BIOESTADÍSTICA BÁSICA

PRUEBAS NO PARAMÉTRICAS
SPEARMAN
CHI2
FISHER

CLASE 6

Cuando existen variables que no siguen las condiciones de parametricidad, por ejemplo: variables cuantitativas, distribución normal de las muestras, varianzas similares y tamaño de las muestras. Sobre todo cuando la normalidad de las distribuciones de la variable en estudio esté en duda, el uso de las pruebas no paramétricas o de distribución libre es el indicado.

Rubio y Berlanga (2012).

VENTAJAS

1) son más fáciles de aplicar

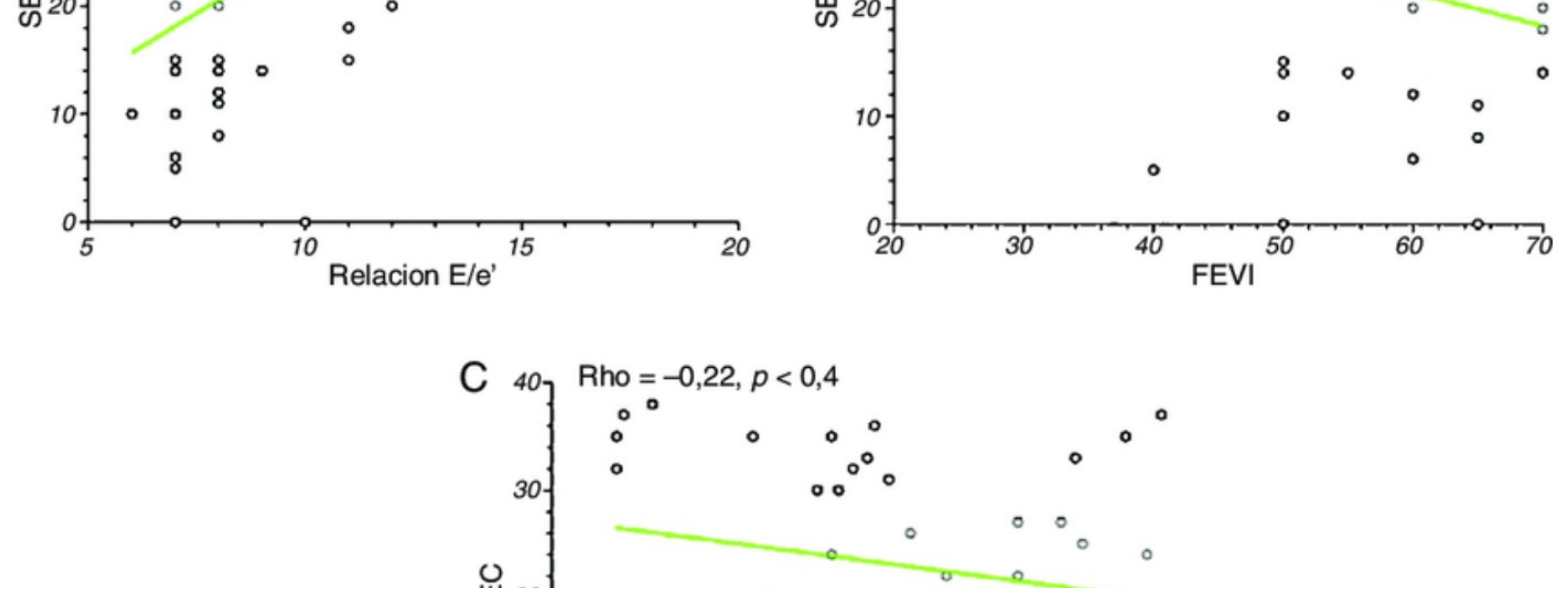
2) son aplicables a los datos jerarquizados

3) se pueden usar cuando dos series de observaciones provienen de distintas poblaciones

4) son útiles a un nivel de significancia previamente especificado

PRUEBAS
NO PARAMÉTRICAS
DISTRIBUCIÓN LIBRE

- ✓ **Kolmogorov**
- ✓ Correlación Spearman
- ✓ Chi2
- ✓ Fisher
- ✓ Mann-Whitney
- ✓ Kruskal/Wallis



1. COEFICIENTES DE CORRELACIÓN

SPEARMAN

- **Relación o Asociación lineal entre dos variables cuantitativas u ordinales**.

- Los coeficientes de correlación de valores absolutos que oscilan entre -1 y 1. Hay que tomar en cuenta que tienen signos.

- Donde -1 y 1 indican correlaciones perfectas, negativas y positivas respectivamente
- 0 indica correlación nula.

RECORDAR

- Los análisis de correlación son herramientas estadísticas que se usan para evaluar el grado en que dos variables se relacionan.

- Los análisis de correlación son pertinentes en los casos que se desea conocer cómo se puede comportar una variable, y al mismo tiempo se conoce el comportamiento de otra variable relacionada.

- En este sentido, como el investigador no manipula de manera intencional las variables de estudio (diseños no experimentales), en los análisis de correlación no es posible predecir cuál de las variables está afectando a otra. Las correlaciones no tienen un alcance causal, sólo asociativo.

https://17jaem.semrm.com/aportaciones/n73.pdf

SI LAS VARIABLES CUANTITATIVAS NO SON NORMALES O NECESITAMOS USAR VARIABLES ORDINALES…

COEFICIENTE SPEARMAN (RANGOS O JERARQUIAS)

- Una o ambas variables que estamos comparando son de escala ordinal o las variables cuantitativas no son normales
- Para calcularlo ordenamos las observaciones de la primera variable de manera ascendente y les damos el valor de su orden.
- Si dos observaciones de la misma variable tienen el mismo valor, se saca el promedio. Hacemos lo mismo para la segunda variable.
- Calculamos la diferencia entre los rangos para cada par de observaciones.
- La correlación Spearman o ρ de la letra griega *rho* se calcula según la definición

$$\boxed{r_s = 1 - \frac{6\sum d^2}{N(N^2-1)}}$$

d = Diferencia entre los rangos de cada par de puntuaciones
N = Número de observaciones

- Toma valores entre -1 y +1
- Cuando el valor absoluto es cercano a 1 indica que hay una FUERTE asociación entre las dos variables, si es cercano a cero indica que hay DÉBIL asociación de las variables.
- No requiere normalidad

$$R_{obs} > R_{cri}$$
$$-R_{obs} < -R_{cri}$$

Se rechaza H_0
hay correlación

Valores críticos de la *rho de Spearman*

	Niveles de significación (α)	
N	*0.05*	*0.01*
4	1.000	
5	0.900	1.000
6	0.829	0.943
7	0.714	0.893
8	0.643	0.833
9	0.600	0.783
10	0.564	0.746
11	0.536	0.709
12	0.506	0.712
13	0.484	0.648
14	0.456	0.645
15	0.446	0.604
16	0.425	0.601
17	0.414	0.566
18	0.399	0.564
19	0.391	0.535
20	0.377	0.534
21	0.370	0.508
22	0.359	0.508
23	0.353	0.486
24	0.343	0.485
25	0.337	0.466
26	0.329	0.465
27	0.324	0.448

	Niveles de significación (α)	
N	*0.05*	*0.01*
28	0.317	0.440
29	0.312	0433
30	0.306	0.425
31	0.301	0.418
32	0.296	0.412
33	0.291	0.405
34	0.287	0.399
35	0.283	0.394
36	0.279	0.388
37	0.275	0.383
38	0.271	0.378
39	0.267	0.373
40	0.264	0.368
41	0.261	0.364
42	0.257	0.359
43	0.254	0.355
44	0.251	0.351
45	0.248	0.347
46	0.246	0.343
47	0.243	0.340
48	0.240	0.336
49	0.238	0.333
50	0.235	0.329

EJERCICIO

Rho Spearman Clase.xlsx

EJERCICIOS EN CASA

1.

Se desea conocer el grado de relación entre las calificaciones que obtuvieron 10 estudiantes de la maestría de epidemiología, en análisis epidemiológicos y bioestadística, los datos no son normales.

Toma como referencia el 95% de confianza. Los resultados se muestran a continuación:

CALIFICACIONES		
Alumno	Análisis ep (X)	Bioestadistica (Y)
1	10	8
2	9	9
3	8	10
4	6	1
5	7	4
6	5	6
7	5	6
8	3	7
9	2	7
10	4	3

HIPÓTESIS

H0: No existe correlación entre las calificaciones de análisis epidemiológicos y bioestadística
(r = 0)

H1: Si existe correlación entre las calificaciones de análisis epidemiológicos y bioestadística
(r =! 0)

Respuesta: 0,39 < 0,564

No hay correlación

2.

En un estudio de la relación entre la edad y los resultados del electroencefalograma- EEG, se recopilaron datos en 20 personas con edades entre 20 y 60 años.

La siguiente tabla muestra las edades y un valor de rendimiento del EEG particular para cada una de esas 20 personas.

Los investigadores pretenden saber si es posible concluir que este rendimiento del EEG particular tiene correlación con la edad a un nivel de signficancia α=0.01. Los datos, no siguen una distribución normal.

Número de individuo	Edad (X)	Valor resultante del EEG(Y)
1	20	98
5	27	99
9	35	85
12	42	66
14	46	62
16	51	54
17	53	63
18	55	52
19	58	67
20	60	55

1. PLANTEA LA HIPÓTESIS
2. REALIZA EL PROCEDIMIENTO
3. TOMA UNA DECISIÓN
4. INTERPRETA TUS RESULTADOS

Respuesta:

$-0.70 < -0,564$

Hay correlación

3.

Se desea saber si existe correlación entre el nivel de instrucción y la actividad física a un nivel de confianza del 95%.

	Puntuaciones	
Individuo	Nivel de instrucción	Nivel de actividad física
1	17	21
2	18	14
3	19	27
4	12	18
5	23	20
6	23	25
7	25	34
8	26	32
9	31	39
10	33	33

1. PLANTEA LA HIPÓTESIS
2. REALIZA EL PROCEDIMIENTO
3. TOMA UNA DECISIÓN
4. INTERPRETA TUS RESULTADOS

Respuesta:

$0.82 > 0.56$

Hay correlación

ASOCIACIÓN

CHI2

FISHER

Asociación estadística, se usa para hacer referencia a la dependencia entre dos variables de cualquier tipo, mientras que el término **correlación** se emplea únicamente para variables cuantitativas (Estepa, 1994).

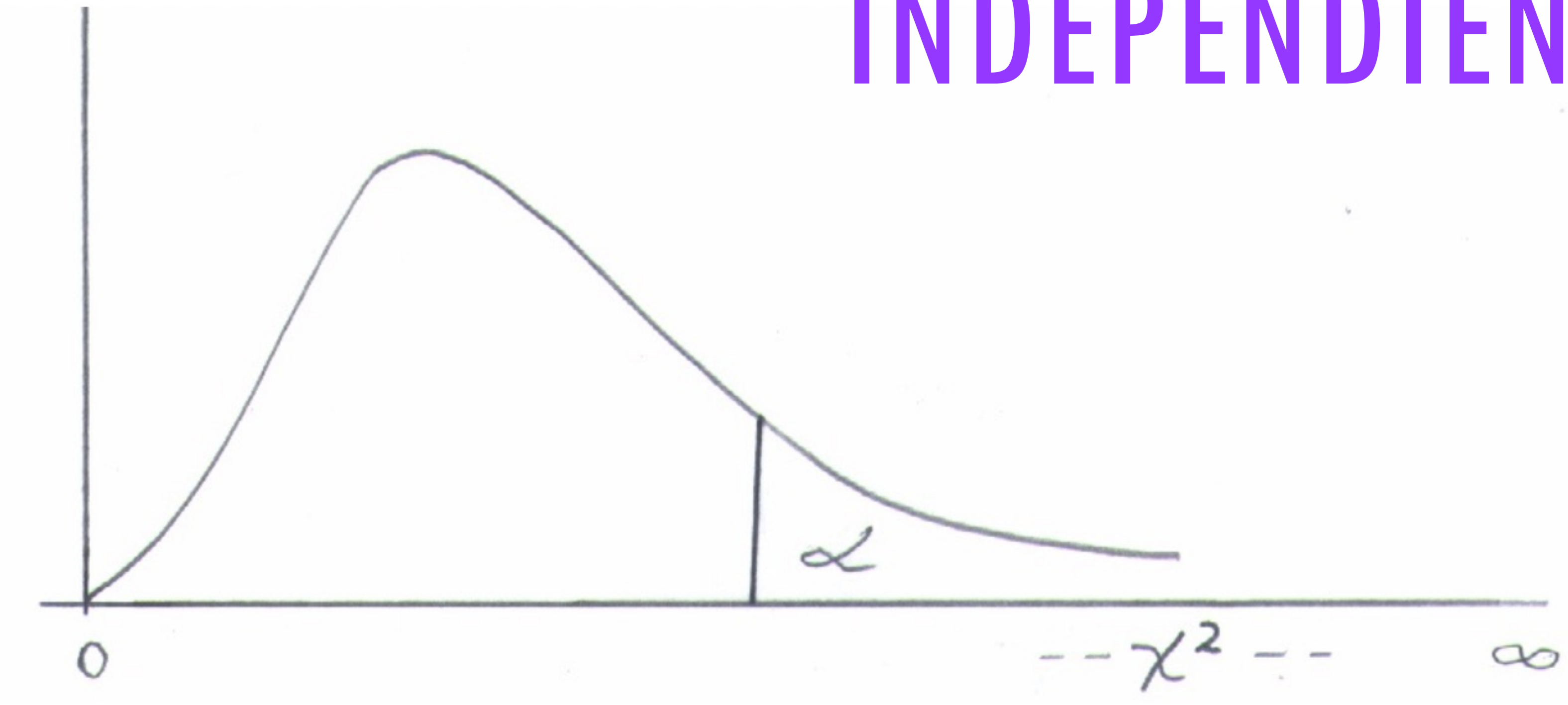
SE REQUIERE EVALUAR ASOCIACIÓN VARIABLES CUALITATIVAS

TIPO DE MUESTRA SP/EP

1. INDEPENDIENTES → Chi 2
   Fisher

2. PAREADAS → McNemar

INDEPENDIENTES



PRUEBA 1: CHI2/JI2

$$\chi^2$$

La distribución es chi2.

Sirve para someter a prueba las hipótesis referidas a distribuciones de frecuencias.

En términos generales, esta prueba contrasta frecuencias observadas con las frecuencias esperadas de acuerdo con la hipótesis nula.

Probar la asociación entre dos variables cualitativas.

Del mismo modo que los estadísticos “z”, con su distribución normal y “t”, con su distribución t de Student, nos han servido para poner a prueba las hipótesis que **involucran a promedios y probabilidades** (%).

El estadístico ji-cuadrado (o chi cuadrado), tiene distribución de probabilidad del mismo nombre, nos servirá para poner a prueba las hipótesis referidas a distribuciones de frecuencias y de esta manera probar **la asociación entre dos variables**.

Como norma general, para el uso de Xi_2 se exige que el 80% de las celdas en una tabla de asociación tengan valores esperados mayores de 5

TABLA DE CONTINGENCIA

La tabla de contingencia contiene la información de dos **variables cualitativas**. Es una tabla de frecuencias de doble entrada.

Las filas y columnas de la tabla corresponden a las variables categóricas.

La tabla incluye los totales para cada nivel de las variables.

La tabla de contingencia para el análisis de correspondencia simple es una tabla de dos factores que cuenta las observaciones de dos variables.

Categoría A_i	Frecuencia Observada O_i	Frecuencia Esperada E_i
A_1	O_1	E_1
A_2	O_2	E_2
...	...	...
A_r	O_r	E_r
TOTAL	n	n

En una determinada muestra se observan una serie de posibles sucesos

E1, E2, E3, . . . , EK,

que ocurren con frecuencias

o1, o2, o3, . . ., oK

FRECUENCIAS OBSERVADAS

Según las reglas de probabilidad, se espera que ocurran con frecuencias **e1, e2, e3, . . . ,eK** llamadas **FRECUENCIAS TEÓRICAS O ESPERADAS**.

Se desea saber si las frecuencias observadas difieren significativamente de las frecuencias esperadas.

FRECUENCIAS ESPERADAS

Es el conteo de observaciones que se espera en una celda, en promedio, si las variables son independientes o la H0 es cierta

Se calculan los conteos esperados como: el producto de los totales de fila y columna, dividido entre el número total de observaciones.

REGLA DE DECISIÓN

$CHI2_{OBS} < CHI2_{CRITICO}$

- SI
  - Acepta H_0
- NO
  - Rechaza H_0

Supongamos que un investigador está interesado en evaluar la asociación entre uso de anticonceptivos y el nivel socioeconómico de las mujeres. Con este objeto se toma una muestra de mujeres a quienes se clasifica en una tabla de asociación, encontrando los siguientes resultados

TABLA DE VALORES OBSERVADOS				
Uso de anticonceptivos	**Nivel socioeconómico bajo**	**Nivel socioeconómico medio**	**Nivel socioeconómico alto**	**Total**
Si	8	15	28	**51**
No	13	16	14	**43**
Total	**21**	**31**	**42**	**94**

Nivel de significancia $\alpha = 0.05$

1. SE DEBE PLANTEAR LAS HIPÓTESIS QUE SOMETEREMOS A PRUEBA

H0: "El uso de anticonceptivos es independiente/no está asociado con el nivel socioeconómico".

H1: "El uso de anticonceptivos depende/está asociado con del nivel socioeconómico".

En esta prueba estadística SIEMPRE la hipótesis nula plantea que las variables analizadas son independientes, es decir, no están asociadas

2. OBTENER (CALCULAR) LAS FRECUENCIAS ESPERADAS

Las frecuencias esperadas se obtendrán de la distribución de frecuencias del total de los casos, **51 mujeres** de un total de 94 que usan anticonceptivos y **43 mujeres** de un total de 94, no lo usan. Esa misma proporción se debería dar al interior de los tres grupos de nivel socioeconómico, de manera que el cálculo responde al siguiente razonamiento:

SI DE 94 MUJERES 51 USAN ANTICONCEPTIVOS; DE 21 MUJERES ¿CUÁNTAS DEBIERAN USARLO?

TABLA DE VALORES OBSERVADOS				
Uso de anticonceptivos	**Nivel socioeconómico bajo**	**Nivel socioeconómico medio**	**Nivel socioeconómico alto**	**Total**
Si	8	15	28	**51**
No	13	16	14	**43**
Total	**21**	**31**	**42**	**94**

$$e=\frac{nitotal1 * nitotal2}{N}$$

Nivel bajo (SI) : (51x21/94)=11,4

Nivel bajo (NO):(43x21/94)=9,6

-Nivel medio (SI): (51x31/94)=16,8
-Nivel medio (NO):(43x31/94)=14,2

-Nivel alto (SI): (51x42/94)=22,8
-Nivel alto (NO):(43x42/94)=19,2

Estas son las frecuencias que debieran presentarse si la hipótesis nula fuera verdadera y, por consiguiente, las variables fueran independientes. Estos valores los anotamos en una tabla con las mismas celdas que la anterior; así tendremos una tabla con los valores observados y una tabla con los valores esperados.

TABLA DE VALORES ESPERADOS				
Uso de anticonceptivos	**Nivel socioeconómico bajo**	**Nivel socioeconómico medio**	**Nivel socioeconómico alto**	**Total**
Si	11,4	16,8	22,8	**51**
No	9,6	14,2	19,2	**43**
Total	**21**	**31**	**42**	**94**

3. OBTENER CHI2

BIOESTADISTICA EJERCICIOS/MODULO 6 ASOCIACIÓN.xlsx

$$\boxed{\chi^2 = 5{,}23}$$

Este es el valor de nuestro estadístico de prueba que ahora, debemos comparar con un valor de la tabla de probabilidades para ji-cuadrado ($\mathbf{x}^2$). Esta tabla es muy parecida a la tabla *t de student*, pero tiene sólo valores positivos porque ji-cuadrado sólo da resultados positivos.

4. OBTENER CHI2 CRÍTICO USANDO LOS GRADOS DE LIBERTAD

$$gl=(n^{o} \text{ de filas}-1)\times(n^{o} \text{ de columnas}-1)$$

Así, en nuestro ejemplo, en que hay 2 filas y 3 columnas, los grados de libertad serán:

TABLA DE VALORES OBSERVADOS				
Uso de anticonceptivos	**Nivel socioeconómico bajo**	**Nivel socioeconómico medio**	**Nivel socioeconómico alto**	**Total**
Si	8	15	28	**51**
No	13	16	14	**43**
Total	**21**	**31**	**42**	**94**

$gl=(2-1)x(3-1)=2$

***Nótese que no se consideran la fila ni la columna de los totales.**

5. REGLA DE DECISIÓN

[BIOESTADISTICA EJERCICIOS/MODULO 6 ASOCIACIÓN.xlsx]

CONCLUSIÓN

Aceptar la hipótesis nula (H_0) que plantea que las variables “uso de anticonceptivos” y “nivel socioeconómico” son independientes

TABLAS DE CONTIGENCIA DE 2X2

- CHI 2
  2X2
  - FO/FE
  - ODDS (OR)

El chi-cuadrado de contingencia 2 X 2 se utiliza para la comparación de dos grupos con una variable dependiente dicotómica. Podríamos comparar hombres y mujeres en una escala de respuesta sí / no.

En la tabla de ODDS, no se requiere la determinación de frecuencias esperadas

Variable A	Variable B		Total
	B1	B2	
A1	a	b	a + b
A2	c	d	c + d
Total	a + c	b + d	n

Variable A	Variable B		Total
	B1	B2	
A1	a	b	a + b
A2	c	d	c + d
Total	a + c	b + d	n

$$\chi^2_{obs} = \frac{n(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)},$$

Sigue un grado de libertad

A un grupo de personas que se quejaba de no dormir bien se les administró dos tipos de tratamientos distintos (A,B). Al concluir el estudio se les preguntó si habían dormido bien o mal y se obtuvieron los siguientes resultados:

	Tratamiento A	Tratamiento B	
Durmieron Mal	338	363	**701**
Durmieron Bien	125	156	**281**
	463	**519**	**982**

RESOLVER MEDIANTE:

FRECUENCIAS

ODDS/RR

1.FRECUENCIAS

$$e=\frac{nitotal1*nitotal2}{N}$$

$E_{11}=\frac{(701)(463)}{982}=330.51$	$E_{12}=\frac{(701)(519)}{982}=370.49$
$E_{21}=\frac{(281)(463)}{982}=132.49$	$E_{22}=\frac{(281)(519)}{982}=148.51$

$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$

$$\chi^2 = \sum \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} = \frac{(338 - 330.51)^2}{330.51} + \frac{(363 - 370.49)^2}{370.49} + \frac{(125 - 132.49)^2}{132.49} + \frac{(156 - 148.51)^2}{148.51}$$

$$= 0.170 + 0.151 + 0.423 + 0.378$$

$$= 1.12$$

2. ODDS

Variable A	Variable B		Total
	B1	B2	
A1	a	b	a + b
A2	c	d	c + d
Total	a + c	b + d	n

	Tratamiento A	Tratamiento B	
Durmieron Mal	338	363	**701**
Durmieron Bien	125	156	**281**
	463	**519**	**982**

$$\chi^2_{obs} = \frac{n(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)},$$

$$\frac{982((338x156)-(363x125))^\wedge 2}{(701) \times (281) \times (463) \times (519)}$$

Comparar Xi2obs y Xicrítico

Tomar una decisión

GL (TABLA 2X2) = 1

Xicritico = 3.84

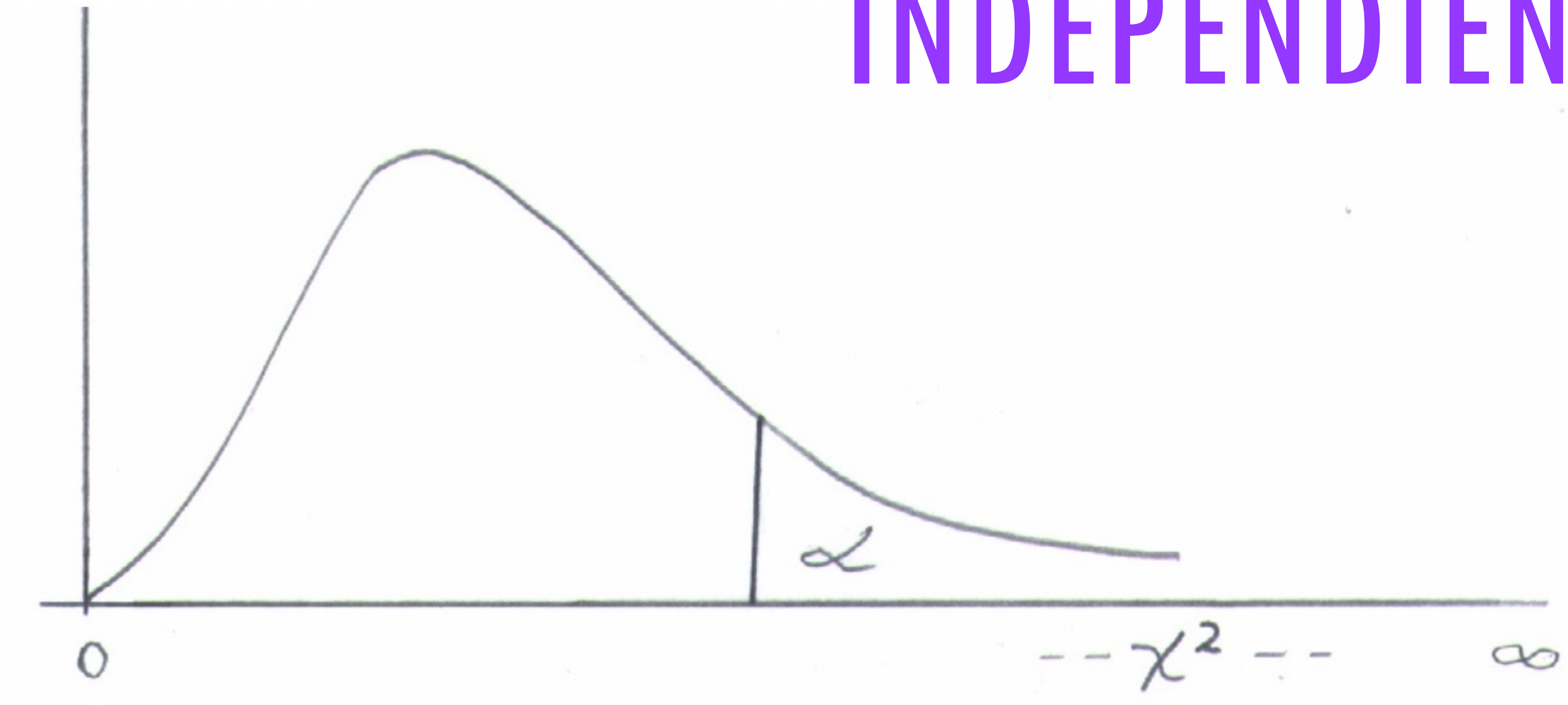
1.12 < 3.84

Acepto H0

EJRCICIOS CON BASES DE DATOS - EXCEL

BIOESTADISTICA EJERCICIOS/MODULO 6 ASOCIACIÓN.xlsx

INDEPENDIENTES



PRUEBA 2: FISHER

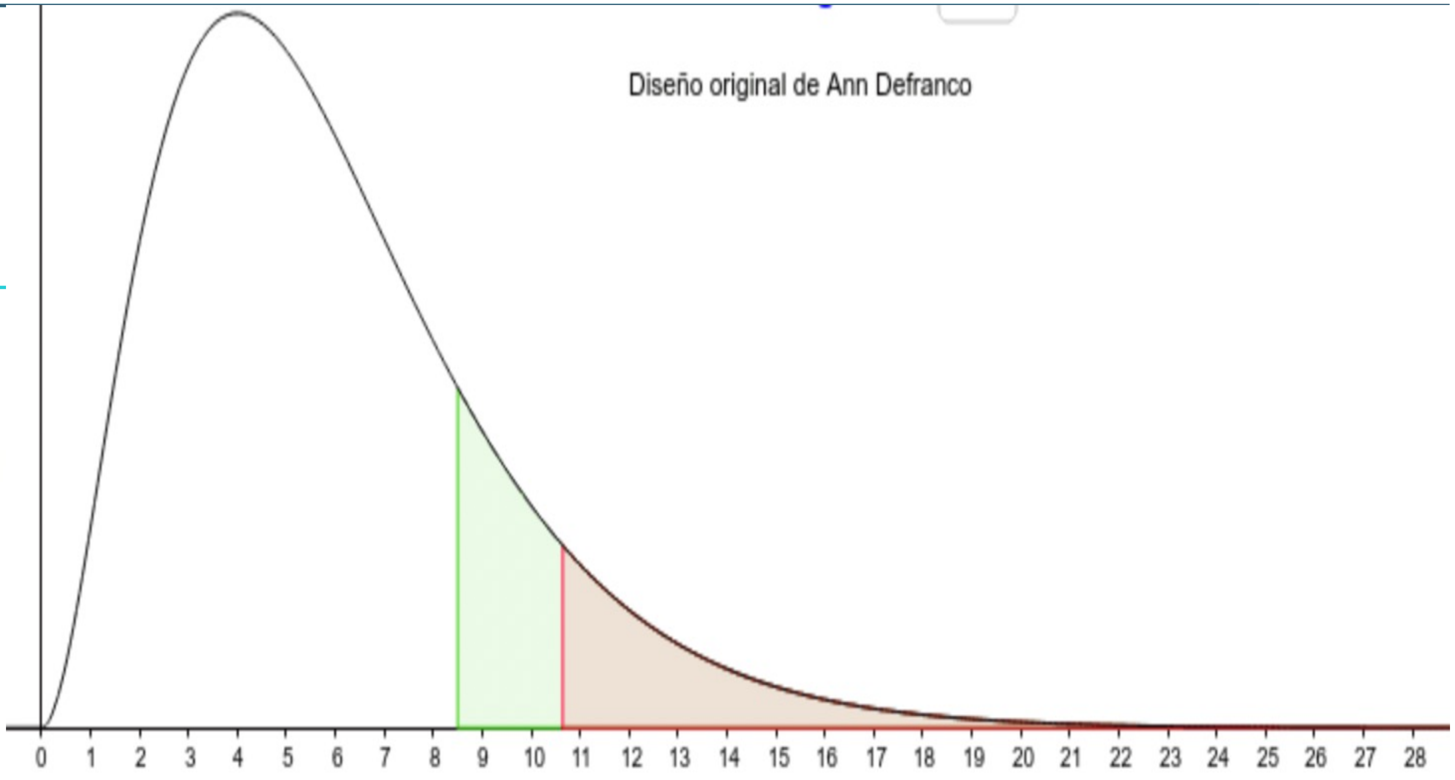
EXACTA

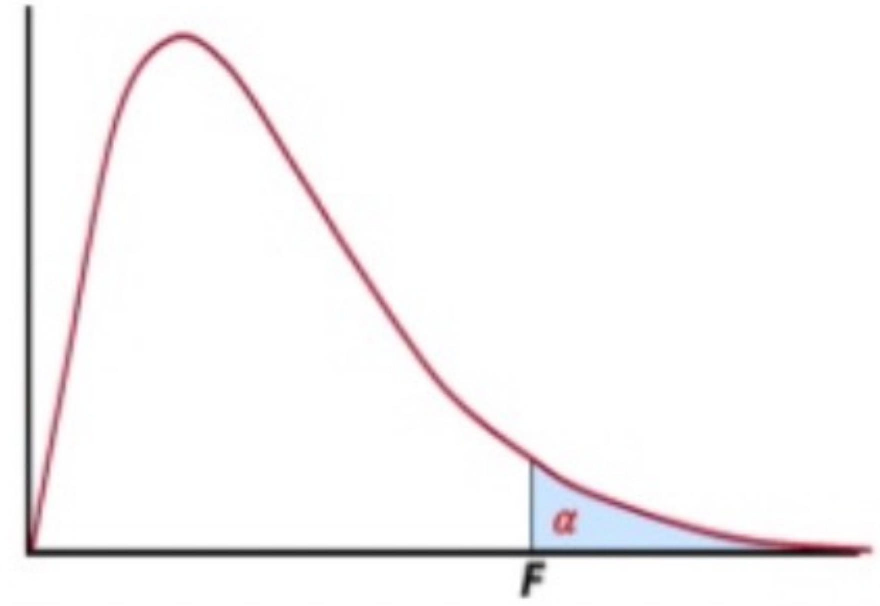
Es una prueba estadística de independencia para variables categóricas.

Lo aplicamos cuando hay dos variables, y cada variable tiene dos valores posibles (dicotómicas), por lo que podemos representar datos en una **tabla de contingencia** 2 **x 2** .

Utilizamos la prueba exacta de Fisher para **determinar si hay *dependencia* entre las dos variables**, es decir, si las proporciones de una variable difieren dependiendo del valor de la otra variable.

La prueba de Fisher es una **prueba exacta** porque conocemos la distribución exacta del muestreo.





La distribución hipergeométrica se utiliza para muestras obtenidas de poblaciones relativamente pequeñas, sin reemplazo. Por ejemplo, la distribución hipergeométrica se utiliza en la prueba exacta de Fisher para probar la diferencia entre dos muestras.

La distribución hipergeométrica se define por 3 parámetros: tamaño de la población, conteo de eventos en la población y tamaño de la muestra.

Dado que tenemos que calcular factores, la prueba exacta de Fisher es difícil de calcular cuando **la muestra es grande o la tabla de contingencia está bien equilibrada**.

Fisher mostró que, bajo el supuesto de que la hipótesis nula se mantiene y los valores para los totales marginales son fijos, la probabilidad de observar la tabla original viene dada por la distribución hipergeométrica, que se ve así:

$$\frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}}$$

donde vemos tres coeficientes binomiales diferentes.
Podemos reescribirlos de la siguiente manera:

$$\boxed{\dfrac{\dfrac{(a+b)!}{a!b!}\cdot\dfrac{(c+d)!}{c!d!}}{\dfrac{n!}{(a+c)!(b+d)!}}}$$

FACTORIAL:

El producto de todos los enteros positivos menores o iguales a un entero positivo dado y denotado por ese entero y un signo de exclamación.

Por lo tanto, el factorial siete está escrito 7!, que significa 1 × 2 × 3 × 4 × 5 × 6 × 7.

El cero factorial se define como igual a 1.

$$\boxed{\frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}}$$

	X	Y	
X1	**a**	**b**	a + b
Y1	**c**	**d**	c + d
	a + c	b + d	a + b + c + d (=n)

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{\binom{a+b}{b}\binom{c+d}{d}}{\binom{n}{b+d}} = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\;\;b!\;\;c!\;\;d!\;\;n!}$$

Se conocen todos los valores, se utiliza 1 grado de libertad

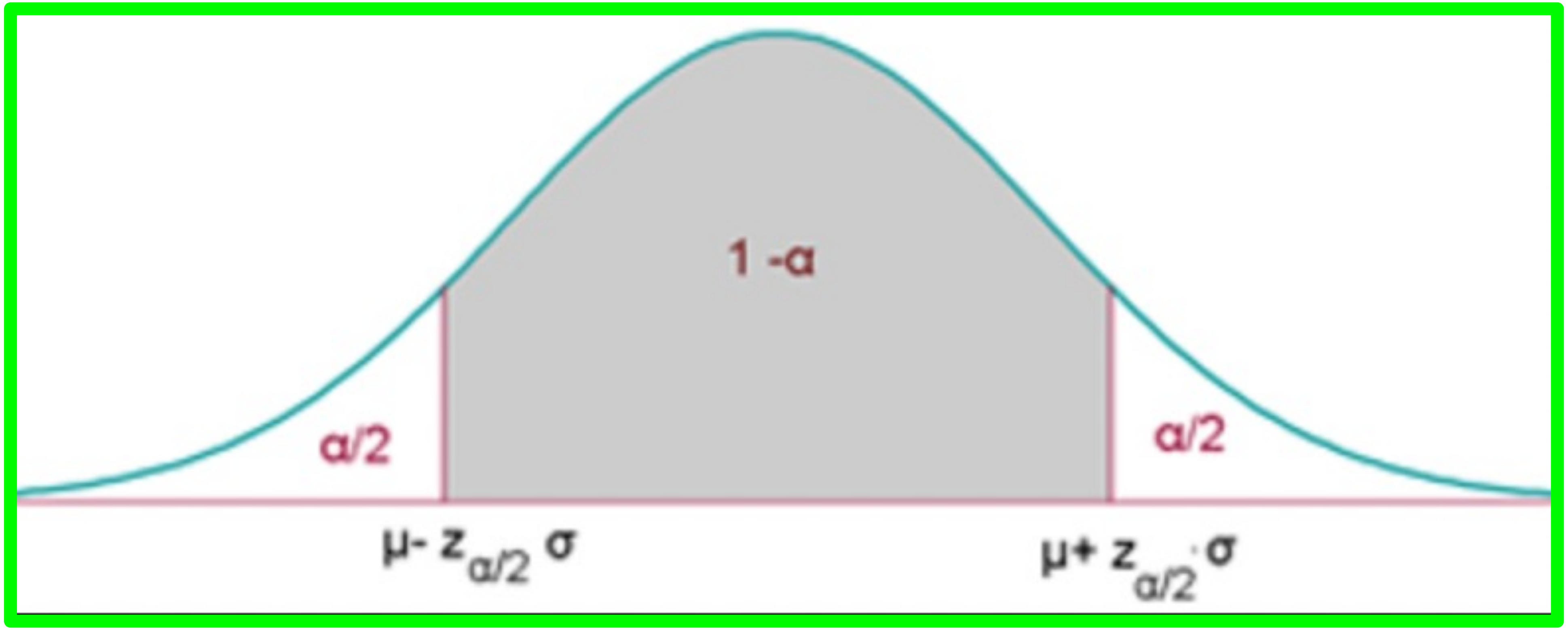
El valor, por ejemplo, de a, es suficiente para deducir todos los demás valores. Por lo tanto, la probabilidad discutida anteriormente depende solo de a!

Para encontrar el valor p para la prueba exacta de fisher, tenemos que encontrar todas las matrices posibles de enteros no negativos con los mismos totales de fila y columna que la tabla original y luego calcular la probabilidad de cada tabla de este tipo.

Trabajaremos con el supuesto de dos colas

VALOR P DE DOS COLAS

Para una prueba de dos colas, se suman las probabilidades de cada tabla que tiene una probabilidad menor o igual que la de la tabla observada.



HIPÓTESIS

H0: Las variables son independientes por lo que una variable no varía entre los distintos niveles de la otra variable.

H1: Las variables son dependientes, una variable varía entre los distintos niveles de la otra variable.

$p < 0{,}05$ (α) rechazo la H_0

Mientras más pequeño sea el p value más evidencia existe para rechazar la H0

EJERCICIO

Se eligen 11 pacientes que probaron 2 tratamientos para curarse de la dependencia de la cocaína: 5 hombres y 6 mujeres fueron sometidos al tratamiento 1 y tratamiento 2. Solo un hombre se curó y solo dos mujeres no se curaron.

-H0: Curarse de la dependencia de la cocaína no se asocia con el sexo del paciente.

-H1: Curarse de la dependencia de la cocaína si se asocia con el sexo del paciente.

	Curado	No curado	Total
Hombre	1	4	**5**
Mujer	4	2	**6**
Total	**5**	**6**	**11**

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

La probabilidad de observar esta tabla es igual a:

5! 6! 5! 6! / (1! 4! 4! 2! 11!) ≈ 0.16234

BIOESTADISTICA EJERCICIOS/MODULO 6 ASOCIACIÓN.xlsx