Godinez Bravo Diego
Inferencia Estadística
Tarea 2

1. Cuando una máquina no se ajusta adecuadamente tiene una probabilidad 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

De acuerdo a la naturaleza del experimento podríamos considerar al número de artículos defectuosos como el número de obsrvaciones que ocurren antes de que se detenga la máquina (ocurrencia de 3 elementos defectuosos). Por lo tanto podríamos pensar que la variable aleatoria X sigue una **distribución binomial negativa** con parámetros $p = 0.15$ y $r = 3$.

$$P(X = x) = \binom{x-1}{r-1}(1-p)^{x-r}p^r \quad \text{con x = r, r + 1, r + 2, ...}$$

En este caso nos interesa la probabilidad de que la máquina produzca 5 o más artículos antes de que sea detenida ($P \geq 5$), por lo que podemos calcular la probabilidad de la siguiente manera:

$$P(X \geq 5) = 1 - \Big[P(X = 4) + P(X = 3)\Big]$$

Se considera desde $P(X = 3)$ ya que este es el número mínimo de observaciones que podríamos observar antes de que se detenga la máquina (3 artículos defectuosos seguidos); calculamos las probabilidades correspondientes:

$$P(X = 3) = \binom{3-1}{3-1}(1-0.15)^{3-3}0.15^3 = 0.15^3 = \underbrace{\binom{2}{2}}_{1}\underbrace{(0.85)^0}_{1}0.15^3 = 0.15^3 = 3.375x10^{-3}$$

$$P(X = 4) = \binom{4-1}{3-1}(1-0.15)^{4-3}0.15^3 = 0.15^3 = \underbrace{\binom{3}{2}}_{3}\underbrace{(0.85)^1}_{0.85}0.15^3 = (3)(0.85)(3.375x10^{-3}) = 8.606x10^{-3}$$

De esta manera tenemos que la probabilidad $P(X \geq 5)$ es igual a:

$$P(X \geq 5) = 1 - \Big[8.606x10^{-3} + 3.375x10^{-3}\Big] = 0.9880$$

El valor esperado de una variable con distibución binomial negativa esta definido como:

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

Reemplazando los valores de r y p tenemos que el valor esperado:

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

2. Los empleados de una compañia de aislantes son sometidos a pruebas para detectar residuos de asbesto en sus pulmones. Se le ha pedido a la compañia 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, encuentre la probabilidad de que deban ser analizados exactamente 10 trabajadores para poder encontrar a 3 con resultado positivo.

Solución

Definiendo la variable aleatoria X = 'número de observaciones (trabajadores) hasta encontrar r positivos', podemos asumir que la variable aleatoria sigue una distribución **binomial negativa** con parámetros $r = 3$ (éxitos) y $p = 0.40$ (probabilidad constante).
Por lo tanto la probabilidad de que tengan que ser analizados **exactamente** 10 trabajadores para poder encontrar 3 positivos esta dada de la siguiente manera:

$$P(X = x) = \binom{x-1}{r-1}(1-p)^{x-r}p^r \quad \text{con x = r, r + 1, r + 2, ...}$$

Reemplazando con los valores dados tenemos que:

$$P(X = 10) = \binom{10-1}{3-1}(1-0.40)^{10-3}0.(40)^3$$

$$= \underbrace{\binom{9}{2}(0.60)^7}_{1.0077}\underbrace{(0.40)^3}_{0.064} = 0.0644$$

Por lo tanto la probabilidad de que deban ser analizados exactamente 10 trabajadores para poder encontrar 3 positivos es:

$$P(X = 10) = 0.0644$$

3. Para el siguiente ejercicio es necesario usar R.

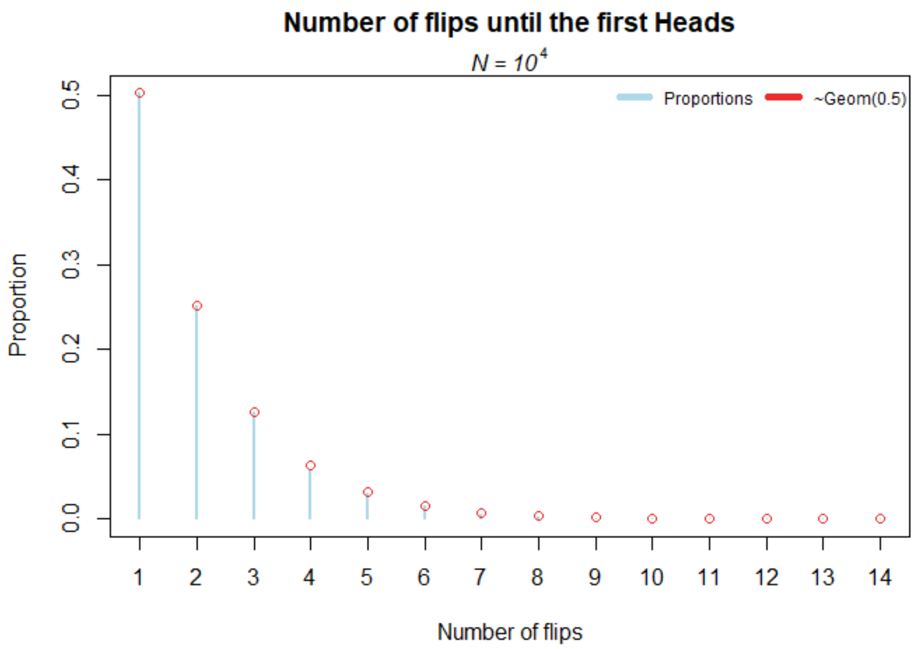
a) Considere una moneda desequilibrada que tiene probabilidad p de obtener águila. Usando el comando **sample**, escriba una función que simule N veces lanzamientos de esta moneda hasta obtener un águila. La función deberá recibir como párametros a la probabilidad p de obtener águila y al número N de veces que se repite el experimento; y tendrá que regresar un vector de longitud N que contenga el número de lanzamientos hasta 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. Sobre esta última figura empalme en rojo la gráfica de la función de masa correspondiente. ¿Qué observa?

Resultados

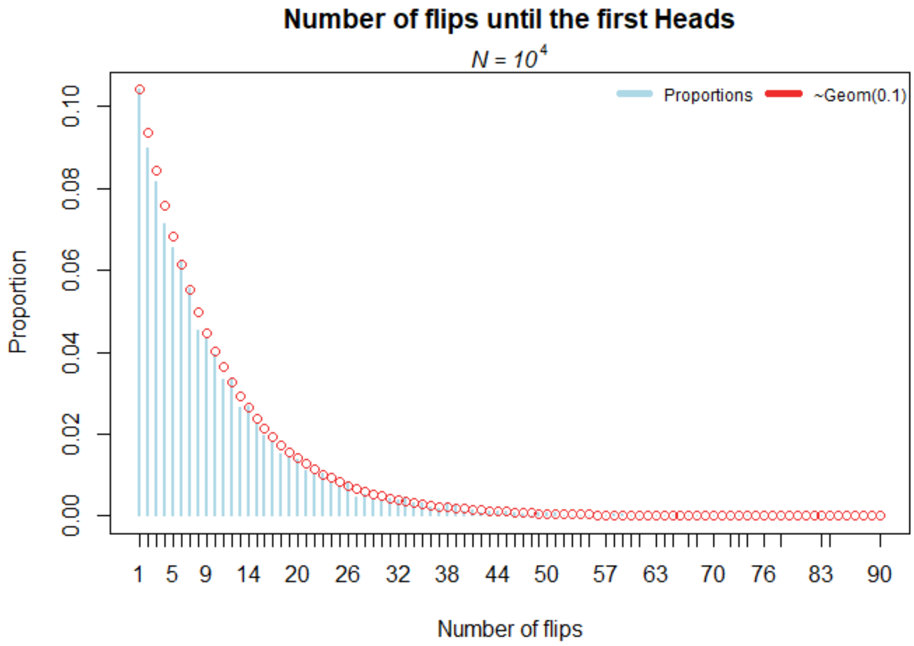
En este experimento definimos la variable aleatoria como el número de observaciones hasta obtener un éxito (águila), por la naturaleza del experimento podríamos pensar que nuestra variable aleatoria X sigue una distribución geométrica $Geom(p)$; con $p = (0.5, 0.1, 0.01)$. Para cada uno de los casos, la distribución de las proporciones del experimento realizado coincide con la distribucíon que sigue una variable aleatoria con distribución $\sim Geom(p)$. Esto lo comprobamos al traslapar la gráfica de proporciones del número de águilas obtenidas junto con la gráfica de la distribución geométrica $\sim Geom(p)$; con $p = (0.5, 0.1, 0.01)$. Además al calcular el valor esperado $E(X)$, este coincide con el valor esperado que tendría una variable aleatoria con distribucíon geométrica, el cual se encuentra definido como $E(X) = \frac{1}{p}$.

Probabilidad $p = 0.5$



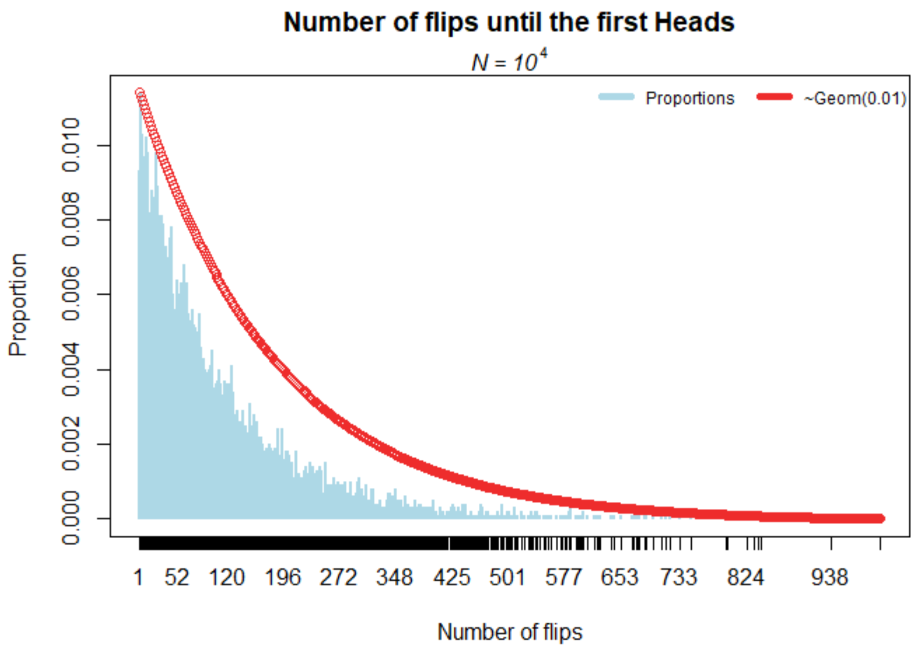
Distribución Geométrica $Geom(0.5)$ y gráfica de frecuencias

Probabilidad $p = 0.1$



Distribución Geométrica *Geom*(0.1) y gráfica de frecuencias

Probabilidad $p = 0.01$



Distribución Geométrica *Geom*(0.01) y gráfica de frecuencias

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

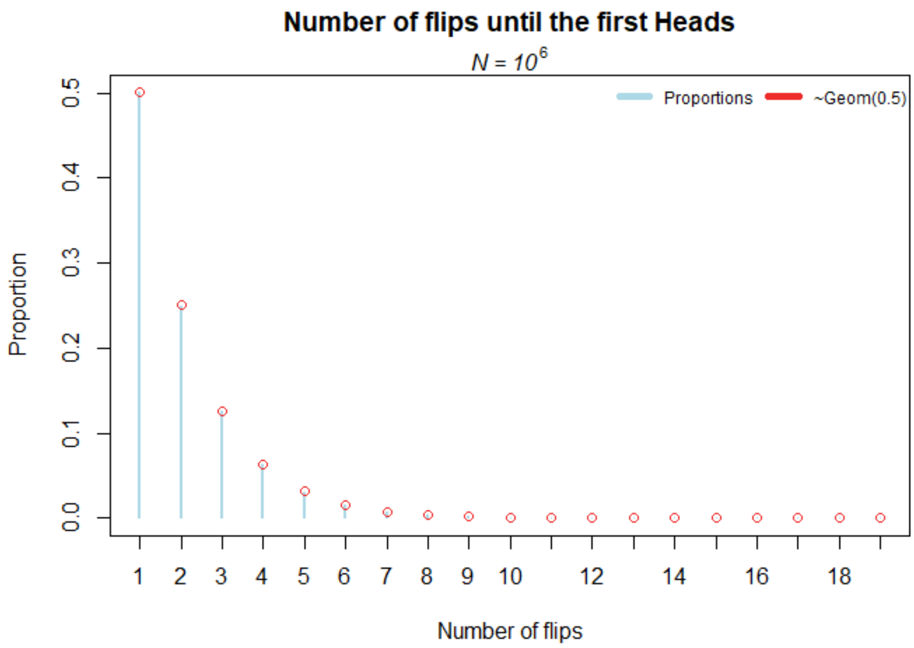
Resultados

De manera similar al caso anterior, nuestra variable aleatoria se ajusta a la curva de distribución que sigue una variable aleatoria con distribución geométrica $\sim Geom(p)$. Es importante resaltar que en el caso $p = 0.01$, la distribucíon de las proporciones se ajusta con la curva de distribución de una variable aleatoria $\sim Geom(0.01)$, esto como consecuencia de aumentar el número de veces que se repite el experimento, a diferencia del caso anterior, donde se realizó el experimento 10^4 veces. Además, en cada uno de los casos el valor esperado $E(X)$ coincide.

Probabilidad $p = 0.5$

Media: 1.99

Desviación Estándar: 1.41

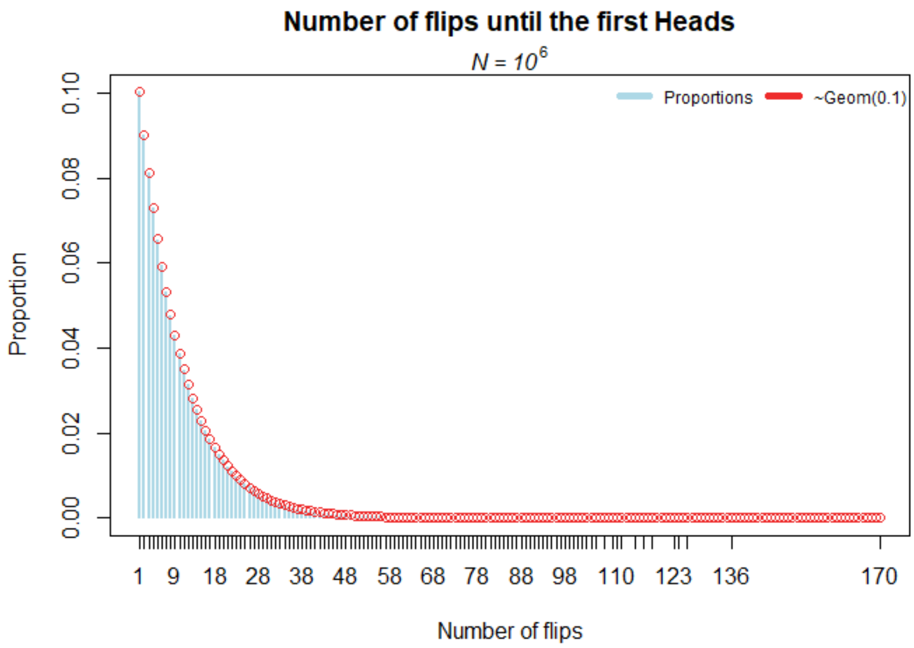


Distribución Geométrica $Geom(0.01)$ y gráfica de frecuencias

Probabilidad $p = 0.1$

Media: 9.99

Desviación Estándar: 9.49

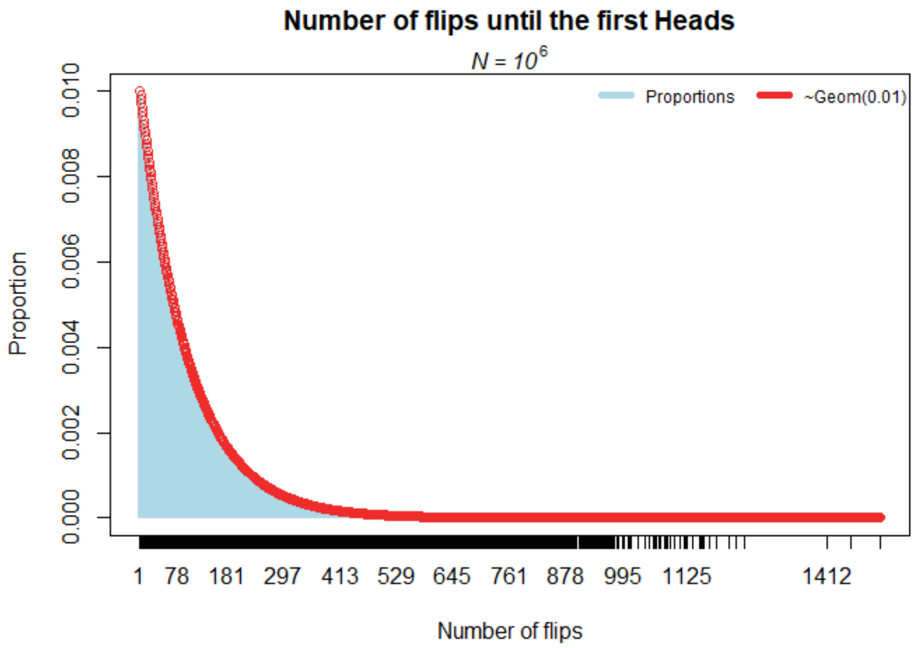


Distribución Geométrica $Geom(0.01)$ y gráfica de frecuencias

Probabilidad $p = 0.01$

Media: 99.92

Desviación Estándar: 99.55



Distribución Geométrica $Geom(0.01)$ y gráfica de frecuencias

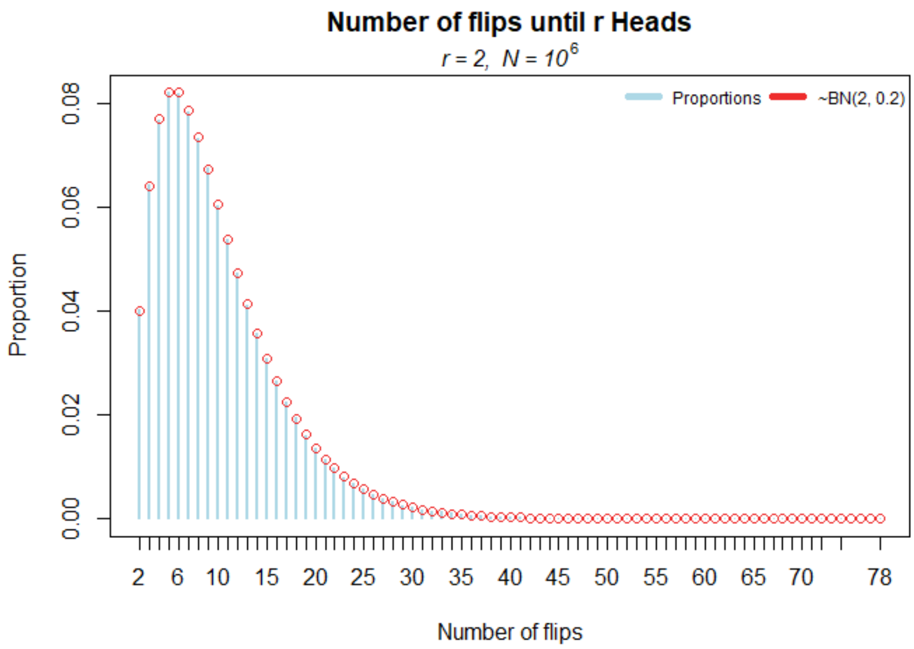
4. Usando las ideas del inciso anterior escriba una función en R que simule N veces los lanzamientos de moneda hasta obtener r águilas. La función deberá recibir como parámetros a la probabilidad p de obtener un águila, al número r de águilas a observar antes de detener el experimento y al número N que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

Resultados

Dadas las condiciones del experimento, siendo la variable aleatoria el número de obsrvaciones hasta conseguir un número finito de águilas r, y apoyandonos con el material visto en clase, podríamos pensar que la variable aleatoria X sigue una distribución **binomial negativa** $\sim BN(r, p)$. Para corroborar esto creamos una función que simula dicho experimento para analizar como se distribuyen las frecuencias de los posibles resultados que toma nuestra variable aleatoria de interés. Al trazar la gráfica de las proporciones del número de águilas obtenidos junto con la función de masa correspondiente, observamos que en efecto nuestra variable aleatoria X sigue una distribución binomial negativa.

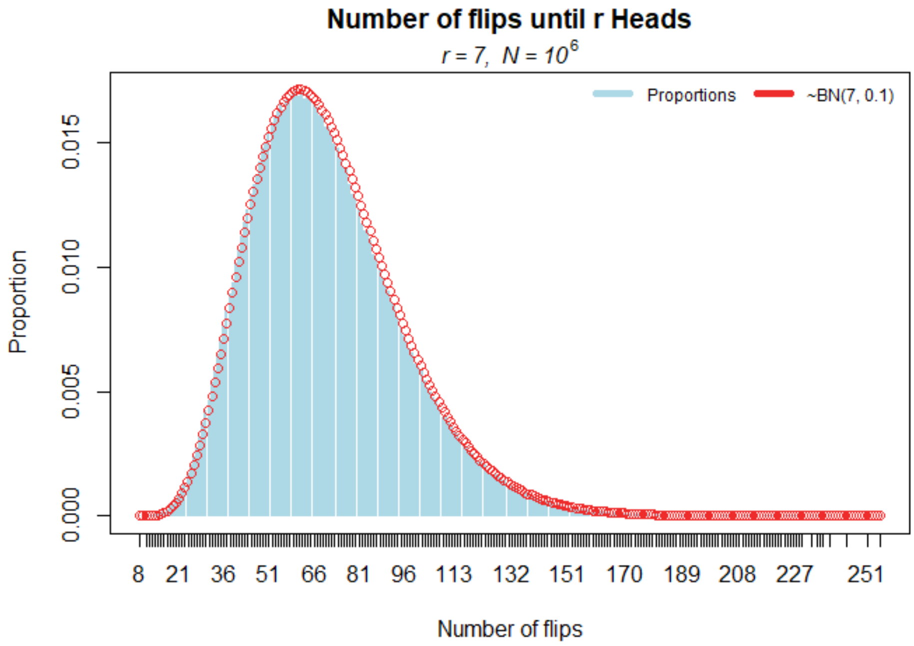
En ambos casos, la distribución de nuestra variable aleatoria se ajusta con la curva generada para una variable aleatoria con distribución binomial negativa; $\sim BN(2, 0.2)$ y $\sim BN(7, 0.1)$, respectivamente.

Probabilidad $p = 0.2$; $\mathbf{r} = 2$



Distribución Binomial Negativa $BN(2,\ 0.2)$ y gráfica de frecuencias

Probabilidad $p = 0.1$; $\mathbf{r} = 7$



Distribución Binomial Negativa $BN(7, 0.1)$ y gráfica de frecuencias

5. Considera X una variable aleatoria con función de distribución F y función de densidad f, y sea A un intervalo de la línea real Re. Definimos la función indicadora $1_A(X)$:

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

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

Solución

Dado un intervalo A; y sea Y definido como $Y = 1_A(X)$, los valores que la variable aleatoria Y puede tomar son $Y = 0$ y $Y = 1$, los cuales pueden representarse como:

$$P(Y = 1) = P(x \in A)$$

y

$$P(Y = 0) = P(x \notin A)$$

La probabilidad $P(Y = 0)$, podría representarse como:

$$P(Y = 0) = P(x \notin A) = 1 - P(x \in A)$$

De esta manera tenemos la función de masa $f(y)$:

$$f(y) = \begin{cases} 0 & \text{si} \quad 1 \text{ - } P(x \in A) \\ 1 & \text{si} \quad P(x \in A) \end{cases}$$

Dado que A es un intervalo, podemos expresar la probabildad $P(x \in A)$ de la siguiente manera:

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

Entonces:

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

Por lo que la probabilidad $P(Y = 0)$ quedaría definida como:

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

Por consecuencia, la función de distribución $F(Y)$ estaría definida de la siguiente manera:

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

Por definición valor esperado de una variable aleatoria continua esta definido como:

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

En este caso:

$$E(Y) = \int I_A(x)f(x)dx = \int_A f(x)dx$$

Sabemos que:

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

Por lo tanto el valor esperado de la variable Y esta definido como:

$$E(Y) = P(x \in A)$$

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

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

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?

Solución

La función de distibución acumulada de una **variable aleatoria continua** se obtiene integrando la función de densidad de probabilidad $f(y)$ entre los límites $-\infty$ y x;

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)dy$$

Por lo tanto:

$$F(x) = P(X \leq 0.4) = \int_{0}^{0.4} 6y(1-y)dy$$

Resolviendo la integral tenemos que:

$$\int 6y - 6y^2 dy = 6\int y - y^2 dy$$

Calculando el valor de la integral definida:

$$6\Big[\frac{y^2}{2} - \frac{y^3}{3}\Big]_0^{0.4}$$

$$6\Big[\frac{0.4^2}{2} - \frac{0.4^3}{3}\Big] - 6\Big[\frac{0^2}{2} - \frac{0^3}{3}\Big]$$

$$6\Big[\frac{0.4^2}{2} - \frac{0.4^3}{3}\Big] - 6\underbrace{\Big[\frac{0^2}{2} - \frac{0^3}{3}\Big]}_{0}$$

$$6\Big[\frac{0.16}{2} - \frac{0.064}{3}\Big] = 6\Big[0.08 - 0.0213\Big]$$

$$6\Big[0.0587\Big] = 0.3522$$

De esta manera la probabilidad de que un estudiante repruebe es:

$$P(X \leq 0.4) = 0.3522$$

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

Solución

En este caso buscamos la probabilidad de que 2 de 6 estudiantes reprueben el examen; tenemos una probabilidad $p = 0.3522$, y el número de estudiantes que presentarán el examen (observaciones), por lo que podríamos considerar una variable aleatoria X (número de estudiantes que reprueban el examen), que sigue una **distribución binomial** con parámetros $\mathbf{n} = 6$ y $p = 0.3522$.

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{para } x = 0,\ 1,\ 2,\ \ldots,\ \mathrm{n}$$

Sustituyendo valores:

$$P(X = 2) = \binom{6}{2} \underbrace{0.3522^2}_{0.1240} \underbrace{(1 - 0.3522)^4}_{0.1761} = 0.3275$$

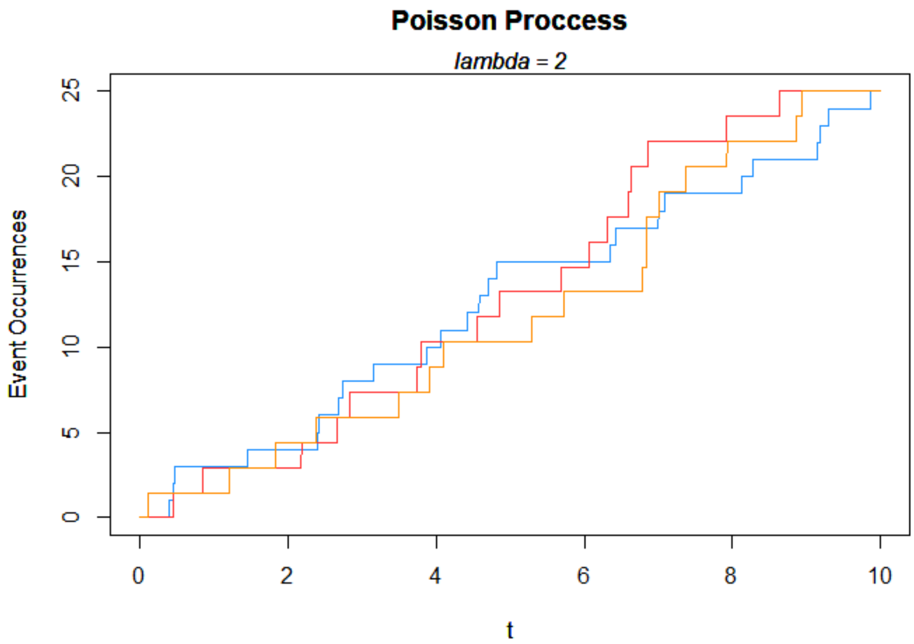
Entonces, si 6 estudiantes realizan el examen, la probabilidad de que 2 reprueben es de:

$$P(X = 2) = 0.3275$$

7. Escriba una funcíon en R que simule una aproximación al proceso Poisson a partir de las 5 hipótesis que usamos en clase para contruir 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.

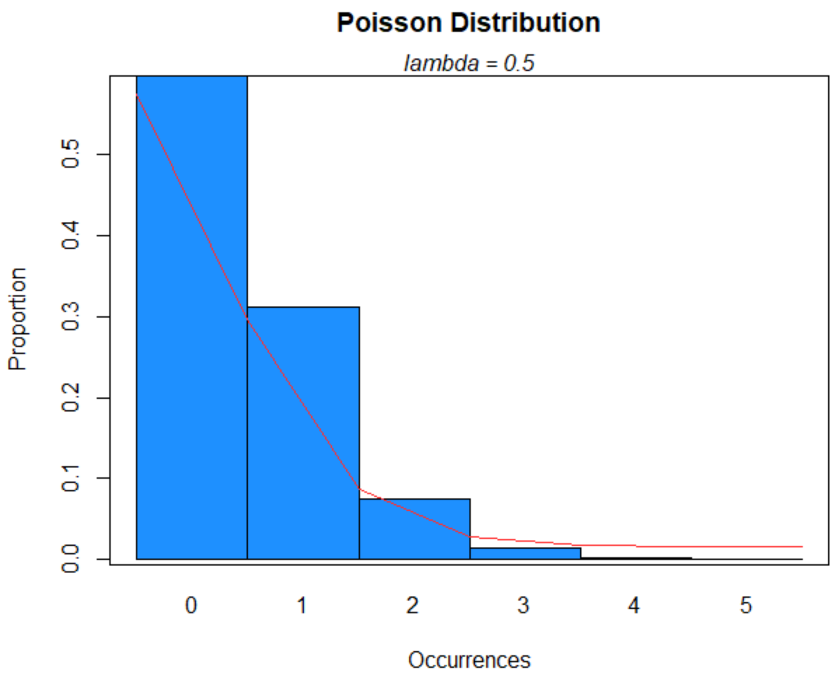
Resultados

Al simular tres trayectorias de Poisson diferentes con $\lambda = 2$, podemos observar que, al ser un proceso que contempla un conjunto de variables aleatorias independientes, cada una de estas trayectorías van a diferir en gran medida entre sí. En cada una de las trayectorias el número de éxitos ocurridos varía de manera notable, esto debido a la naturaleza del proceso de Poisson en sí, donde el número de ocurrencias durante el intervalo de tiempo es aleatorio. De esta manera podemos concluir que no hay patrón específico entre cada una de las trayectorias, cada trayectoria es única, lo cual coincide con la definición del proceso de Poisson.



Tres trayectorías diferentes de un Proceso Poisson con $\lambda = 2$

Al realizar el experimento 10^6 veces y graficar las proporciones de las ocurrencias de los eventos, podemos ver que la distribución coincide con la distribución que sigue una variable aleatoria con distribución poisson $\sim Poisson(\lambda)$; con $\lambda = \frac{1}{2}$. Siendo el valor 0 y 1 los más frecuentes, esto por la naturaleza del proceso en sí, un proceso estocástico donde la probabilidad de ocurrencia de éxito en cada uno de los subintervalos es pequeña.



Distribución de resultados de 10^4 Procesos Poisson

8. Considere la siguiente función.

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

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

Solución

Una función $F(x)$ es una función de distribución si cumple con las siguientes propiedades:

1. $F(x)$ debe ser no negativa para todos los valores de x (i.e. $F(x) \geq 0 \ \forall \ x$).
2. $F(x)$ es continua por la derecha para todo x (i.e. $F(x) = F(x^+)$ con $F(x^+) = \lim_{y \to x} F(y)$ donde $y > x$).
3. $F(x)$ debe ser no decreciente para todos los valores de x (i.e. si $x_1 < x_2$ entonces $F(x_1) \leq F(x_2)$).
4. La función de distribución tiende a cero cuando x tiende a menos infinito; $\lim_{x \to -\infty} F(x) = 0$.
5. La función de distribución tiende a uno cuando x tiende a más infinito; $\lim_{x \to \infty} F(x) = 0$.

Para este caso la función $F(x)$ cumple las siguientes condiciones:

1. $F(x)$ es no negativa para todos los valores de x; todas las expresiones que conforman la función son no negativas.
2. $F(x)$ debe ser continua en el intervalo $(0, \frac{3}{4})$, para esto evaluamos en $x = 0$ y $x = \frac{3}{4}$;

$$F(0) = 0.1$$

Por otro lado:

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

Entonces:

$$\lim_{x \to 0^+} F(x) = F(0)$$

Por último:

$$F(\frac{3}{4}) = 1$$

De manera similar:

$$\lim_{x \to \frac{3}{4}^+} F(x) = \lim_{x \to \frac{3}{4}^+} 1 = 1$$

$$\lim_{x \to \frac{3}{4}^+} F(x) = F(\frac{3}{4})$$

La función $F(x)$ es continua por la derecha.

3. Para comprobar que $F(x)$ es no decreciente nos basamos en cada una de las expresiones que conforman $F(x)$, de la siguiente manera:

Sean $x_1 \in (-\infty, 0)$ y $x_2 = 0$, donde $x_1 < x_2$:

$$\begin{aligned} x_1 &< x_2 \\ 0(x_1) &< 0(x_2) \\ 0 \leq 0 &< 0.1 \end{aligned}$$

De manera similar se cumple que:

$$F(x_1) \leq F(x_2)$$

Sean $x_1 = 0$ y $x_2 \in (0, \frac{3}{4})$, donde $x_1 < x_2$:

$$\begin{aligned} x_1 &< x_2 \\ 0 &< 0.8x_2 \\ 0.1 + 0 &< 0.1 + 0.8x_2 \\ 0.1 &< 0.1 + 0.8x_2 \end{aligned}$$

Por lo tanto:

$$F(x_1) \leq F(x_2)$$

Sean x_1 y $x_2 \in (0, \frac{3}{4})$, donde $x_1 < x_2$:

$$\begin{aligned} x_1 &< x_2 \\ 0.8x_1 &< 0.8x_2 \\ 0.1 + 0.8x_1 &< 0.1 + 0.8x_2 \end{aligned}$$

Por lo tanto para este caso se cumple con:

$$F(x_1) \leq F(x_2)$$

Por último consideramos a $x_1 \in (0, \frac{3}{4})$ y $x_2 \in (\frac{3}{4}, \infty)$, donde $x_1 < x_2$, y $F(x)$ al ser continua en el intervalo $(0, \frac{3}{4})$, evaluamos en el límite del intervalo, por lo que:

$$\begin{aligned} F(x_1) &= 0.1 + 0.8(\frac{3}{4}) = 0.7 \\ F(x_2) &= 1 \end{aligned}$$

Entonces:

$$F(x_1) \leq F(x_2)$$

Por lo tanto la función $F(x)$ es no decreciente para todos los valores de x.

4. $\lim\limits_{x \to -\infty} F(x) = \lim\limits_{x \to -\infty} 0 = 0$

5. $\lim\limits_{x \to \infty} F(x) = \lim\limits_{x \to -\infty} 1 = 1$

De esta manera comprobamos que la función $F(x)$ **es una función de distribución**. Observando cada una de las expresiones que conforman la funcion $F(x)$ podemos observar que la función es continua en todos los puntos, excepto cuando x toma valores $x = 0, x = \frac{3}{4})$. Por lo que podríamos considerar que la variable aleatoria que se encuentra asociada a la función $F(x)$ corresponde a una variable aleatoria continua la cual toma valores discretos en dichos puntos. De esta manera podemos concluir que se trata de una variable aleatoria mixta; toma algunos valores del conjunto discreto y otros valores del conjunto continuo.

9. El archivo Delitos.csv contiene información sobre los delitos denunciados en la ciudad de Aguascalientes, para el período comprendido entre enero de 2011 a junio del 2016. Dicho archivo contiene 5 columnas: la primera contiene la fecha de denuncia del delito; la columna TIPO muestra una descripción del tipo de delito; la segunda columna CONCATENAD presenta una descripción más amplia del delito; la columna SEMANA contiene la semana del año a la que corresponde la fecha de la denuncia; y la columna SEMANA_COMPLETAS indica la semana a lo largo del estudio en la cual se presentó la denuncia. A través de métodos gráficos (e.g. boxplots) traten de determinar el comportamiento semanal de los delitos y discutan alternativas de modelos para describir los delitos cometidos en forma relativamente apropiadas.

Resultados

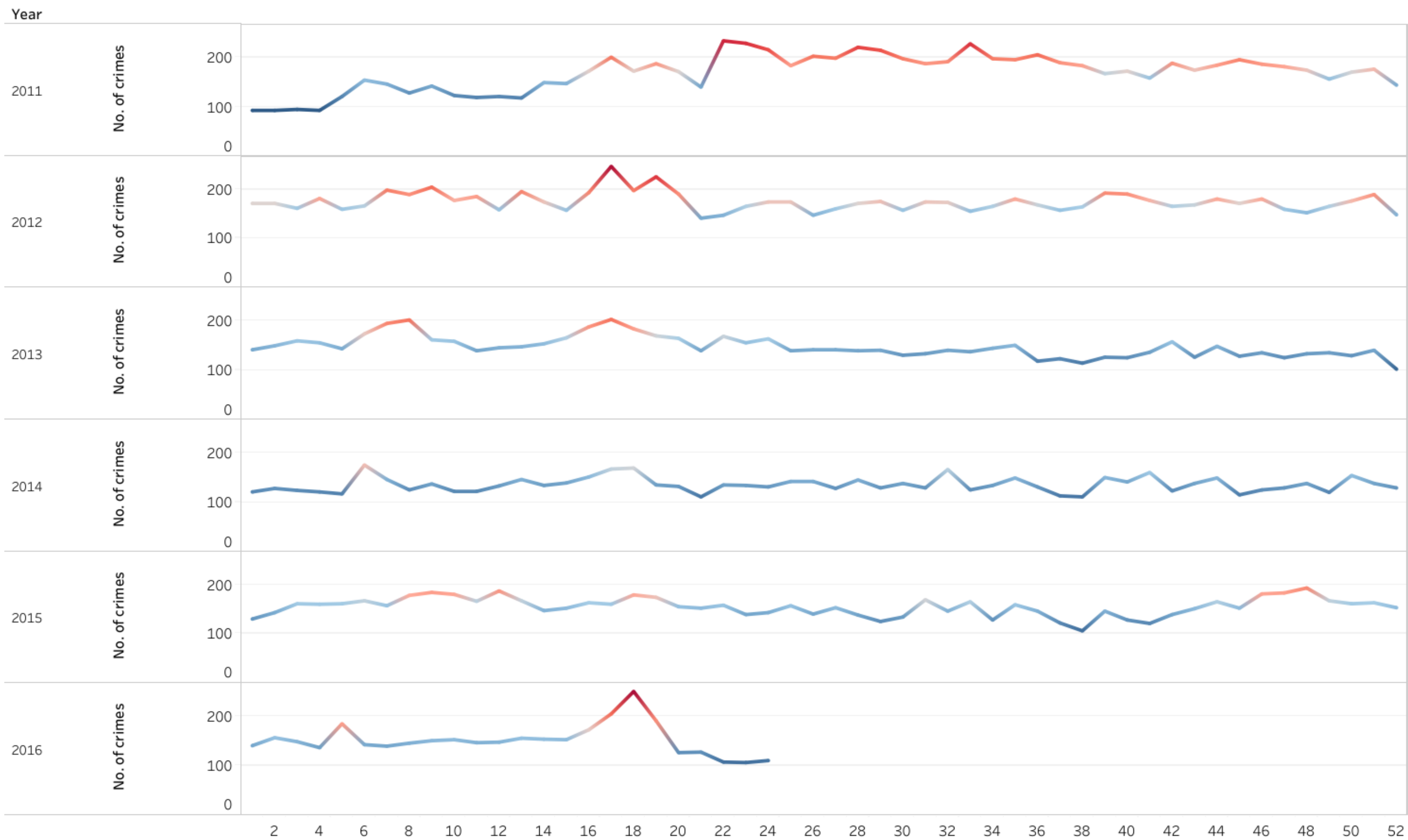
Utilizando la herramienta de visualización de datos Tableau pudimos generar diferentes visualizaciones de datos con el fin de describir el comportamiento de los delitos a lo largo de las semanas entre los años 2011-2016, en la ciudad de Aguascalientes.

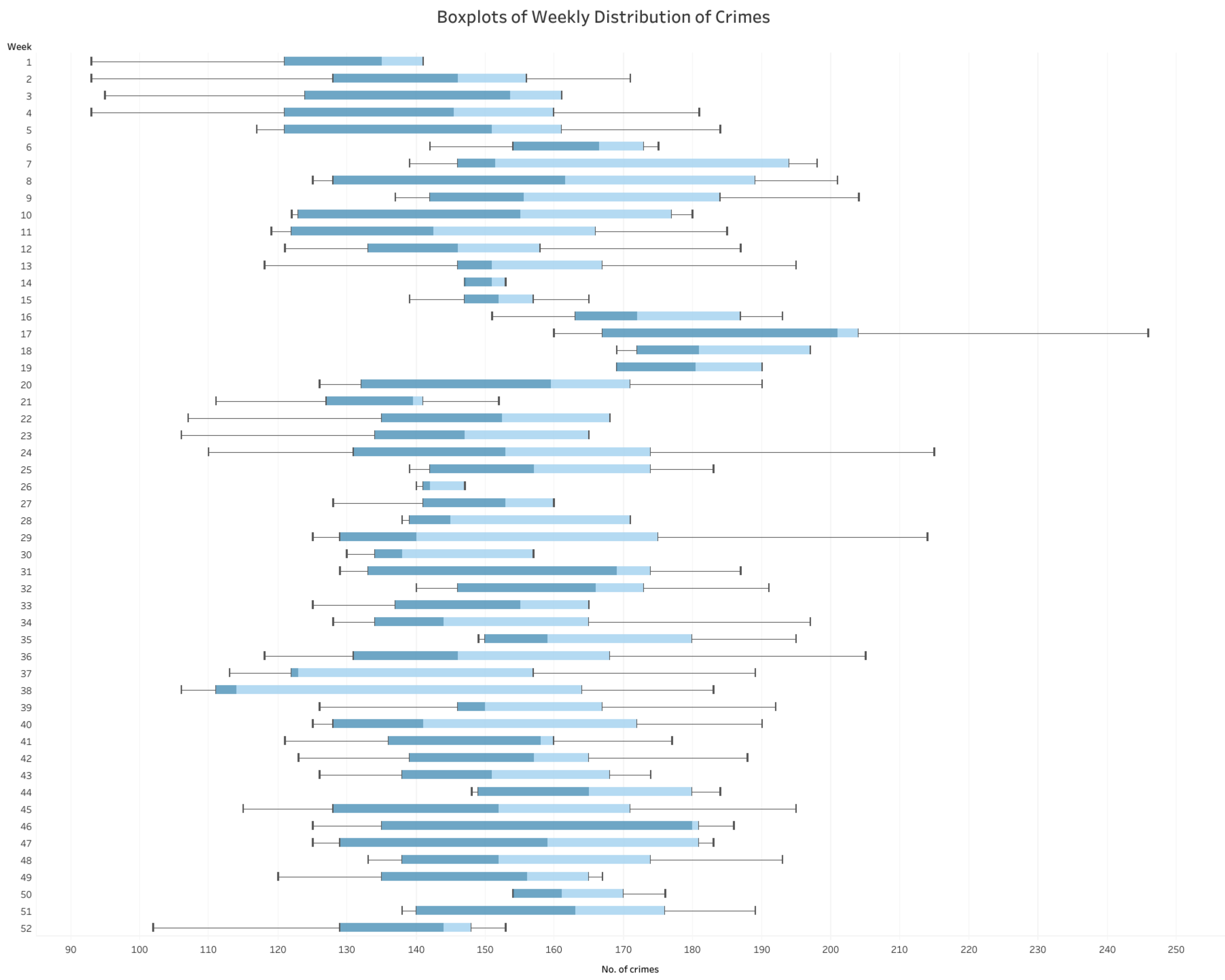
El primer gráfico generado corresponde a un gráfico de líneas, el cual nos muestra el número de delitos que ocurrieron en cada una de las semanas para cada uno de los años en que se colectaron los datos. Se decidió generar un gráfico de este tipo para poder identificar con facilidad patrones existentes en el comportamiento de los delitos cometidos a lo largo de las semanas para cada uno de los años de interés. De manera general encontramos que durante las semanas 16-20 hay un ligero aumento en la ocurrencia de los delitos en cada uno de los años. Esto podría convertirse en un hallazgo significativo realizando análisis exhaustivos que permitan identificar y relacionar los tipos de delitos que predominen en ese período. Además se podría investigar si es que en ese período se realiza un evento local que promueva la ocurrencia de algún tipo específico de delito.

El segundo gráfico corresponde a un gráfico de cajas, donde cada una de las cajas corresponde a cada una de las semanas presentes en un año. En este caso cada caja incluye todos los delitos cometidos en cada una de las semanas para todos los años (i.e. la caja correspondiente a la semana número i incluye el número de delitos cometidos durante la semana i de cada uno de los años en que se recolectaron los datos). A primera vista podemos observar que el número mínimo de delitos ocurridos se presenta en la última y en las primeras semanas del año, por otro lado, el número máximo de delitos esta registrado durante la semana número 17, lo cual coincide con el aumento general de delitos entre las semanas 16 a 20, observado en el gráfico anterior. Sin embargo, fuera de este período observamos que el número de delitos presenta una variabilidad similar.

Dentro de los modelos que podemos proponer para describir el comportamiento de los delitos se encuentra el modelo de regresión lineal. Esto con el objetivo de encontrar posibles variables que relacionen el número de delitos cometidos o su tipo, con alguna variable de interés, como puede ser la hora del día, la época del año, o la tasa de desempleo para un año en particular.

Weekly distribution of crimes in Aguascalientes by year




Boxplots of Weekly Distribution of Crimes
Week
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
90
100
110
120
130
140
150
160
170
180
190
200
210
220
230
240
250
No. of crimes