PRUEBA DE LA PROBABILIDAD EXACTA DE FISHER

El test exacto de Fisher permite analizar si dos variables están asociadas cuando la muestra a estudiar es demasiado pequeña.

Se utiliza cuando los valores esperados son menores a 5 y $n \leq 100$.

Esta prueba consiste en fijar las frecuencias marginales y calcular la probabilidad de ocurrencia de las frecuencias de las celdas asumiendo independencia entre las variables.

Se consideran las configuraciones más extremas que pueden ocurrir en los datos y se calculan las probabilidades exactas p_i para cada tabla escogiendo la celda con el menor de los valores.

El test exacto de Fisher se basa en evaluar la probabilidad asociada a cada una de las tablas 2 x 2 que se pueden formar manteniendo los mismos totales de filas y columnas que los de la tabla observada.

Hipótesis a contrastar:

H_0: Las variables son independientes, no están asociadas

H_1:Las variables no son independientes, están asociadas

Nivel de significación :α

Cálculo del p-valor asociado al estadístico de prueba

Para calcular el estadístico de contraste, se construye en primer lugar la tabla de contingencia de dimensiones 2 x 2 con las frecuencias absolutas observadas, con la notación siguiente:

Tabla de contingencia general para la comparación de dos variables dicotómicas en el caso de grupos independientes.

	Característica A		
Característica B	Presente	Ausente	Total
Presente	a	b	a + b
Ausente	c	d	c + d
Total	a + c	b + d	n

Tabla de contingencia general para la comparación de dos variables dicotómicas en el caso de grupos independientes.

	Característica A		
Característica B	Presente	Ausente	Total
Presente	a	b	a + b
Ausente	c	d	c + d
Total	a + c	b + d	n

Tabla de contingencia general para la comparación de dos variables dicotómicas en el caso de grupos independientes.

	Característica A		
Característica B	Presente	Ausente	Total
Presente	a’	b’	a + b
Ausente	c’	d’	c + d
Total	a + c	b + d	n

Si asumimos que “a” es el menor de los valores de la celda entonces:

se construyen todas las tablas de contingencia 2 x 2 posibles con celdas a’, b’, c’, d’, donde:

$0 \leq a' \leq \text{mín}[(a+c),(a+b)]$:

$b' = (a+b) - a'$,
$c' = (a+c) - a'$ y
$d' = (c+d) - c'$.

A partir de dichas tablas se calcula las probabilidades asociadas a cada una de ellas de la siguiente forma:

$$p_{a'} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a'!b'!c'!d'!}$$

El p-valor bilateral resultante es p= $\sum_{p_{a'} \leq p_a} p_{a'}$

p_a=probabilidad de la tabla con los datos observados (tabla original)

a= valor de la casilla “a” en la tabla original.

Ejemplo para la generación de las tablas a partir de la tabla (original) y las probabilidades asociadas a cada tabla

	F1	F2	
C1	2	3	5
C2	16	21	37
	18	24	42

Solución:

Construiremos tablas del ejemplo considerando: $0 \leq a' \leq 5$ (mín[(a+c),(a+b)])

1º Calculamos la tabla para a´=0

	F1	F2	
C1	0	5	5
C2	18	19	37
	18	24	42

$$p_{a_0'} = \frac{5!37!18!24!}{42!0!5!18!19!} = 0{,}049$$

2º Calculamos la tabla para a´=1

	F1	F2	
C1	1	4	5
C2	17	20	37
	18	24	42

entonces

$$p_{a_1'} = \frac{5!37!18!24!}{42!1!4!17!20!} = 0,224$$

3º Calculamos la tabla para a´=2

	F1	F2	
C1	2	3	5
C2	16	21	37
	18	24	42

entonces

$$p_{a_2'} = \frac{5!37!18!24!}{42!2!3!16!21!} = 0,364$$

4º Calculamos la tabla para a´=3

	F1	F2	
C1	3	2	5
C2	15	22	37
	18	24	42

Entonces

$$p_{a_3'} = \frac{5!37!18!24!}{42!3!2!15!22!} = 0,264$$

5º Calculamos la tabla para a´=4

	F1	F2	
C1	4	1	5
C2	14	23	37
	18	24	42

entonces

$$p_{a_4'} = \frac{5!37!18!24!}{42!4!1!14!23!} = 0,086$$

6º Calculamos la tabla para a´=5

	F1	F2	
C1	5	0	5
C2	13	24	37
	18	24	42

Entonces

$$p_{a_5'} = \frac{5!37!18!24!}{42!5!0!13!24!} = 0{,}01$$

Los valores de "p" para cada a'

a'	$p_{a'}$	
0	0.049	
1	0.224	
2	0.364	p_a
3	0.264	
4	0.086	
5	0.01	

El valor p bilateral es p= $\sum_{p_{a'} \le p_a} p_{a'} = 0{,}049 + 0{,}224 + 0{,}364 + 0.264 + 0.086 + 0.01 = 0.997$

p = probabilidad asociada con la ocurrencia de H_0

Regla de Decisión:

- Prueba Bilateral:
  Se rechaza H_0, si :

$$p \leq \alpha/2$$

p = probabilidad asociada con la ocurrencia de H_0

Aplicativo:

En una determinada población se desea identificar si la obesidad de las personas están relacionadas con su sexo (hombres y mujeres). Tras ser observada una muestra de 14 personas se obtuvieron los resultados que se muestran en la siguiente tabla:

Resultados de la información recogida.			
	Obesidad		
Sexo	Sí	No	**Total**
Mujeres	1 (a)	4 (b)	5 (a+b)
Hombres	7 (c)	2 (d)	9 (c+d)
Total	8 (a+c)	6 (b+d)	14 (n)

Planteamos las hipótesis:

H_0: las variables sexo y obesidad son independientes

H_1: las variables sexo y obesidad no son independientes

Establecemos un α=0.05

Calculamos el estadístico de prueba

Posibles combinaciones de frecuencias con los mismos totales marginales de filas y columnas

		Obesidad Si	Obesidad No	
(i)	Mujeres	0	5	5
	Hombres	8	1	9
		8	6	14
(ii)	Mujeres	1	4	5
	Hombres	7	2	9
		8	6	14
(iii)	Mujeres	2	3	5
	Hombres	6	3	9
		8	6	14
(iv)	Mujeres	3	2	5
	Hombres	5	4	9
		8	6	14
(v)	Mujeres	4	1	5
	Hombres	4	5	9
		8	6	14
(vi)	Mujeres	5	0	5
	Hombres	3	6	9
		8	6	14

p_a

calculado la probabilidad exacta de ocurrencia bajo la hipótesis nula, según:

$$p_{a'} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a'!b'!c'!d'!}$$

Entonces calculando tenemos:

Probabilidad exacta asociada con cada una de las disposiciones de frecuencias

	a	b	c	d	p
(i)	0	5	8	1	0,0030
(ii)	1	4	7	2	0,0599
(iii)	2	3	6	3	0,2797
(iv)	3	2	5	4	0,4196
(v)	4	1	4	5	0,2098
(vi)	5	0	3	6	0,0280

p_a

Cálculo del p-valor asociado al estadístico de prueba

Calculamos el p-valor para una prueba bilateral:

El valor p bilateral es $p=\sum_{p_{a'} \le p_a} p_{a'} = 0{,}003 + 0{,}028 + 0{,}0599 = 0{,}0909$

Regla de Decisión:

Prueba Bilateral:

Se rechaza H_0, si $p \le \alpha/2=0.025$, de los datos $p=0.0909>0.025$, por lo tanto con un nivel de significancía del 5% no rechazamos H_0.

Conclusión:

Con un nivel de significancia del 5% podemos concluir que la obesidad no se encuentra relacionada con el sexo de las personas en la población en estudio.

PRUEBA DE DOS MUESTRAS DE KOLMOGOROV-SMIRNOV

La Prueba de dos muestras de Kolmogorov-Smirnov puede confirmar que dos muestras independientes han sido extraídas de la misma población (o de poblaciones con la misma distribución).

Esta prueba está construida, teniendo como base detectar las diferencias existentes entre las frecuencias relativas acumuladas de las dos muestras objeto de estudio.

La prueba admite que los tamaños de las muestras no sean iguales.

La prueba de una muestra de kolmogorov-Smirnov examinaba los puntos de coincidencia de la distribución de un conjunto de valores muestrales y una distribución teórica. La prueba de dos muestras examina los puntos de coincidencia de dos conjuntos de valores muestrales.

Si las muestras han sido extraídas de la misma distribución de población, puede esperarse que las distribuciones acumulativas de ambas muestras sean próxima entre sí, ya que debería mostrar solamente desviaciones debido a la aleatoriedad de la muestra.

METODO DE APLICACION DE LA PRUEBA

1) Planteamiento de Hipótesis y determinación del nivel de significancia. Entre las hipótesis tenemos:

Hipótesis

a) Bilateral

H_0:- Las distribuciones poblacionales son iguales
- Las muestras proceden de la misma población

H_1:- Las distribuciones poblacionales son distintas
- Las muestras proceden de poblaciones diferentes

b) Unilateral (podría ser derecha o izquierda)

H_0: Los valores de la población de la que se extrajo una de las muestras es estocásticamente menor o igual que de los de la población de la que se sacó la otra. (puede cambiar la dirección de acuerdo al análisis).

H_1: Los valores de la población de la que se extrajo una de las muestras es estocásticamente más grande que de los de la población de la que se sacó la otra. (puede cambiar la dirección de acuerdo al análisis).

2) Hacemos una tabla de distribución de frecuencia acumulativa particionada en "k" categorías o intervalos, para cada muestra de observaciones (se usan tantos intervalos como sea factible), usaremos los mismos intervalos para ambas distribuciones.

3) Se determinan las diferencias entre las frecuencias acumuladas de las dos muestras en cada punto registrado. Se analiza entonces en la columna de las diferencias de las frecuencias, en qué clases se obtiene la más grande de las diferencias (valor máximo) denotado por "D".

Para una prueba de una cola (debe considerarse la dirección establecida en el estudio):

$$D_c = máxima(S_{n_1} - S_{n_2})$$

Para una prueba de dos colas será la diferencia máxima en valor absoluto.

$$D_c = máxima\left|S_{n_1} - S_{n_2}\right|$$

4) Determinación de los valores críticos para la toma de decisión

a) Cuando $\mathbf{n_1=n_2=n,\ \ n>40}$, se usa la tabla de Kolmogrov-Smirnov, si son diferentes usaremos la tabla para Muestras de distinto tamaño, según sean de una o dos colas.

Entonces: rechazamos H_0, si: Dc≥ Dtabla.

b) Cuando **n_1 y n_2 son mayores a 40 haciendo caso omiso de que sean iguales o no**, el estadístico de prueba a utilizarse es:

$$\chi_c^2 = 4D_c^2 \frac{n_1 n_2}{n_1 + n_2} \approx \chi^2_{(2g.l.)}$$

Este estadístico de prueba es también útil para muestras pequeñas con $n_1 \neq n_2$, no tabulados.

Entonces para una prueba de una cola, rechazamos H_0, si: $\chi_c^2 \geq \chi^2_{tabla,(2)\alpha}$

5)Conclusión

Aplicativo:
Se muestran las pérdidas en peso (medidas en kilogramos), de dos grupos de personas que han sido sometidas a dos tipos diferentes de medicamentos, designado por Grupo1 y Grupo2. Los resultados obtenidos se muestran en la siguiente tabla:

GRUPO1	GRUPO2
5.49	3.76
3.08	4.22
4.13	4.17
5.03	5.03
7	4.85
6.03	2.09
4.45	4.45
5.13	3.58
4.26	3.86
4.62	4.13
	4.4
	2.81

Con un nivel de significancia del 5%, ¿podemos afirmar que existe diferencia significativa entre las poblaciones de las cuales se extrajeron las muestras?.

Solución:

En cuanto a la prueba a aplicar, observamos que los grupos 1 y 2 son independientes.

Entonces bajo estas condiciones una prueba de Kolmogorov-Smirnov será la adecuada y como nos interesa decidir si existe diferencias o no entre las poblaciones, entonces la aplicación será de una prueba de 2 colas.

Planteamos las hipótesis:

H_0: No existe diferencia significativa entre las poblaciones de donde fueron extraídas las muestras.

H_1: Existe diferencia significativa entre los grupos poblacionales en estudio.

Luego del enunciado tenemos α=0.05, calculamos la tabla de distribución correspondiente:

$n=n_1+n_2$	22
Máximo	7
Mínimo	2.09
Rango	4.91
N° de clases o intervalos	5.46
$(1+3.32\log_{10}(n))$	5
Ancho de clase	1.0

Intervalos - medidas de perdida de pesos	frecuencia grupo 1	frecuencia grupo 2	F-acuma1	F-acuma2
2.09 a 3.09	1	2	1	2
3.10 a 4.10	0	3	1	5
4.11 a 5.11	5	7	6	12
5.12 a 6.12	3	0	9	12
6.13 a 7.13	1	0	10	12
	10	**12**		

Para aplicar la prueba de Kolmogorov-Smirnov, reorganizamos estos datos en dos distribuciones de frecuencias acumulativas, que se muestra a continuación y por simple sustracción encontramos las diferencias entre las distribuciones en los diferentes intervalos de las 2 muestras:

Distribución de frecuencia acumulativa 1	Distribución de frecuencia acumulativa 2	Diferencia $\|S_{10}-S_{12}\|$
0.1	0.17	**0.07**
0.1	0.42	**0.32**
0.6	1	**0.40**
0.9	1	**0.10**
1	1	**0.00**

Luego calculamos el estadístico: $D_c = máxima\left|S_{n_1} - S_{n_2}\right|$

Dc=0.4, como no se tiene el dato tabulados calculamos el siguiente estadístico de prueba:

$$\chi_c^2 = 4D_c^2 \frac{n_1 n_2}{n_1 + n_2} \approx \chi^2_{(2g.l.)}$$

		p (Unilateral)				
		0'9	0'95	0'975	0'99	0'995
		p (Bilateral)				
n_1	n_2	0'80	0'90	0'95	0'98	0'99
10	15	2/5	7/15	1/2	17/30	19/30
	20	2/5	9/20	1/2	11/20	3/5
	40	7/20	2/5	9/20	1/2	
12	15	23/60	9/20	1/2	11/20	7/12
	16	3/8	7/16	23/48	13/24	7/12
	18	13/36	5/12	17/36	19/36	5/9
	20	11/30	5/12	7/15	31/60	17/30

Reemplazando:

$$\chi_c^2 = 4(0.4)^2 \frac{(10)(12)}{10+12} = 3.49 \approx \chi^2_{(2g.l.)}$$

$$\chi^2_{tabla(0.05,2gl)} = 5.99$$

Entonces: $\chi_c^2 < \chi^2_{tabla}$ por lo tanto, No rechazamos H0.

Concluimos, bajo un nivel de significancia del 5%, no existen diferencias significativas entre las perdidas de peso de las personas sometidas a los medicamentos en estudio.