Reporte de gases contaminantes

Diego Andai Castilla

28 de septiembre de 2016

1. Resumen

En el siguiente informe se presenta el análisis estadístico de la cantidad de gases contaminantes en 59 ciudades de EE.UU. y su relación con distintas variables como mortalidad, temperatura, densidad poblacional, ingresos medios, entre otras.

Los gases contaminantes a estudiar son:

- NOx: óxidos de nitrógeno, generados por combustión a altas temperaturas.
- HC: hidrocarburos, compuestos de carbono e hidrógeno derivados del petróleo.
- SO2: dióxido de azufre, tambien causado por la combustión.

Todos estos son perjudiciales para la salud de los ciudadanos y están relacionados con la utilización de derivados del petróleo en procesos de combustión, como por ejemplo lo que ocurre en los motores de los autos. La abreviación que se muestra arriba será utilizada en este informe.

2. Variables a estudiar

A continuación se presenta una descripción estadística de las distintas variables a considerar en el estudio.

2.1. Gases contaminantes

2.1.1. NO_x

Variable	Unidad
NO_x	ppm

Medidas descriptivas de centro:

Promedio ponderado	Mediana	Moda
22.9661	9	4

Como podemos observar, la cantidad de NOx en el aire tiene gran asimetría. Sus medidas centrales están muy separadas por lo que probablemente existan outliers que estén arrastrando el promedio hacia valores mayores. Esperamos encontrar ciertas ciudades que se escapen de las demás, estas son San Francisco (NOx: 171[ppm]) y Los Ángeles - Long Beach (NOx: 319[ppm]).

Medidas descriptivas de posición:

Mínimo valor	Primer cuartil	Segundo cuartil	Tercer cuartil	Máximo valor
1	4	9	24.5	319

Comprobamos la idea anterior, el 75 % de los datos se encuentra antes de los 24.5[ppm].

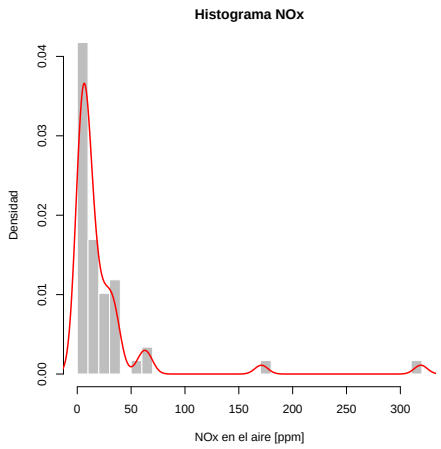
Medidas descriptivas de dispersión y forma:

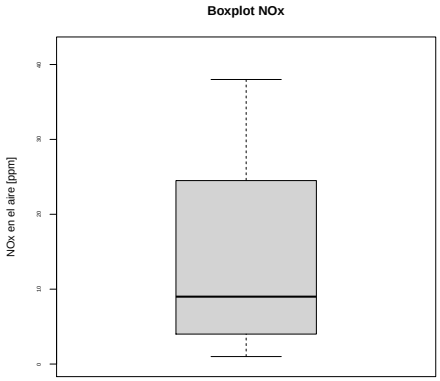
Desviación estándar	Varianza	Skewness	Kurtosis
46.6657	2177.689	4.8685	26.2062

Por último, encontramos que las medidas de dispersión confirman nuestra hipótesis, la medida de skewness es positiva asi que la mayor densidad de datos está a la izquierda. La kurtosis a su vez nos indica que la forma de la distribución será puntiaguda, ya que tenemos la cola derecha con valores muy lejanos a la media.

Gráficos de distribución:

En el histograma de densidad, podemos observar todo lo que esperábamos. Un gráfico cargado a la izquierda, por esto la medida de skewness nos arrojó un valor positivo. Existe un par de valores que se escapan de la densidad general, justamente los que mencionamos en las medidas centrales. La mayoría de los datos se encuentran antes de los 50[ppm]. Podemos apreciar gráficamente la medida de kurtosis, en la punta del gráfico.





Por último, el boxplot de los datos acota la mayor parte de su distribución antes de los 40[ppm]. Esto será algo a considerar más adelante cuando comparemos estos datos con otras variables.

En definitiva, la variable de NOx si bien tiene algunos outliers, se concentra entre 1[ppm] y 40[ppm]

2.1.2. HC

Variable	Unidad
HC	ppm

Medidas descriptivas de centro:

Promedio ponderado	Mediana	Moda
38.4746	15	6

Al igual que los niveles de NOx, los niveles de HC en el aire presentan asimetría, juzgando los valores de medida central. De nuevo observamos que San Francisco (HC: 311[ppm]) y Los Ángeles-Long Beach (HC: 648[ppm]) se escapan bastante de los valores medios, y pueden ser considerados como outliers.

Medidas descriptivas de posición:

Mínimo valor	Primer cuartil	Segundo cuartil	Tercer cuartil	Máximo valor
1	7	15	30.5	648

Comprobamos la idea anterior, el 75 % de los datos se encuentra antes de los 30.5[ppm].

Medidas descriptivas de dispersión y forma:

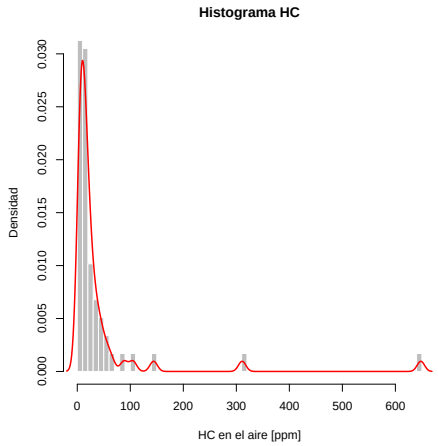
Desviación estándar	Varianza	Skewness	Kurtosis
92.6388	8581.943	5.2732	30.0754

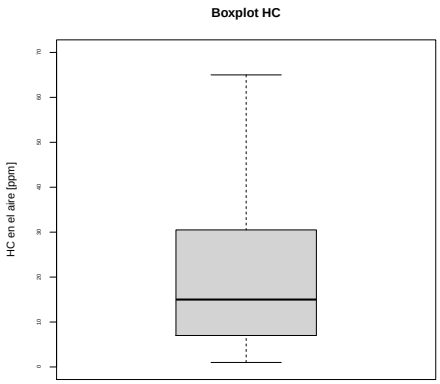
Encontramos que las medidas de dispersión confirman nuestra hipótesis, esta variable tiene una dispersión y asimetría incluso mayor a la anterior, esto obviamente debido a que

los valores que se escapan son incluso mayores a los que se escapan en la cantidad de NOx. La kurtosis es también mayor y por lo tanto de nuevo esperamos una forma puntiaguda en la línea de densidad.

Gráficos de distribución:

El histograma confirma el análisis anterior, muy cargado a la izquierda, puntiagudo y podemos observar los outliers de San Francisco y Los Ángeles-Long Beach. La mayoría de los datos se encuentran entre 1[ppm] y 100[ppm], esto se confirma al estudiar el boxplot de la variable.





En el boxplot de los datos, en el que no se muestran los outliers, la mayoría de mediciones se encuentra bajo los 70[ppm].

Con este análisis se desprende que la densidad de cantidades de HC se encuentra entre 1[ppm] y 70[ppm], esto debe considerarse al momento de relacionar la variable con otras, puesto que los outliers no entregan información estadísticamente concluyente.

2.1.3. SO_2

Variable	Unidad
SO_2	ppm

Medidas descriptivas de centro:

Promedio ponderado	Mediana	Moda
54.661	32	1

Como en los dos gases anteriores, se observa una gran diferencia entre las medidas descriptivas centrales de la variable SO_2. La moda es uno, por lo que este parece ser un gas menos común que los anteriores.

Medidas descriptivas de posición:

Mínimo valor	Primer cuartil	Segundo cuartil	Tercer cuartil	Máximo valor
1	13	32	70	278

A pesar de las medidas de tendencia central, esta variable tiene una distribución más equitativa, según el estudio de sus cuartiles. También tiene un máximo menor al de los otros dos gases, confirmando que este aparece en menores cantidades

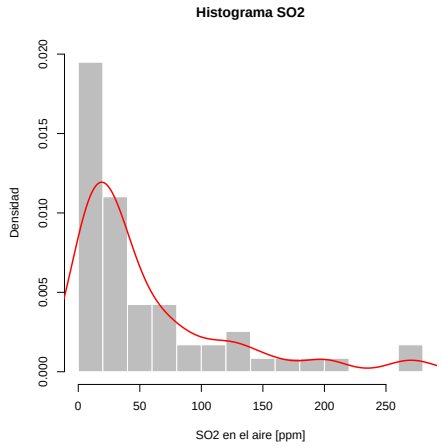
Medidas descriptivas de dispersión y forma:

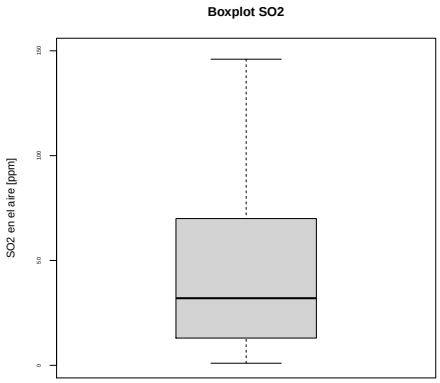
Desviación estándar	Varianza	Skewness	Kurtosis
63.5517	4038.814	1.8022	2.8894

Las medidas de dispersión confirman que esta variable esta distribuida de manera más equitativa. Tiene, sin embargo, asimetría, otra vez inclinándose a la izquierda.La kurtosis es muho menor a las otras dos variables, lo que denota una concentración alrededor de la mediana también menor, un gráfico menos puntiagudo y más suave.

Gráficos de distribución:

El histograma es, efectivamente, más equitativo. La kurtosis se nota menor pues la linea de densidad es mucho más suave que la de los otros dos gases. Si bien la medida de skewness resulto menor que en los otros dos, el gráfico parece estar igual de cargado a la izquierda. En realidad este valor fue menor porque el rango de valores de esta variable es menor, y los valores de la derecha son más comunes. La gran mayoría de los datos se encuentra antes de los 150[ppm].





Se confirma con el boxplot que la distribución de los valores tiene mayor densidad antes de los 150[ppm]

Las cantidades de SO_2 son menores que las de los otros gases, y se acota su distribución más densa entre 1[ppm] y 150[ppm]. A pesar de lo anterior, los valores extremos no están tan alejados, así que de todas formas vale la pena observar su comportamiento.

2.2. Variables de población

2.2.1. Población total

Variable	Unidad
Población total	Número de personas

Medidas descriptivas de centro:

Promedio ponderado	Mediana	Moda
1.438.037	914.427	Ninguna

Se observa una diferencia entre la esperanza y la mediana que indica asimetría en los datos. Al ser una variable de gran magnitud y valores muy específicos es dificil qe se produzca una moda, hay una probabilidad muy baja de que eligiendo 59 valores entre ∼8.000.000 valores (diferencia entre el mínimo y máximo que se expone más adelante) uno se repita.

Medidas descriptivas de posición:

Mínimo valor	Primer cuartil	Segundo cuartil	Tercer cuartil	Máximo valor
124.833	566.515	914.427	1.717.201	8.274.961

Vemos que, de nuevo, existe una tendencia a que aparezcan outliers mayores al dataset general, concentrándose la muestra en la porción que es menor a la media.

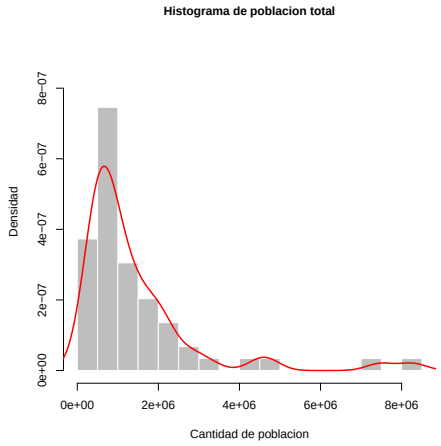
Medidas descriptivas de dispersión y forma:

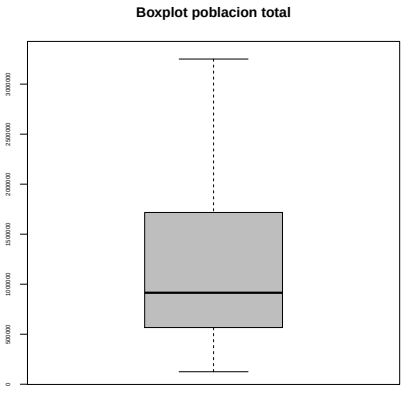
Desviación estándar	Varianza	Skewness	Kurtosis
1,54x10^6	2.38x10^8	2.7438	8.2747

Si bien los valores de la desviación estándar y varianza son grandes, esto se debe en parte a la magnitud de la variable. La media de esta es 1,48x10^6, asi que un valor de esta magnitud nos indica una desviación mayor, pero dentro de lo esperable. La medida de skewness confirma que la muestra es más densa a la izquierda del gráfico, por último la kurtosis nos hace esperar una concentración mayor en la media y por lo tanto un gráfico ni tan suave ni tan puntiagudo.

Gráficos de distribución:

El histograma, como esperábamos, está cargado a la izquierda. Podemos ver los outliers y que la mayor parte de los datos se encuentra antes de 2x10^6. Se observa la punta que sugería la kurtosis.





Por último el boxplot muestra que los datos se concentran bajo los 3.000.000, mostrando una densidad muy cargada alrededor del millón de personas. Los datos corresponden a ciudades grandes pero no masivas (exceptuando los outliers). Como punto de referencia Santiago tiene alrededor de 5 millones de personas (datos no tan confiables dado nuestro último censo).

2.2.2. Población por hogar

Variable	Unidad
Población por hogar	Número de personas

Esta variable es un promedio de la cantidad de personas que viven en un domicilio particular para cada ciudad.

Medidas descriptivas de centro:

Promedio ponderado	Mediana	Moda
3.2466	3.27	3.21 y 3.32

Las medidas de tendencia central de esta variable son muy cercanas, por lo que esperamos simetría. Era de esperar este resultado pues en nuestra cultura occidental el número de personas en cada casa no varía mucho más allá del rango 1-6.

Medidas descriptivas de posición:

Mínimo valor	Primer cuartil	Segundo cuartil	Tercer cuartil	Máximo valor
2.65	3.21	3.27	3.36	3.53

A pesar de la simetría que sugieren las medidas de tendencia central, vemos que el valor mínimo se escapa más de la media de lo que lo hace el valor máximo, lo que sugiere que los datos están un poco cargados a la derecha.

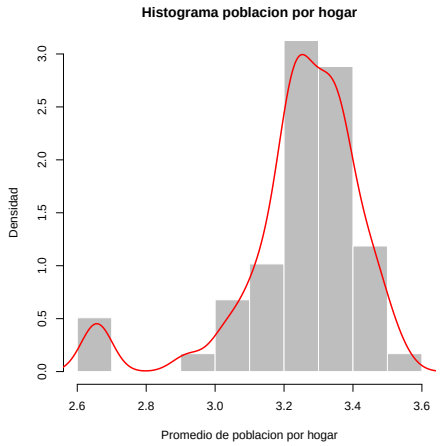
Medidas descriptivas de dispersión y forma:

Desviación estándar	Varianza	Skewness	Kurtosis
0.1829	0.0335	-1.6032	3.1184

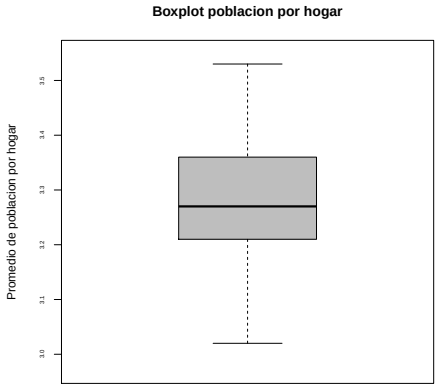
La desviación estándar es pequeña, lo que era de esperarse por la magnitud de diferencia entre los valores y por la distribución simétrica de estos. El skewness confirma que el gráfico está un poco cargado a la derecha y la kurtosis nos hace esperar una forma más bien suave.

Gráficos de distribución:

El histograma efectivamente está cargado a la derecha, pero podemos observar que esto se debe a un outlier ubicado a la izquierda. Sin este, la distribución sería muy simétrica, y es por esto que los valores de medida central son tan cercanos. La forma es suave, como predijo la kurtosis y parece tener dos puntas, probablemente causado por la naturaleza multimodal de la variable.



El boxplot, que deja afuera al outlier, presenta la simetría esperada.



2.2.3. Ingresos medios

Variable	Unidad
Ingresos medios anuales	Miles de USD

Medidas descriptivas de centro:

Promedio ponderado	Mediana	Moda
33246.66	32452	Ninguna

Al igual que en la variable de población total, es difícil que se presentara una moda. Las otras dos medidas son cercanas asi que debiese existir simetría en los datos.

Medidas descriptivas de posición:

Mínimo valor	Primer cuartil	Segundo cuartil	Tercer cuartil	Máximo valor
25782	30004.5	32452	35496	47966

La distribución presenta una naturaleza simétrica, concentrándose los datos entre el primer y tercer cuartil, y los extremos están separados equitativamente.

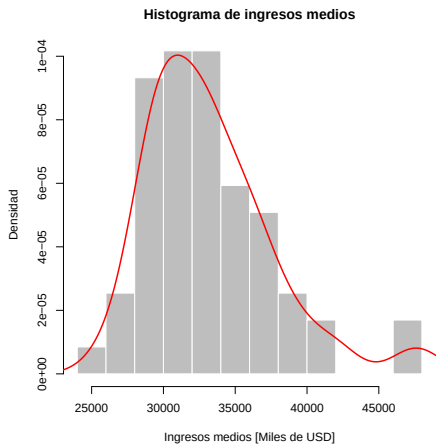
Medidas descriptivas de dispersión y forma:

Desviación estándar	Varianza	Skewness	Kurtosis
$4.47 \text{x} 10^3$	$2 \text{x} 10^7$	1.2131	1.6983

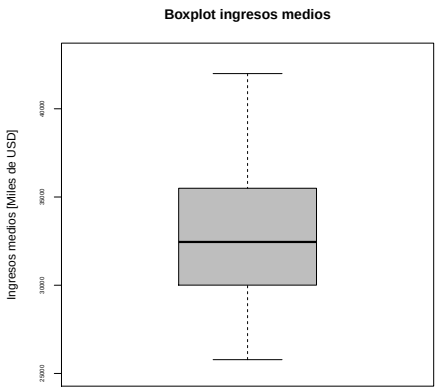
La desviación estándar no es tanta (magnitud 10^3 en una variable de magnitud 10^4) pero muestra que los datos están distribuidos, por lo mismo la kurtosis es pequeña. El skewness es tampoco es muy elevado pero de todas formas existe una inclinación de los datos hacia la izquierda.

Gráficos de distribución:

Hay un outlier a la derecha que es el que 'arruina' la simetría de los datos, de todas maneras la parte densa de estos se inclina un poco a la izquierda. La kurtosis define esa forma suave.



El boxplot es simétrico, con tendencia a la baja.



2.2.4. Mortalidad

Variable	Unidad
Mortalidad (por mil habitantes)	Cantidad de personas

Medidas descriptivas de centro:

Promedio ponderado	Mediana	Moda
941.1731	946.19	Ninguna

La mortalidad parece ser simétrica, la mediana y media son cercanas sobre todo considerando la magnitud de la variable.

Medidas descriptivas de posición:

Mínimo valor	Primer cuartil	Segundo cuartil	Tercer cuartil	Máximo valor
790.73	899.395	946.19	984.12	1113.16

La distribución presenta una naturaleza simétrica, concentrándose los datos entre el primer y tercer cuartil, y los extremos están separados equitativamente.

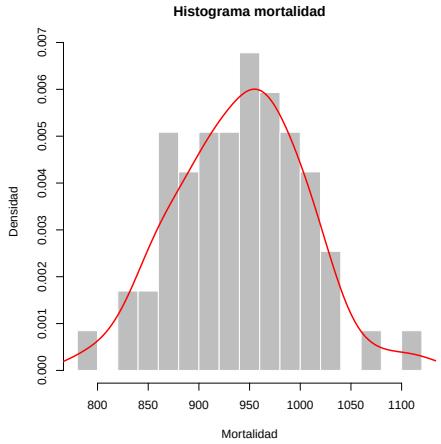
Medidas descriptivas de dispersión y forma:

Desviación estándar	Varianza	Skewness	Kurtosis
62.4213	3896.423	0.0629	-0.0495

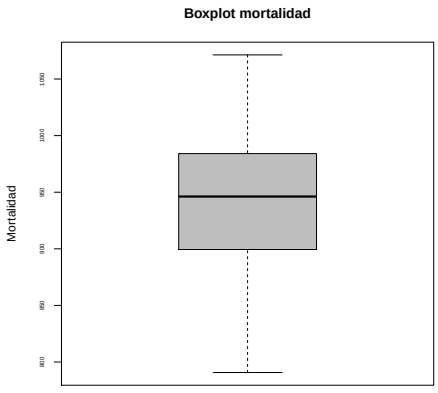
La desviación estándar es mediana, por lo que los datos están distribuidos a lo largo del soporte. El skewness es muy bajo por lo que esperamos bastante simetría. La kurtosis es negativa así que el gráfico es mas bien suave, platicúrtico.

Gráficos de distribución:

Como se puede apreciar en el histograma, la variable presenta una naturaleza simétrica.



El boxplot es simétrico, con tendencia a la alza. Abarca la mayoría de los datos.



2.3. Variables mediambientales

2.3.1. Cantidad de lluvia

Variable	Unidad
Lluvia caída	Pulgadas

Medidas descriptivas de centro:

Promedio ponderado	Mediana	Moda
38.5085	38	36, 35 y 42

Esta es una variable multimodal, esto puede deberse a que algunas ciudades son cercanas. Se denota una cercanía entre los valores que hace pensar en simetría de la variable.

Medidas descriptivas de posición:

Mínimo valor	Primer cuartil	Segundo cuartil	Tercer cuartil	Máximo valor
10	33.5	38	44	65

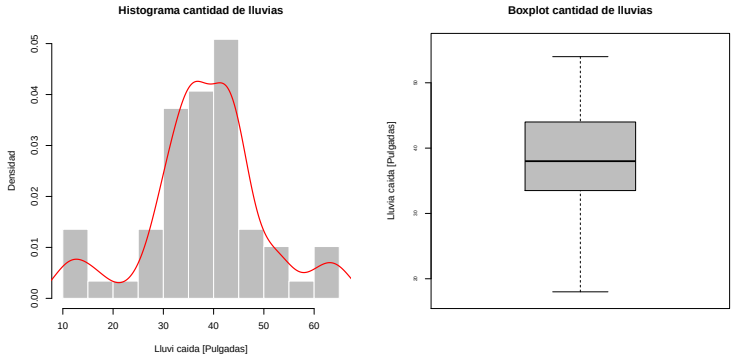
La distribución presenta una naturaleza simétrica, concentrándose los datos entre el primer y tercer cuartil, y los extremos están bastante separados (con respecto al centro) por lo que podemos esperar que la línea de densidad tenga colas más largas.

Medidas descriptivas de dispersión y forma:

Desviación estándar	Varianza	Skewness	Kurtosis
11.57	133.94	-0.1724	0.6749

La desviación estándar es considerable, lo que reafirma la idea de datos más dispersos y no tan concentrados, por eso la kurtosis es tan pequeña. El skewness es bajo también lo que denota simetría.

Gráficos de distribución:



Se puede apreciar la simetría del histograma y del boxplot, aunque este último tiene una tendencia al alza que se observa también en las barras. Es esta pequeña asimetría la que nos entrega ese valor de skewness negativo. La doble punta probablemente se deba a la naturaleza multimodal.

2.3.2. Humedad

Variable	Unidad
Humedad	Porcentaje de humedad en el aire

Medidas descriptivas de centro:

Promedio ponderado	Mediana	Moda
57.7458	57	56

Esta variable presenta simetría.

Medidas descriptivas de posición:

Mínimo valor	Primer cuartil	Segundo cuartil	Tercer cuartil	Máximo valor
38	55.5	57	60	73

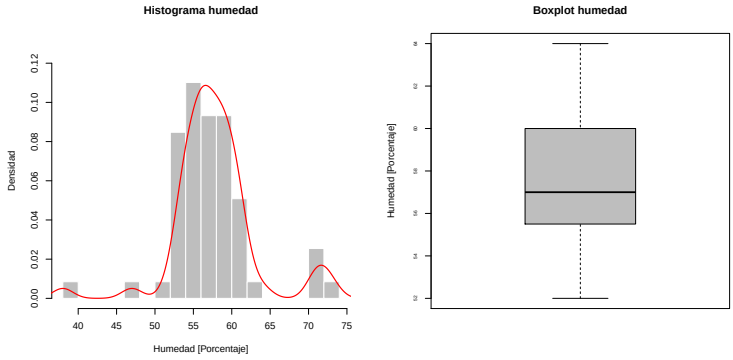
Se presenta una mayor concentración alrededor de la mediana y una pequeña tendencia hacia los valores menores.

Medidas descriptivas de dispersión y forma:

Desviación estándar	Varianza	Skewness	Kurtosis
5.38	28.95	0.1959	3.652

La desviación estándar es mediana. El skewness confirma que los datos están un poco cargados a la izquerda y la kurtosis se debe a la densidad de datos en el medio del soporte.

Gráficos de distribución:



Podemos observar claramente en el histograma que la mayor densidad de datos se concentra en el medio, por esto la punta que sugiere la kurtosis. El boxplot nos muestra la tendencia a la baja que esperábamos por el valor del skewness.

2.3.3. Temperaturas en Enero

Variable	Unidad
Temperatura en el mes de Enero	Fahrenheit

Se presenta el promedio de las temperaturas registradas en el mes de Enero.

Medidas descriptivas de centro:

Promedio ponderado	Mediana	Moda
33.7966	31	30 y 24

Es una variable multimodal, que presenta una pequeña tendencia hacia los valores menores.

Medidas descriptivas de posición:

Mínimo valor	Primer cuartil	Segundo cuartil	Tercer cuartil	Máximo valor
12	27	31	39.5	67

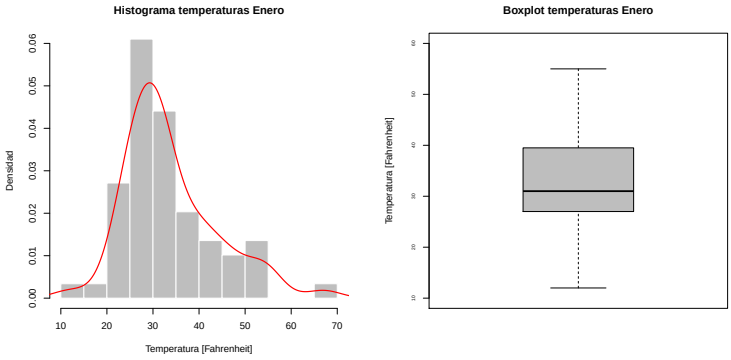
Confirmamos que los datos se cargan hacia la izquierda, con un máximo que probablemente sea considerado como outlier.

Medidas descriptivas de dispersión y forma:

Desviación estándar	Varianza	Skewness	Kurtosis
10.15	103.06	0.9659	0.9072

La desviación estándar es mediana. El skewness y la kurtosis son bajos, la variable es bastante simétrica con una inclinación hacia la izquierda y debiese tener forma suave.

Gráficos de distribución:



En los gráficos se observa claramente el analisis anterior, se ve la inclinación, el outlier y la forma relativamente suave.

2.3.4. Temperaturas en Julio

Variable	Unidad
Temperatura en el mes de Julio	Fahrenheit

Se presenta el promedio de las temperaturas registradas en el mes de Julio.

Medidas descriptivas de centro:

Promedio ponderado	Mediana	Moda
74.4	74	72

Variable simétrica, con una pequeña tendencia a la baja.

Medidas descriptivas de posición:

Mínimo valor	Primer cuartil	Segundo cuartil	Tercer cuartil	Máximo valor
63	72	74	77	85

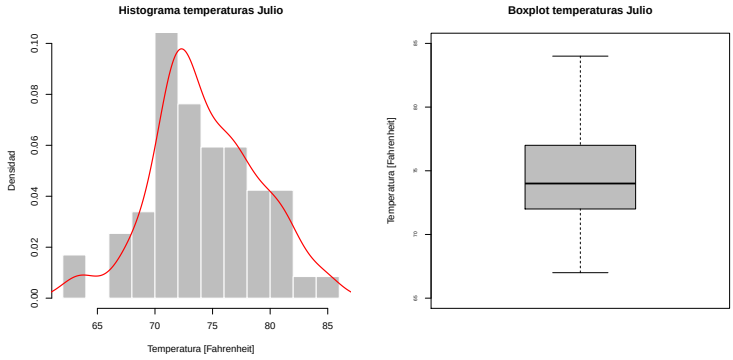
Confirmamos que los datos se cargan hacia la izquierda, con extremos equitativamente distribuidas.

Medidas descriptivas de dispersión y forma:

Desviación estándar	Varianza	Skewness	Kurtosis
4.6	21.18	0.0632	-0.1576

La desviación estándar es pequeña. El skewness y la kurtosis son bajos, la variable es bastante simétrica y debiese tener forma suave.

Gráficos de distribución:



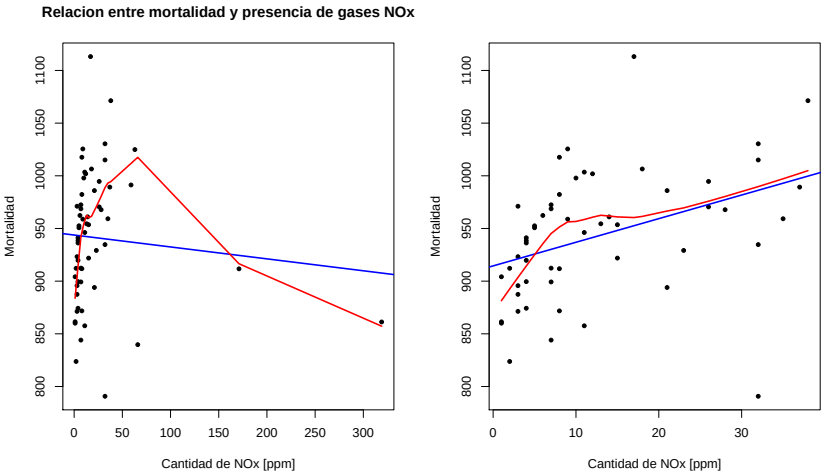
Si bien el histograma está cargado al medio, no tenía tanta simetría como esperábamos. El boxplot es bastante simétrico con una pequeña tendencia a la baja.

3. Mortalidad y gases contaminantes

Los gases contaminantes descritos al principio del informe podrían tener cierta relación con la mortalidad de la zona. Para estudiar esto, analizaremos gráficos de mortalidad vs. presencia del gas

3.1. NOx

Gráfico Mortalidad vs. Presencia de NOx:



En estos gráficos las lineas rojas y azules ajustan la relación entre los puntos, las primeras con una función aproximada y las segundas con una regresión lineal.

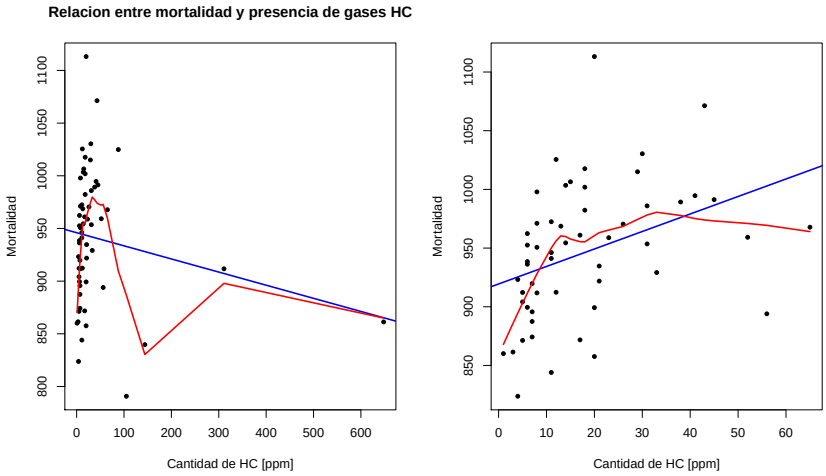
Observamos dos gráficos, el primero es la relación entre todos los valores de NOx y la mortalidad, este nos entrega una relación inversa entre las variables. Esto no es esperado porque NOx es dañino para la salud, el problema es que, como vimos en la sección anterior, la variable de NOx tiene outliers que no deben ser considerados ya que no representan conclusiones estadísticas sólidas. Si acotamos la cantidad de NOx a 40[ppm] obtenemos el segundo gráfico. En este, como vimos antes, si existen suficientes datos como para hacer conclusiones. Deducimos entonces que a mayor cantidad de NOx efectivamente, aumenta la mortalidad.

La pendiente de la segunda recta es de aproximadamente 2.2, lo que significa un aumento en la mortalidad de 2.2 por cada [ppm] de NOx adicional.

La acotación anterior debe hacerse puesto que la mortalidad depende de muchas variables, y para relacionarla con alguna de estas debemos obtener una tendencia que se sostenga en varios casos.

3.2. HC

Gráfico Mortalidad vs. Presencia de HC:

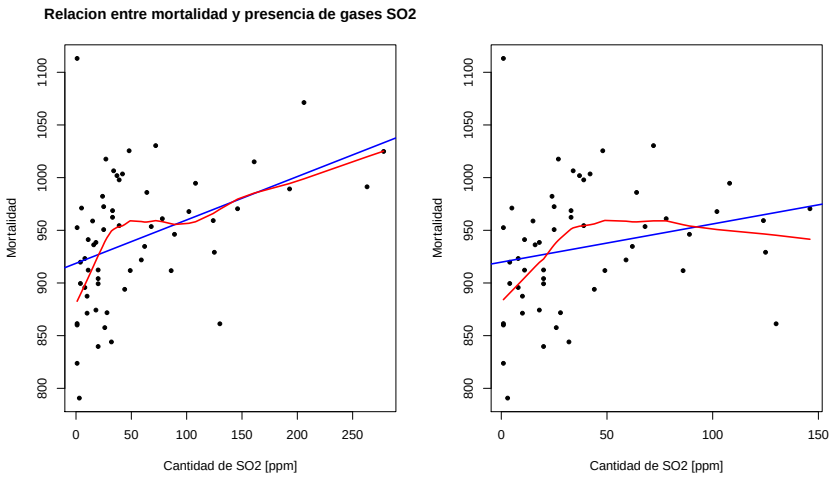


Al igual que en NOx, los outliers alteran el gráfico si tomamos todos los valores de HC, pero si tomamos hasta 70[ppm] que es, como vimos en la sección 2.1.2, donde se concentra la mayor cantidad de datos, obtenemos de nuevo que a mayor presencia de HC aumenta la mortalidad.

La pendiente de la segunda recta es de aproximadamente 1.5, lo que significa un aumento en la mortalidad de 1.5 por cada [ppm] de HC adicional.

3.3. SO_2

Gráfico Mortalidad vs. Presencia de SO_2 :



Con el SO_2 no encontramos el problema de antes, aun con todos los datos este gas presenta una relación de proporcionalidad directa con la mortalidad, de todas maneras graficamos la zona de alta densidad obtenida en 2.1.3 para confirmar nuestro resultado.

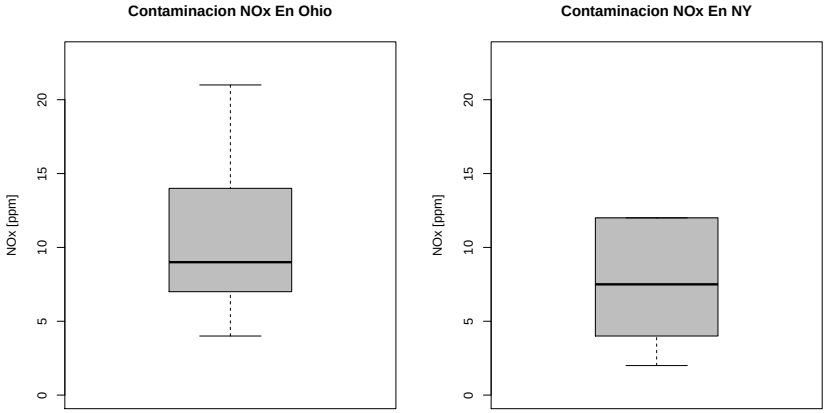
La pendiente de las rectas es aproximadamente 0.35, lo que significa un aumento de la mortalidad en 0.35 por cada [ppm] de SO_2 que aumenta en el ambiente.

3.4. Conclusión

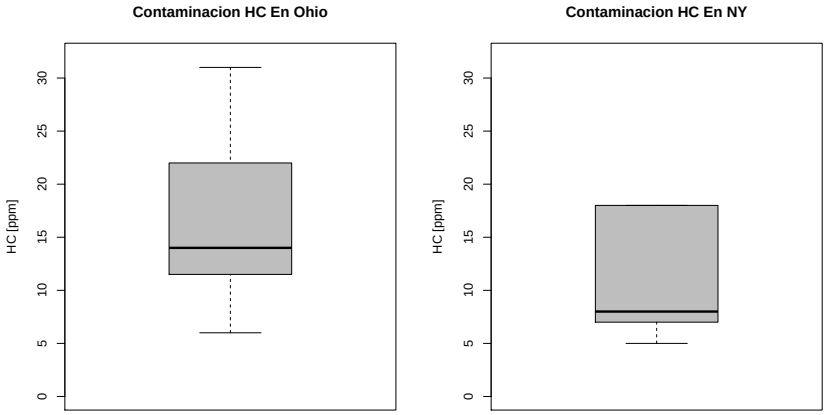
Los tres gases presentan un efecto agravante en la mortalidad de la población, siendo NOx que tiene la mayor tasa de aumento de mortalidad por cada [ppm] adicional, y SO_2 el con menor tasa de aumento.

4. Contaminación en NY y OH

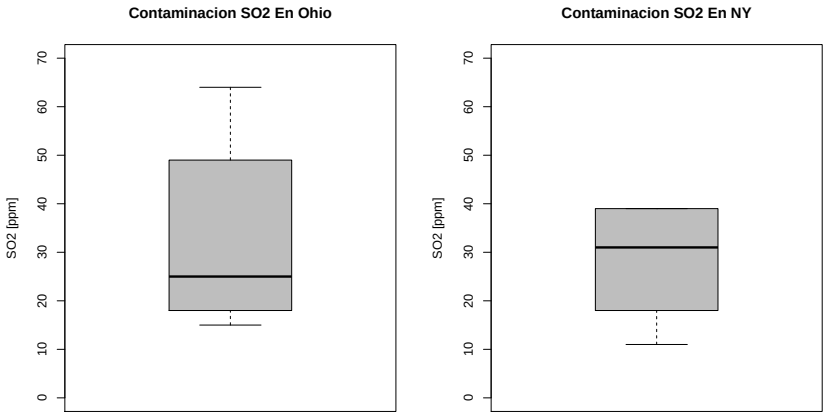
Podemos comparar la contaminación entre ambos estados, a continuación se presentan gráficos de cajas para cada gas lado a lado (con la misma escala para que sea más fácil comprarar visualmente):


Contaminacion NOx En Ohio
NOx [ppm]
0
5
10
15
20

Contaminacion NOx En NY
NOx [ppm]
0
5
10
15
20


Contaminacion HC En Ohio
HC [ppm]
0
5
10
15
20
25
30

Contaminacion HC En NY
HC [ppm]
0
5
10
15
20
25
30

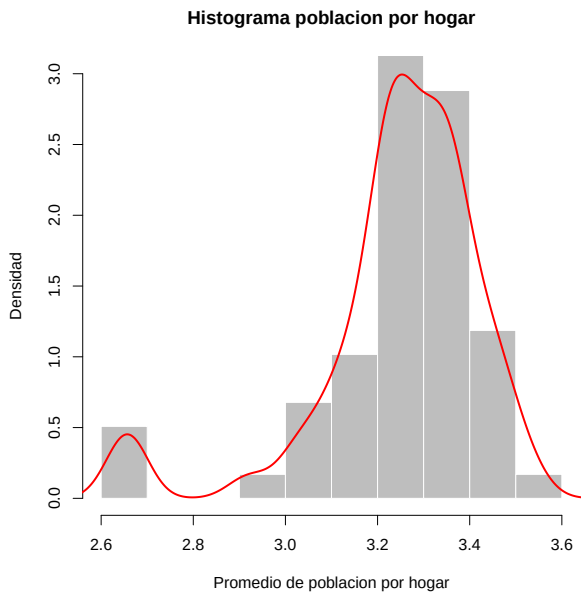

Contaminacion SO2 En Ohio
SO2 [ppm]
0
10
20
30
40
50
60
70

Contaminacion SO2 En NY
SO2 [ppm]
0
10
20
30
40
50
60
70

Notamos como Ohio, para los tres gases, tiende a presentar valores mayores de contaminación

5. Ajuste de modelo para población por hogar

Por último, buscamos ajustar un modelo de probabilidad a la variable de población por hogar estudiada en la sección 2.2.2, recordamos su histograma de distribución:

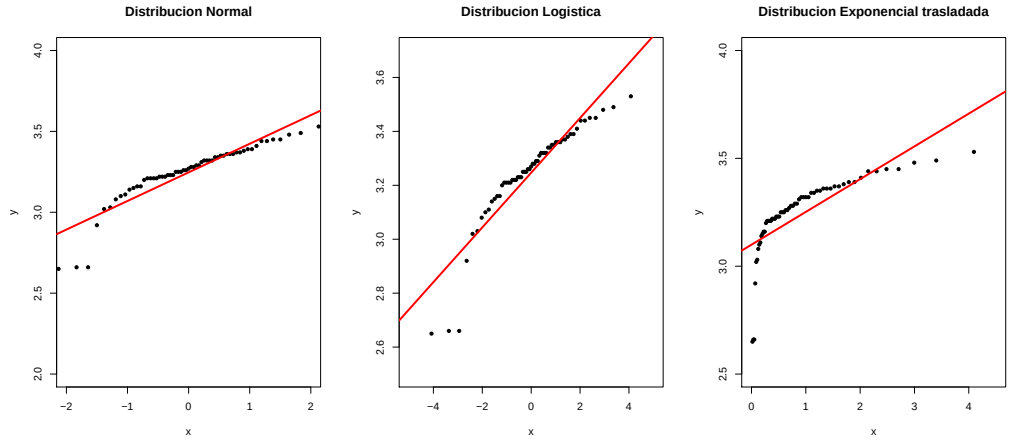


Buscamos un modelo en la familia de localización y escala, o sea alguno de:

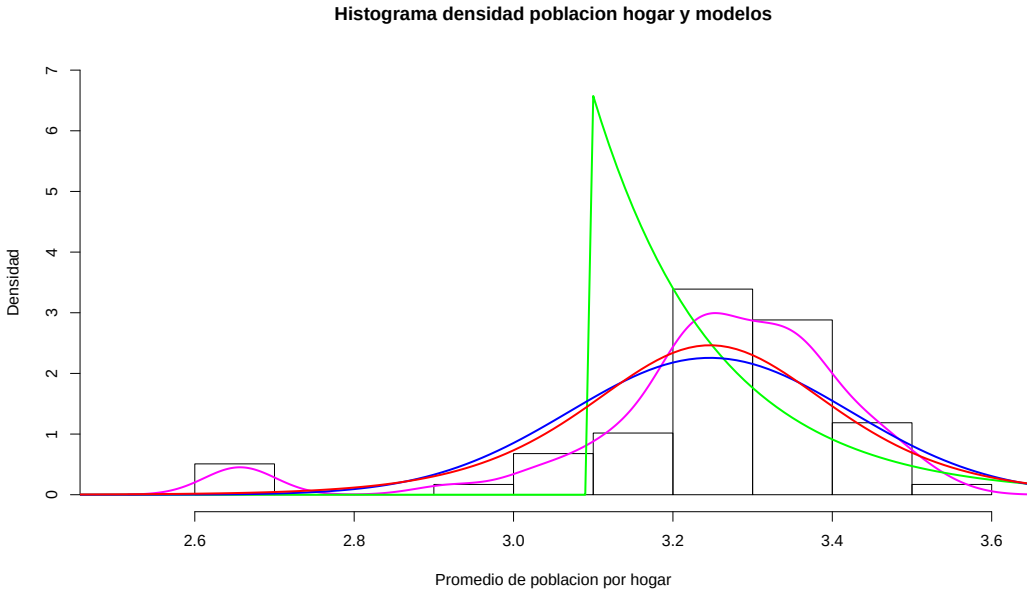
- Distribución Exponencial trasladada
- Distribución Normal
- Distribución Logística

Recordando que los modelos de localización y escala son aquellos que tienen un parametro de escala no negativo y uno de localización.

Para hacer esto ordenamos la variable de forma creciente con respecto a su probabilidad, hacemos una regresión lineal para intentar alinearla con la función de distribución y podemos analizar gráficamente qué modelo se adapta mejor, siendo este cuya recta se acerque más al gráfico de los valores.



Notamos que, la recta más ajustada es la de la distribución logística, seguido de cerca por la normal. Podemos confirmar esto con el gráfico de densidad y los tres modelos:



Donde la línea magenta representa la densidad de la variable, la roja es la distribución logística, la azul la normal y la verde la exponencial trasladada.

Se concluye entonces que:

$$\text{Población por hogar} \sim \text{Logística}(3.25,\ 0.1)$$