Simulación

Trabajo Práctico Nº2

Curso: 4k4

GRUPO B:

- 79420 – Donalisio Juan Pablo
- 80405 – Medina, Juan Cruz

Docentes:

- Castro, Sergio Horacio
- Pedroni, Juan Pablo
- Barale Lorena Natalia

Fecha de Presentación:

29 de agosto del 2021

Introducción

El presente archivo tiene como finalidad explicar el trabajo práctico número 2 de la materia Simulación que fue desarrollado en Excel.

Este trabajo consiste en tomar muestras, esto es, realizar la toma de datos de un sistema real y luego intentar encontrar la distribución a la cual se ajusta dicha muestra con lo aprendido en la cátedra de Simulación. Esto se logra luego de haber realizado una hipótesis acerca del comportamiento probabilístico de la muestra y posteriormente haber realizado dos pruebas de bondad (Ji cuadrado y Kolmogorov - Smirnov) que nos indican si la serie numérica que tenemos se ajusta o no a la distribución que planteamos en nuestra hipótesis.

Desarrollo

Se tomaron dos muestras para llevar a cabo el trabajo.
Comencemos por la primera muestra:

Muestra 1

Fuente: https://www.kaggle.com/mrisdal/open-exoplanet-catalogue

Esta muestra se basa en las masas de Júpiter (1,898 x 10^27 kg) de 219 exoplanetas elegidos al azar en nuestra galaxia.
Entonces si tenemos 219 valores, este es el ***tamaño de nuestra muestra 1*** (recibe el nombre de N en el archivo Excel presentado).
Ahora podremos definir:

- ***Media:*** como sumatoria de todos los valores dividido el tamaño de la muestra. Esto es:

Media	2,2932

- ***Cantidad de intervalos:*** para definir los mismos se respetó la sugerencia de, siendo k los intervalos del histograma, $k = \sqrt{n}$. Recordemos que este n hace referencia a nuestro N (es decir, el tamaño de la muestra).

Cantidad de intervalos	14

- ***Mínimo:*** el valor más pequeño de nuestra muestra, este es:

Mínimo	0,000338

- ***Máximo***: el valor más grande de nuestra muestra, este es:

Máximo	21

- ***Ancho***: representa el ancho de cada intervalo en el histograma. Este se calcula como:

  ((máximo - mínimo) / cantidad de intervalos) + una pequeña cifra

  Se suma esta pequeña cifra para incluir el extremo superior del intervalo, ya que si tenemos un valor de nuestra muestra que sea igual al extremo superior del último intervalo, este valor quedará excluido del histograma porque como sabemos, el extremo superior de cada intervalo es abierto.

Ancho	1,5

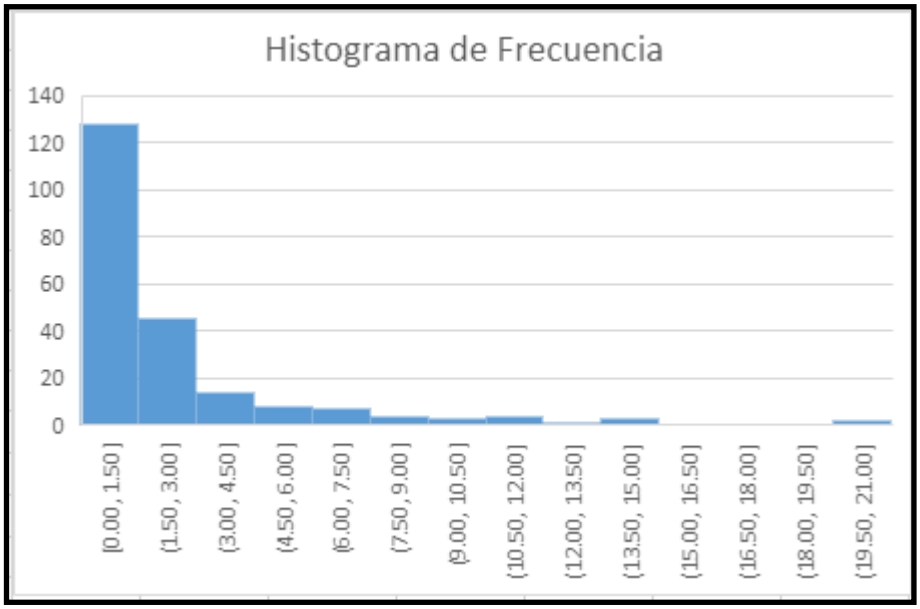
- ***Lambda***: este es un dato empírico, y se obtiene a través de la siguiente fórmula:

$$\lambda = \frac{1}{\mu}$$

  Siendo μ la media obtenida anteriormente.

Lambda	0,4361

Ahora bien, procedemos a graficar el histograma y resulta la siguiente gráfica:



Además, se obtienen las frecuencias observadas en cada intervalo (esto es, la cantidad de valores que me caen en cada intervalo).

Es en este punto, en el cual establecemos la hipótesis nula H0, es decir, ¿Qué distribución se ajusta a la forma obtenida en el histograma?

H0: Los valores tienen una distribución Exponencial Negativa
H1: Los valores no tienen una distribución Exponencial Negativa

Procedemos entonces a realizar el siguiente cuadro, donde podemos ver los intervalos definidos (desde-hasta), la frecuencia observada de cada intervalo y la frecuencia esperada de cada intervalo a partir de la función acumulada de la distribución estadística propuesta. Recordemos que la frecuencia esperada es aquella que me dice cuántos números se supone que deben caer en cada intervalo.

Y recordemos también las fórmulas correspondientes a la función acumulada de la exponencial negativa:

Acumulada	$F(x) = 1 - e^{-\lambda x}$

(esta es la columna **P()**)

Y la fórmula que corresponde a la frecuencia esperada:

P() * N

Desde	Hasta	fo	P()	fe
0	1.5	128	0.4801	105.1405
1.5	3.0	43	0.2496	54.6632
3	4.5	16	0.1298	28.4197
4.5	6.0	7	0.0675	14.7756
6	7.5	8	0.0351	7.6819
7.5	9.0	4	0.0182	3.9939
9	10.5	3	0.0095	2.0764
10.5	12.0	4	0.0049	1.0796
12	13.5	1	0.0026	0.5613
13.5	15.0	3	0.0013	0.2918
15	16.5	0	0.0007	0.1517
16.5	18.0	0	0.0004	0.0789
18	19.5	0	0.0002	0.0410
19.5	21.0	2	0.0001	0.0214
		219		218.9769

A partir de aquí, tenemos las dos pruebas:

Prueba de Chi-Cuadrado

Teniendo en cuenta que: “Las frecuencias esperadas para cada intervalo de clase deben ser de 5 o más. De no alcanzar esta cifra, se deberá agrupar clases o intervalos adyacentes.” procedemos a agrupar los intervalos adyacentes y sumar las frecuencias esperadas y observadas de los mismos.

Posteriormente, se calcula el estadístico de prueba:

$$c = \sum_{i=1}^{k} \frac{(fe_i - fo_i)^2}{fe_i}$$

Y luego se lo acumula.

Entonces tenemos:

Desde	Hasta	fo	fe	c	c(AC)		
0.0	1.5	128	105.1405	4.9701	4.9701		
1.5	3.0	43	54.6632	2.4885	7.4586		
3.0	4.5	16	28.4197	5.4276	12.8862		
4.5	6.0	7	14.7756	4.0919	16.9781		
6.0	7.5	8	7.6819	0.0132	16.9912		
7.5	10.5	7	6.0703	0.1424	17.1336		
10.5	21.0	10	2.2257	27.1561	44.2897	<-	Calculado
					11.1000	<-	Tabulado

Para obtener el valor tabulado (o valor crítico), se definen primero los grados de libertad:

Grados de libertad	5	<-	Intervalos (7) - Cantidad de datos empiricos (1) - 1	
	k - 1 - m		k = 7	m = 1

y luego se define el nivel de confianza: usamos **0,95**. Finalmente con estos dos valores podemos buscar el ***valor crítico o tabulado en la tabla, que es 11.1000***.

Como el estadístico (calculado) es mayor al valor crítico (tabulado) entonces decimos que se rechaza la hipótesis nula.

Prueba de Kolmogorov-Smirnov

Se procede a calcular la probabilidad observada en cada intervalo, siendo:

$Po = Fo / n$ (recordemos que n es el tamaño de la muestra) y luego se las acumula.

Además, se calcula la probabilidad esperada acumulada para cada intervalo a partir de la función acumulada de la exponencial negativa. Es decir, la columna P() = Pe().
Y acumulamos la Pe().

Ahora, procedemos a calcular el estadístico de prueba:

$$c = máx|PeA_i - PoA_i| \text{ con i=1,2,3,...,k}$$

Quedando nuestra tabla de la siguiente manera:

Prueba de Kolmogorov-Smirnov											
Desde	Hasta	fo	fe	Po()	Pe()	Po() AC	Pe() AC	|Po(AC)-Pe(AC)|	MAX		
0	1.5	128	105.1	0.584475	0.4801	0.584475	0.4801	0.104381482	0.104381		
1.5	3.0	43	54.7	0.196347	0.2496	0.780822	0.7297	0.051124896	0.104381		
3	4.5	16	28.4	0.073059	0.1298	0.853881	0.8595	0.005586255	0.104381		
4.5	6.0	7	14.8	0.031963	0.0675	0.885845	0.9269	0.0410913	0.104381		
6	7.5	8	7.7	0.03653	0.0351	0.922374	0.9620	0.03963893	0.104381		
7.5	9.0	4	4.0	0.018265	0.0182	0.940639	0.9803	0.039611008	0.104381		
9	10.5	3	2.1	0.013699	0.0095	0.954338	0.9897	0.035393867	0.104381		
10.5	12.0	4	1.1	0.018265	0.0049	0.972603	0.9947	0.022058514	0.104381		
12	13.5	1	0.6	0.004566	0.0026	0.977169	0.9972	0.020055175	0.104381		
13.5	15.0	3	0.3	0.013699	0.0013	0.990868	0.9986	0.007688998	0.104381		
15	16.5	0	0.2	0	0.0007	0.990868	0.9992	0.008381749	0.104381		
16.5	18.0	0	0.1	0	0.0004	0.990868	0.9996	0.008741915	0.104381		
18	19.5	0	0.0	0	0.0002	0.990868	0.9998	0.008929167	0.104381		
19.5	21.0	2	0.0	0.009132	0.0001	1	0.9999	0.000105395	0.104381	<-	Calculado
		219	219.0						0.0919	<-	Tabulado

Finalmente, se establece el nivel de confianza de 0.95, por lo que el nivel de significancia sería 0.05. Con este valor, junto con el tamaño de la muestra, procedemos a buscar en la tabla ***el valor crítico, el cual es 0.0919.***

Como el estadístico (calculado) es mayor al valor crítico (tabulado) entonces decimos que se rechaza la hipótesis nula.

Muestra 2

Fuente: *https://www.kaggle.com/neuromusic/avocado-prices*

Esta segunda muestra se basa en los precios en dólares de una unidad de palta (orgánica y convencional). Muestra tomada durante el año 2018 en Estados Unidos. (Hay que considerar que los precios se encuentran en dólares)
Tenemos 288 valores (N), este es el ***tamaño de nuestra muestra 2.***

Definiendo:

Cantidad de intervalos: Si bien se recomienda raíz de N (tamaño de la muestra), consideramos ideal para la visualización de los gráficos utilizar 9 intervalos.

Cantidad de valores	288
Cantidad de intervalos	9

Valor máximo y mínimo de la muestra: Esto se calculó mediante las funciones MAX() y MIN() que provee el Excel, y el objetivo de su cálculo es poder establecer los límites superior e inferior del total de intervalos.

Mínimo	0.56
Máximo	2.06

Ancho:

Representa el ancho de cada intervalo en el histograma. Este se calcula como:

((máximo - mínimo) / cantidad de intervalos) + una pequeña cifra

Se suma esta pequeña cifra para incluir el extremo superior del intervalo, ya que si tenemos un valor de nuestra muestra que sea igual al extremo superior del último intervalo, este valor quedará excluido del histograma porque como sabemos, el extremo superior de cada intervalo es abierto.

Ancho	0.17

Media: Representa el promedio, o sea, la sumatoria de todas las muestras dividido la cantidad de muestras. Esto se realizó con la función de Excel SUMA() y luego una simple división.

Media	1.3479

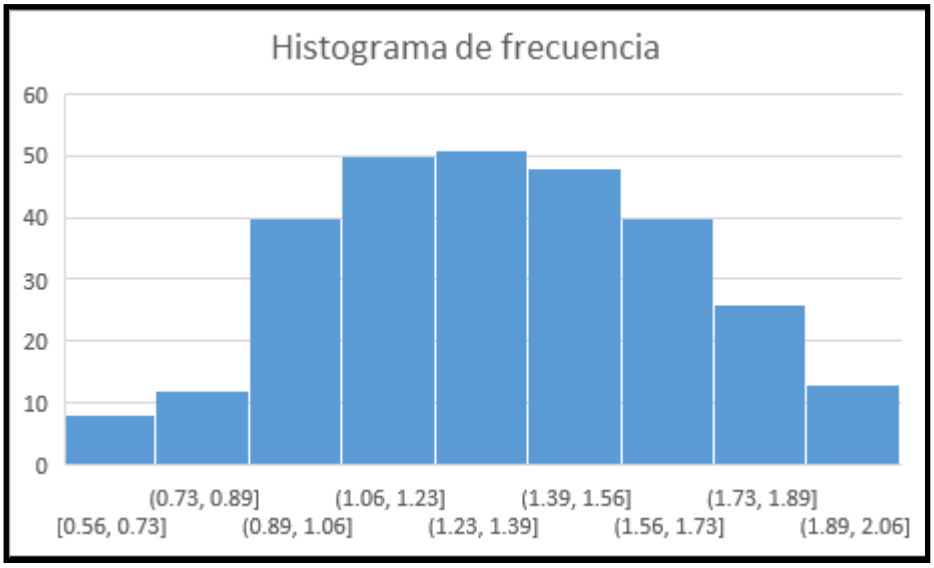
Desviación: Viene dada por la siguiente ecuación

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\mu)^2$$

Pero, al ser la anterior la fórmula de Varianza para una distribución Normal, despejamos para calcular la desviación pasando el término al cuadrado hacia el otro lado de la ecuación como la raíz cuadrada.
Quedando entonces, la raíz cuadrada de la media de los cuadrados de las puntuaciones de **desviación, en este caso:**

Desviación	0.3195

Ahora bien, procedemos a graficar el histograma y resulta la siguiente gráfica:



Obteniendo las frecuencias observadas en cada intervalo (la cantidad de valores que pertenecen a cada intervalo), se puede observar en esta imagen la semejanza a una distribución normal y considerando visualmente que la media está entre el intervalo 1.23-1.39, y siendo la media calculada=1.34 podemos plantear una primer hipótesis o hipótesis nula (HO):

H0: Los valores tienen una distribución Normal

Y a su vez planteando una hipótesis alternativa o (H1):

H1: Los valores no tienen una distribución Normal

Para poder realizar las pruebas de bondad que nos rechacen o no la hipótesis planteada, primero es necesario obtener las frecuencias esperadas. Para el caso de la distribución normal no tenemos una función de acumulación que nos de la frecuencia esperada exacta, es por esto que nos aproximamos a la misma utilizando la marca de clase, que se obtiene haciendo el promedio entre los extremos de cada intervalo.

Desde	Hasta	Marca
0.56	0.73	0.64
0.73	0.89	0.81
0.89	1.06	0.98
1.06	1.23	1.14
1.23	1.39	1.31
1.39	1.56	1.48
1.56	1.73	1.64
1.73	1.89	1.81
1.89	2.06	1.98

Teniendo en cuenta esta marca de clase, la ecuación de la función de densidad es:

$$f(x) = \frac{1}{\sigma.\sqrt{2.\pi}}.e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Entonces calculamos las probabilidades y multiplicamos cada una por el tamaño de la muestra (N) para obtener las frecuencias esperadas

Desde	Hasta	Marca	fo	P()	fe
0.56	0.73	0.64	8	0.02	5.27
0.73	0.89	0.81	12	0.05	14.55
0.89	1.06	0.98	40	0.11	30.56
1.06	1.23	1.14	50	0.17	48.89
1.23	1.39	1.31	51	0.21	59.56
1.39	1.56	1.48	48	0.19	55.25
1.56	1.73	1.64	40	0.14	39.03
1.73	1.89	1.81	26	0.07	21.00
1.89	2.06	1.98	13	0.03	8.60

Prueba de Chi-Cuadrado

Una vez conseguidas las frecuencias esperadas, podemos proseguir a efectuar la prueba de chi cuadrado. No fue necesario agrupamiento de intervalos debido a que todas las frecuencias esperadas son mayores a 5. El valor calculado de esta prueba viene dado por la siguiente ecuación:

$$\lambda^2 = \sum_{1}^{k} \frac{(f_0 - f_e)^2}{f_e}$$

Entonces lo que tenemos que hacer es calcular $\lambda^2 = C(AC)$, y para ello calculamos cada $\frac{(fo-fe)^2}{fe}$ individual en una columna y los vamos acumulando en otra. El último valor acumulado corresponde con el valor calculado del test.

Desde	Hasta	fo	fe	c	c(AC)		
0.56	0.73	8.00	5.27	1.4085	1.4085		
0.73	0.89	12.00	14.55	0.4467	1.8552		
0.89	1.06	40.00	30.56	2.9135	4.7687		
1.06	1.23	50.00	48.89	0.0251	4.7938		
1.23	1.39	51.00	59.56	1.2304	6.0242		
1.39	1.56	48.00	55.25	0.9521	6.9763		
1.56	1.73	40.00	39.03	0.0239	7.0002		
1.73	1.89	26.00	21.00	1.1911	8.1913		
1.89	2.06	13.00	8.60	2.2477	10.4391	<-	Calculado
					12.6	<-	Tabulado

Por otro lado, y como se ve en la imagen, el valor tabulado de chi cuadrado es 12.6 (se utilizaron 6 grados de libertad, debido a que vienen dados por k - 1 - m, donde k son los 9 intervalos y m son los 2 datos empíricos que se utilizaron. También usamos un grado de confianza del 0.95). Entonces, al ser el valor calculado menor al tabulado, **no se puede rechazar la hipótesis nula**.

Prueba de Kolmogorov-Smirnov

Finalmente, para efectuar la prueba de Ks, sabemos que el valor calculado viene dado por:

$$KS = \max\left(|P(f_o)_{AC} - P(f_e)_{AC}|\right)$$

Entonces, debemos obtener las probabilidades esperadas (que ya las tenemos) y las observadas (que se obtienen haciendo Fo / N), acumularlas, calcular los valores absolutos de las diferencias de estas y obtener el máximo de ellos. Este valor máximo será representativo del valor calculado de Ks

Desde	Hasta	fo	fe	Po()	Pe()	Po() AC	Pe() AC	| Po(AC) - Pe(AC) |	MAX		
0.56	0.73	8.00	5.27	0.0278	0.02	0.0278	0.02	0.0095	0.0095		
0.73	0.89	12.00	14.55	0.0417	0.05	0.0694	0.07	0.0006	0.0095		
0.89	1.06	40.00	30.56	0.1389	0.11	0.2083	0.17	0.0334	0.0334		
1.06	1.23	50.00	48.89	0.1736	0.17	0.3819	0.34	0.0372	0.0372		
1.23	1.39	51.00	59.56	0.1771	0.21	0.5590	0.55	0.0075	0.0372		
1.39	1.56	48.00	55.25	0.1667	0.19	0.7257	0.74	0.0177	0.0372		
1.56	1.73	40.00	39.03	0.1389	0.14	0.8646	0.88	0.0143	0.0372		
1.73	1.89	26.00	21.00	0.0903	0.07	0.9549	0.95	0.0030	0.0372		
1.89	2.06	13.00	8.60	0.0451	0.03	1.0000	0.98	0.0183	0.0372	<-	Calculado
									0.0801	<-	Tabulado

El valor tabulado de Ks, con un nivel de confianza del 0.95 y 288 grados de libertad (tamaño de la muestra), se obtiene de la tabla y nos dice que es igual a $1.36 \,/\, \sqrt{288} = 0.0801$. Vemos que el valor tabulado es mayor al calculado, por ende, y en concordancia con la prueba de Chi Cuadrado, no se puede rechazar la hipótesis nula.