Inferencia Estadística | Tarea 02

Aguirre Calzadilla César Miguel

12 de septiembre de 2024

Códigos

Todo el código escrito para esta tarea será anexado en un archivo de RStudio. Dentro se encuentran las rutinas escritas para la tarea así como comentarios sobre las mismas.

Problema 1

Una pareja decide tener hijos hasta el nacimiento de la primera niña. Calcule la probabilidad de que tengan más de cuatro hijos. Suponga que las probabilidades de tener niño o niña son iguales. ¿Cuál es el tamaño esperado de la familia?

Primero, intentemos determinar qué nos está diciendo el problema. Lo que dice el enunciado es que hay independencia de eventos en cada observación, es decir, el nacimiento de un hijo o hija no afectará el sexo de la siguiente hija o hijo que nazca. Además, la probabilidad de que haya un éxito es constante, esto significa que la probabilidad de que nazca una niña es de $p = 0.5$. Sumado a ello, consideremos, por definición, que $q = 1 - p = 0.5$, la probabilidad de que nazca un niño. Asimismo, cada ensayo es de Bernoulli, i.e., hay solo dos opciones posibles, que el bebé sea niño o que sea niña.

Por lo tanto, usamos distribución geométrica.

La distribución geométrica (DG), modela el número de ensayos X hasta el primer éxito. La función de masa de probabilidad (PMF) de la DG está dada como:

$$P(X = k) = (1 - p)^{k-1} p$$

Con $k :=$ el número de ensayos y $p :=$ probabilidad de éxito, para el caso, iguala 0.5.

Como queremos conocer la probabilidad de que la pareja tenga más de cuatro hijos antes de tener una niña, entonces queremos conocer $P(X > 4)$.

Por identidad, sabemos que $P(X > 4) = 1 - P(X \leq 4)$, y sabemos que $P(X \leq 4) = \sum_{i=1}^{4} P(X = i)$. Por lo tanto, usando la función de masa de probabilidad de la distribución geométrica, nos queda:

$$P(X = k) = (1 - 0.5)^{k-1}(0.5) = (0.5)(0.5)^{k-1} = (0.5)^k \quad \forall \quad k = 1, 2, 3, 4$$

Entonces:

$$P(X \leq 4) = 0.5^1 + 0.5^2 + 0.5^3 + 0.5^4 = 0.9375$$

$$\Rightarrow P(X > 4) = 1 - P(X \leq 4) = 1 - 0.9375 = 0.0675$$

Por lo tanto, la probabilidad de tener más de cuatro hijos antes del primer éxito, es decir, antes de tener una niña es del 6.25 %.

Cuando nos preguntan por el tamaño esperado de la familia hasta el nacimiento de la primera niña, en realidad nos están preguntando pro la media (i.e. el valor esperado) que X sigue en una distribución geométrica. Esto es:

$$E(X) = \frac{1}{p} = \frac{1}{0.5} = 2$$

Como el valor esperado es 2, esto significa que, en promedio, se esperan 2 hijos hasta que nazca la primera niña.

Problema 2

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

Nuevamente, tratemos de identificar qué distribución de probabilidad es la más adecuada para el problema. Por el enunciado, se nos está pidiendo modelar el número total de artículos hasta obtener tres defectuosos.

Sabemos que la distribución binomial negativa[1] (DBN) modela el número de ensayos necesarios hasta obtener un número fijo de éxitos. Para este problema, se satisfacen los supuestos:

- **Independencia de observaciones**: el resultado de éxito o fracaso de un artículo no afecta al resultado del siguiente.
- **Probabilidad constante de éxito**: la probabilidad de obtener un artículo defectuosos es constante para cada prueba. Aquí, igual a $p = 0.15$.
- **Número fijo de éxitos**: el proceso continúa hasta alcanzar un número fijo de éxitos deseados, i.e, artículos defectuosos.

Tenemos que la función de masa de probabilidad (PMF) de la binomial negativa es:

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

Con $k :=$ el número total de ensayos hasta el nuestro r deseado; $r :=$ el número de éxitos (artículos defectuosos) deseados; $p :=$ probabilidad de éxito.

Para nuestro caso, $p = 0.15$ y $r = 3$. Ahora, queremos saber la probabilidad de que la máquia produzca cinco o más artículos antes de ser detenida, i.e., $P(X \geq 5)$. Por identidad, sabemos que: $P(X \geq 5) = 1 - P(X < 5)$. Para $P(X < 5)$ sumamos la probabilidad de que la máquina produzca tres o cuatro artículos antes de ser detenida, entonces $P(X < 5) = P(X < 3) + P(X < 4)$.

Entonces: Para $X = 3$:

$$P(X = 3) = \binom{3-1}{3-1}(1-p)^{3-3}p^3 = \binom{2}{2}(1-0.15)^0(0.15)^3 = 1 \cdot 1 \cdot (0.15)^3 = 0.003375$$

[1] La diferencia clave entre el problema actual y el problema anterior está en la naturaleza de las distribuciones utilziadas.

La distribución geométrica se modela el número de ensayos necesarios hasta el primer éxito. En otras palabras, te dice cuántos intentos se necesitan antes de obtener el primer evento de interés (por ejemplo, el primer artículo defectuoso). La DG se utiliza cuando queremos conocer el número de ensayos hasta que ocurra un único éxito.

Por otra parte, la distribución binomial negativa (DBN) modela el número total de ensayos necesarios para alcanzar un número fijo de éxitos. No importa cuántos ensayos sean necesarios para alcanzar el primer éxito; lo que cuenta es el total de ensayos requeridos para obtener un número predefinido de éxitos (en este caso, tres artículos defectuosos).

Para $X = 4$:

$$P(X = 4) = \binom{4-1}{3-1}(1-p)^{4-3}p^3 = \binom{3}{2}(1-0.15)^1(0.15)^3$$

$$\Rightarrow P(X = 4) = 3 \cdot (1-0.15) \cdot 0.03375 = 3 \cdot 0.85 \cdot 0.03375 = 0.0086$$

Entonces:

$$P(X < 5) = 0.003375 + 0.0086 = 0.0119$$

$$\Rightarrow P(X \geq 5) = 1 - P(X < 5) = 1 - 0.0119 = 0.98$$

Por lo tanto, la probabilidad de que la máquina produzca cinco o más artículos antes de que se detenga es del 98 %.

El número promedio de artículos producidos antes de que la máquina sea detenida corresponde a la media de la distribución binomial negativa:

$$E(X) = \frac{r}{p} = \frac{3}{0.15} = 20$$

Por lo tanto, el promedio de artículos producidos antes de ser detenida será de 20 artículos.

Problema 3

Los empleados de una compañía de aislantes son sometidos a pruebas para detectar residuos en sus pulmones. Se le 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 10 trabajadores para poder encontrar a tres con resultado positivo.

Para este problema, queremos modelar el número total de empleados que deben ser analizados hasta dar con tres que resulten positivos, i.e., con presencia de asbesto en sus pulmones.

Al inicio, dudé de si debía utilizar binomial o binomial negativa. Sin embargo, la distribución binomial nos pide un problema que solicite cierto número de éxitos con una cantidad de observaciones fija. Para este caso, tenemos lo siguiente:

- Número de éxitos fijo: haremos un estudio hasta encontrar un número fijo de éxitos. Para nuestro problema, igual a tres.
- Independencia de eventos: cada observación es independiente de las demás, i.e., no importa que salió en la prueba n, su resultado no afecta al de la prueba $n + 1$.
- Resultado por observación: cada resultado es de Bernoulli, solo hay éxito (con residuos) o fracaso (sin residuos).
- Probabilidad de éxito constante: para toda prueba, se sabe que la probabilidad de obtener un éxito (trabajador con residuos de asbesto) es $p = 0.4$..

Por lo tanto, debemos trabajar con una distribución binomial negativa (DBN). Esto se decide así por los supuestos anteriores, que se ajustan a los de la DBN pues mientras la binomial fija el número de ensayos, la binomial negativa fija el número de éxitos sin importar cuántas observaciones se hagan. En otras palabras, estamos viendo el número de fracasos hasta dar con la cantidad de éxitos fijos.

Sabemos que la DBN está representada por:

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

Donde r := el número de éxitos deseado, para el caso igual a tres; k := al número total de empleados analizados hasta obtener r; p := la probabilidad de éxito tras cada prueba, aquí igual a 0.4.

Entonces, se nos pide encontrar la probabilidad de que deban ser analizados 10 empleados para encontrar exactamente tres empleados con asbesto. Esto es, $P(X = 10)$. Entonces:

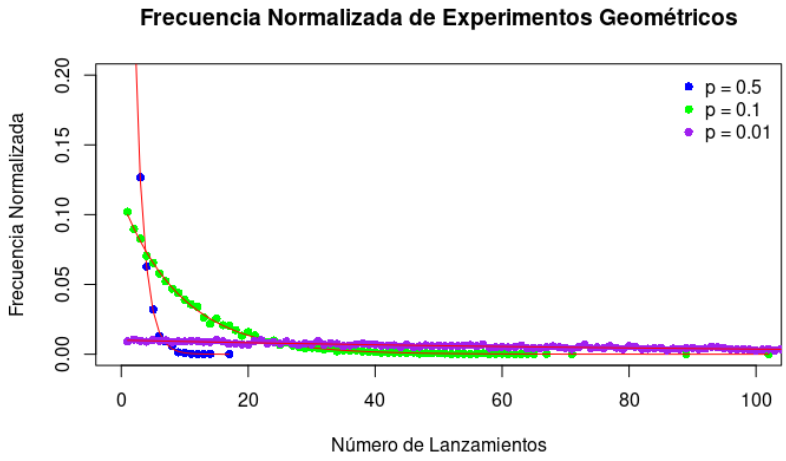
$$P(X = 10) = \binom{10-1}{3-1}(1-0.4)^{10-3}(0.4)^3 = 36 \cdot 0.0279 \cdot 0.064 = 0.06428$$

Por lo tanto, la probabilidad de que deban ser analizados 10 trabajadores para encontrar exactamente tres casos positivos es del 6.42 %.

Problema 4

Considere una moneda desequilibrada que tiene probabilidad p de obtener águila. Usando el comando *sample*, escribe 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 un águila y al número N que contenga el número de lanzamientos hasta obtener un águila en cada uno de los N experimentos.

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. Sibre esta última figura, empalme en rojo la gráfica de la función de masa correspondiente. ¿Qué observa?



Viendo la gráfica, podeos llegar a algunas conclusiones. Para comenzar, la gráfica nos está mostrando la frecuencia normalizada del número de lanzamientos hasta obtener el primer éxito, i.e., el número de águilas a distintas probabilidades de p. Aquí, están marcadas en colores: azul para $p = 0.5$, verde para $p = 0.1$, y morado para $p = 0.01$.

Creo que es buena idea dejar claro que, en el eje X estamos representando el número de lanzamientos, es decir, la cantidad necesaria de lanzamientos para obtener el primer éxito. En cuanto al eje Y, este representa la frecuencia normalizada, es decir, representa la proporción de experimentos donde el primer éxito ocurrió en cierta cantidad de lanzamientos. Así, cada punto indica la frecuencia con la que un número particular de lanzamientos fue necesario para obtener el primer éxito. Si tenemos el punto $(5, 0.05)$, esto significa que para el 5 % de los experimentos se necesitaron exactamente cinco lanzamientos para obtener el primer éxito.

En cualquiera de los tres casos, se logra ver una cierta tendencia de la distribución geométrica más clásica. Quizá solo para la frecuencia morada es la más complicada de

ver qué está sucediendo. Por otra parte, la mayoría de los éxitos ocurren en los primeros lanzamientos, esto notorio sobre todo para $p = 0.5$ y $p = 0.1$.

A medida que la probabilidad va disminuyendo su valor, la distribución comienza a aplanarse, extendiéndose. Esto nos indica que se necesitan más lanzamientos para obtener el primer éxito. Esto tiene sentido, pues nuestra probabilidad “p"va disminuyendo en cada caso (azul, verde y morado).

Hablando de las líneas rojas, estas representan la función de masa de probabilidad teórica de la distribución geométrica para cada una de nuestras p. Podemos notar como las curvas teóricas se ajustan bien a nuestros valores empíricos (los puntos), por lo que la simulación muestra de manera correcta la forma de la distribución. Además, para cada uno de nuestros casos, a medida que aumenta el número de lanzamientos, la frecuencia de normalización tiende a cero. Esto es consistente con lo que podría esperarse par aun experimento de este tipo, pues lanzar una moneda, aunque sea injusta, no debería llevar una cantidad infinita de lanzamientos para que finalmente caiga águila, nuestro primer éxito.

Finalmente, es claro como a medida que la probabilidad de éxito disminuye, son necesarios más lanzamientos para llegar a nuestro primer éxito.

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

Probabilidad p	Promedio	Desviación Estándar
0.5	1.998857	1.41243
0.1	10.00283	9.496298
0.01	100.124	99.51468

De aquí, podemos llegar a varias conclusiones. Primero que nada, para el caso de una variable aleatoria geométrica con parámetro p, el promedio del número de lanzamientos hasta el primer éxito se modela como $\frac{1}{p}$.

Entonces, para $p = 0.5$, tenemos un promedio teórico de $\frac{1}{0.5} = 2$. La simulación para $N = 10^6$ lanzamientos devuelve un promedio de 1.998857, algo muy cercano a lo esperado teóricamente. Esto se repite para $p = 0.1$ y $p = 0.01$, con promedios teóricos de $\frac{1}{0.1} = 10$ y $\frac{1}{0.01} = 100$, respectivamente, y promedios teóricos de 10.00283 y 100.124 en cada caso. Por lo tanto, nuestra simulación fue bastante correcta y refleja lo esperado por la teoría.

En el caso de la desviación estándar de una variable aleatoria geométrica, esta se modela como $\frac{\sqrt{1-p}}{p}$. Esto eso, la raíz cuadrada de la diferencia entre 1 y p, sobre el valor p. Para $p = 0.5$, la desviación estándar teórica es igual a 1.4142, para $p = 0.1$ es igual a 9.4868, y para $p = 0.01$ es igual a 99.4987. Nuestra simulación nos devuelve valores aproximados muy similares, con 1.4124, 9.4962 y 99.5146 respectivamente. Esto confirma que la simulación está trabajando bien.

Ahora bien, me gustaría dejar claro qué representan estos valores bajo este contexto. En este experimento, el promedio nos está indicando cuántos lanzamientos son necesarios para obtener nuestro primer éxito. Por otro lado, la desviación estándar nos indica cuánto varía el número de linimientos necesarios para obtener dicho primer éxi-

to, es decir, nos está indicando qué tan dispersos son los resultados alrededor del valor esperado. Tomando como ejemplo $p = 0.5$, tenemos que la desviación estándar es de 1.414, entonces, aunque en promedio podemos esperar dos lanzamientos antes de obtener la primera águila, la cantidad de lanzamientos puede variar por 1.4 lanzamientos alrededor de nuestros promedio. En física, creo que esto puede ser representado como el error. Tenemos una medida promedio y para este caso tenemos que en la realidad tenemos una incertidumbre de $\pm$ 1.4 lanzamientos alrededor de los 2 lanzamientos necesarios.

De ese modo, podemos asegurar que nuestra simulación tiene consistencia con la teoría. Los resultados simulados del promedio y desviación estándar se comportan bien. También podemos hacer mención a que, en simulaciones, mientras más pruebas o iteraciones se hagan, si se hacen bien, los valores van a tender a converger hacia los resultados teóricos. Además, cuando p es más pequeño, el número promedio de lanzamientos necesarios para obtener le primer éxito aumenta, y la variabilidad (desviación estándar), también aumentan. De acuerdo a cómo se comporta la distribución geométrica, los eventos con menor probabilidad de éxito requieren un mayor número de intentos promedio apra alcanzar el primer éxito.

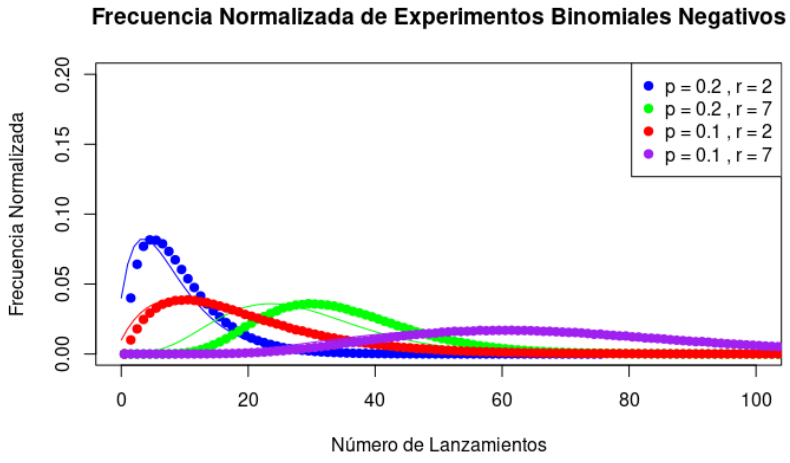
Problema 5

Usando las ideas del inciso anterior, escriba una función en R que simule N veces los lanzamientos de moneda hasta obtener r águilas. La función deberá recibir como parámetros a la probabilidad p de obtener águila, al número r de águilas a observar antes de detener el experimento y al número N que contenga el número de lanzamientos hasta obtener r águilas en cada uno de los N experimentos. Grafique las frecuencias normalizadas de los experimentos para $N = 10^6$, $p = 0.1, 0.2$ y $r = 2, 7$. Compárelos contra la función más adecuada para modelar este tipo de experimentos.

Para este problema intenté implementar la función como se pedía, pero mi máquina se tarda muchísimo en procesarla. Buscando alternativas, utilicé una rutina con la función rbinom() para simular los lanzamientos.

La función de rbinom() genera números aleatorios siguiendo la distribución binomial negativa. Esta distribución es ideal para este problema pues estamos interesados en el número total de intentos necesarios para obtener un número fijo de éxitos (que la moneda caiga águila). Los supuestos que se cumplen para ello son:

- **Independencia de lanzamientos**: cada resultados de lanzamiento de moneda es independiente, su resultado no afecta al de la siguiente tirada.
- **Número fijo de éxitos**: queremos llegar a un r dado de éxitos.
- **Observación de Bernoulli**: cada lanzamiento tiene solo dos posibles resultados: éxito (águila) o fracaso (sol) y la probabilidad de éxito permanece constante a lo largo del experimento.



Ahora bien, las observaciones respecto al problema. Esta rutina simula el experimento de lanzar una moneda aun n cantidad de veces hasta obtener un número fijo de

éxitos (águilas). Dicho número total de lanzamientos necesarios para obtener nuestros éxitos sigue una distribución binomial negativa que depende de nuestras variables p (probabilidad de obtener águila) y r (número de águilas deseado).

Lo que estamos haciendo es realizar 10^6 simulaciones del experimento para diferentes combinaciones de p y de r. En cada una de ellas se revisa cuántos lanzamientos fueron necesarios hasta dar con los r éxitos. Tras cada simulación, se normalizan las frecuencias dividiendo entre N para obtener una frecuencia normalizada, la proporción de simulaciones en las que se obtiene el número de lanzamientos. Las líneas suaves en el histograma representan la distribución teórica para cada caso.

En este caso, cada punto muestra la frecuencia normalizada para un número de lanzamientos fijo. En otras palabras, si el eje X marca n, y el punto tiene una altura de 0.1, entonces significa que el 10 % de las simulaciones necesitaron exactamente n lanzamientos para obtener r éxitos. Por ejemplo, si lancé la moneda hasta obtener 2 águilas, y para 50,000 de las 10^6 simulaciones se necesitaron solo 10 lanzamientos, la frecuencia normalizada para n=10 será de 0.05, es decir, $\frac{50{,}000}{10^6}$. Este ejemplo nos devuelve un punto en x=10 y y=0.05.

Utilizando dbinom() fue que pudimos hacer la comparación entre la simulación y la curva teórica que debería describir a la distribución binomial negativa de este problema. Esta curva muestra la probabilidad teórica de que se necesiten exactamente n lanzamientos hasta obtener nuestros r éxitos deseados. Podemos notar como la curva se alinea bastante bien a nuestros resultados empíricos, por lo que podemos concluir que nuestro modelo simula bien lo esperado pro la teoría. Entonces, la simulación refleja bien un comportamiento binomial negativo.

Entrando en detalle, cuando tenemos una p grande (de 0.2 en lugar de 0.1) la distribución tiende a concentrarse en un menor número de lanzamientos. Esto hace sentido si pensamos que con mayor probabilidad de éxito, entonces se necesitarán menos lanzamientos para obtener nuestra cantidad deseada de éxitos.

Por otra parte, si tenemos valores más grandes de r, es decir, una mayor cantidad de éxitos deseados, la distribución tiende a extenderse más hacia números de lanzamientos mayores. Nuevamente, esto hace sentido, pues queremos más éxitos, por lo tanto, hay cierta tendencia a necesitar mayor cantidad de lanzamientos para llegar al r deseado.

Asimismo, las distribuciones con p=0.1 muestran una dispersión más grande. Esto significa que las curvas se extienden más a lo largo del eje X (eje de los lanzamientos). La razón detrás de esto radica en que con una menor probabilidad de éxito por lanzamiento, entonces será más probable que se necesiten más intentos para obtener el núemro deseado de éxitos. En el caso de las simulaciones con r=7, podemos notar como las curvas se extienden más horizontalmente, mientras que las que tienen r=2 presentan un pico ligeramente más claro, sobre todo para la combinación p=0.2 y r=2, esto también hace sentido, pues tenemos una probabilidad mayor de éxito con una menor cantidad de éxitos necesarios.

Problema 6

Considere X una variable aleatoria con función de distribución F y función de densidad f, y sea A un intervalo de la línea real R. Definimos la función indicadora $I_A(X)$ como se muestra en la siguiente ecuación.

$$I_A = \begin{cases} 1 & \text{si } x \in \text{A} \\ 0 & \text{cualquier otro caso} \end{cases}$$

Sea $Y = I_A(X)$, encuentre una expresión para la distribución acumulada de Y.

Lo primero que necesitamos es encontrar la función de masa de probabilidad (PMF) de Y. La variable aleatoria Y toma el valor de 1 si $X \in A$, y toma el de 0 si $X \notin A$, por lo tanto, Y es una variable aleatoria tipo Bernoulli.

Entonces, la PMF de Y está definida por:

$$f(Y) = \begin{cases} P(Y = 1) & \text{si } P(X \in A) \\ P(Y = 0) & \text{si } P(X \notin A) \end{cases}$$

Utilizando la distribución X:

$$P(X \in A) = \int_A f(x)dx$$

$$P(X \notin A) = 1 - P(X \in A) = 1 - \int_A f(x)dx$$

Por lo tanto:

$$P(Y = 1) = \int_A f(x)dx$$

$$P(Y = 0) = 1 - \int_A f(x)dx$$

Ahora, la función de distribución acumulada (CDF) de Y está dada por:

$$F(y) = P(Y = y)$$

Como Y solo puede tomar valor de 1 o de 0, necesitamos evaluar para los casos: $y < 0$ & $0 \leq y < 1$ & $y \geq 1$.

Para $y < 0$ tenemos:

En este caso, Y solo puede ser 0 o 1, entonces la probabilidad de que $Y \geq y$ cuando $Y < 0$ es 0, pues ni 0 ni 1 son menores que un $y < 0$.

Por lo tanto:

$$F(y) = P(Y \geq y) = 0$$

Para $0 \leq y < 1$ tenemos: Este rango es un poco más complejo que el anterior. Aquí, Y solo puede tomar valor $Y = 0$, ya que $Y = 1$ no es menor que y en este intervalo, i.e., $P(1 = 1)$ no aplica, pero $P(Y = 0)$ sí. Entonces:

$$F(y) = P(Y = y) = 1 - \int_A f(x)dx$$

Para $y \leq 0$ tenemos: Cuando $y \leq 1$, entonces Y puede ser 0 o 1, pues ambos valores son menores o iguales a 1. En otras palabras, $P(Y \leq 1)$ puede ser $P(Y = 0)$ o $P(Y = 1)$. Entonces:

$$F(y) = P(Y = y) = P(Y = 0) + P(Y = 1) = \left(1 - \int_A f(x)dx\right) + \int_A f(x)dx = 1$$

Por lo tanto, la función de distribución acumulada para la variable aleatoria $Y = I_A(X)$ quedaría como:

$$F(y) = \begin{cases} 0 & \text{si } y < 0 \\ 1 - \int_A f(x)dx & \text{si } 0 \leq y < 1 \\ 1 & \text{si } y \geq 1 \end{cases}$$

Problema 7

Entre las más famosas de todas las lluvias de meteoros están las Perseidas, que ocurren cada año a principios de agosot. En algunas áreas, la frecuencia de Perseidas visibles promedian 15 por cada cuarto de hora. El modelo de probabilidad que describe Y, el número de meteoros que una persona ve en un cuarto de hora, tiene la función de probabilidad asociada a Poisson con $\lambda = 6$. Encuentre la probabilidad de que una persona vea en un cuarto de hora determinando, al menos la mitad de los meteoros que esperaría ver.

Antes que nada, el enunciado nos indica que la función de probabilidad de estos eventos es la de Poisson, es decir:

$$f_y(Y) = \frac{e^{-6}6^y}{y!} \quad \forall \quad y = 0, 1, 2, ...$$

Ahora bien, ya desde el enunciado tenemos que la expresión propuesta para modelar el problema corresponde con la Distribución de Poisson (DP). Para recapitular, sabemos que la DP es una manera adecuada para modelar eventos que ocurren en determinado intervalo (espacial o temporal). Esta distribución es sobre todo utilizada para eventos que se pueden considerar aleatorios "raros", que son independientes y que ocurren bajo una tasa en promedio constante.

No viene mal recordar que la Distribución de Poisson está, en general, descrita de la siguiente manera:

$$f_Y(y) = \frac{e^{-\lambda}\lambda^y}{y!} \quad \forall \quad y = 0, 1, 2, ...$$

En la DP, λ corresponde a la intensidad (o tasa) de eventos esperada por intervalo. Los supuestos que se deben cumplir son:

- **Eventos independientes**: la aparición de un evento (en este caso, un meteoro) es independiente de la aparición de otros.
- **Existe un número promedio de éxitos llamado λ sobre un cierto intervalo**: como mencionamos anteriormente, esta λ se mantiene constante a lo largo del experimento.
- **Éxitos aleatorios**: los éxitos deben aparecer de manera aleatoria en intervalos de la misma magnitud. Estos intervalos no se traslapan y solo se espera que en promedio se tenga el mismo λ.

Regresando al problema, se nos está pidiendo responder: ¿Cuál es la probabilidad de que cierta persona vea "al menos" la mitad de los meteoritos que esperaría ver en un cuarto de hora?

Para esto, sabemos por el enunciado que la tasa esperada es de $\lambda = 6$ meteoritos, por lo tanto, si queremos obtener al menos la mitad de los esperados, esto es $\frac{\lambda}{2} = 3$. Entonces, queremos ver qué pasa con $Y \geq 3$.

Para ello, debemos encontrar entonces:

$$P(Y \geq 3) = 1 - P(Y < 3) = 1 - [P(Y = 0) + P(Y = 1) + P(Y = 2)]$$

Calculamos las probabilidades:

Para $y = 0$:

$$P(Y = 0) = \frac{e^{-6}6^0}{0!} = e^{-6}$$

Para $y = 1$:

$$P(Y = 1) = \frac{e^{-6}6^1}{1!} = 6e^{-6}$$

Para $y = 2$:

$$P(Y = 2) = \frac{e^{-6}6^2}{2!} = \frac{36}{2}e^{-6} = 18e^{-6}$$

Entonces, sumamos $P(Y < 3) = e^{-6} + 6e^{-6} + 18e^{-6} = 25e^{-6}$.

Por lo tanto, la probabilidad de ver al menos tres meteoritos en un intervalo de 15 minutos es de $P(Y \geq 3) = 1 - 25e^{-6} \approx 0.9380$. Hay un chance del 93.8 % de ver al menos tres meteoritos en un cuarto de hora.

Problema 8

Las calificaciones de un estudiante de primer semestre en el examen de Química se describen por cierta densidad de probabilidad $f_Y(y) = 6y(1-y)$ para $0 \leq y \leq 1$. Aquí, y representa la proporción de preguntas que el estudiante responde correctamente. Cualquier calificación menor a 0.4 se considera reprobatoria. Responde: ¿Cuál es la probabilidad de que un estudiante reprueba? Si seis estudiantes toman el examen, ¿Cuál es la probabilidad de que justo dos reprueben?

Comenzaremos respondiendo "¿Cuál es la probabilidad de que un estudiante reprueba?". Sabemos que cualquier calificación menor a 0.4 es reprobatoria, por lo cual necesitamos calcular la integral de la función de densidad de probabilidad desde 0 hasta 0.4. Entonces:

$$P(Y < 0.4) = \int_0^{0.4} 6y(1-y)dy = \int_0^{0.4} (6y - 6y^2)dy = \int_0^{0.4} 6ydy - \int_0^{0.4} 6y^2dy$$

Tenemos:

$$\int_0^{0.4} 6ydy = 6\int_0^{0.4} ydy = 3\left[y^2\right]_0^{0.4} = 3(0.4)^2 - 3(0)^2 = 0.48$$

Y también que:

$$-\int_0^{0.4} 6y^2dy = -6\int_0^{0.4} ydy = -2\left[y^3\right]_0^{0.4} = -3(0.4)^3 + 3(0)^3 = -0.128$$

Así:

$$P(Y < 0.4) = 0.48 - 0.128 = 0.352$$

Por lo tanto, la probabilidad de que un estudiante repruebe es del 35.2 %.

Ahora, respondemos "Si seis estudiantes toman el examen, ¿Cuál es la probabilidad de que justo dos reprueben?". Para ello, primero hay que analizar nuestro problema. Se nos pide encontrar la probabilidad de que dos estudiantes reprueben dado que seis presentan el examen. Tenemos las siguientes condiciones:

- **Tenemos un número fijo de ensayos**: en este caso $n = 6$.
- **Hay dos posibles resultados**: aprobado ($y > 0.4$) o reprobado ($y \leq 0.4$).
- **Tenemos probabilidad constante de éxito**: la probabilidad de que uno de los seis estudiantes repruebe es la misma y no cambia entre observaciones. Aquí igual a 0.352.
- **Existe independencia**: la calificación de un estudiante no afecta al de los demás.

Por lo tanto, se cumplen los supuestos para utilizar la Distribución Binomial:

$$P(X = k) = \binom{n}{k}(1 - p)^{n-k}p^k$$

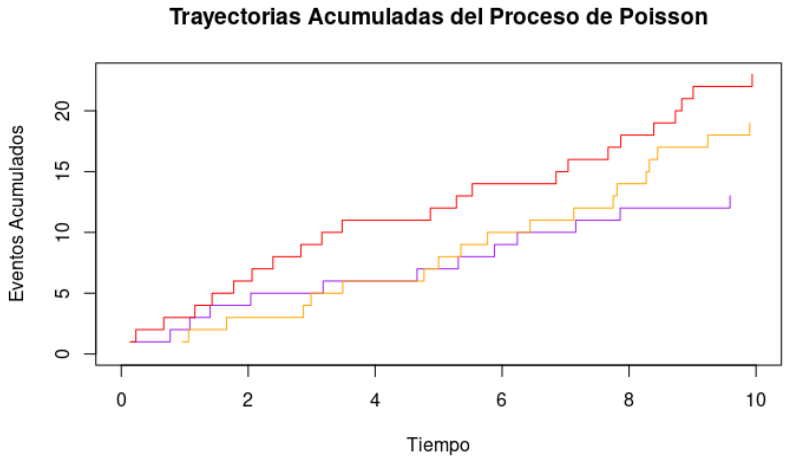
Con $n = 6$, $k = 2$ y $p = 0.352$. Así:

$$P(X = 2) = \binom{6}{2}(1 - p)^{6-2}p^2 = 15 \cdot 0.65^4 \cdot 0.352^2 = 0.3277$$

Esto nos indica que la probabilidad de que exactamente dos estudiantes de seis que presentaron el examen sean reprobados es igual al 32.77 %.

Problema 9

Escriba un a función en R que simule una aproximación al proceso Poisson a partir de las cinco hipótesis que usamos en clases para construir tal proceso. Usando esta función, simule tres trayectorias de un proceso Poisson $\lambda = 2$ sobre el intervalo $[0, 10]$ y grafíquelas. Además, simule 10^4 veces un proceso de Poisson N con $\lambda = \frac{1}{2}$, y hasta el tiempo $t = 1$. Haga un histograma de $N(1)$ en su simulación anterior y compare contra la distribución de Poisson correspondiente [2]



Las trayectorias mostradas en este gráfico representa la evolución de un proceso Poisson con parámetro $\lambda = 2$, sobre el intervalo de tiempo $[0, 10]$. Este proceso modela la ocurriencia de ciertos eventos aleatorios a lo largo del tiempo con:

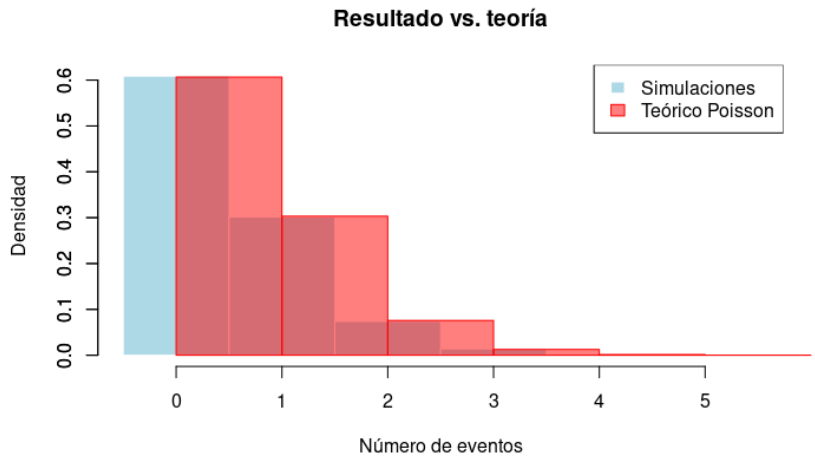
- $\lambda = 2$, la tasa de ocurrencia de eventos sobre la unidad de tiempo, por lo que, en promedio, se esperan dos eventos cada unidad de tiempo.
- Los eventos ocurren en intervalos disjuntos, son independientes entre sí.
- Trabajamos con un incremento estacionario, lo que significa que la probabilidad de que ocurra un número determinado de eventos en un intervalo depende solo de la longitud del intervalo, no de su posición en el tiempo.

En la gráfica se muestran tres trayectorias del proceso de Poisson homogéneo. Cada una de las trayectorias representa la acumulación de eventos a lo largo del tiempo en el intervalo $[0, 10]$. Como es de esperar en un proceso de Poisson, los eventos ocurren

[2] Hint: considere el intervalo $[0, T]$, y un número real positivo dt que sea mucho más pequeño que la longitud $[0, T]$ y que divida dicha longitud, digamos $\frac{T}{dt} = 1000$ veces. Divida el intervalo en intervalitos de longitud dt que tengan forma $(k \cdot dt, (k+1) \cdot dt]$ con $k = 0, 1, 2, ..., (T/dt - 1)$. Para cada uno de estos intervalitos simula una v.a. Bernoulli $(\lambda dt + 10^{-6})$ y guarde su resultado en un vector de tamaño adecuado.

de manera discreta y aleatoria, lo que se refleja en los saltos que tienen lugar en puntos específicos del tiempo. Estos saltos indican que en esos momentos ocurrió un evento.

Aunque las tres trayectorias son distintas, esto es coherente con la naturaleza estocástica del proceso de Poisson, que introduce variabilidad en la ocurrencia de eventos. Sin embargo, a pesar de esta aleatoriedad, todas las trayectorias exhiben una estructura común, caracterizada por un incremento acumulado de eventos a lo largo del tiempo, lo cual es esperado dado que la tasa de eventos λ es constante.



La segunda parte de nuestro problema involucra la simulación del proceso Poisson, pero ahora con $\lambda = 0.5$, sobre un intervalo de $[0, 1]$. Se pide comparar nuestros resultados con los de una distribución de Poisson Teórica, de tal manera que confirmemos que nuestra simulación sigue la tendencia de Poisson.

El código simula el proceso $N(T)$ para $10,000$ observaciones, se cuenta el número de eventos ocurridos en el intervalo solicitado tras cada simulación. Nuestro gráfico muestra como el pico de las simulaciones es que no ocurran eventos (hay más valores en 0), esto puede explicarse si tomamos en cuenta que nuestra intensidad λ es pequeña, por lo tanto es raro que ocurra un evento en el intervalo.

Podemos notar como la distribución teórica (en rojo) tiene una tendencia muy parecida a la simulada, esto nos peude dar la seguridad de que lo que estamos haciendo está bien. Esto queda confirmado tanto para valores grandes de k, como para valores pequeños.

Además, hacer una gran cantidad de simulaciones en nuestro proceso nos ayuda a aproximarnos de la mejor manera a Poisson Teórico. Esto demuestra que el código simula correctamente un proceso de Poisson bajo las hipótesis dadas.

Problema 10

En una oficinia de correos, los paquetes llegan según un proceso de Poisson de intensidad λ. Hay un costo de almacenamiento de c pesos por paquete por unidad de tiempo. Los paquetes se acumulan en el local y se despachan en grupos cada T unidades de tiempo (i.e. se acumulan en el local y se despachan en $T, 2T, 3T, ...$). Hay un costo por despacho fijo de K pesos (es decir, el costo es independiente del número de paquetes que se despachen). (a) ¿Cuál es el costo promedio por paquete por almacenamiento en el primer ciclo $[0, T]$? (b) ¿Cuál es el costo promedio por paquete por almacenamiento y despacho en el primer ciclo? (c) ¿Cuál es el valor de T que minimiza este costo promedio?

Solución (a):

Sabemos que los paquetes llegan de acuerdo a un proceso de Poisson con intensidad λ. Esto significa que el número de paquetes que llegan en un intervalo dt es λdt. Si suponemos que cierto paquete llega al tiempo t, y permanece en despacho hasta el momento T, entonces el tiempo que dicho paquete permanece almacenado es igual a la diferencia $T - t$, y si el costo es c pesos por paquete por unidad de tiempo, el costo de almacenamiento por paquete es $c(T - t)$.

Por lo tanto, el costo total de almacenamiento en el intervalo $[0, T]$ deberá ser la suma de todos los paquetes que llegan en dicho ciclo. Entonces, hay que integrar:

$$C_a(T) = \int_0^T c\lambda t \, dt = c\lambda \int_0^T t \, dt = \frac{c\lambda T^2}{2}$$

Con C_a := costo total de almacenamiento. Para el costo promedio por paquete, tenemos que dividir este resultado entre el número esperado de paquetes que llegan en cada intervalo de longitud T, i.e.:

$$E[N(T)] = \lambda T$$

Por lo tanto, el costo promedio por paquete en el intervalo $[0, T]$ se ve como:

$$C_{pa}(T) = \frac{\frac{c\lambda T^2}{2}}{\lambda T} = \frac{1}{2}cT$$

Solución (b):

Para el costo total en el primer ciclo hay que sumar el costo de almacenamiento y el costo fijo de despacho (llamémosle $C_T(T)$), el cual de antemano el enunciado nos menciona que dicho costo es igual a K pesos. Entonces:

$$C_T(T) = C_a(T) + K = \frac{c\lambda T^2}{2} + K$$

Por lo tanto, el costo promedio por paquete es:

$$\frac{C_T(T)}{\lambda T} = \frac{1}{2}cT + \frac{K}{\lambda T} = \mathbb{C}(T)$$

De ese modo, $\mathbb{C}(T)$ será el costo promedio por paquete por almacenamiento y despacho en el primer ciclo.

Solución (c):
Ahora, se nos pide encontrar T tal que dicha T minimice el costo promedio. Entonces, hay que derivar respecto de T e igualar a cero, para después comprobar que ese punto crítico minimice nuestra función.

Por lo tanto:

$$\frac{d}{dT}\mathbb{C}(T) = \frac{d}{dT}\left(\frac{1}{2}cT\right) + \frac{d}{dT}\left(\frac{K}{\lambda T}\right) = \frac{1}{2}c - \frac{K}{\lambda T^2}$$

Igualamos la derivada a cero:

$$\frac{d}{dT}\mathbb{C}(T) = \frac{1}{2}c - \frac{K}{\lambda T^2} = 0$$

Entonces:

$$T = \left(\frac{2K}{c\lambda}\right)^{1/2}$$

Para asegurarnos de que nuestra solución es para una T que minimiza la función, podemos recurrir al criterio de la segunda derivada:

- Si $f'' < 0$, entonces f tiene un máximo relativo en $(x, f(x))$.
- Si $f'' > 0$, entonces f tiene un mínimo relativo en $(x, f(x))$.
- Si $f'' = 0$, entonces el criterio no puede aplicarse. Entonces, quizás f tenga un máximo relativo, un mínimo relativo o ninguno de los dos en $(x, f(x))$.

Por lo tanto, obtenemos la segunda derivada de nuestra $\mathbb{C}(T)$:

$$\frac{d^2}{dT^2}\mathbb{C}(T) = \frac{2K}{\lambda T^3}$$

Podemos asegurar que $\frac{2K}{\lambda T^3} > 0$, pues λ, T y K son siempre positivos. Así, $\frac{d^2}{dT^2}\mathbb{C}(T) > 0$.