Trabajo Practico especial: Distribución de datos muestrales

Valentin Diaz, Pedro Salomone

Modelos y Simulación 2023

Índice

1. Resumen

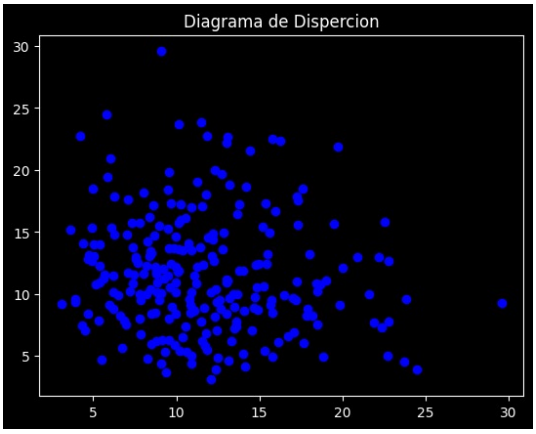
En este informe desarrollaremos el estudio de la distribución una muestra. Por un lado, haremos un análisis exploratorio de datos utilizando estadística descriptiva, y por otro lado utilizaremos la estadistica inferencial para poder estimar ciertos parámetros y poder concluir si la muetra proviene o no de cierta distribución.

En un primer lugar generaremos un diagrama de dispersión para poder interpretar que la muestra efectivamente sea independiente. En la parte descriptiva utilizaremos las típicas medidas de posición central y de dispersión para obtener una mirada genereal de los datos. Por último, propondremos 2 familias de distribuciones y sus respectivos parametros, y con ayuda de la estadística inferencial podremos desarrollar de manera mas precisa una conclusion con respecto a cual de estas no pertence. Para esto utilizaremos el Test de Pearson con aproximación chi-cuadrado y el test de Kolmogorov-Smirnov con el fin de estimar los p-valores de las pruebas de hipótesis.

2. Desarrollo

2.1. Independencia Estadística

El siguiente gráfico corresponde al diagrama de dispersión de los datos de la muestra:



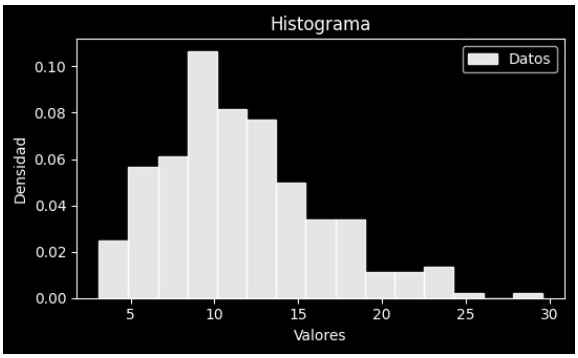
En este podemos ver que no hay indicios de algún patrón que los datos sigan; no se observan comportamientos lineales, ni cíclicos, ni tampoco de agrupamiento. En general no podemos asegurar ninguna tendencia en la muestra, es decir que no observamos ninguna dependencia directa de los datos de la muestra.

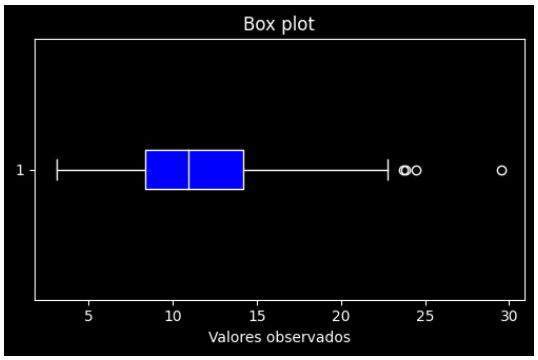
2.2. Análisis Descriptivo

En la siguiente tabla observamos algunas medidas que calculamos de la muestra:

	Valor
Valor Mínimo	3.093
Valor Máximo	29.568
Media	11.601
Varianza	22.067
Desviación Estándar	4.698
Skewness	0.443
Mediana	10.908
Q1	8.38
Q3	14.225

Estas métricas nos dan una idea de como se ve la distribución de la muestra. Uno puede intuir por la media y la desviación estándar que la mayoría de los datos deberían estar entre 8 y 15 aproximadamente (media±desv). Otro parámetro a tener en cuenta es "skewness", dado que al ser positivo nos indicaría que la cola derecha de la distribución tiende a ser mas pesada que la izquierda. Más aún, esto podemos verificarlo con el siguiente histograma y boxplot:



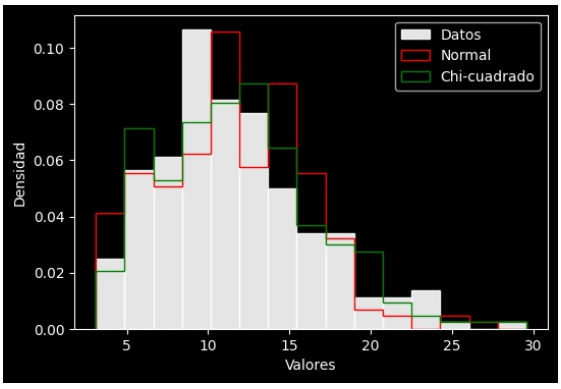


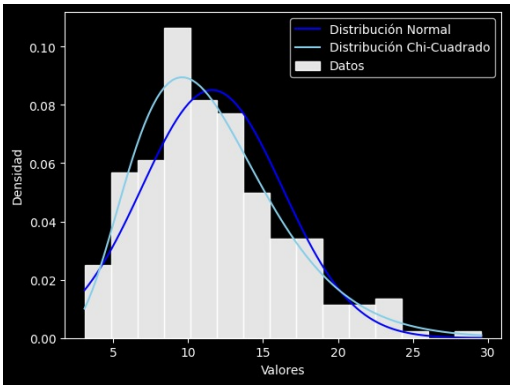
2.3. Propuesta de Distribuciones

En base a lo analizado anteriormente podemos confiar en que la distribucion de la muestra tenga forma de campana. Otra característica que podría tener la distribución es ser asimétrica debido a que, como dijimos antes, presenta una cola mas pesada a derecha. En base a esto proponemos 2 tipos de distribuciones:

- Distribución Normal
- Distribución Chi-Cuadrada

En la elección procuramos que ambas respeten la forma de campana, dado que es la característica mas obvia de la muestra. En cuanto a la asimetría, decidimos elegir una distribución con tal caracteristica y otra que no.
Luego tuvimos que definir los parámetros de dichas distribuciones. En el caso de la normal lo hicimos por máxima verosimilitud, y en el de Chi-Cuadrada por el método de los momentos. En el caso de la Normal utilizamos la media muestral para μ y la varianza muestral para σ^2. En el caso de la Chi-Cuadrada utilizamos la media muestral para sus grados de libertad.
De esa forma graficamos el histograma de la muestra en conjunto con, por un lado, los histogramas de ambas distribuciones generados de manera aleatoria, y por otro lado, las funciones de densidad.





En base a las imagánes, podríamos apuntar que las propuestas son razonablemente aceptables. Los gráficos se plegan de manera acorde con el histograma, y más aún, la Chi-Cuadrada respeta la caida de cola a la derecha.

2.4. Estimaciones del p-valor

Por último utilizaremos el test de bondad de ajuste para intentar rechazar o no las distribuciones propuestas. Planteamos lo siguiente:

Normal

H_0:Los datos de la muestra son independientes y provienen de una distribución Normal.

H_1:Los datos de la muestra no provienen de una distribución Normal.

Chi-cuadrada

H_0:Los datos de la muestra son independientes y provienen de una distribución Chi-cuadrada.

H_1:Los datos de la muestra no provienen de una distribución Chi-cuadrada.

Luego evaluaremos los test de hipótesis mediante el p-valor en cada caso. A continuación explicaremos como utilizamos cada método de manera general. Notar que cada test lo realizamos con la distribución Normal o Chi-Cuadrada

Test de Pearson

N_i: Para obtener las frecuencias acumuladas subdividimos el rango de la muestra en 21 intervalos.

p_i: Resultados teóricos de la probabilidad de que la v.a de distribución F esté en cada intervalo.

Paso siguiente obtenemos t valuando el estadístico de prueba en la muestra original. Por último simulamos una muestra de distribución F y comparo el estadístico valuado en esa muestra con t para poder estimar el p-valor

Test Kolmogorov-Smirnov

Primero ordenamos los datos de la muestra para poder calcular el estadístico de K-S según la distribución F. Seguimos con n simulaciones en las que se genera una muestra en cada una para calcular un nuevo estadístico y compararlo con el que calculamos al principio para poder estimar el p-valor.

Finalmente los resultados de los p-valores generados fueron los siguientes:

	K-S	**Pearson**
Normal	0.01	0.0342
Chi-Cuadrada	0.49	0.2545

3. Conclusión

Para el caso de la distribución Normal con parametros estimados por máxima verosimilitud, obtuvimos que los p-valores simulados para ambos test son pequeños y por lo tanto rechazaríamos

la hipótesis nula, entendiendo así que la muestra de datos no proviene dicha distribución con tales parámetros.
Por otro lado, los p-valores obtenidos para la distribución Chi-Cuadrada con parámetros estimados por el método de los momentos están en una situacion diferente. Estos valores son relativamente buenos, por lo tanto no indicaría rechazar la hipótesis nula. Si bien esto no significaría aceptarla, podemos concluir que la segunda distribución propuesta es más adecuada que la primera.

4. Referencias

Referencias

[1] A. M. Law , W. D.Kelton. *Simulation, Modeling and Analysis* .

[2] Kisbye, Patricia. *Modelos y Simulación.* FAMAF,2023.