UNIVERSIDAD TECNOLÓGICA NACIONAL

FACULTAD REGIONAL CÓRDOBA

Trabajo Práctico N°2:

Toma de datos

Cátedra: Simulación

Comisión: 4K1

Grupo G - Integrantes:

- Cargnelutti, Clever Lautaro 82376
- Cañete Julio, Federico Pedro 83184
- Farace, Florencia Candelaria 82043
- Luque, Mariano Nicolás 84777
- Nóbile Cudós, Valentina Celeste 82186
- Torres, Ezequiel Daniel 83205

Docentes:

- Sanchez, Daniel Mario
- Berrotaran, Juan Jose
- Carena, Gonzalo Ezequiel

Fecha de Presentación:14/04/2022

Introducción

El presente trabajo tiene por objetivo analizar muestras tomadas desde distintas fuentes, en este caso serán muestras de la duración de partidas en el juego “League of Legends” y de los ingresos de los deportistas mejor pagos en en el mundo según Forbes. A partir de estas muestras planteamos una hipótesis para cada una de ellas, en las cuales planteamos que la primera muestra tiene una distribución normal y la segunda una distribución exponencial negativa. Con los métodos de Chi Cuadrado y Kolmogorov - Smirnov se puede determinar la veracidad de las hipótesis planteadas con anterioridad, determinando si estas pueden ser rechazadas o no.

Desarrollo

MUESTRA 1

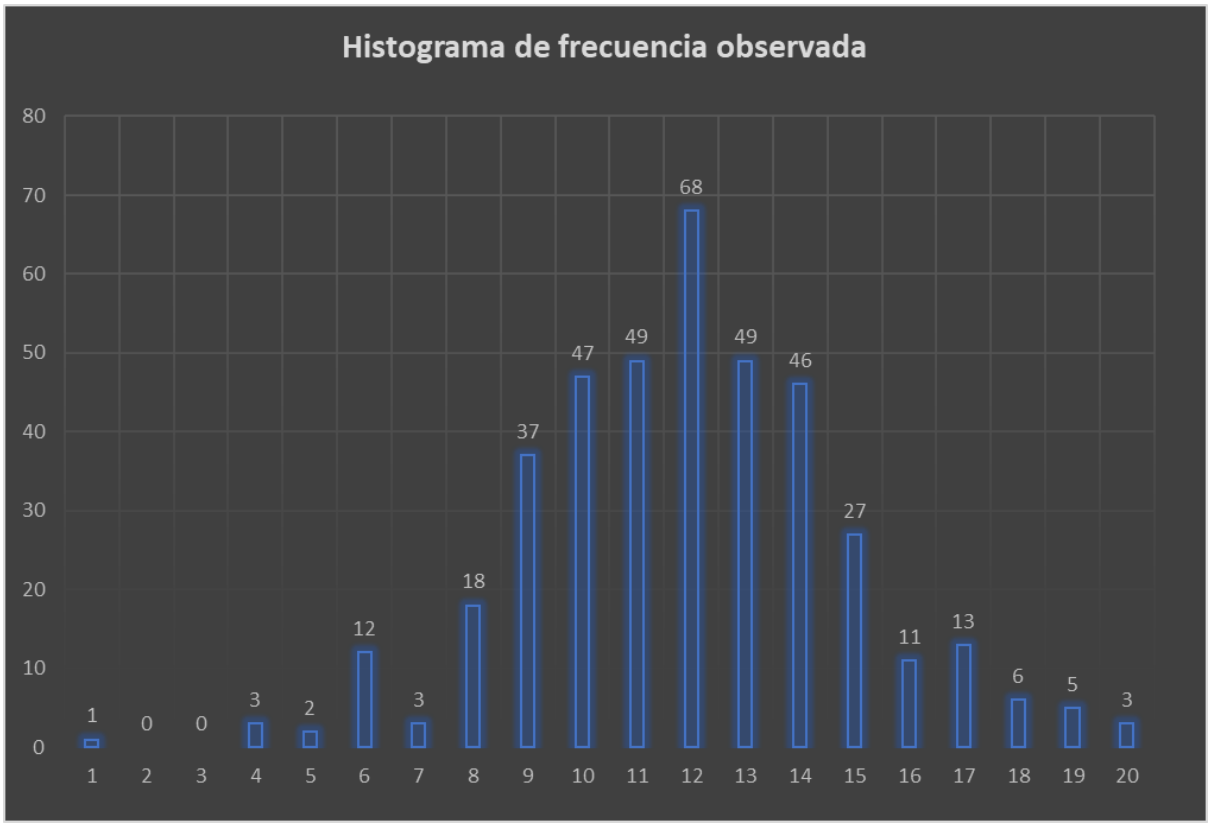
Como primera muestra tomamos la duración en minutos de 400 partidas de League of Legends en la liga de oro, un videojuego muy popular entre los jóvenes.

Fuente:
https://www.kaggle.com/datasets/gbolduc/league-of-legends-lol-gold-ranked-games?select=matches_meta_4_4.csv

Entonces el tamaño de la muestra 1 es N=400. Con esto definimos:

Cantidad de Intervalos (k)	Para calcular la cantidad de intervalos k, utilizamos $k = \sqrt{N}$ lo que resulta una cantidad de intervalos $k = 20$
Mínimo (min)	El valor más pequeño de la muestra, en este caso 3,3
Máximo (max)	El valor más grande de la muestra, en este caso 47,58
Ancho	El ancho de cada intervalo se calcula como $(max - min)/k + 0,01$ que nos dá igual a $2,22$ de ancho. El $0,01$ se suma a fin de incorporar los números que caen sobre el extremo superior del intervalo que quedarían excluidos.
Media ($\bar{x}$)	Calculada como la suma de todos los valores dividido por el tamaño de la muestra, lo que nos da igual 28,68
Desviación (σ)	Se calcula como la raíz cuadrada de la varianza, siendo la varianza $\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\mu)^2$ Esto es igual a una desviación $\sigma = 6,54$

Graficamos el histograma de frecuencias de la muestra y obtenemos:



Esta distribución de frecuencias visualmente se asemeja a una distribución normal. Además, la media calculada es 28,68 y observando el histograma podemos ver que la media se encuentra en el intervalo 12, intervalo que va desde 27,77 a 29,98.

Dadas estas condiciones, podemos plantear la hipótesis: los valores tienen una distribución normal. Para saber si esta hipótesis es rechazada o no, se ejecutan las pruebas de bondad de ajuste Chi-Cuadrado y Kolmogorov-Smirnov.

Para ello, se obtendrá en primer lugar la Frecuencia esperada. Al trabajar con la Distribución Normal, debemos obtener una Marca de Clase que se aproxima a la Frecuencia Esperada y se obtiene haciendo el promedio de los extremos de cada intervalo.

Desde	Hasta	Marca Clase
3,30	5,51	4,41
5,52	7,74	6,63
7,75	9,96	8,86
9,97	12,19	11,08
12,20	14,41	13,30
14,42	16,64	15,53
16,65	18,86	17,75
18,87	21,08	19,98
21,09	23,31	22,20
23,32	25,53	24,42
25,54	27,76	26,65
27,77	29,98	28,87
29,99	32,20	31,10
32,21	34,43	33,32
34,44	36,65	35,55
36,66	38,88	37,77
38,89	41,10	39,99
41,11	43,33	42,22
43,34	45,55	44,44
45,56	47,77	46,67

Una vez obtenida la Marca de Clase de cada uno de los intervalos, se utilizará la ecuación de la función de densidad para obtener la probabilidad en dicho punto::

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

Siendo *X* la marca de clase. Finalmente multiplicamos el valor obtenido para obtener la probabilidad en el intervalo completo. Dicha probabilidad multiplicada por el tamaño de muestra nos permite conocer la frecuencia esperada para dicho intervalo.

Calculo de probabilidades y frecuencias					
Desde	**Hasta**	**Marca Clase**	**Fo**	**P()**	**fe**
3,30	5,51	4,41	1	0,00	0,06
5,52	7,74	6,63	0	0,00	0,18
7,75	9,96	8,86	0	0,00	0,55
9,97	12,19	11,08	3	0,00	1,45
12,20	14,41	13,30	2	0,01	3,41
14,42	16,64	15,53	12	0,02	7,16
16,65	18,86	17,75	3	0,03	13,38
18,87	21,08	19,98	18	0,06	22,29
21,09	23,31	22,20	37	0,08	33,07
23,32	25,53	24,42	47	0,11	43,71
25,54	27,76	26,65	49	0,13	51,47
27,77	29,98	28,87	68	0,13	53,99
29,99	32,20	31,10	49	0,13	50,45
32,21	34,43	33,32	46	0,10	41,99
34,44	36,65	35,55	27	0,08	31,14
36,66	38,88	37,77	11	0,05	20,57
38,89	41,10	39,99	13	0,03	12,11
41,11	43,33	42,22	6	0,02	6,35
43,34	45,55	44,44	5	0,01	2,96
45,56	47,77	46,67	3	0,00	1,23

Una vez realizado esto, seguimos por el **cálculo de Chi Cuadrado**. El valor calculado de esta prueba viene dado por la siguiente ecuación:

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

Para hacer la prueba de Chi-Cuadrado debemos reagrupar los intervalos adyacentes y sumar las frecuencias de los mismos de ser necesario, ya que para cada intervalo la frecuencia esperada debe ser de mínimo 5.

Para calcular λ^2 realizamos el cálculo de $(f_0 - f_e)^2/f_e$ para cada intervalo, en la columna *c* y en la columna $C_{(AC)}$ acumulamos los valores de la columna c. El último valor de la columna $C_{(AC)}$ es la acumulación de todos los c calculados, que corresponde a λ^2 (chi calculado), en este caso 23,27.

Prueba de Ji-Cuadrada					
Desde	Hasta	fo	fe	c	C(ac)
3,30	14,41	6,00	5,65	0,02	0,02
14,42	16,64	12,00	7,16	3,28	3,30
16,65	18,86	3,00	13,38	8,06	11,35
18,87	21,08	18,00	22,29	0,83	12,18
21,09	23,31	37,00	33,07	0,47	12,65
23,32	25,53	47,00	43,71	0,25	12,89
25,54	27,76	49,00	51,47	0,12	13,01
27,77	29,98	68,00	53,99	3,64	16,65
29,99	32,20	49,00	50,45	0,04	16,69
32,21	34,43	46,00	41,99	0,38	17,07
34,44	36,65	27,00	31,14	0,55	17,62
36,66	38,88	11,00	20,57	4,45	22,08
38,89	41,10	13,00	12,11	0,07	22,14
41,11	47,77	14,00	10,54	1,13	23,27

Este valor se compara con el chi tabulado correspondiente a la fila v columna de nivel de significancia 0,05 de la tabla de valores percentiles para la distribución Chi-Cuadrado. Siendo $v = k - 1 - m = 14 - 1 - 2 = 11$, la celda correspondiente indica un chi tabulado de 19,7 (*m* corresponde a la cantidad de datos empíricos, en este caso la desviación estándar y la media).

Si el valor de chi calculado es menor o igual al chi tabulado, entonces no se rechaza la hipótesis. En este caso el valor de chi calculado (23,27) es mayor a chi tabulado (19,7), por ende la hipótesis se rechaza. Con el nivel de significancia elegido se rechaza la hipótesis pero con un nivel de significancia de 0,99 no se puede rechazar la misma debido a que el valor tabulado (24,70) es mayor al calculado.

Grados de libertad= 14-1-2 = 11		
Nivel de confianza = 0,95	Valor tabulado =	19,70
Nivel de confianza = 0,99	Valor tabulado =	24,70
	Valor calculado =	23,27
Tomando un nivel de confianza de 0,95 la hipótesis es rechazada. Tomando un nivel 0,99 la hipótesis no puede ser rechazada		

Para la **prueba de Kolmogorov-Smirnov** sabemos que el valor calculado está dado por:

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

Ya obtuvimos las probabilidades esperadas (columna $Pe()$) y ahora nos queda calcular las probabilidades observadas (columna $Po()$), que se calculan como frecuencia observada dividido tamaño de muestra F_o/N. Una vez obtenidas, las acumulamos en sus columnas

correspondientes $Pe()\,AC$ y $Po()\,AC$, y calculamos el valor absoluto de la diferencia de ambas en la columna $|Po()\,AC - Pe()\,AC|$. Por último, buscamos el valor máximo de esa columna (en la columna MAX) y ese valor será nuestro valor calculado, en este caso 0,03.

Po()	Pe()	Po() AC	Pe() AC	|Po(AC)-Pe(AC)|	MAX
0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00
0,01	0,00	0,01	0,01	0,00	0,00
0,01	0,01	0,02	0,01	0,00	0,00
0,03	0,02	0,05	0,03	0,01	0,01
0,01	0,03	0,05	0,07	0,01	0,01
0,05	0,06	0,10	0,12	0,02	0,02
0,09	0,08	0,19	0,20	0,01	0,02
0,12	0,11	0,31	0,31	0,01	0,02
0,12	0,13	0,43	0,44	0,01	0,02
0,17	0,13	0,60	0,58	0,02	0,02
0,12	0,13	0,72	0,70	0,02	0,02
0,12	0,10	0,84	0,81	0,03	0,03
0,07	0,08	0,91	0,89	0,02	0,03
0,03	0,05	0,93	0,94	0,00	0,03
0,03	0,03	0,97	0,97	0,00	0,03
0,02	0,02	0,98	0,98	0,00	0,03
0,01	0,01	0,99	0,99	0,00	0,03
0,01	0,00	1,00	0,99	0,01	0,03

Para encontrar el valor tabulado, nos dirigimos a la tabla de ks y buscamos la columna de nivel de significancia 0,05 y los grados de libertad en este caso serán el tamaño de muestra: 400. Entonces el valor tabulado es $1,36/\sqrt{n} = 1,36 / \sqrt{400} = 1,36/20 = 0,068$

El valor calculado (0,03) es menor al valor tabulado (0,068), por lo que la hipótesis no puede ser rechazada.

MUESTRA 2

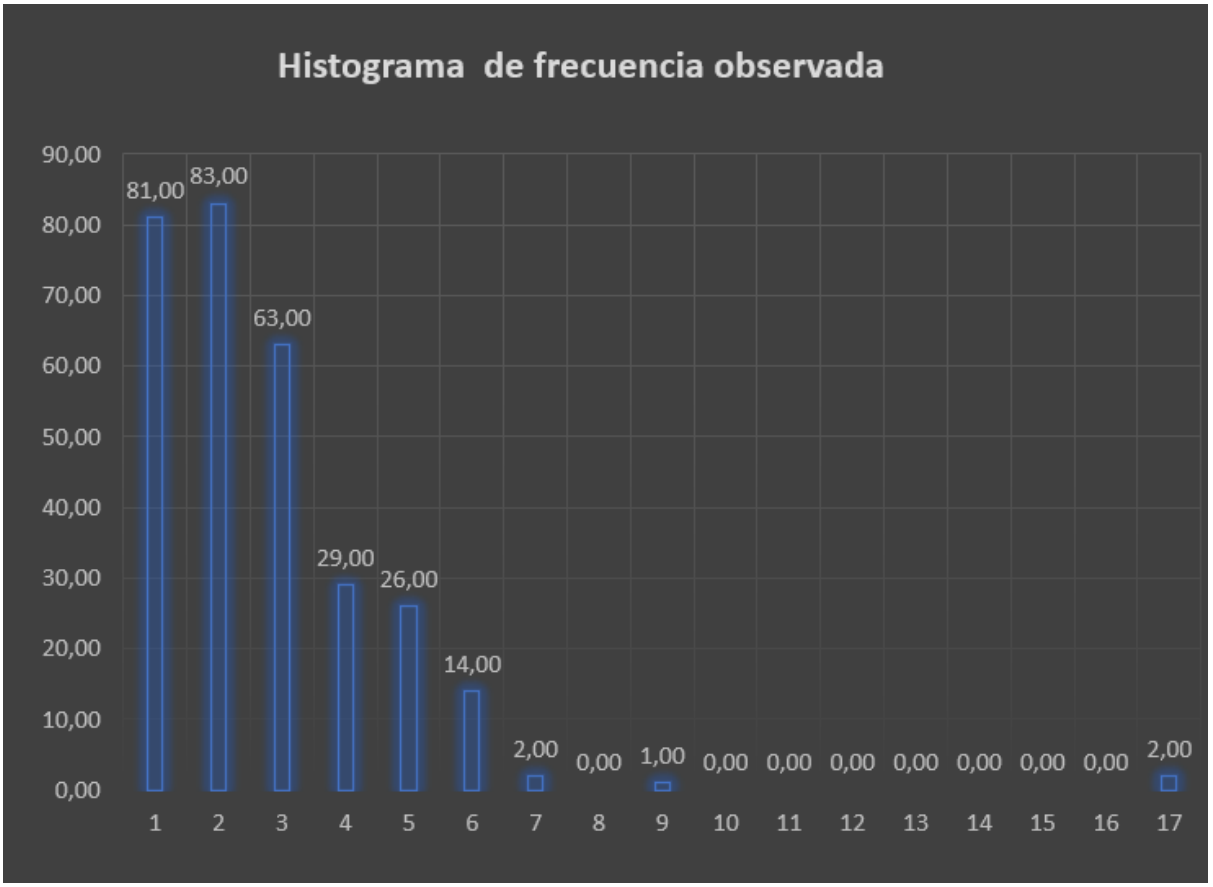
Como segunda muestra tomamos los ingresos en millones de dólares de los atletas mejor pagados entre 1990 y 2020.

Fuente:
https://www.kaggle.com/datasets/parulpandey/forbes-highest-paid-athletes-19902019

A continuación los valores calculados en base a la muestra de 301 valores.

Media ($\bar{x}$)	Calculada como la suma de todos los valores dividido por el tamaño de la muestra, lo que nos da igual a 45,516
Cantidad de Intervalos (k)	Para calcular la cantidad de intervalos k, utilizamos $k = \sqrt{N}$ lo que resulta una cantidad de intervalos $k = 17$
Mínimo (min)	El valor más pequeño de la muestra, en este caso 8,1
Máximo (max)	El valor más grande de la muestra, en este caso 300
Ancho	El ancho de cada intervalo se calcula como $(max - min)/k + 0,01$ que nos dá igual a $17,18$ de ancho. El $0,01$ se suma a fin de incorporar los números que caen sobre el extremo superior del intervalo que quedarían excluidos
Lambda (λ)	Se calcula como la inversa de la media $(\lambda = \frac{1}{\mu})$ y es el dato empírico que luego nos permitirá calcular los grados de libertad. En este caso es igual a 0,022

Graficamos el histograma de frecuencias de la muestra y obtenemos:



La forma obtenida en el histograma se asemeja a una distribución exponencial negativa, por lo que establecemos la Hipótesis Nula (H0) : Los valores tienen una distribución exponencial negativa.

Luego procedemos a realizar las siguiente tabla, definiendo los intervalos, la frecuencia observada en ellos y calculando la frecuencia esperada a partir de la función acumulada de la distribución exponencial negativa: $F(x) = 1 - e^{-\lambda.x}$. Al darnos el valor de la densidad desde el origen hasta el punto *x*, debemos calcular la densidad tanto para el extremo superior como para el inferior y luego restarlos obteniendo la probabilidad requerida del intervalo en cuestión. A la función acumulada la simbolizamos como $P() \; c/ \; Pac$ en la tabla.

La frecuencia esperada se calcula como el valor de $P() \; c/ \; Pac$ multiplicado por el tamaño de la muestra (N).

Calculos de probabilidades y frecuencias								
Desde	Hasta	Marca Clase	Fo	P() c/mc	P() c/Pac	fe	Fr	FR Ac
8,10	25,18	16,64	81,00	0,26	0,26	78,83	0,27	0,27
25,28	42,36	33,82	83,00	0,18	0,18	54,04	0,28	0,54
42,46	59,54	51,00	63,00	0,12	0,12	37,05	0,21	0,75
59,64	76,72	68,18	29,00	0,08	0,08	25,40	0,10	0,85
76,82	93,90	85,36	26,00	0,06	0,06	17,42	0,09	0,94
94,00	111,08	102,54	14,00	0,04	0,04	11,94	0,05	0,98
111,18	128,26	119,72	2,00	0,03	0,03	8,19	0,01	0,99
128,36	145,44	136,90	0,00	0,02	0,02	5,61	0,00	0,99
145,54	162,63	154,09	1,00	0,01	0,01	3,85	0,00	0,99
162,73	179,81	171,27	0,00	0,01	0,01	2,64	0,00	0,99
179,91	196,99	188,45	0,00	0,01	0,01	1,81	0,00	0,99
197,09	214,17	205,63	0,00	0,00	0,00	1,24	0,00	0,99
214,27	231,35	222,81	0,00	0,00	0,00	0,85	0,00	0,99
231,45	248,53	239,99	0,00	0,00	0,00	0,58	0,00	0,99
248,63	265,71	257,17	0,00	0,00	0,00	0,40	0,00	0,99
265,81	282,89	274,35	0,00	0,00	0,00	0,27	0,00	0,99
282,99	300,07	291,53	2,00	0,00	0,00	0,19	0,01	1,00

Para hacer la prueba de Chi-Cuadrado debemos reagrupar los intervalos adyacentes y sumar las frecuencias de los mismos de ser necesario, ya que para cada intervalo la frecuencia esperada debe ser de mínimo 5.

Procedemos a calcular el estadístico de prueba c y el estadístico de prueba acumulado $c(AC)$ a través de la siguiente fórmula:

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

Luego de obtener el valor calculado debemos compararlo con el valor tabulado. Para este último necesitamos determinar los grados de libertad, que se calculan como el número de intervalos menos la cantidad de datos empíricos (en este caso Lambda) menos uno. Además, debemos determinar el nivel de confianza a utilizar. Con estos dos valores buscamos el valor tabulado en la tabla de valores percentiles de la distribución Chi Cuadrado.

Prueba de Ji-Cuadrada					
Desde	Hasta	fo	fe	c	C(ac)
8,10	25,18	81,00	78,83	0,06	0,06
25,28	42,36	83,00	54,04	15,51	15,57
42,46	59,54	63,00	37,05	18,17	33,75
59,64	76,72	29,00	25,40	0,51	34,25
76,82	93,90	26,00	17,42	4,23	38,49
94,00	111,08	14,00	11,94	0,36	38,84
111,18	128,26	2,00	8,19	4,68	43,52
128,36	145,44	0,00	5,61	5,61	49,13
145,54	179,81	1,00	6,49	4,64	53,77
179,91	300,07	2,00	5,34	2,09	55,86

Grados de libertad= 10-1-1 = 8		
Nivel de confianza = 0,95	Valor tabulado =	15,50
Nivel de confianza = 0,99	Valor tabulado =	20,10
	Valor calculado =	55,86
Tomando un nivel de confianza de 0,95 o un nivel 0,99 la hipótesis es rechazada		

De acuerdo al nivel de significancia y a los grados de libertad que tomemos, que en nuestro caso es 0,05 y v, respectivamente, siendo $v = k - 1 - m = 10 - 1 - 2 = 7$ vamos a poder seleccionar el valor de tabulado que es 15,50. Tanto con un nivel de significancia de 0,95 como con un nivel de 0,99 la hipótesis se rechaza, ya que el calor calculado (55,86) es mayor que los valores tabulados.

Para la **prueba de Kolmogorov-Smirnov** sabemos que el valor calculado está dado por

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

Se deben obtener las probabilidades esperadas y las observadas (que se obtienen haciendo Fo / N), acumularlas y calcular los valores absolutos de las diferencias y obtener el máximo de estos valores. Este valor máximo será representativo del valor calculado de Ks

Ya obtuvimos las probabilidades esperadas (columna $Pe()$) y ahora nos queda calcular las probabilidades observadas (columna $Po()$), que se calculan como frecuencia observada dividido tamaño de muestra F_o/N. Una vez obtenidas, las acumulamos en sus columnas correspondientes $Pe()\ AC$ y $Po()\ AC$, y calculamos el valor absoluto de la diferencia de ambas en la columna $|Po()\ AC - Pe()\ AC|$. Por último, buscamos el valor máximo de esa columna (en la columna MAX) y ese valor será nuestro valor calculado, en este caso 0.24

Prueba de Kolmogorov-Smirnov									
Desde	Hasta	Fo	Fe	Po()	Pe()	Po() AC	Pe() AC	|Po(AC)-Pe(AC)|	MAX
8,10	25,18	81,00	78,83	0,27	0,26	0,27	0,26	0,01	0,01
25,28	42,36	83,00	54,04	0,28	0,18	0,54	0,44	0,10	0,10
42,46	59,54	63,00	37,05	0,21	0,12	0,75	0,56	0,19	0,19
59,64	76,72	29,00	25,40	0,10	0,08	0,85	0,65	0,20	0,20
76,82	93,90	26,00	17,42	0,09	0,06	0,94	0,71	0,23	0,23
94,00	111,08	14,00	11,94	0,05	0,04	0,98	0,75	0,24	0,24
111,18	128,26	2,00	8,19	0,01	0,03	0,99	0,77	0,22	0,24
128,36	145,44	0,00	5,61	0,00	0,02	0,99	0,79	0,20	0,24
145,54	162,63	1,00	3,85	0,00	0,01	0,99	0,81	0,19	0,24
162,73	179,81	0,00	2,64	0,00	0,01	0,99	0,81	0,18	0,24
179,91	196,99	0,00	1,81	0,00	0,01	0,99	0,82	0,17	0,24
197,09	214,17	0,00	1,24	0,00	0,00	0,99	0,82	0,17	0,24
214,27	231,35	0,00	0,85	0,00	0,00	0,99	0,83	0,17	0,24
231,45	248,53	0,00	0,58	0,00	0,00	0,99	0,83	0,16	0,24
248,63	265,71	0,00	0,40	0,00	0,00	0,99	0,83	0,16	0,24
265,81	282,89	0,00	0,27	0,00	0,00	0,99	0,83	0,16	0,24
282,99	300,07	2,00	0,19	0,01	0,00	1,00	0,83	0,17	0,24

Tomamos dos niveles de confianza, cuando el nivel es de 0.99 es rechazada y con un nivel de confianza del 0.95 la hipótesis también es rechazada.

Grados de libertad = N = 301		
Nivel de confianza = 0,95	Valor tabulado =	0,08
Nivel de confianza = 0,99	Valor tabulado =	0,09
	Valor calculado =	0,24
Tomando un nivel de confianza de 0,95 o un nivel 0,99 la hipótesis es recahzada		