Inferencia Estadística (Fecha: 14/09/2023)

Tarea #2

Estudiante: Juan Javier Monsivais Borjón *ID:*

Problema 1

Cuando una máquina no se ajusta adecuadamente, tiene una probabilidad del 0.15 de producir un artículo defectuoso. Diariamente, la máquina trabaja hasta que se producen 3 artículos defectuosos. Se detiene la máquina y se revisa para ajustarla. ¿Cuál es la probabilidad de que una máquina mal ajustada produzca 5 o más artículos antes de que sea detenida? ¿Cuál es el número promedio de artículos que la máquina producirá antes de ser detenida?

(Solución)

En este ejemplo podemos darnos cuenta que estamos ante una binomial negativa, esto debido aque nos preguntan probabilidad de obtener un r-ésimo éxito un conjunto de intentos, veamos:

Se nos dice que la probabilidad de un defectuosos es de 0.15, dado que la máquina está mal ajustada, se nos especifica además que estamos trabajando dentro de ese espacio de "máquinas mal ajustadas", lo cual nos permite ignorar alguna probabilidad condicional y así tomar las probabilidades de éxito $P(A) = 0.15$, donde A es el evento de un defectuoso.

Ahora bien, se nos pregunta por la probabilidad de que X tome valores mayores o iguales a 5, tendremos entonces:

$$P(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-r} \tag{1}$$

Para los valores:

$$\begin{aligned} p &= 0.15 \text{ Probabilidad de un defectuoso} \\ 1-p &= 0.85 \text{ Complemento de la probabilidad} \\ x &\geq 5 \text{ Número de intentos} \\ r &= 3 \text{ Número de éxitos} \end{aligned}$$

El primer acercamiento al problema podría parecer que debemos realizar una suma al infinito:

$$P(X \geq 5) = \sum_{x=5}^{\infty} \binom{x-1}{2} (0.15)^3 (0.85)^{x-3}$$

Sin embargo, el saber que estamos trabajando en un espacio de probabilidades nos permite obtener el complemento total, y al valor total restarle esta complemento:

$$P(x) = 1 - P(X < 5)$$

Nótese que al ser una variable discreta es muy importante denotar entre "$<$" y "$\leq$", en este problema debemos usar el menor estricto para el complemento.

Tendremos entonces:

$$P(X < 5) = P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4)$$

De lo anterior hay que observar que si X toma valores menores que tres, la probabilidad es nula, debido a que requerimos mínimo 3 intentos para obtener los 3 defectuosos.

$$\begin{aligned} P(X < 5) = P(X = 3) + P(X = 4) = \\ = \binom{2}{2}(0.15)^3(0.85)^0 + \binom{3}{2}(0.15)^3(0.85)^1 = \\ = 0.011981 \end{aligned}$$

Con lo anterior debemos obtener el complemento:

$$\begin{aligned} P(X \geq 5) = 1 - P(X < 5) = 1 - 0.011981 = \\ = \boxed{0.988019} \end{aligned}$$

(b)

Ahora también se nos pregunta el número promedio de artículos defectuosos, por lo que debemos obtener la esperanza de nuestra variable binomial negativa:

$$E[X] = r \cdot \frac{q}{p} = 3 \cdot \frac{0.85}{0.15} = \boxed{17} \tag{2}$$

Problema 2

Los empleados de una compañía de aislantes son sometidos a pruebas para detectar residuos de asbesto en sus pulmones. Se les ha pedido a la compañía que envíe a tres empleados, cuyas pruebas resulten positivas, a un centro médico para realizarles más análisis. Si se sospecha que el 40 % de los empleados tienen residuos de asbesto en sus pulmones, encuentra la probabilidad de que deban ser analizados exactamente 10 trabajadores para poder encontrar a 3 con resultado positivo.

(Solución)

Para el problema número dos tenemos un problema similar, que de hecho la edescripción encaja con una binomial negativa, con los parámetros:

$$\begin{aligned} r &= 3 \\ p &= 0.4 \quad \text{(Sospecha de probabilidad)} \\ 1-p &= 0.6 \\ x &= 10 \end{aligned}$$

Debido a que nos preguntan por un intento en particular, simplemente calculamos con este valor de x.

$$P(X=10)\binom{10-1}{3-1}(0.4)^3(0.6)^7 = \\ = \boxed{0.064497}$$

Problema 3

Para el siguiente ejercicio es necesario usar R.

a) Considere una moneda desequilibrada que tiene una probabilidad p de obtener águila. Usando el comando sample, escriba una función que simule N veces lanzamientos de esta moneda hasta obtener un águila. La función deberá recibir como parámetros la probabilidad p de obtener águila y el número N de veces que se repite el experimento; y deberá regresar un vector de longitud N que contenga el número de lanzamientos necesarios para obtener un águila en cada uno de los N experimentos.

b) Usando la función anterior, simule $N = 10^4$ veces una variable aleatoria Geom(p) para $p = 0.5, 0.1, 0.01$. Grafique las frecuencias normalizadas en color azul. Luego, sobre esta última figura, empalme en rojo la gráfica de la función de masa correspondiente. ¿Qué observa?

c) Repita el inciso anterior para $N = 10^6$. Además, calcule el promedio y la desviación estándar de las simulaciones realizadas. ¿Qué observa?

(Solución)

(b)

En el problema tres se nos pide hacer una función que simule el lanzamiento de una moneda de manera iterativa hasta obtener un águila, variando las probabilidades de la moneda (moneda cargada), con valores: $0.5, 0.1, 0.01$

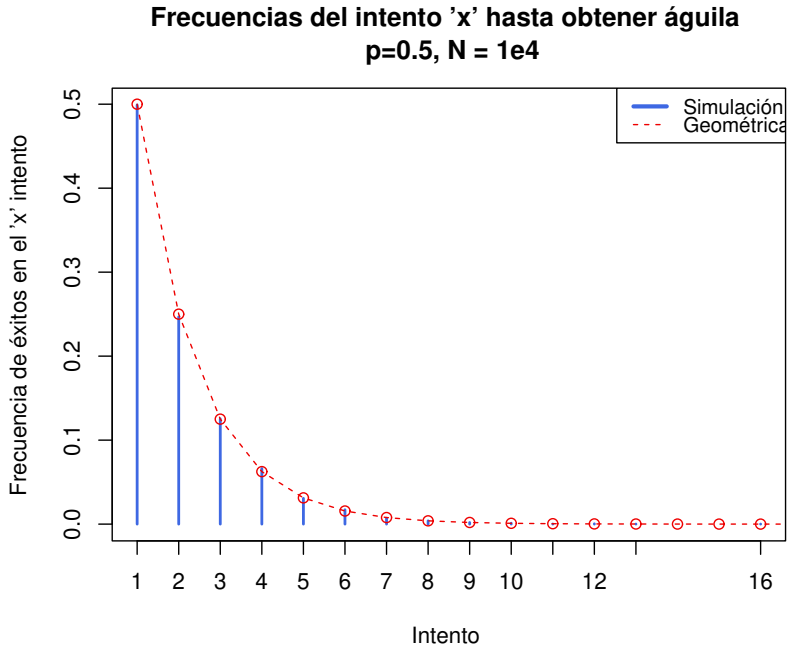

Figura 1: Frecuencias del número de intentos "x", necesarios para obtener un águila, con $p = 0.5$ y $N = 10000$

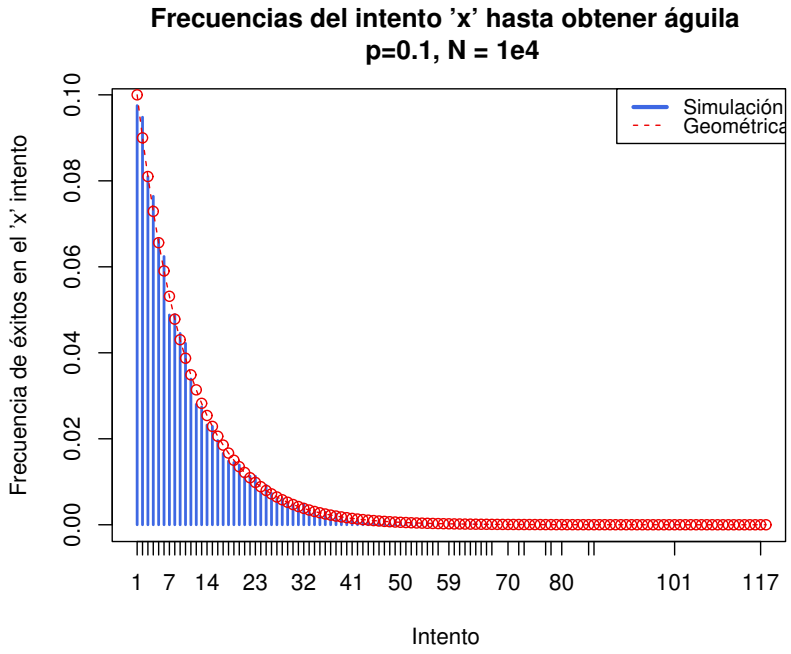

Figura 2: Frecuencias del número de intentos "x", necesarios para obtener un águila, con $p = 0.1$ y $N = 10,000$

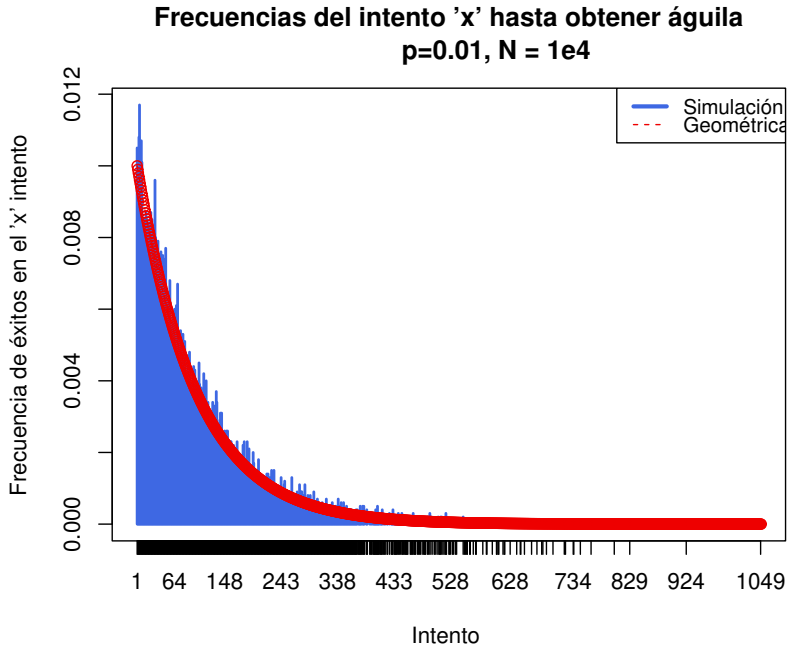

Figura 3: Frecuencias del número de intentos "x", necesarios para obtener un águila, con $p = 0.01$ y $N = 10000$

Una vez se presentan los gráficos debemos notar algo muy interesante, que la misma distribución geométrica nos "advierte". Cuando se obtuvieron las expresiones para la varianza de la distribción geométrica se obtuvo:

$$Var(X) = \frac{1-p}{p^2} \tag{3}$$

Observando el término del denominador aumenta/disminuye de forma cuadrada, por lo que a probabilidades más bajas la variación será mucho mayor.

Con ello en mente debemos observar que en efecto para $p = 0.5$ la predicción es perfecta, conforme disminuimos la probabilidad (Las dos gráficas siguientes), empuezan a observarse picos que no concuerdan con la predicción. Lo anterior podemos atribuirlo a que, al ser menor la probabilidad, el número de veces que tendré que lanzar la moneda será naturalmente mayor, por ende más posibles "casillas" en las que podría caer. Observe el número máximo para cada experimento, para $p = 0.01$ hubo casos que hasta el intento 548 se obtuvo un águila, pero pudo haber caído en otros 547, lo que nos habla que el poder predictivo de la geométrica para p pequeños es pobre.

(c)

La anterior discusión nos permite introducir las simulaciones para N muy grandes y ver si el hacer más pruebas nos ayuda a ajustar mejor nuestros experimentos con la geométrica:

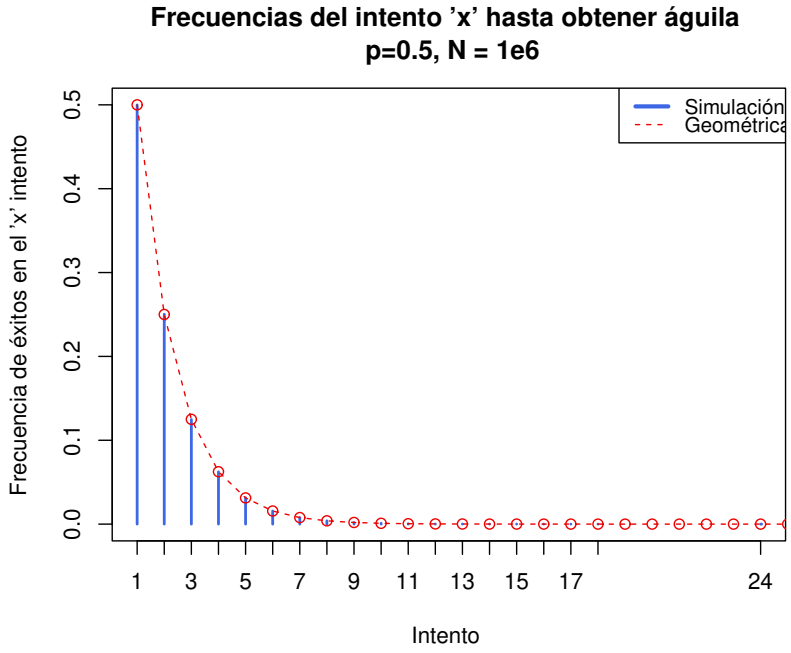

Figura 4: Frecuencias del número de intentos "x", necesarios para obtener un águila, con $p = 0.5$ y $N = 1,000,000$

Para el primer caso, no deberíamos observar un comportamiento peor que para $N = 10,000$, por lo que no comentaremos mucho de este experimento. A continuación presentamos los otros dos experimentos:

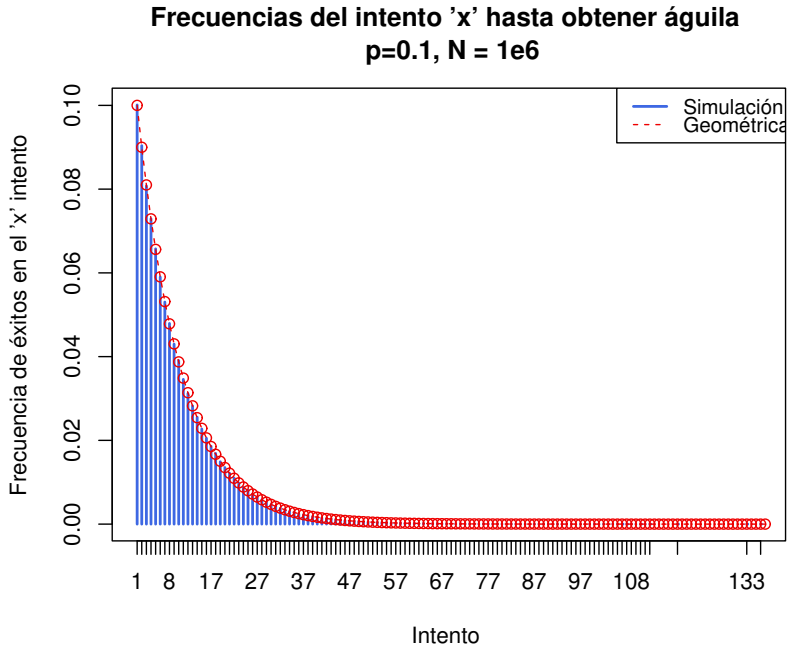

Figura 5: Frecuencias del número de intentos "x", necesarios para obtener un águila, con $p = 0.1$ y $N = 1,000,000$

Y finalmente tenemos el último experimetno que no pudo ajustarse de manera decuada con $N = 10,000$, veamos lo que ocurre con dos órdenes de magnitud más de pruebas.

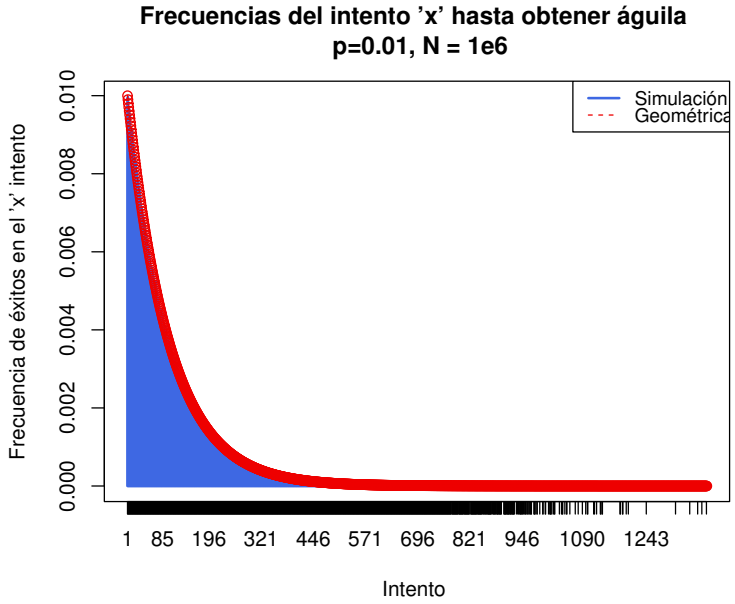

Figura 6: Frecuencias del número de intentos "x", necesarios para obtener un águila, con $p = 0.01$ y $N = 1,000,000$

En primera instancia podríamos decir que efectivamente nuestra predicción es mucho mejor, en términos relativos es claro que sí, nuestra geométrica se ha ajustado mucho mejor, observemos un zoom para verlo con mayor claridad:

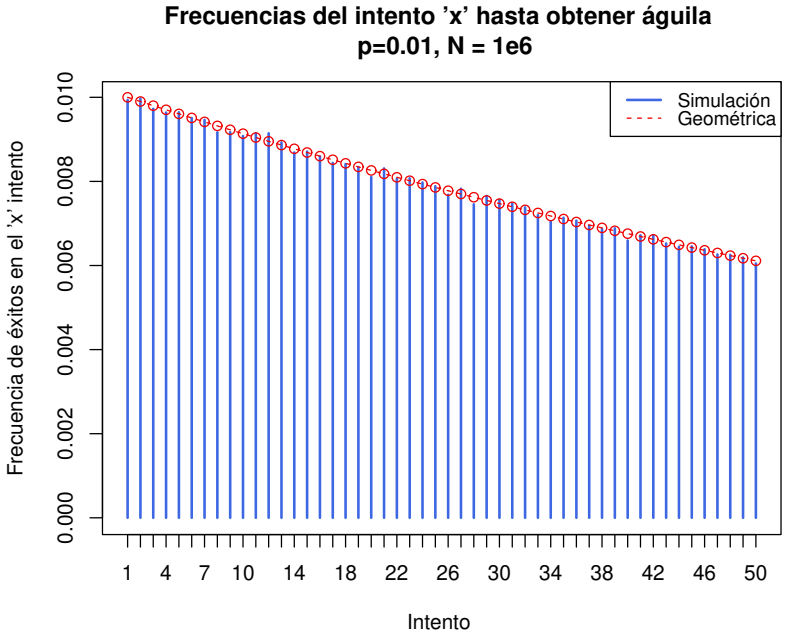

Figura 7: Zoom de 6, con $p = 0.01$ y $N = 1,000,000$

Las medias para cada caso se presentan a continuación:

$$\mu_1 = \frac{1}{p} = \frac{1}{0.5} = 2 \quad \text{Para p} = 0.5$$
$$\mu_2 = \frac{1}{0.1} = 10 \quad \text{Para p} = 0.1$$
$$\mu_2 = \frac{1}{0.01} = 100 \quad \text{Para p} = 0.01$$

Y sus respectivas desviaciones:

$$\sigma_1 = \sqrt{(1-p)/p^2} = \sqrt{2}$$
$$\sigma_2 = \sqrt{90}$$
$$\sigma_3 = \sqrt{9900}$$

Es claro que aumenta mucho más rápido la varianza que la media conforme p disminuye.

Problema 4

Siguiendo las ideas del problema anterior, escriba una función en R que simule N veces los lanzamientos de una moneda hasta obtener r águilas. La función deberá recibir como parámetros la probabilidad p de obtener águila, el número r de águilas a observar antes de detener el experimento y el número N de veces que se repite el experimento. Deberá regresar un vector de longitud N que contenga el número de lanzamientos necesarios para obtener las r águilas en cada uno de los N experimentos. Luego, grafique las frecuencias normalizadas de los experimentos para $N = 10^6$, $p = 0.2$ y 0.1, y $r = 2$ y 7. Finalmente, compare estos resultados contra la función de masa de la distribución más adecuada para modelar este tipo de experimentos.

(Solución)

Para este problema tenemos un caso similar al anterior, salvo la modificación de que el intento se detendrá hasta obtener el r-ésimo éxito, ya no a la primer águila, ello lleva al uso de la binomial negativa, que representa una extención de la geométrica. A continuación se presentan los experimentos:

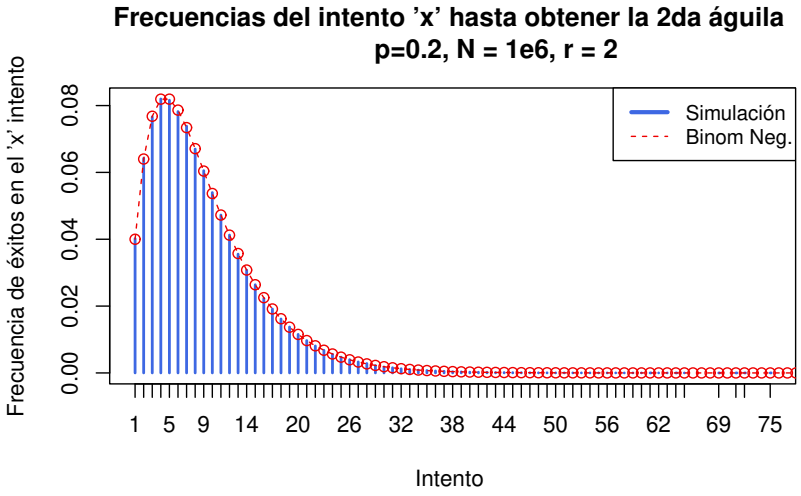

Figura 8

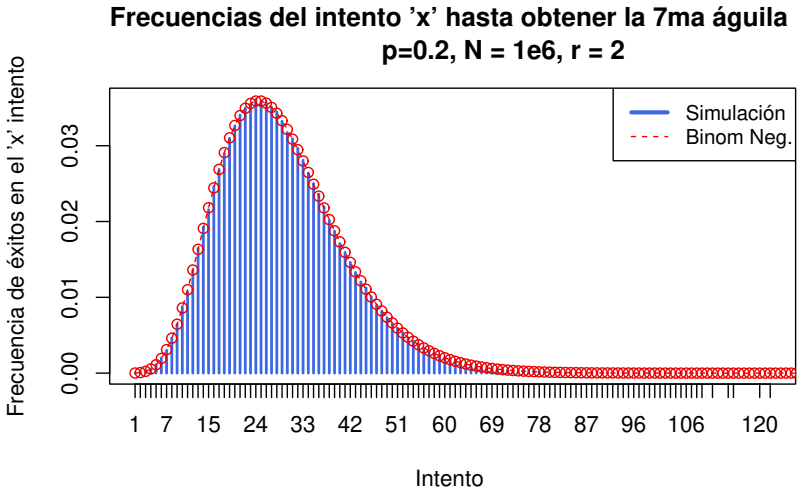

Figura 9

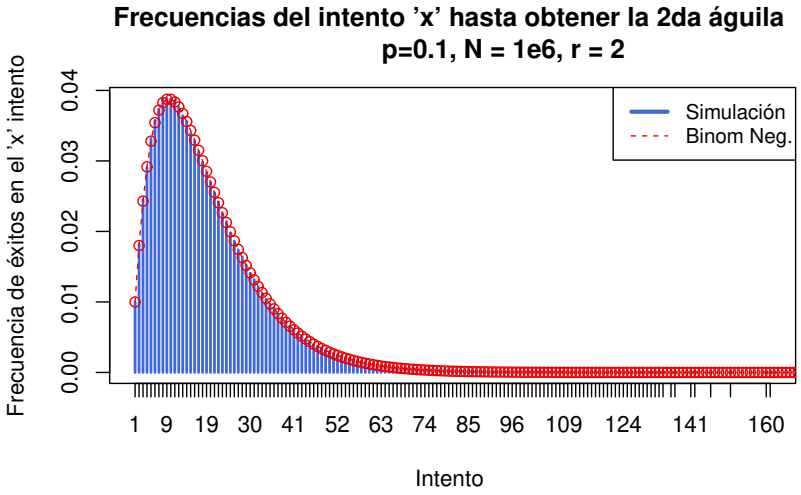

Figura 10

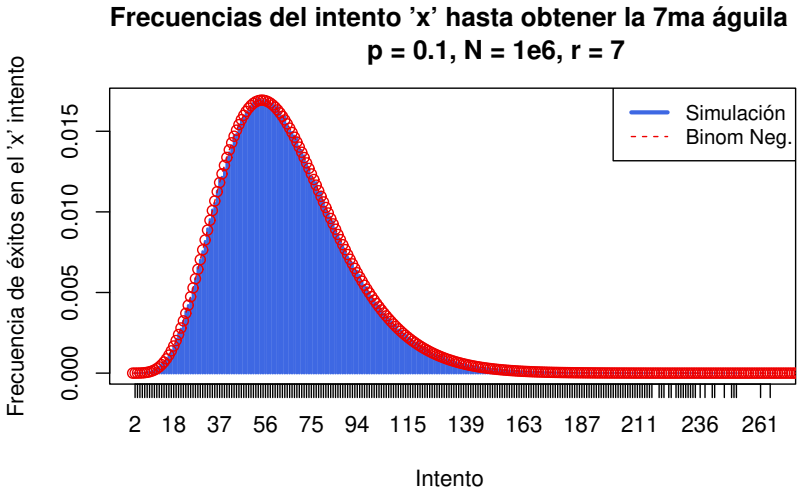

Figura 11

Algo que debe recalcarse es que, a pesar de que el problema no lo pida, si la simulación se ejecutara para $10,000$ pruebas, la variabilidad sería enorme incluso más que para la geométrica debido a que la binomial negativa también se ve afectada en la varianza por el factor r, veamos el ejemplo de $N = 10,000$, $p = 0.02$, $r = 7$

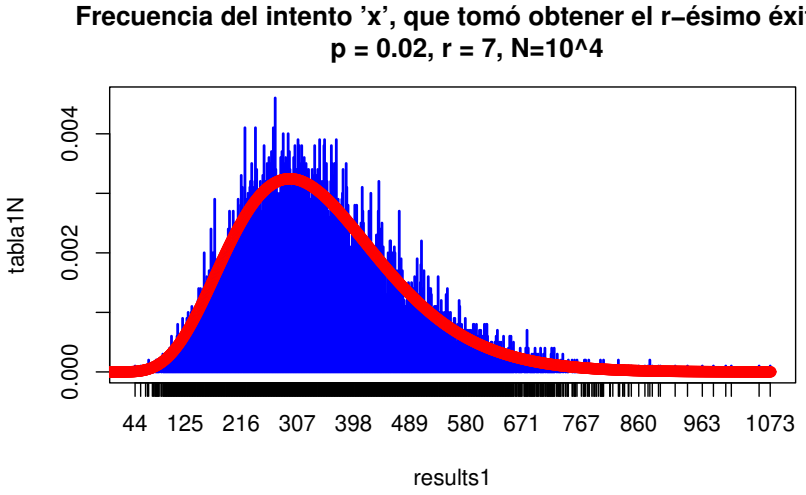

Figura 12

Se observa de manera clara que para p y N pequeños hay una sub-estimación de la simulación.

Como se mencionó al principio, la geométrica es un caso especial de la binomial negativa con $r = 1$, por lo que su varianza directamente nos queda:

$$Var(X) = r \cdot \frac{(1-p)}{p^2}$$

Problema 5

Consideremos X como una variable aleatoria con función de distribución F y función de densidad f. Sea A un intervalo en la línea real R. Definimos la función indicadora $1_A(x)$:

$$1_A(x) = \begin{cases} 1 & \text{si } x \in A \\ 0 & \text{en otro caso} \end{cases} \tag{4}$$

Sea $Y = 1_A(X)$. Encuentre una expresión para la distribución acumulada y el valor esperado de Y.

(Solución)

(a)

Se nos pide hacer una cambio de variable de la forma:

$$Y = 1_A(x) \tag{5}$$

Y se nos pide encontrar la función acumulada $F_Y(y)$, recordemos que la podemos expresar como la probabilidad hasta ese punto:

$$F_Y(y) = P(Y \leq y) = F(1_A(x) \leq y)$$

Lo anterior podemos subdividirlo en los casos de la función $1_A(x)$.

Para $y = 1$ significa que 'x' cae en A, el hecho de que caiga en 'A', nos permite usar la variable X y recalcular en terminos de la misma la acumulada:

$$F_Y(1) = P(1_A(x) = 1) = P(X \in A)$$

Notemos que no se ha considerado el caso $y < 0$, ya que $Y = 1, 0$. Por lo que la probabilidad es 0.

Para $y = 0$

$$F_Y(0) = P(1_A(x) = 0) = P(X \notin A)$$

Lo que es igual a:

$$F_Y(0) = P(1_A(x) = 0) = 1 - P(X \in A)$$

Para obtener el valor esperado se tiene:

$$E(Y) = E(r(X)) = \int r(x) dF_X(x) \tag{6}$$

$$E(Y) = \int 1_A(x) f_x(x) dx$$

El valor esperado solo "sobrevive", para $1_A = 1$ por lo que:

$$\int_A 1 \cdot f_X(x) dx = \int_A f_X(x) dx = P(X \in A).$$

Problema 6

Las calificaciones de un estudiante de primer semestre en un examen de química se describen por la densidad de probabilidad:

$$f_y(y) = 6y(1-y) \quad 0 \leq y \leq 1 \tag{7}$$

Donde 'y' representa la proporción de preguntas que el estudiante contesta correctamente. Cualquier calificación menor a 0.4 es reprobatoria. Responda lo siguiente:

a) ¿Cuál es la probabilidad de que un estudiante repruebe?

b) Si 6 estudiantes toman el examen, ¿cuál es la probabilidad de que exactamente 2 reprueben?

(Solución)

Primero debemos verificar que sí sea una distribución de probabilidad, para lo cual debemos verificar que vale:

$$\int_{-\infty}^{\infty} f(x) = 1 \tag{8}$$

Dado que la función solo está definida entre 0 y 1:

$$\int_0^1 6y(1-y) = (3y^2 - 2y^3)\Big|_0^1 = \\ = 1$$

Por lo que vale como función de densidad de probabilidad, lo siguiente es:

(a)

Se nos dice que una calificación menor a 0.4 es reprobatoria, ($y < 0.4$ considerando que todas las preguntas valen lo mismo).

$$P(X < 0.4) = F_y(0.4) = 3(0.4)^2 - 2(0.4)^3 = \boxed{0.352}$$

(b)

Para el inciso b, debemos notar que ya tenemos la probabilidad de que alguien en general repruebe, la parte continua ya no la usaremos, ahora tenemos un problema discreto con una probabilidad obtenida mediante una distribución continua.

Ahora bien, si consideramos independencia entre alumnos, es fácil ver que debemos usar una binomial: $Y \sim Binom(0.352, 6)$;

$$P(y = 2) = \binom{6}{2}(0.352)^2(0.648)^4 = \boxed{0.3277}$$

Problema 7

Aquí tienes el texto corregido:

Escriba una función en R que simule una aproximación al proceso Poisson a partir de las 5 hipótesis que usamos en clase para construir dicho proceso. Usando esta función, simule tres trayectorias de un proceso Poisson con $\lambda = 2$ sobre el intervalo $[0, 10]$ y grafíquelas. Además, simule 104 veces un proceso de Poisson N con $\lambda = \frac{1}{2}$ hasta el tiempo $t = 1$. Luego, haga un histograma de $N(1)$ en su simulación anterior y compárelo con la distribución de Poisson correspondiente.

Sugerencia: Considere el intervalo $[0, T]$ y un número real positivo dt que sea mucho más pequeño que la longitud de $[0, T]$, de manera que divida dicha longitud en, digamos, $\frac{T}{dt} = 1000$ intervalos. Divida el intervalo $[0, T]$ en intervalos de longitud dt que tengan la forma $(k \cdot dt, (k+1) \cdot dt]$, donde $k = 0, 1, 2, \ldots, \left(\frac{T}{dt} - 1\right)$. Para cada uno de estos intervalos, simule una variable aleatoria Bernoulli con parámetro $(\lambda \cdot dt + 10^{-6})$ y guarde su resultado en un vector del tamaño adecuado.

(Solución)

Haciendo uso de la sugerencia se obtuvieron las siguientes trayectorias para T de 0 a 10, y $\lambda = 2$

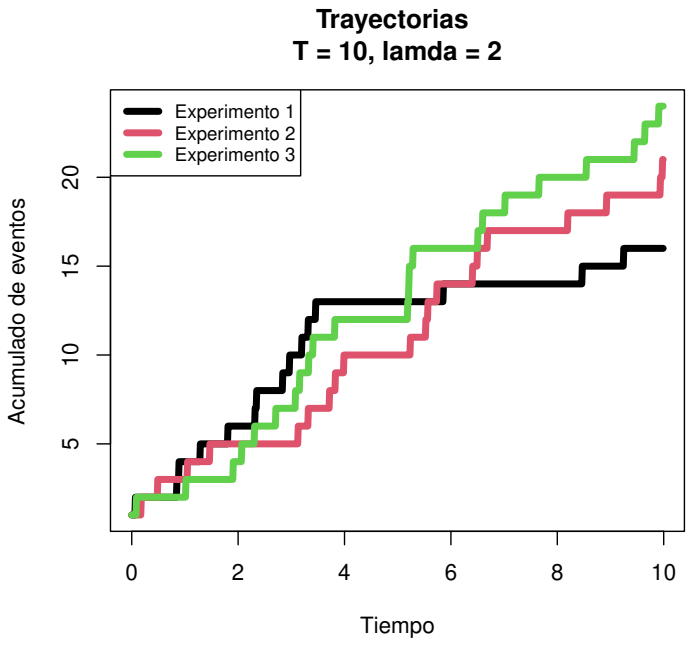

Figura 13

Lo que observamos en la gráfica son procesos de Poisson, en la que al avanzar en el tiempo la detección de un "fenómeno" se registra, y se va acumulando según se encuentra el siguiente, lo anterior para intervalos de una milésima del intervalo grande T. Por ende se puede reducir a

un proceso de Bernulli con dt muy pequeños. El proceso termina hasta alcanzar T, Y ese será el número de eventos registrados para el intervalo. Esta idea nos lleva a intentar simular muchos procesos de Poisson y así comparar el número de eventos registrados para cada intervalo.

Se muestran a continuación los resultados de la simulación comparados con la distribución Poisson correspondiente.

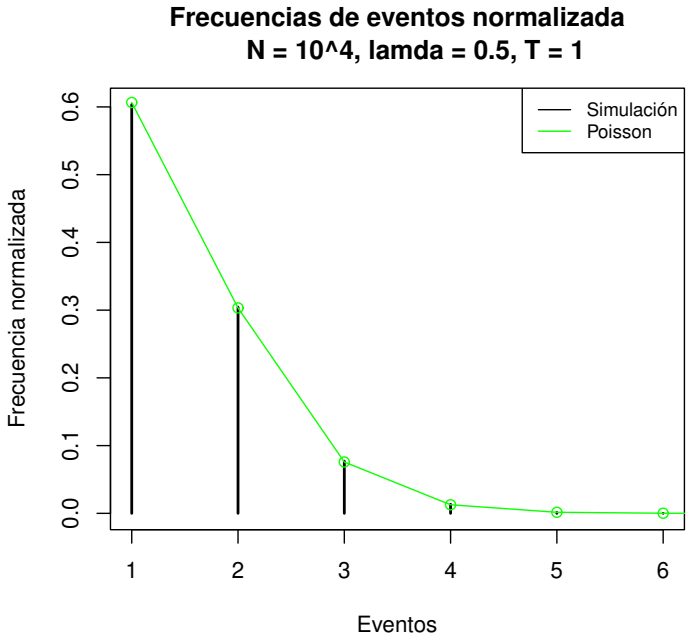

Figura 14: En verde se muestra la Poiss(1,$\lambda = 0.5$)

Problema 8

Considere la función:

$$F(x) = \begin{cases} 0 & \text{para } x < 0 \\ 0.1 & \text{para } x = 0 \\ 0.1 + 0.8x & \text{para } 0 < x < 3/4 \\ 1 & \text{para } 3/4 \leq x \end{cases} \tag{9}$$

¿Es una función de distribución? Si lo es, ¿corresponde a una variable aleatoria continua o discreta?

(Solución)

Debemos demostrar que se cumple el teorema 2.8 [1].

Lo primero que debemos demostrar es que en ningún punto es decreciente, el intervalo que podría preocupar es el de la recta:

$$0.1 + 0.8x \quad 0 < x \leq 3/4 \tag{10}$$

Para demostrar su que no decrece podemos derivar la función, si la función es positiva o cero en el intervalo entonces cumple con lo primero:

$$\frac{d}{dx}(0.1 + 0.8x) = 0.8 > 0$$

Lo siguiente que debe cumplir es que esté normalizada:

$$\lim_{x \to -\infty} F(x) = 0 \tag{11}$$

$$\lim_{x \to \infty} F(x) = 1 \tag{12}$$

En el primer caso tomamos $F(x) = 0$ ya que vara x que tiende a menos infinito le corresponderá esta función.

Para el segundo caso tomamos $F(x) = 1$, para valores mayores a $3/4$.

Y finalmente debemos demostrar que nuestra función es continua por la derecha:

$$F(x^+) = \lim_{y \to x} F(y) \quad y > x$$

Analizando los puntos en los que hay corte, por ejemplo en $x = 0$:

$$\lim_{y \to 0} F(y) = 0.1 + 0.8(0) = 0.1$$

Que debe ser igual a $F(0) = 0.1$ lo cual es cierto, por la misma definición de la función.

Y el otro punto que debemos analizar es $x = 3/4$,

$$\lim_{y \to 3/4} F(y) = 1$$

Que debe coincidir con $F(3/4) = 1$, con todo lo anterior queda demostrado que en efecto representa una función de de distribución acumulada.

(b)

Ahora, para responder a la pregunta de si pertenece a una v.a discreta o continua, podemos comenzar analizando el caso $0 < x < 3/4$, dado que es una recta no hay manera de definir una función de masa, es decir, una función discreta, ya que en todo caso debería ser infinitesimal, lo que la definiría como una funcion continua, pero nos encontramos con que al graficar la función, tenemos saltos que de ninguna manera podrían dar a una función continua en su totalidad.

La única posibilidad es que sea **mixta** en los puntos de salto de la función, para, de esta manera poder tener definida da probabilidad en puntos sin ser 0.

$$P(X = 3/4) = 1 - \lim_{x \to (3/4)^{-1}} F(x) = 1 - 0.7 = 0.3 \tag{13}$$

$$P(X = 0) = 0.1 - \lim_{x \to (0)^{-1}} F(x) = 0.1 - 0 = 0.1 \tag{14}$$

Tendremos pues dos puntos con masa de probabilidad diferente de 0.

$$P(X = 0) = 0.1$$

y

$$P(X = 1) = 0.3$$

De otra manera, lo que pasaría si la tratamos como una variable aleatoria continua, en teoría solo deberíamos derivar la función y nos quedaría $f_x(x) = 0.8 \quad 0 < x < 3/4$, si hiciéramos la integral en el intervalo definido no nos daría la unidad. debemos sumar los puntos de masa para que esto se cumpla en los puntos de corte.

Problema 9

El archivo *Delitos.csv* contiene información sobre los delitos denunciados en la ciudad de Aguascalientes durante el período que abarca desde enero de 2011 hasta junio de 2016. Este archivo consta de 5 columnas:

1. La primera columna muestra la fecha en la que se denunció el delito.
2. La columna *TIPO* proporciona una descripción del tipo de delito.
3. La columna *CONCATENADO* ofrece una descripción más detallada del delito.
4. La columna *SEMANA* indica la semana del año a la que corresponde la fecha de denuncia.
5. La columna *SEMANA COMPLETAS* señala la semana específica en la que se realizó la denuncia durante el período de estudio.

Se busca analizar el comportamiento semanal de los delitos utilizando métodos gráficos, como los *boxplots*, y discutir posibles modelos que permitan describir de manera adecuada los delitos cometidos en esta área.

(Solución)

Este último problema resulta ser más libre en la forma de abordarlo, nosotros tenemos dos formas de comenzar el análisis, en primera instancia debemos agrupar según la categoría, tenemos 23 categorías

Hemos de notar que puede existir cierta relación entre cada tipo delito, cosa que no podemos corroborar hasta seguir ciertos pasos.

Lo primero que deberíamos hacer es ver la cantidad de datos que tenemos para cada categoría, muchos de ellos son despreciables, que incluso podrían causar ruido en el análisis, por mencionar algunos, a a continuación se muestran los Boxplots de las categorías, primero lo haremos por semana del año.

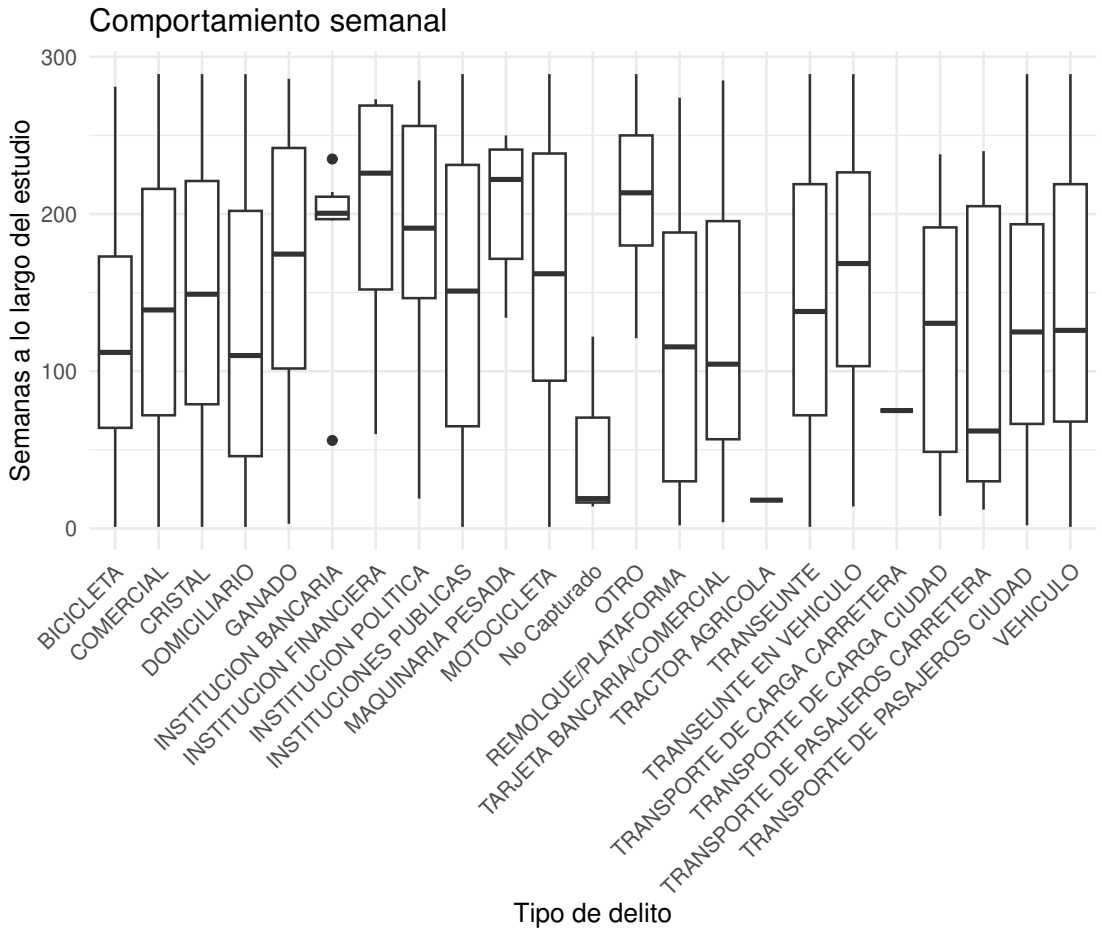

Figura 15: Caption

Al observar el gráfico vemos que para algunos como se mencionaba ni siquiera existe la cantidad de datos suficientes para dar una estimación, tal es el caso de "tractor agricola", "transporte de carga a carretera", etc.

Lo más tentador es sin duda alguna anilizar los datos por categoría como se muestra en la gráfica anterior, sin embargo dadas las condiciones para algunas categorías, el panorama general del número de delitos por semana (sin categorizar) al contener más datos, probablemente nos de una mejor estimación del fenómeno. Veamos el Boxplot:

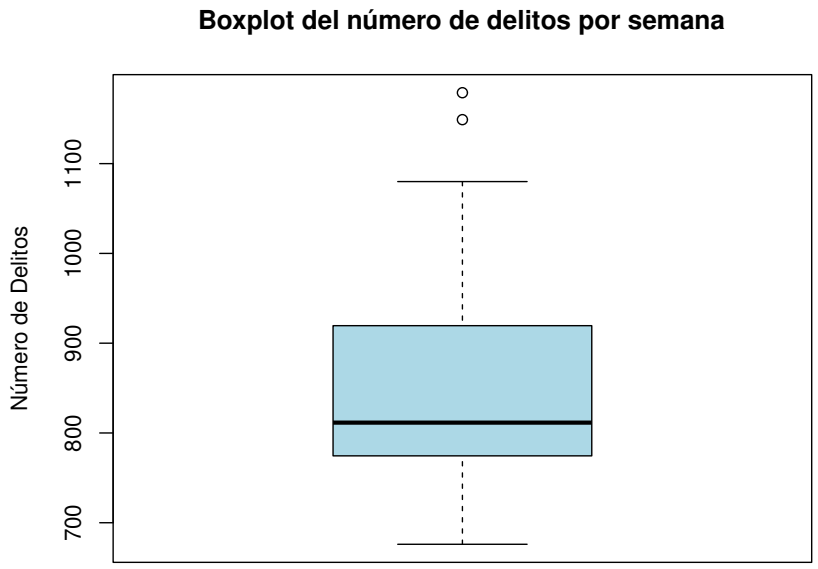

Figura 16

Hay dos cosas que resaltan a la vista, el primero que tenemos 2 valores atípico correspondientes a las semanas 17 y 18 del año en la que hubo considerablemente más delitos. Lo siguiente un notable sesgo hacia la izquierda, veamos el histograma de frecuencias:

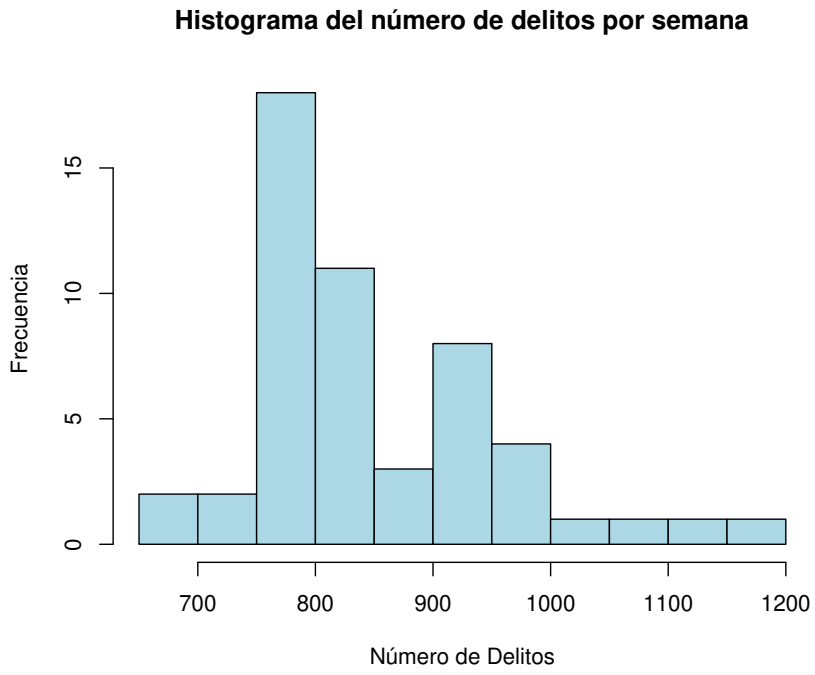

Figura 17

La propuesta final dada la descripción y la aperente geometría del gráfico lleva a pensar en una Poisson, con intervalos semanales, y una media de 850 delitos por semana.

Al visualizar la acumulada empírica tenemos:

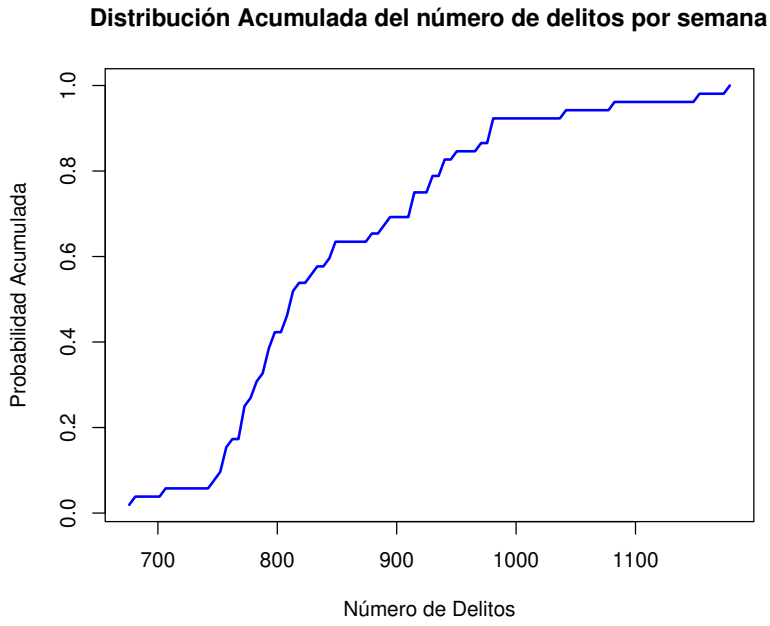

Figura 18

Una vez más, podría en la primera parte sugerirnos una Poisson, pero realmente se tiene que tener cuidado, empíricamente tenemos:

$$\mu = 850$$
$$\sigma = 109$$

La diferencia tan grande entre la media y la desviación estándar no nos permiten hacer uso de la Poisson, además tenemos una tasa de delitos muy altos, cosa que tampoco concuerda con la Poisson en la que se toman en cuenta sucesos muy raros.

Otra opción que podría proponerse es la normal, pero dada la asimetría no parece encajar del todo. No debemos olvidar que estamos considerando que la cantidad de delitos se comporta igual entre un año y otro, asumir comportamiento constante entre cada año puede no ser lo correcto.

Referencias

[1] L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference*. Springer, 2008.