores de D_n tales que $D_n > D_{n,crit}$ forman la zona crítica. Si y sólo si el estadístico observado cae dentro de la zona crítica, la H_o se rechaza con el nivel de significancia α.

Los valores $D_{n,crit}$ se encuentran tabulados en el Frodesen (tablas de Kolmogorov) para distintos valores de α y n. En particular, para $\alpha = 0{,}05$ y $n \geq 100$ (que es el caso con el que se trabajará), se

tiene que $D_{n,crit} = 1{,}36/\sqrt{n}$.

3.2. ¿Uniformemente distribuidas?

En esta sección, se estudiará si las patentes observadas representan una variable aleatoria con distribución uniforme utilizando el Test de Kolmogorov-Smirnov sobre la muestra.

Antes de salir a la calle en busca de patentes, es razonable pensar que la probabilidad de observar cualquier patente que existe hasta el día de hoy es equiprobable. Obviamente, la probabilidad de observar una patente que todavía no existe es cero. Se testeará como hipótesis nula si las patentes existentes hasta el día de hoy representan una variable aleatoria con distribución uniforme:

H_o : La distribución de la variable aleatoria x está uniformemente distribuida.

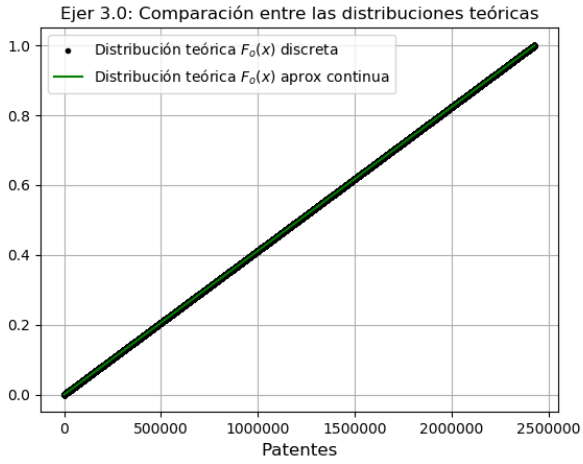
Entonces, en el problema de las patentes, se tiene que la cdf teórica es la cdf $F_o(x)$ de una distribución uniforme entre la patente más vieja (a=1, AA000AA) y la más nueva (patmax=2428510, de la patente AD592MF) que existen (para asociar cada patente a un número se utiliza la ecuación 1). Se recuerda que no se pueden estimar parámetros de la $F_o(x)$ a partir de las mediciones y utilizar las tablas de Kolmogorov. Por ende, a y b no pueden ser las patentes más vieja y más nueva observadas (respectivamente) porque se estarían utilizando las mediciones para estimar los parámetros de $F_o(x)$. Tampoco pueden ser estimaciones (bayesianas, por ejemplo) basadas en las patentes más vieja y más nueva observadas.

La $F_o(x)$ que se debería utilizar es la cdf de una distribución uniforme discreta ya que la probabilidad es no nula sólo en los naturales entre la patente más vieja y la más nueva (hay probabilidad cero de hallar una patente con un número no natural asociado). Entonces la $F_o(x)$ debería ser una función escalonada $F_o^d(x)$ con saltos de discontinuidad en cada entero (hay una cantidad finita de discontinuidades). Sin embargo, se mencionó en la primera ventaja del test de Kolmogorov-Smirnov (sección 3.1) que la $F_o(x)$ debe ser continua para poder utilizar las tablas de Kolmogorov. Por lo tanto, para poder utilizar dichas tablas, se aproximó la cdf $F_o^d(x)$ de la distribución uniforme discreta (la que mejor modelaría al problema de las patentes) por la cdf $F_o^c(x)$ de la distribución uniforme continua.

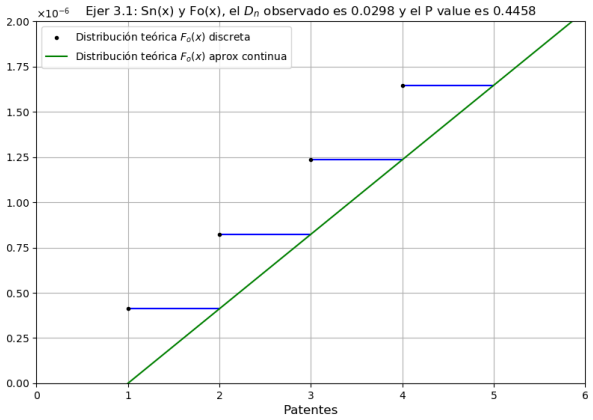
Utilizando una versión de la desigualdad triangular, se puede ver que el error cometido en el cálculo del estadístico de Kolmogorov al utilizar la aproximación continua $F_o^c(x)$ de la discreta $F_o^d(x)$ es menor o igual que la diferencia entre ambas distribuciones:

$$\Big|\,||F_o^d(x) - S_n(x)|| - ||F_o^c(x) - S_n(x)||\,\Big| \leq ||F_o^d(x) - S_n(x) - F_o^c(x) + S_n(x)|| = ||F_o^d(x) - F_o^c(x)||$$

Entonces se compararon las distribuciones $F_o^d(x)$ y $F_o^c(x)$ para observar su diferencia:



(a)



(b)

Figura 1: $F_o^d(x)$ discreta y $F_o^c(x)$ continua (a la izquierda sin zoom y a la derecha con zoom).

Para ver estos gráficos, poner al principio del código: comparacion_discreta_continua=1

Se puede observar en la figura 1 con zoom que al utilizar la $F_o(x)$ continua se comete un error de $1/\text{patmax} \sim 10^{-7}$, siendo patmax la última patente existente hasta el día de hoy, en cada patente. Se considera un error despreciable y a partir de ahora se utilizará como $F_o(x)$ su aproximación continua:

$$F_o(x) = \text{unif\_cdf}(x) = \frac{x}{\text{patmax} - 1} - \frac{1}{\text{patmax} - 1}$$

En el código se define a función Kolmogorov que depende de la muestra de los datos y de la $F_o(x)$ (la cdf del modelo teórico) para calcular el estadístico de Kolmogorov D_n. Para el cálculo de D_n se ordena la muestra de datos de menor a mayor y se utiliza la fórmula 10 del apéndice para optimizar su cálculo computacional. La función Kolmogorov devuelve el estadístico del test D_n observado (Dn_obs), es decir, el que se obtiene con las mediciones. El mismo se muestra como un segmento negro, junto con la cdf teórica y empírica, en la figura 2. Además devuelve el número

asociado a la patente en la cual se observa el estadístico (Patente_Dn_obs). Para construir el segmento negro de la figura 2 la función Kolmogorov también devuelve los puntos del máximo de la fórmula 10 cuya distancia determina el estadístico observado (cdf_obs=F_o(Dn_obs) y ecdf_obs)).

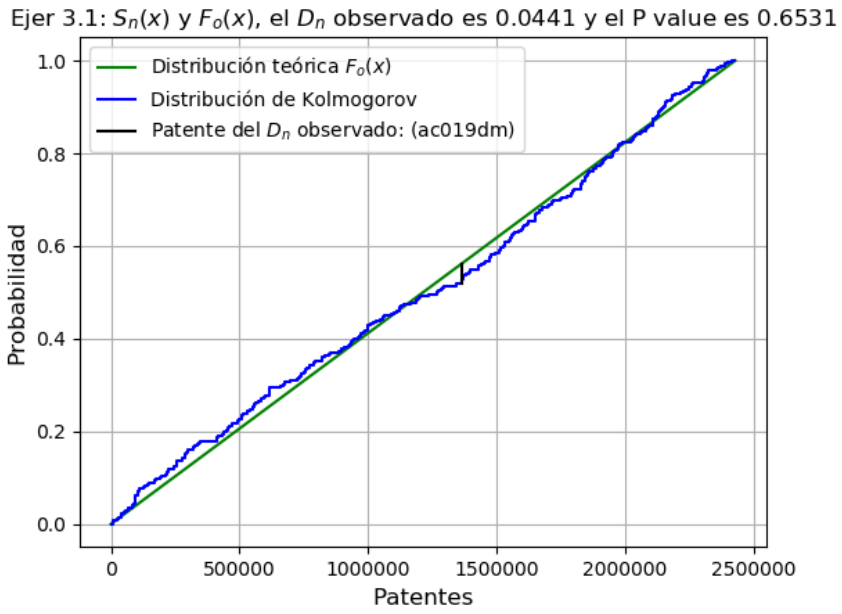


Figura 2: El estadístico observado Dn_obs es 0,0298 y el P value (obtenido con la distribución del estadístico) es 0,4458.

El P value de la figura se obtuvo a partir de la distribución acumulada del estadístico (fórmula 3) evaluada en el estadístico observado, con la ecuación 5. Dicha distribución acumulada se observa en rojo en la siguiente figura:

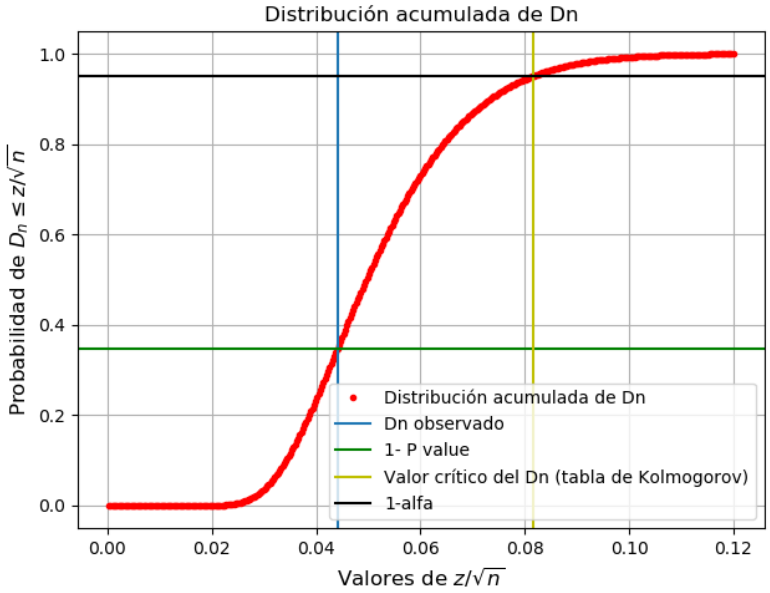


Figura 3: Distribución acumulada del estadístico (fórmula 3).

Para observar la figura anterior, poner al principio del código distribucion_acumulada_Dn=1.

Se observa en la figura, la fórmula 5: 1-P value= $F_D(Dn_{Obs})$. Además se observa la directa relación entre el valor crítico para el D_n (zona crítica a partir de la cual se rechaza H_o) con el α determinado.

Dado que se conoce la distribución del estadístico, no es necesario simularla para obtener el P value (el cual se obtuvo con la fórmula 5).

Sin embargo, si se desea simular la distribución del estadístico de Kolmogorov, hay que generar N listas de números randoms entre 1 y patmax y aplicar la función Kolmogorov a cada una de esas N listas. Con los N=1000 valores del estadístico de Kolmogorov se construyó el siguiente histograma:

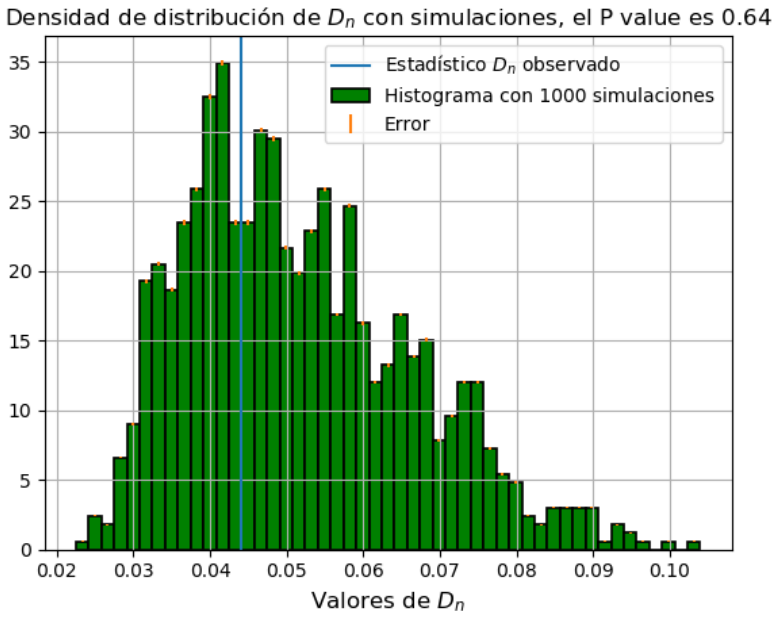


Figura 4: Distribución de densidad del estadístico con N=1000 simulaciones.

El P value obtenido con las simulaciones de la pdf del estadístico es parecido al P value obtenido con la cdf del estadístico. Al aumentar el valor de N, el P value obtenido con las simulaciones se parece más al otro.

3.3. ¿Exponencialmente distribuidas?

Se simuló la distribución del estadístico D_n dado que la hipótesis alternativa H_1 es cierta. Se considera como hipótesis alternativa que la distribución de patentes es exponencial de parámetro $\lambda = 4 \cdot 10^{-7}$. Para ello, se crearon N listas de mediciones aleatorias con dicha distribución exponencial y la longitud de cada una de esas listas fue el número de patentes observadas (ya que el estadístico D_n depende de la cantidad de datos empíricos n). El histograma de la distribución del estadístico D_n realizado con N=1000 simulaciones se muestra a continuación:

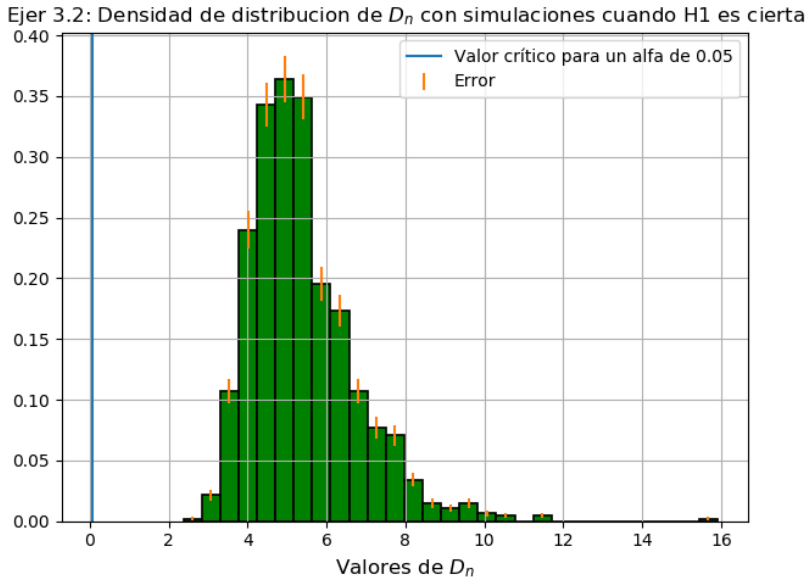


Figura 5: Densidad de distribución del estadístico D_n realizado con N=1000 simulaciones cuando H_1 es cierta.

Se puede observar en la figura que todos los valores de D_n obtenidos de la simulación son mayores al valor crítico del estadístico de la tabla Kolmogorov (el mínimo de todos los valores de D_n fue 4,09 y el valor crítico de la tabla de Kolmogorov para $\alpha = 0{,}05$ es igual a $1{,}36n^{-1/2}$). La $F_o(x)$ sigue siendo la cdf de la distribución uniforme continua, lo que cambió fue suponer que la distribución de la muestra de los datos es exponencial.

El poder del test (considerando sólo dicha hipótesis alternativa H_1) se define como la probabilidad de rechazar H_o con el test dado que la hipótesis alternativa H_1 es verdadera. El mismo es igual a $1 - \beta$, donde β es el error tipo 2. Se calculó computacionalmente con las $N = 1000$ simulaciones realizadas como la cantidad de veces que se rechazó el test (equivalentemente la cantidad de veces que el estadístico resultó mayor al valor crítico), dividido N. Puesto que todos los estadísticos resultaron mayores al valor crítico el test se rechazó en las N=1000 simulaciones y se obtuvo un poder del test igual a 1.

4. La patente del auto más nuevo

En los histogramas de esta sección, a diferencia de los otros, para el error de los bines se utilizó $N \cdot k$ intentos y no N. Lo anterior se debe a que en la definición de experimento(N,k,a,b) se generan $N \cdot k$ datos pseudoaleatorios.

4.1. Distribución $P(m|k,n)$

La P(m;k,n) se definió en el código por recurrencia de k para disminuir el tiempo computacional:

$$P(m|k,n) = \begin{cases} 1/n \quad \text{si} \quad k = 1 \\ \dfrac{k}{k-1} \cdot \dfrac{m-k+1}{n-k+1} \cdot P(m|k-1,n) \quad \text{si} \quad k \neq 1 \end{cases} \tag{6}$$

Se definió la función experimento(N,k,a,b) la cual N listas, cada una con k números naturales aleatorios entre a y b, y guarda el máximo de cada lista (se obtienen N máximos). Se tomó a=1 (primera patente) y como b a la última patente existente hasta el momento: patmax. Se simuló la distribución de m, la máxima patente observada, utilizando la función experimento(N,k,a,b). Además se comparó la simulación con la distribución teórica de m de la fórmula 6:

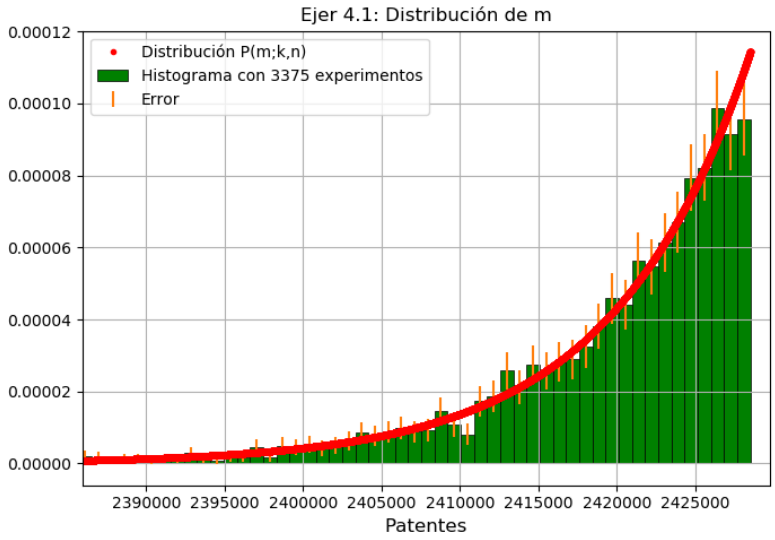


Figura 6: Rojo: Distribución teórica de $P(m|k,n)$. Verde: Distribución simulada de $P(m|k,n)$ con 3375 experimentos.

Se puede observar en la figura 6 que las distribuciones teórica (rojo) y simulada (verde) se parecen.

También se puede utilizar como distribución teórica la fórmula de la guía 3 ejercicio 8 (aproximación al continuo):

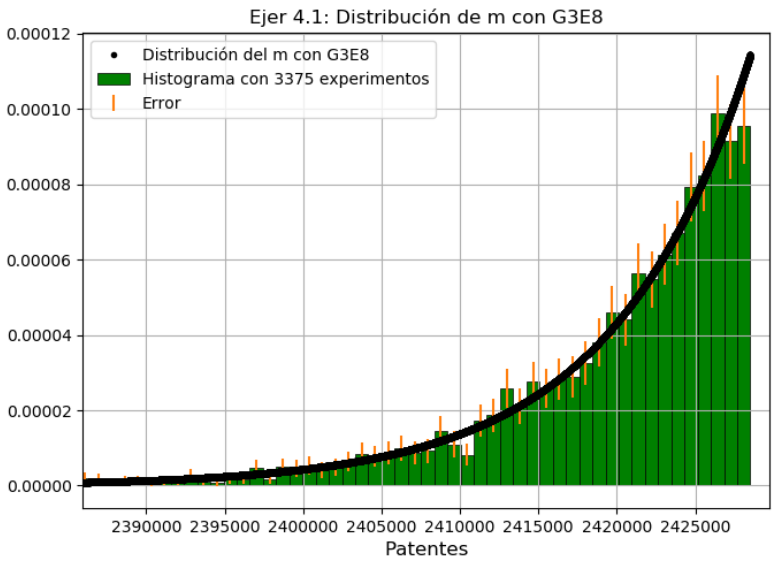


Figura 7: Distribución $P(m|k,n)$ con la distribución del máximo de G3E8.

Para observar la figura 7 poner al principio del código dist_de_max=1. Se puede observar en la figura 7 que las distribuciones teórica (rojo) y simulada (verde) se parecen.

Dado que las patentes de interés son números grandes, se puede utilizar la aproximación de Stirling en la fórmula 6. Esa aproximación evita realizar factoriales de números grandes: $\log n! \sim n \log n - n$. A continuación se muestra una comparación de las tres fórmulas para P(m—k,n) mencionadas:

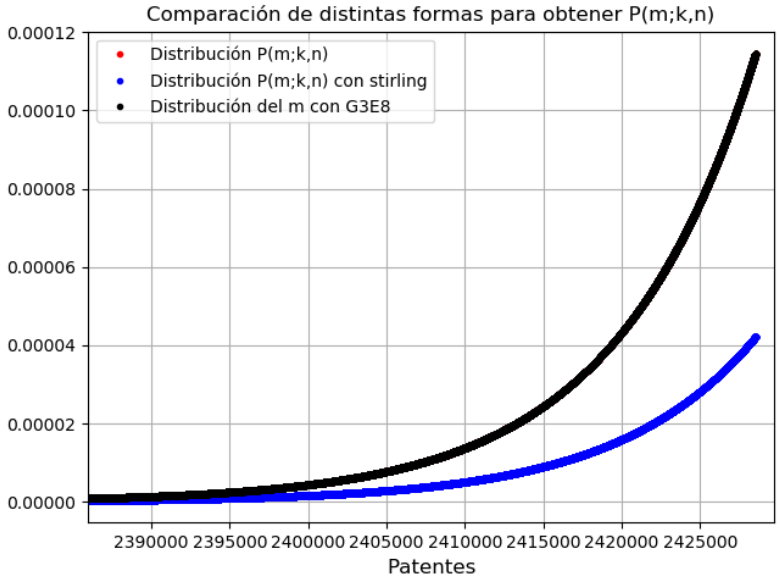


Figura 8: Error cometido al utilizar la aproximación de Stirling.

Se puede observar el error cometido en la figura 8 al utilizar la aproximación de Stirling. El tiempo computacional que conlleva hacer la distribución de m con la fórmula 6 es muy similar al tiempo

computacional de Stirling, por lo tanto no hay mucho ahorro en utilizar la aproximación mencionada. Se decidió no utilizar la aproximación de Stirling en las siguientes secciones.

La curva roja de la figura 8 no se observa porque coincide con la curva negra.

4.2. Distribución de $P(n|k,m)$

La definición de $P(n|k,m)$ se construyó utilizando la definición de $P(m|k,n)$. Se fijó un valor de m y se obtuvo $P(m|k,n)$ para varios valores de n. Los valores de n iban desde el m que se fijó hasta una cota superior. Al verdadero valor de n se lo llamó patmax (la última patente existente). La cota superior utilizada fue patmax$+5 \cdot \left[\frac{m-k}{k}\right]$ (los corchetes indican parte entera).

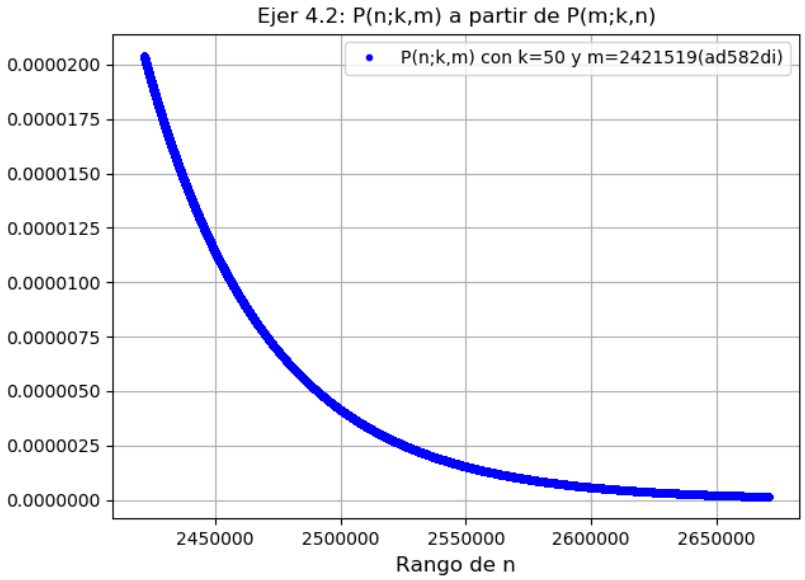


Figura 9: Distribución de P(n;k,m) a partir de P(m;k,n) con k=228 y m=2416519.

Se utilizaron valores de m y k distintos a los de las mediciones para comparar con la siguiente sección.

El valor real de n es patmax (máxima patente existente) es 2428510. El intervalo bayesiano obtenido fue (2423955,89; 2515749,79) mientras que el teórico fue (2420456,33; 2523476,30) [5]. El valor real de n cayó dentro de la estimación bayesiana obtenida.

4.3. Distribución de $P(n|k,m)$ evaluada en mis k y m

Esta sección es prácticamente idéntica a la anterior sólo que hay que evaluar en los k y m que se obtuvieron de las patentes medidas.

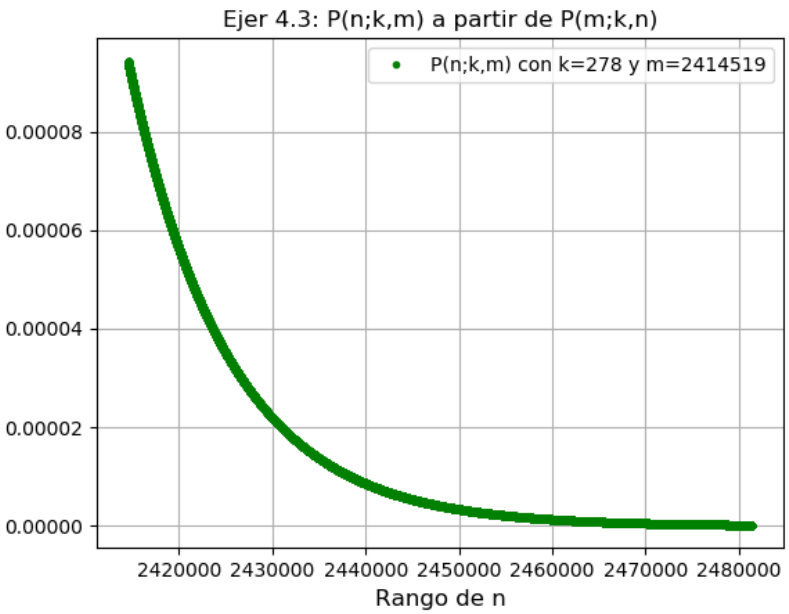


Figura 10: Distribución de P(n;k,m) a partir de P(m;k,n) utilizando los datos de mis mediciones (k=largo de mi muestra y el m = máxima patente observada).

El valor real de n es patmax (máxima patente existente) es 2428510. El intervalo bayesiano obtenido fue (2416807,30; 2437296,99) mientras que el teórico fue (2416471,09; 2437950,03) [5]. El valor real de n cayó dentro de la estimación bayesiana obtenida.

Se observó que para dos valores de m y de k (10 y 9) el valor real se encontraba dentro de la estimación bayesiana obtenida.

4.4. Probabilidad de obtener una peor estimación para n

Se define la función Estimación_n (la estimación bayesiana del valor n, la patente máxima existente) pero, en lugar de utilizar la distribución de probabilidad para m de la fórmula 6, se utilizó la distribución del máximo de la guía 3 ejercicio 8 para disminuir el tiempo computacional (además ya se observó en la sección 4.1 que coincide más la distribución del G3E8 que la aproximación por Stirling).

La máxima patente observada ahora es una variable: Se generan N valores aleatorios del máximo observado entre 1 y patmax, utilizando la función experimento(N,k,a=1,b=patmax). Las estimaciones bayesianas realizadas para un m fijo en las sección 4.3 y 4.2 ahora se realizan para muchos valores de m. Con cada uno de esos N valores de m se calcula la Estimación_n. Se obtienen N estimaciones de P(n;k,m) para N valores de m distintos. Se calcula la probabilidad de obtener una estimación de manera frecuentista: la variable estpeor cuenta la cantidad de veces que mi estimación fue igual

o peor que la obtenida (usando mi m) y luego se divido por el total de intentos N. Si la distancia entre la estimación de n y patmax es mayor que la distancia entre la estimación de n para mi m hallado y patmax entonces a estpeor le sumo 1. A continuación se muestra un gráfico que representa lo mencionado anteriormente:

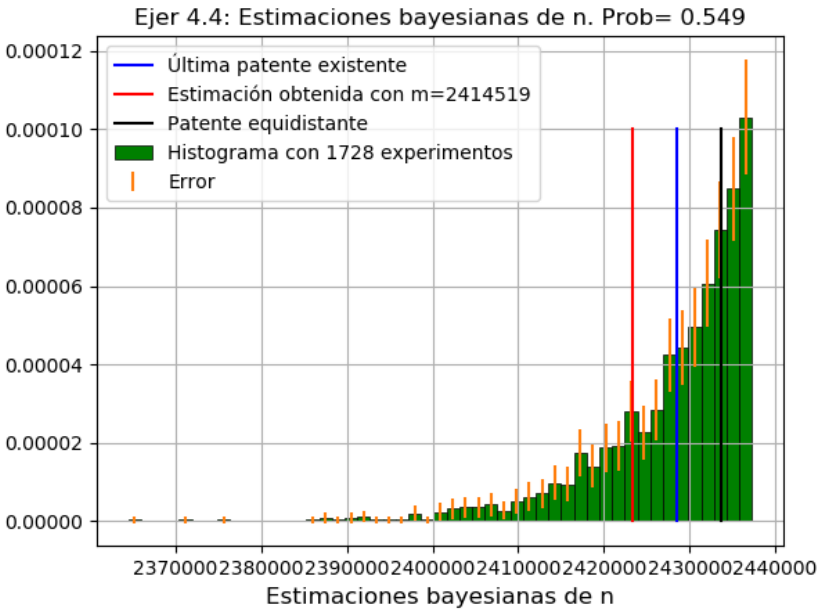


Figura 11: Estimaciones de n

La línea roja marca la estimación obtenida con el m de mis datos y la azul, la última patente existente (patmax). Se añade una línea negra tal que la distancia entre la azul y la negra sea la misma que la distancia entre la azul y la roja. Una estimación peor a la obtenida es una estimación menor a la que marca la línea roja y mayor a la que marca la línea negra.

5. ¿Independiente del barrio?

Los dos test de esta sección son a dos colas. El P value de un test a dos colas está bien definido cuando la distribución del estadístico es simétrica. Sin embargo, se observará que el histograma obtenido en la sección 2 no es perfectamente simétrico (esperable por las fluctuaciones estadísticas). De todos modos, para los dos tests de la sección (a dos colas) se interpretará la fórmula 4 del P value de la siguiente forma [4]:

$$\text{P value}|_{\text{dos colas}} = 2 \cdot \text{mín}\{\Pr(X \leq x|H_o), \Pr(X \geq x|H_o)\} = 2 \cdot \text{mín}\{F_t(t_{obs}), 1 - F_t(t_{obs})\} \qquad (7)$$

En el test de Wilcoxon, el estadístico tiene distribución normal (simétrica).

5.1. Test de Wilcoxon

El test de Wilcoxon permite testear si dos muestras provienen de poblaciones idénticas [3].

Se definió la función Wilcoxon(lista1,lista2) que devuelve el estadístico z del test de Wilcoxon (que si las cantidades $n \leq m$ de elementos de las listas son suficientemente grandes, posee distribución normal $\mathcal{N}(0,1)$ [3]) y el pvalor del test.

Primero, la función se construye una lista total listatot que une las dos listas lista1 y lista2 de forma que los primeros len(lista1) elementos sean de la lista1 y los últimos sean de la lista2. Luego se construye una listatot_sort que es la lista total listatot ordenada de menor a mayor.

Después, para cada elemento listatot_sort[i] de la listatot_sort que es distinto a su anterior (para considerar sólo los elementos no repetidos), se calculan todas las posiciones j_1 de la lista listatot y j_2 de la listatot_sort tales que listatot[j_1]=listatot_sort[i]=listatot_sort[j_2] (la lista de los j_1 es la lista posiciones, y la lista de los j_2 es la lista posicionessort). Notemos que el promedio de la lista posicionessort (para hacer este promedio hay que sumarle 1 a cada posición porque las posiciones en Python comienzan desde el 0) es el rango (definido en el código como sumandorank) que le va a corresponder a cada elemento de la lista listatot que sea igual a listatot_sort[i] [3] (es decir, a los elementos listatot[j_1] con $j_1 \in$ posiciones, o equivalentemente a los elementos listatot_sort[j_2] con $j_2 \in$ posicionessort). Para cada uno de dichos elementos (para cada $j_1 \in$ posiciones), el elemento listatot[j_1] pertenece a lista1 o lista2 si j_1 es menor estricto o mayor o igual a len(lista1), respectivamente; en el primer caso se le suma sumandorank a la variable rank1 y en el segundo se le suma sumandorank a la variable rank2.

El w del test de Wilcoxon es la suma de los rangos de la lista con menor cantidad de elementos (n), por lo que, se define w =rank1, n =len(lista1) y m =len(lista2) si len(lista1)≤len(lista2) y en otro caso se define w =rank2, m =len(lista1) y n =len(lista2). Se calcula la esperanza Ew, la varianza Vw y la desviación estándar sw teóricas de w según Frodesen [3]. Luego, se calcula el valor $z = (w - Ew)/sw$ que, según Frodesen [3], para valores grandes de n,m (mayores a 10) la distribución de z es una normal stándard $\mathcal{N}(0,1)$. Por lo tanto se puede usar z como estadístico del test (ya que tenemos su distribución) y podemos calcular el P-valor como P value= $2\,\text{mín}\{F(z_o), 1 - F(z_o)\}$ (considerando que el test es a dos colas, se utilizó la fórmula 7) donde z_o es el estadístico observado y $F(z) = (1 + \text{erf}(z/\sqrt{2}))/2$ es la función de distribución acumulada (cdf) de la distribución normal stándard $\mathcal{N}(0,1)$. Finalmente, la función Wilcoxon(lista1,lista2) devuelve el valor z_o y el pvalue pv.

Una vez definido el test de Wilcoxon se aplicó el mismo a la lista observada de patentes de recoleta y a otra lista patentesanmartin de patentes de San Martín. Para la lista patentesanmartin se eliminaron las patentes repetidas porque se consideró que no tenía sentido incluirlas.

Se obtuvieron los siguientes estadísticos y el pvalor del test de Wilcoxon para los barrios Recoleta y San Martín: $z = -0{,}91$; P value $= 0{,}36$.

No se rechaza la hipótesis nula para $\alpha = 0{,}05$ dado que el P value obtenido resultó mayor a α.

5.2. Test G8E4

Se define la función U(lista1,lista2) según el estadístico que aparece en el ejercicio 4 de la guía 8. Dado que las listas en cuestión (las patentes observadas) no poseen distribución gaussiana, entonces no es válido que el estadístico U tenga distribución t-student. Las mismas poseen (como hipótesis nula) distribución uniforme.

Para obtener la distribución de U se simularon N pares de listas lista1 y lista2 con distribución uniforme entre 1 y patmax (cuyas longitudes m y n son las longitudes de las listas patentesrecoleta y patentesanmartin), y se guardaron los N valores U(lista1,lista2) calculados en la lista valores_de_U. Realizando un histograma de dicha lista obtenida, se halló computacionalmente la distribución teórica de U (ver figura 12).

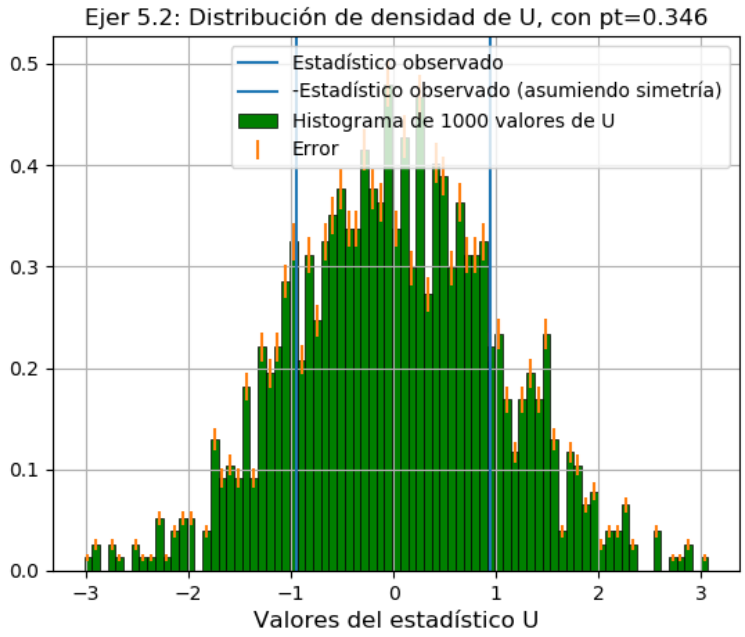


Figura 12: Distribución del estadístico U del Test G8E4 hallado con 1000 simulaciones.

Se aplicó el test G8E4 a las listas patentesrecoleta y patentesanmartin utilizando como estadístico observado a U_{obs} =U(patentesrecoleta,patentesanmartin), cuya distribución se obtuvo computacionalmente. El estadístico observado (y su reflexión respecto al cero) se observa en la figura 12).

Dado que el test en cuestión es un test a dos colas, se calculó el P value del test utilizando la fórmula 7. Para calcularlo se asumió que la distribución de U es simétrica. Con los $N = 1000$ valores

simulados de U se calculó computacionalmente el P value como la cantidad de valores simulados de U que se encuentran más alejados del cero que el U_{obs} dividido la cantidad total N de valores simulados. Ésto a su vez es equivalente (asumiendo que la distribución es simétrica) al doble del mínimo entre: la cantidad de valores simulados de U mayores a U_{obs} dividido la cantidad total N de valores simulados, y la cantidad de valores simulados de U menores a U_{obs} dividido la cantidad total N de valores simulados.

El estadístico y el pvalue obtenidos resultaron iguales a U_{obs} = y P value=.

6. Combinando los tests

Se combinan los dos P values anteriores mediante el estadístico T:

$$T(p_w,p_t) = -2\log(p_w \cdot p_t) = -2\left(\log(p_w) + \log(p_t)\right) \tag{8}$$

En esta sección se calcula la correlación de dos listas mediante corr $= \dfrac{\mathrm{E}\left((X-\bar{X})\cdot(Y-\bar{Y})\right)}{\sigma_x \cdot \sigma_y}$

6.1. Tests independientes

Sea la variable $y = -2\log(x)$ y x una variable con distribución uniforme entre 0 y 1. Entonces $x = e^{-y/2} \rightarrow \left|\dfrac{\partial x}{\partial y}\right| = \left|-\dfrac{1}{2}e^{-y/2}\right| = \dfrac{1}{2}e^{-y/2}$, por lo tanto y posee distribución exponencial. Entonces T tiene la distribución de la suma de **dos** exponenciales con $\lambda = \dfrac{1}{2}$ y eso es una distribución gamma(α =**2**, $\beta = \frac{1}{\lambda} = 2$) [7]. Ver apéndice B.

Cuando H_o es cierta, el P value tiene una distribución uniforme entre 0 y 1. Por ende, para obtener P values, se generaron de manera aleatoria dos listas (lista_pw y lista_pt) con N elementos cada una y distribución uniforme entre 0 y 1. Se obtuvieron N valores de T reemplazando los elementos de las listas lista_pw, lista_pt en la fórmula 8. El siguiente gráfico se construyó utilizando los N=1000 valores del estadístico T(p_w,p_t) y la función pdf de gamma(2,2):

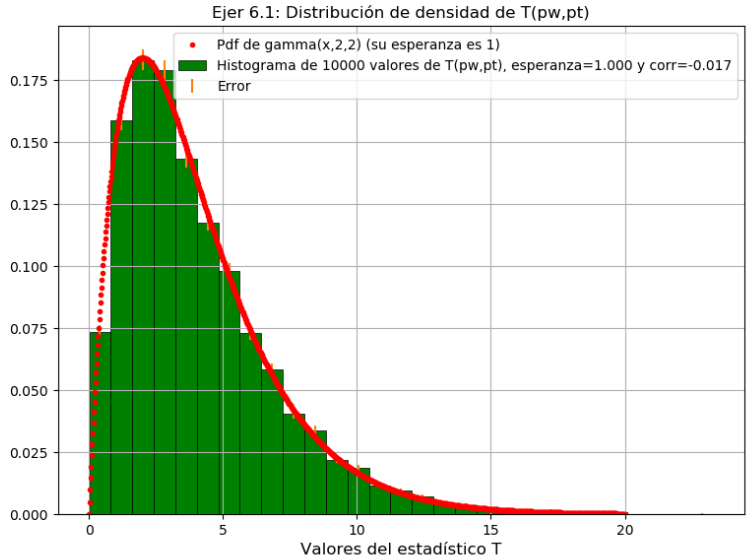


Figura 13: Distribución del estadístico T hallado con 1000 simulaciones.

Se puede observar en la figura que la función gamma(2,2) y la distribución del estadístico T son similares. Además la esperanza del histograma y la esperanza de la gamma(2,2) coinciden.

En este caso, las listas de p_w y p_t fueron generadas de manera independiente, es decir que son variables independientes. Su independencia no implica que la correlación entre ellas sea nula, sin embargo se calculó su correlación para compararla con la correlación de la próxima sección. Uno espera que la correlación de las listas en esta sección sea menor que la correlación de las listas en la próxima sección.

6.2. P values correlacionados

Para obtener la distribución de T (dado que la hipótesis nula de que las listas provienen de las mismas distribuciones uniformes es verdadera) se simularon N pares de listas lista1 y lista2 con distribución uniforme entre 1 y patmax (cuyas longitudes m y n son las longitudes de las listas patentesrecoleta y patentesanmartin), se calcularon los P values p_w y p_t de los tests de Wilcoxon y G8E4 (respectivamente) aplicados a cada uno de los N pares de listas generadas lista1 y lista2, y finalmente se guardaron los N valores T(p_w,p_t) calculados en la lista T_Ho_cierta. También se guardaron los N P values p_w del test de Wilcoxon en la lista pwtot y los N P values p_t del test G8E4 en la lista pttot. Nótese que para calcular cada valor de T se calcularon los P values que surgen de aplicar los dos tests previamente mencionados a un mismo par de listas, por lo que se espera que exista una cierta correlación entre ambos P values.

Realizando un histograma de dicha lista obtenida T_Ho_cierta, se halló computacionalmente la

distribución de T, y se comparó con la distribución teórica de T gamma(2,2) que poseería T si los tests fueran independientes (ver figura 14).

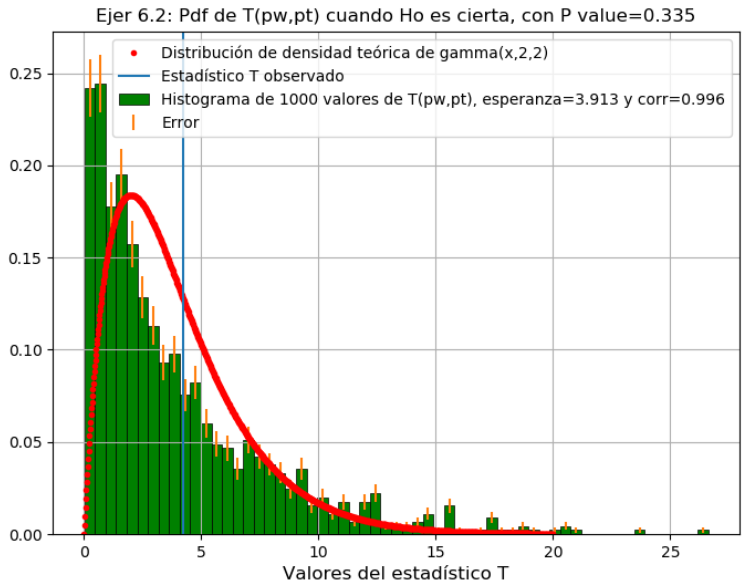


Figura 14: Distribución del estadístico T hallado con 1000 simulaciones cuando H_o es cierta.

A diferencia de la simulación que se muestra en la figura 13, la simulación de la figura 14 no es similar a la distribución gamma(2,2).

Se puede observar en la figura 14 que la correlación entre las listas de P values simuladas en el ejercicio 6.2 es considerablemente mayor a la correlación que se muestra en la figura 13, entre las listas de P values simuladas en el ejercicio 6.1. En particular, la primera correlación es mucho mayor que 0, lo que implica que ambas listas pwtot y pttot no son independientes. En consecuencia, la distribución de T puede no ser la gamma(2,2) (ya que no se cumple la hipótesis de que los tests sean independientes). Es por eso que el histograma de la figura 14 no coincide con la distribución gamma(2,2).

Dado que se calculó computacionalmente la distribución de T para tests aplicados al mismo par de listas (no independientes), se puede utilizar dicha variable como estadístico de un test T que combine los tests de Wilcoxon y G8E4.

Para ello, se calculo el estadístico observado T_{obs} evaluando la función T en los P values obtenidos en el Ejercicio 5: T_{obs}=T(pw_obs5,pt_obs5). Se puede observar el mismo en la Figura 14. Luego se calculó el P value del test T considerando que el mismo es un test a cola derecha (dado que un estadístico más extremo que el observado tendría P values p_w y p_t menores a pw_obs5 y pt_obs5 y en consecuencia un valor de T mayor a T_{obs}). Se calculó computacionalmente el P value del test T como la cantidad de valores simulados de T mayores a T_{obs} dividido la cantidad total N de valores

simulados.

El estadístico y el P value obtenidos resultaron iguales a $T_{obs} = 4{,}21$ y P value= 0,335.

7. Sé tu propia verduga

Se desea testear la veracidad de las patentes informadas:

H_o=Las patentes informadas no fueron inventadas

Por lo tanto, en esta sección se calculará el P value de un test a cola izquierda, para el cual un estadístico es más extremo que el observado si es menor a éste. La fórmula del P value para un test a cola izquierda es [4]:

$$\text{P value}|_{\text{cola izquierda}} = \Pr(t \leq t_{obs}|H_o) = F_t(t_{obs}) \tag{9}$$

Donde F_t es la cdf del estadístico t evaluada en el estadístico observado t_{obs}.

En esta sección, se utilizó la distribución pdf P_m(m;k,n) de la fórmula 6 con la cual se halló su distribución cdf: P_m_cdf(m;k,n). Utilizar la cdf de la distribución de m para calcular el P value conlleva menor tiempo computacional ya que sólo se debe evaluar en un punto de la cdf para calcular el P value. La deducción de la distribución cdf de P_m(m;k,n), utilizando la *Hockey-Stick identity* [6], se muestra a continuación:

$$\sum_{m=m_o+1}^{m_F} P(m;k,n) = \sum_{m=m_o+1}^{m_F} \binom{m-1}{k-1}\binom{n}{k}^{-1} = \binom{n}{k}^{-1}\left[\sum_{\tilde{m}=\tilde{k}}^{\tilde{m}_F}\binom{\tilde{m}}{\tilde{k}} - \sum_{\tilde{m}=\tilde{k}}^{\tilde{m}_o}\binom{\tilde{m}}{\tilde{k}}\right]$$

$$= \binom{n}{k}^{-1}\left[\binom{\tilde{m}_F+1}{\tilde{k}+1} - \binom{\tilde{m}_o-1}{\tilde{k}+1}\right] = \binom{n}{k}^{-1}\left[\binom{m_F}{k} - \binom{m_o}{k}\right]$$

$$P_{cdf} = 1 - \sum_{m=m_o+1}^{n} P(m;k,n) = 1 - \binom{n}{k}^{-1}\left[\binom{n}{k} - \binom{m_o}{k}\right] = \binom{m_o}{k}\binom{n}{k}^{-1}$$

$$P_{cdf} = \frac{m_o}{k}P(m_o;k,n)$$

Si se calcula el P value con una simulación de la densidad de distribución del estadístico m, se utiliza la fórmula del P value del test a cola izquierda 9. Este test es un test a cola izquierda, ya que una peor estimación es obtener una estimación para m más chica que la real.

El P value para la patente máxima como estadístico que se obtuvo fue: P_m_cdf(m,k,b)=0,2 (evaluando en m,k y b datos obtenidos). Se elige una significancia $\alpha = 0{,}05$ porque es la más usada y permite aceptar la hipótesis H_o (las patentes no fueron inventadas).

Además de realizar un test para el m (patente más nueva observada) se puede realizar un test para la mínima patente. Por una cuestión de simetría, sólo se debe reordenar la muestra haciendo : patentes al reves = patmax - lista de las patentes + 1. Ahora a la patente mínima le corresponde el número máximo que existe de patente así que podemos utilizar la distribución cdf del máximo m: P_m_cdf(minimo,k,b)=0,39. Dado que les lectores dieron libre elección para el valor de α, se determina un $\alpha = 0{,}05$ y deberán aceptar que la mínima patente informada no fue inventada.

A. Apéndice: Cálculo del estadístico de Kolmogorov

El estadístico de Kolmogorov se define como $D_n \coloneqq \sup\{|F(x) - S_n(x)|/x \in \mathbb{R}\})$, donde $F(x)$ es la función de distribución acumulada teórica (cdf teórica) y S_n (ecdf, cdf empírica) se define como $S_n(x) = 0$ si $x < x_1$, $S_n(x) = i/n$ si $x_i \leq x < x_{i+1}$ (con $1 \leq i \leq n-1$), y $S_n(x) = 1$ si $x \geq x_n$. Se asume que la función $F(x)$ es creciente (ya que es una cdf) y continua en todo $\mathbb{R}$, con $\lim_{x\to-\infty} F(x) = 0 \leq F(x) \leq \lim_{x\to+\infty} F(x) = 1$ (ya que es una cdf). En particular, la cdf de una distribución uniforme continua verifica dichos requisitos.

Nótese que calcular computacionalmente el supremo de dicha función en todos los reales puede tardar mucho en ejecutarse. Para optimizar su cálculo computacional, se puede plantear lo siguiente:

$$D_n \coloneqq \sup\{|F(x) - S_n(x)|/x \in \mathbb{R}\} = \text{máx}\{\sup\{|F(x) - S_n(x)|/x < x_1\}, \sup\{|F(x) - S_n(x)|/x_1 \leq x < x_n\}, \sup\{|F(x) - S_n(x)|/x \geq x_n\}\}$$

El primer supremo se puede calcular considerando que $S_n(x) = 0$ si $x < x_1$ y que $F(x)$ es creciente, continua y positiva: $\sup\{|F(x) - S_n(x)|/x < x_1\} = \sup\{|F(x)|/x < x_1\}) = \sup\{F(x)/x < x_1\} = F(x_1)$.

El último supremo se puede calcular considerando que $S_n(x) = 1$ si $x \geq x_n$ y que $F(x)$ es creciente, continua y menor o igual a 1: $\sup\{|F(x) - S_n(x)|/x \geq x_n\} = \sup\{|1 - F(x)|/x \geq x_n\} = \sup\{1 - F(x)/x \geq x_n\} = 1 - F(x_n)$.

El supremo del medio se puede calcular considerando

$$\sup\{|F(x) - S_n(x)|/x_1 \leq x < x_n\} = \text{máx}\{\sup\{|F(x) - S_n(x)|/x_i \leq x < x_{i+1}\}/1 \leq i \leq n-1\}$$

Y para calcular cada supremo de $|F(x) - S_n(x)|$ en cada intervalo $[x_i, x_{i+1})$ se puede considerar que en dicho intervalo $S_n(x) = i/n$. Por lo que:

$$\sup\{|F(x) - S_n(x)|/x_i \leq x < x_{i+1}\} = \sup\{|F(x) - i/n|/x_i \leq x < x_{i+1}\} = \text{máx}\{\underbrace{\sup\{F(x) - i/n/x_i \leq x < x_{i+1}\}}_{\coloneqq S_i^+}, \underbrace{\sup\{i/n - F(x)/x_i \leq x < x_{i+1}\}}_{\coloneqq S_i^-}\}$$

Se calcula S_i^+ considerando que $F(x) - i/n$ es creciente y continua en $[x_i,x_{i+1}]$ por lo que: $S_i^+ = F(x_{i+1}) - i/n$. Y análogamente, como $i/n - F(x)$ es decreciente y continua en $[x_i,x_{i+1}]$: $S_i^- = i/n - F(x_i)$. Luego $\sup\{|F(x) - S_n(x)|/x_i \leq x < x_{i+1}\} = \text{máx}\{F(x_{i+1}) - i/n, i/n - F(x_i)\}$. En consecuencia, se tiene que

$$\sup\{|F(x) - S_n(x)|/x_1 \leq x < x_n\} = \text{máx}\{\text{máx}\{F(x_{i+1}) - i/n, i/n - F(x_i)\}/x_i \leq x < x_{i+1}\}/1 \leq i \leq n-1\} = \text{máx}\{F(x_{i+1}) - i/n, i/n - F(x_i) \ / \ 1 \leq i \leq n-1\}$$

Y finalmente, se tiene que el estadístico de Kolmogorov-Smirnov es igual a

$$D_n = \text{máx}\{F(x_1), \text{máx}\{F(x_{i+1}) - i/n, i/n - F(x_i) \ / \ 1 \leq i \leq n-1\}, 1 - F(x_n)\} = \text{máx}\{F(x_1) - (1-1)/n, F(x_2) - 1/n, 1/n - F(x_1), F(x_3) - 2/n, 2/n - F(x_2), F(x_4) - 3/n, 3/n - F(x_3), (...), F(x_n) - (n-1)/n, (n-1)/n - F(x_{n-1}), n/n - F(x_n)\} = \text{máx}\{F(x_i) - (i-1)/n, i/n - F(x_i) \ / \ 1 \leq i \leq n\}$$

La fórmula

$$D_n = \text{máx}\{F(x_i) - (i-1)/n, i/n - F(x_i) \ / \ 1 \leq i \leq n\} \tag{10}$$

permite calcular computacionalmente el estadístico de Kolmogorov mediante el cálculo del máximo de una lista de $2n$ elementos (recordar que n es el número de datos empíricos). Este cálculo computacional es más simple y rápido en comparación con su definición original como el supremo de una función en todos los reales (habría que rellenar con muchísimos puntos y calcular la resta).

B. Apéndice: Suma de dos exponenciales es una gamma

Se puede demostrar por inducción la siguiente propiedad [7]:

$$\sum_{r=0}^{k-1} \frac{(\lambda \cdot t)^r \, e^{-\lambda t}}{r!} = \int_{\lambda \cdot t}^{\infty} \frac{z^{k-1} \, e^{-z}}{(k-1)!} \text{dz} \tag{11}$$

En nuestro caso, sólo se necesita la suma de dos exponenciales. Probar que la fórmula anterior vale para k-1=1 (dos exponenciales) es trivial:

$$\sum_{r=0}^{1} \frac{(\lambda \cdot t)^r \, e^{-\lambda t}}{r!} = e^{-\lambda t} + \lambda \cdot t \cdot e^{-\lambda t} = \int_{\lambda \cdot t}^{\infty} z \, e^{-z} \text{dz}$$

La integral sale por partes.

Referencias

[1] A.G.Frodesen, O.Skjeggestad *Probability and statistics in particle physics* pág 424-427

[2] https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test#Discrete_and_mixed_null_distribution

[3] A.G.Frodesen, O.Skjeggestad *Probability and statistics in particle physics* pág 450-452

[4] https://en.wikipedia.org/wiki/P-value#Definition_and_interpretation

[5] https://epub.ub.uni-muenchen.de/2094/1/paper_499.pdf sección 3

[6] https://en.wikipedia.org/wiki/Hockey-stick_identity

[7] A.G.Frodesen, O.Skjeggestad *Probability and statistics in particle physics* pág 95-98