Tema 7 RELACION ENTRE VARIABLES

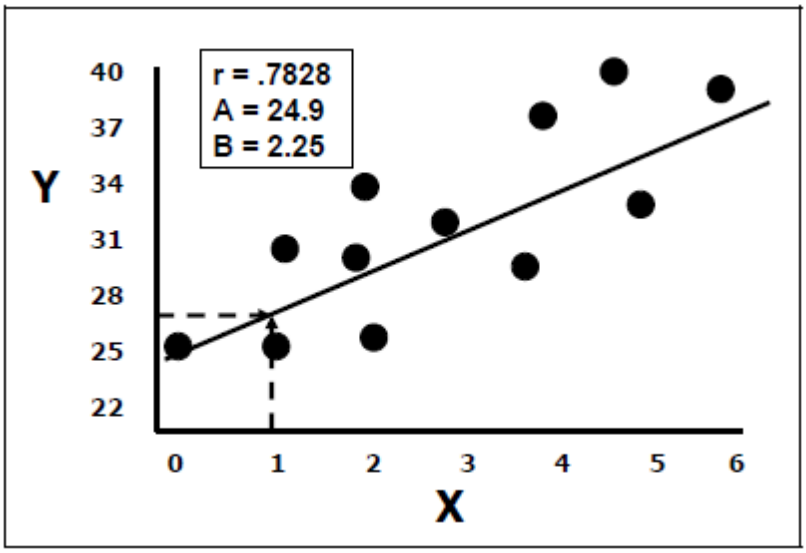
En educacion es interesante encontrar la relation entre dos o mas variables y poder aventurar (predecir) las puntuaciones de una avariable conociendo los de otras (vamos que si saca siempre un 8 en estadistica y un 7 en las actividades de estadistica, si tenemos un 7 en la actividad sabemos que ha sacado un 8 en general porque siempre a pasado lo mismo. Y estudiando la correlacion (que si pasa una cosa pasa la otra) como son la Pearson , la de Spearman , el de contingencia , la correlacion biserial-puntual, la phi , la tetracortica y la biserial.(vamos siete formulitas este tema es algo durillo)

7.2. El concepto de correlaccion

Para los estudios de educacion es fundamental establecer la relacion existente entre dos variables, el grado de acercamiento o distanciamiento entre variables o sea su variabilidad. Para conocerla se estudia la covarianza entre variables, por ejemplo si existe una relacion entre las notas sacadas en matematicas y en estadistica . Y siempre dara valores las formulitas entre 0 y +1

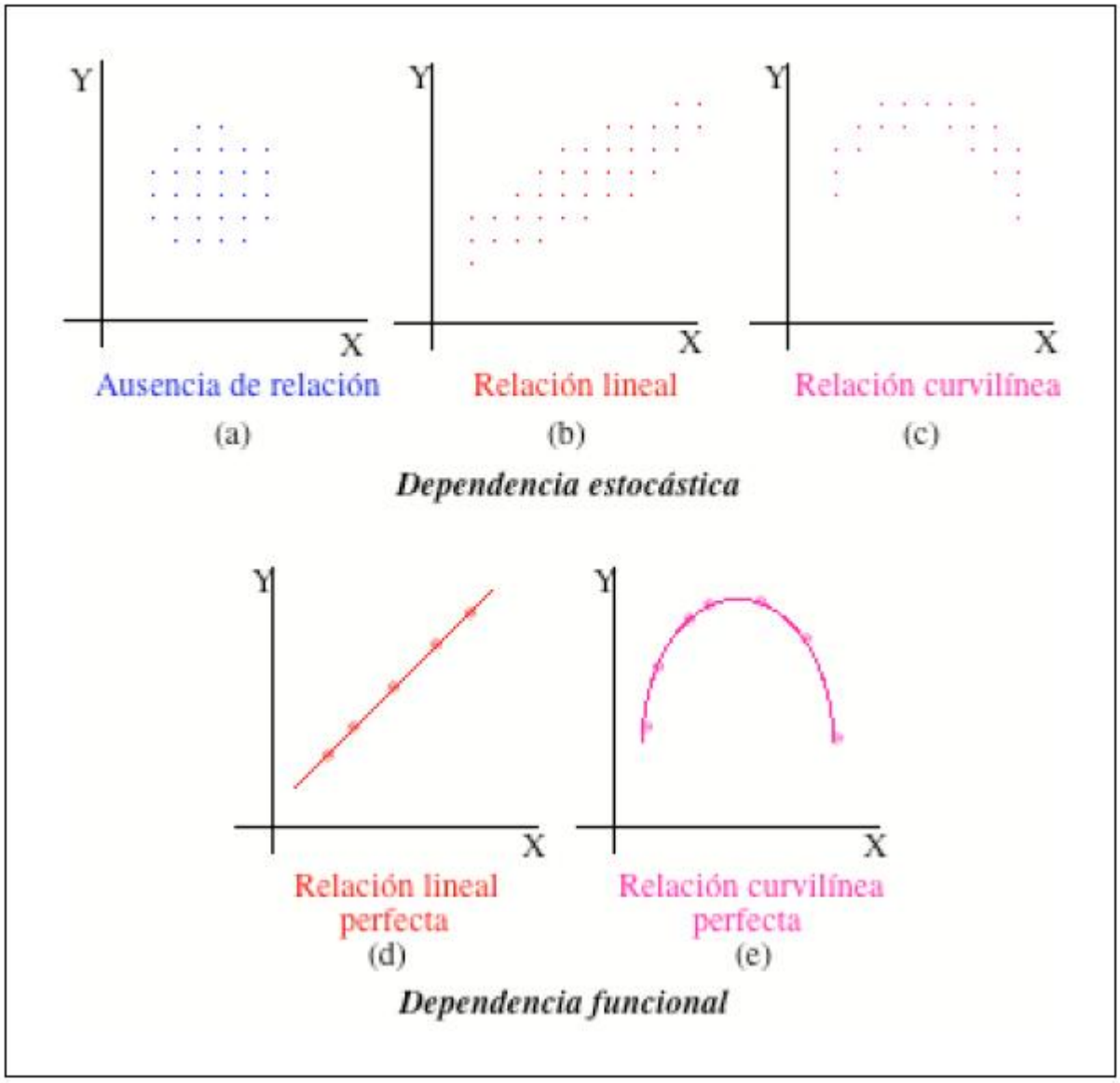
a) Relacion perfecta positiva

Cuando al aumentar los valores de una variable aumenta los de la otra en la misma proporcion . La correlacion es +1



b) La relacion imperfecta positiva , o relacion directa de variables, a valores elevados de una variable le corresponden valores elevados de la otra y al reves . Y a lo vemos en la vidad diaria quien saca buena nota en mates, saca buena nota en fisica y quimica y quien saca mala nota en mates saca mala nota en en fisica y quimica. La correlacion numerica es entre 0 y 1+.
c) Relacion perfecta negativa, es una relacio inversa entre variables, al aumentar el valor de una disminuye el de la otra proporcionalmente. Su exprsion cuantitativ es -1

d) Relacion imperfecta negativa. Las puntuaciones de una variable alta le corresponden las de otra variable baja o en un momento puntuan alto y en otro bajo. Se expresa numericamente entre 0 y -1 (Las variables modifican sus valores en sentidos opuestos por lo que la nube de puntos se concentra en torno a una linea decreciente , pero no coinciden con ella
e) Relacion nula o ausencia de relacion . Se da cuando dos variables son independientes una de otra . Las puntuaciones de las dos variables son aleatorias , no hay ninguna relacion aparente y cuantitativamentes se expresa con el 0.

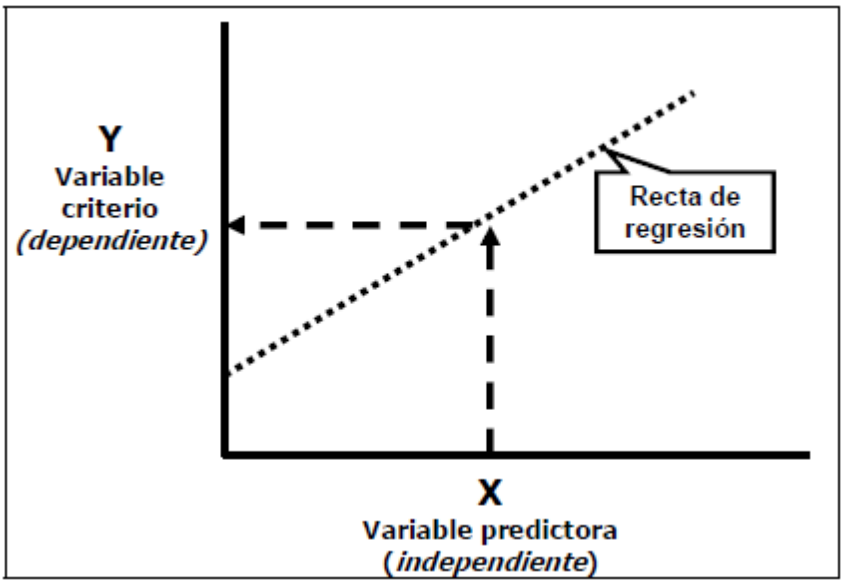


Ahora esta parte que voy a decir no aparece en el libro pero creo que lo aclara todo mucho.

El nombre y concepto de recta de regresion tiene su origen en las investigaciones sobre genetica de Francis Galton (primo de Darxin) sobre la altura de padres e hijos. El eje llamado de abcisas x horizontal es la variableindependiante o predictora y el que pone las cordenadas de la linea y perpendicualr a x es la variabledependiente o predictora. A una determinada puntuacion en Y le

corresponde una determinada puntuacion en x. La variable X es la que buscamos explicar y la variable Y dependiente es la que controlamos y conocemos desde un principio.

La existencia de correlacion entre dos variables no implica causalidad (que saque buenas notas en mates y estadistica no significa que halla causa efecto entre las notas de una y otra)



7.3 El coefficiente de correlacion simple y su interpretacion

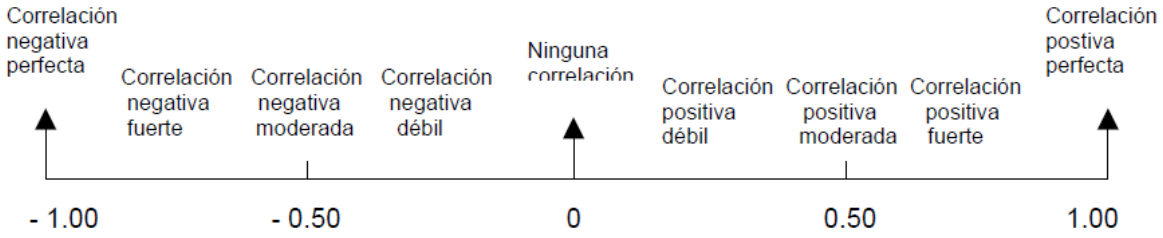
El coefficiente , de correlacion nos mide el valor de covariacion o variacion conjunta de dos series de datos (o sea traducido la relacion o no relacion entre dos variables representadas de manera numerica). Puede ser una relacion directa (valores positivos) o inversa (valores negativos). La relacion depende de diversos factores, del procedimento de calculo , de la calidad de medida de las variables. Y la correlacion (numerica) tambien depende del numero de personas de la muestra. Para interpretar los datos por norma general se utiliza unas reglas:

- Relacion entre variables cuando en otros estudios tambien hay valores similares del coefficiente.
- La variabilidad del grupo , los grupos que no son iguales no pueden interpretarse igual la relacion y cuanto mas homogeneo es el grupo mas correlacion hay , y la homogeneidad se veia con la varianza y la desviacion tipica.
- La fiabilidad a la que se destina el coefficiente depende de para que se quiera. Se necesita un coefficiente mayor para aceptar un instrumento de medida , que ha de ser al menos de un 0,85 , y un coeficiente menor para dar por validad una relacion con un 0,6 llega.

En la mayoria de las occasione se dice que

- Un coeficiente de 0,00 a 0,20 +- es una correlacion baja indiferente, despreciable.
- Un coeficiente de 0,20 a 0,40 positivo o negativo es una correlacion baja
- Un coeficiente entre 0,41 a 0,70 es una correlacion media , marcada , o notable

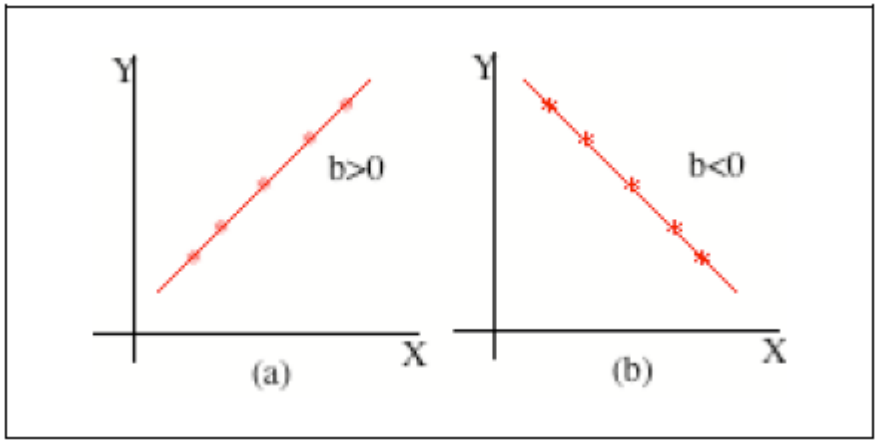
- Un coeficiente entre 0,71 a 0,90 tanto positivo como negativo es una correlacion alta , elevada fuerte.
- Un coeficiente entre 0,91 y 1 tanto positivo como negativo es una correlacion muy alta , elevada



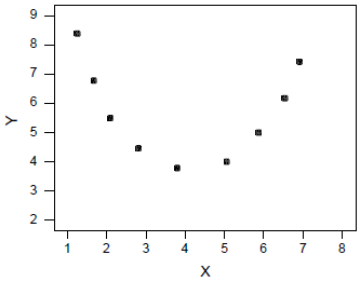
Este coeficiente como ocurre con mas puntuaciones, se suele transformar para que se vea mas guapo con numeros mas grandes y sin comas, por eso se transforma en el coeficiente de determinacion (d) donde el coeficiente de correlacion r se eleva al cuadrado (conviertiendose en positivo si es negativo) y se moltiplica por 100 (elimina comas)

7.4 El coeficiente de correlacion Perarson (r)

Tambien llamado correlacion producto momento. Es el mas famoso de todos los coeficientes (por tanto hay un gran riesgo que caiga en el examen como cayo en la PEC) Y se utiliza cuando dos variables estudiadas son cuantitativas (numericas) , continuas (por intervalo) o discreta(enteros) , con una distribucion normal y estando linearmente relacionada. Vamos si el grafico de dispersione es parecido a esto



Aunque no tiene que ser tan lineal los datos pueden estar algo mas alejados de la linea no encima encima como en este ejemplo. . Porque todo este rollo porque hay variables relacionadas pero con forma de U , que no nos servira para hacer una pearson, o sea que si tiene su grafico de dispersion esta forma no nos sirve



Hay diferentes formulas segun el tipo de puntuacion con las que se trabaja , puntuaciones directas (valor real), diferenciales (indica distancia de un valor de la media) y tipica (indica lo mismo que la diferencial pero tomando como unidad de medida la desviacion tipica).

Las formulas son las siguientes

Para puntuaciones directas

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

Para puntuaciones diferenciales

$$r_{xy} = \frac{\sum xy}{\sqrt{\sum x^2 \sqrt{\sum y^2}}}$$

Ss valores deben estar entre 0 y 1 y se epresa para dos variables x e y con el simbolo rxy.

En resumen es una formula complicada no hace falta saberla de memoria solo entenderla y saber aplicarla , para lo cual tendremos qeu crear una tabla con tres columnas nuevas (nunca vistas) y por supuesto continuar haciendo la fila del sumatorio∑ que tanto nos gusta(je ,je). Las columnas seran X^2 Y^2 XY , y recordar estimo hablando de variables una sera la X y otra la Y, no hay frecuencia absoluta , para el primer alumno saco un 105 del Ci y un 9 en mates , otro alumnos saca un 95 de CI y un 10 en mates , otro saca un 105 (se pueden repetir los valores de la variable) y un 6 en mates, por eso es aconsejable al aplicarla crear una tabla nueva solo con estos datos X Y X^2 Y^2 XY porque sino es imposible introducirla en la tabla que ya hemos hecho para las frecuencias y desviaciones.

Otra cosa cuando en la formula aparece $-(\sum x)^2$ significa que todo el resultado de la suma de x lo pones al cuadrado y luego lo restas , no se puede confundir con $\sum x^2$ que el la suma de todos

los resultados de x^2 , aunque parece lo mismo no lo es. Un ejemplito para ver como se hace no viene mal o sea que aqui va

Tenemos estos datos

X:	105	116	103	124	137	126	112	129	118	105
Y:	4	8	2	7	9	9	3	10	7	6

Creamos esta tablita

X	Y	X^2	Y^2	XY
105	4	11025	16	420
116	8	13456	64	928
103	2	10609	4	206
124	7	15376	49	868
137	9	18769	81	1233
126	9	15876	81	1134
112	3	12544	9	336
129	10	16641	100	1290
118	7	13924	49	826
105	6	11025	36	630
1175	65	139245	489	7871

Y ahora la formulita

$$Rxy = \frac{10x\ 7871-1175x65}{\sqrt{[(10x\ 139245)-(1175)^2][10x489-(65)^2]}} = 0{,}8326....$$

La dificultad de esta formula no es su realizacion en si es sencilla pero el problema son los numeracos que salen , mientras la realizas , no te preocupes si salen numeros muy grandes es lo normal en esta formula lo unico apuntalo todo y compruebalo un par de veces y seguro que te sale bien

7.5 Coeficiente de correlacion de los rangos de Spearman

Es una prueba estadistica para medir la asociacion (correlacion 9 lineal de ldos variables y se aplica cuando las mediciones se realizan en una escala ordinal, aprovechando la clasificacion por rangos. Se rige por las mismas regras que la Pearson. La medicion de este indice tiene que estar entre ±1 y 0 . Donde 0 es la no correlacion entre variables y 1 es la maxima correlacion . En si es muy sencillo solo tenemos que tener en cuenta el rango (traducido la posicion de un valor con respecto a los otros ordenandolos de menor a mayor) El coeficiente Spearman es recomendable cuando los datos presentan alores externos o ante distribuciones no normales. Este coeficiente nos da un rango que nos permite identifica facilmente el grado de correlacion de dos variables , con esta formula

$$1-\left(\frac{6\Sigma d^2}{n(n^2-1)}\right)$$

que es lo mismo que esto $r_s = 1 - \frac{6\ \Sigma\ d^2}{N3 - N}$ pero con menos pasos. Porque los parentesis $n(n^2-1)$ si desacemos el parentesis nos quedara n^3-n que es lo que aparece en la otra formula.

Ya vereis como haciendo un ejemplo lo encontrareis super sencillo ademas va por pasos y son siempre los mismos

1°- Necesitaremos hacer una tabla de 6 columnas , con el siguiente encabezado

Data 1	Data 2	Rank 1	Rank 2	d	d^2

Las filas seran las necesarias tantas como miembros hay en la muestra.

2° LLena las primeras columnas con los valores de cada individuo

Data 1	Data 2	Rank 1	Rank 2	d	d^2
6	2				
4	9				
7	3				

3° En la tecera columna clasifica los datos de la primera columna del 1 hasta n (el numero de individuos que hay) Comienza con el mas bajo , que se le da el valor 1,….. Y lo mismo se hace con el rango dos de la variable 2

Data 1	Data 2	Rank 1	Rank 2	d	d^2
6	2	2	1		
4	9	1	3		
7	3	3	2		

Si hay dos datos iguales se les dara a cada uno un valor y se hara la media aritmetica y ese sera el valor que se pone ejemplo

Data 1	Rank 1
4	1
5	2
5	3
6	4

Becomes

Data 1	Rank 1
4	1
5	2.5
5	2.5
6	4

4° En la columna d ponemos la diferencia o sea la resta entre el rango 1 y el rango 2 , sin importar el signo negativo

Data 1	Data 2	Rank 1	Rank 2	d	d^2
6	2	2	1	1	
4	9	1	3	2	
7	3	3	2	1	

5° Rellenar la columna de d^2 haciendo el cuadrado de cada casilla de d

Data 1	Data 2	Rank 1	Rank 2	d	d^2
6	2	2	1	1	1
4	9	1	3	2	4
7	3	3	2	1	1

6° paso suma todos los valores de la columna d^2 y eso sera $\sum d^2$ de la formula o sea 1+4+1=6

El resultado sera

$$1-\left(\frac{6\Sigma d^2}{n(n^2-1)}\right)=1-\left(\frac{6\times 6}{n(n^2-1)}\right) \quad 1-\left(\frac{6\times 6}{3(3^2-1)}\right)=-0.5$$

=

Asi vemos que la unico que exige este coeficiente de correlacion es transformar las puntuaciones en rangos (mayor o menor)

7.6 El coeficiente de contingencia C

Esta prueba estadistica es una alternativa para cuando las variables no pueden ser ordenadas sino unicamente clasificadas (nominales o atributivas) . Es habitual en el campo de la educacion. Este coeficiente necesita para hallarlo tres pasos pirmero hallar las frecuencias esperadas, despues con estas frecuencias esperadas hallar el chi cuadrado y por ultimo hallar el coeficiente utilizando el chi. Por tanto este coeficiente no se hace con una sola formula sino con tres en un orden concreto.

Primero la frecuencia esperada su formula es la sigiente

$$Fe = \frac{f_f \times f_c}{F_t}$$

Donde f_f es el numero de sujetos de la fila

Donde f_c es el numero de sujetos de la columna

Donde ft es el numero total de sujetos

La segunda formula es la del Chi cuadrado

$$\chi^2 = \sum \left[\frac{(f_0 - f_e)^2}{f_e} \right]$$

Donde Fo = a la frecuencia observada

Fe= la frecuencia esperada que hallamos antes

Por ultimo esa la formula del coeficiente de contingencia C

$$C = \sqrt{\frac{X^2}{X^2 + N}}$$

El chi cuadrado mide el grado de discrepancia que se manifesta entre frecuencias observadas o empiricas fo con las qeu realmente se ven en la muestra. Para ello es necesario colocar las dos variables y hacer el sumatorio de cada fila no solo de cada columna. Porque necesitaremos hallar la fe frecuencia esperada que es la f_f la suma de los sujetos de la fila para moltiplicarle el

numero de sujetos de la columna y dividirlo entre la ft que es el numero de sujetos totales. Y esto se ha de hacer para cada casilla de la tabla

Mejor se ve con este ejemplo

aprobado	matematicas	lengua	total∑
si	45	24	69
no	15	16	31
total∑	60	40	100

Fe1=69x60/100=41,4

Fe2=69x40/100=27,6

Fe3 =31x60/100=18,6

Fe4 =31x40/100 =12,4

Con estas operaciones conseguiremos todas las frecuencias esperadas y una vez determinadas podremos hacer el ji o chi cuadrado sumando todas las fo(frecuencias observadas) restandoles las fe (frecuencias esperadas) el resultado poniendolo al cuadrado y dividiendolo entre la frecuencia esperada. Es muy largo y un poco complicado pero con paciencia se hace si os consuela pensar que mas dificil es escribirlo con word como lo tengo que hacer aqui abajo)

$$\text{Chi cuadrado} = \frac{(45 - 41{,}4)^2}{41{,}4} + \frac{(24 - 27{,}6)^2}{27{,}6} + \frac{(15-18{,}6)^2}{18{,}6} + \frac{(16- 12{,}4)^2}{12{,}4} = 2{,}51$$

Vamos pues esta operacion tan larga es el chi cuadrado con el que porfin hallaremos el coeficiente de correlacion c , vamos lo que nos interesa en el fondo y cuya formula os recuerdo era esta $C = \sqrt{\frac{X^2}{X^2 + N}}$ o sea que es igual a la division del chi cuadrado entre el chi cuadrado mas el total de la muestra , y a cuyo resultado le haremos una raiz cuadrada (porque el chi es cuadrado se ve en el signo por eso para hallar el coeficiente le tenemos que quitar el cuadrado que solo se puede hacer por medio de una raiz) . Vamos ya vereis como con el ejemplo es mas facil

$$C = \sqrt{\frac{2{,}51}{2{,}51 + 100}} = 0{,}156$$

Como es muy complicado pondre otro ejemplo sencillo para que practiqueis

Queremos determinar si existe relación entre el sexo y la especialidad cursada por alumnos que estudian Magisterio, a partir de los datos correspondientes a 349 alumnos de una Escuela de Magisterio. La distribución conjunta de frecuencias para ambas variables aparece en la tabla

Distribución conjunta de frecuencias para sexo y especialidad

	Ciencias	**Humanas**	**Lenguas**	**Preescolar**	

Hombres	70	60	36	12	178
Mujeres	40	54	39	38	171
	110	114	75	60	349

Ahora haremos las frecuencias esperadas para no poner todas las operaciones y no volverme tarumba del todo pondre solo el resultado en una tabla junto a su valor real entre parentesis.

Frecuencias observadas y esperadas para sexo y especialidad

	Ciencias	**Humanas**	**Lenguas**	**Preescolar**	
Hombres	70 (56.1)	60 (58.1)	36 (38.3)	12 (25.5)	178
Mujeres	40 (53.9)	54 (55.9)	39 (36.7)	38 (24.5)	171
	110	114	75	50	349

A partir de las frecuencias observadas y esperadas podremos aplicar la fórmula de cálculo para χ^2 y obtener un valor que puede ser tomado como medida de independencia entre las dos variables. Si las frecuencias empíricas (las observadas en este caso) resultaran ser iguales que las frecuencias teóricas (las que aparecen entre paréntesis), diremos que no existe relación entre las variables sexo y especialidad. Cuanto más se alejen las frecuencias teóricas de las observadas, mayor será la relación entre las dos variables. El valor χ^2 se construye a partir de la distancia entre las frecuencias observadas y las frecuencias esperadas, es decir, indica en qué medida la distribución de frecuencias se aleja de los valores que cabría esperar en el caso de que no hubiera relación entre las dos variables.

$$\chi^2=\frac{(70-56.1)^2}{56.1}+\frac{(60-58.1)^2}{58.1}+\frac{(36-38.3)^2}{38.3}+\ldots+\frac{(38-24.5)^2}{24.5}=22.006$$

El valor de χ^2 presenta problemas como medida de correlación, puesto que su cuantía depende del número de sujetos considerados. A medida que se incrementa n, crece también el valor de χ^2. Si dispusiéramos del doble de alumnos en cada celda de la tabla de contingencia, el valor de χ^2 sería también el doble.

Precisamente, para evitar el efecto del tamaño de la muestra, utilizamos como coeficiente de correlación el coeficiente de contingencia C:

$$C=\sqrt{\frac{\chi^2}{n+\chi^2}}=\sqrt{\frac{22.006}{349+22.006}}=0.24$$

La interpretación de la correlación habrá de hacerse examinando la tabla de contingencia. Puesto que en las celdas hombres-Ciencias, hombres-Humanas, mujeres-Lenguas y mujeres-Preescolar se observan frecuencias por encima de lo esperado, la posible relación entre las dos variables se concretaría en una tendencia a que estas parejas de modalidades se den conjuntamente. Es decir, parece existir una asociación entre alumnos y las especialidades de Ciencias y, en menor medida, Humanas, así como entre alumnos y las especialidades de Lenguas y, sobre todo, Preescolar.

E l coeficiente de contingencia c por el contrario que el resto de coeficiente nunca puede llegar a 1 porque el denominador(lo que divide el chi mas la muestra) es mayor que el numerador (solo el chi lo que es dividido). Que significa esto que a la hora de explicar los resultados no podremos comparalo por lo que se acerca o aleja a 1 sino por lo que se acerca o aleja del c maximo. Que es

el numero maximo al que puede llegar el contingente. Por eso para interpretarlo recurrimos al Cmax. Que solo se usa en las tablas de contingencias cuadradas (o sea cuando en la tabla el numero de filas y columnas son iguales , para los ejemplos que he puesto solo podria ser utilizado en el primero no serviria para el segundo porque tiene cuatro columnas y solo dos filas) y la formula para hallar este C maximo es

Cmax= $\sqrt{\frac{c-1}{C}}$. c es el numero de columnas y de filas

Este deficiente de correlacion maximo seria el maximo que puede alcanzar el coeficiente de contingencia c ya que como dijimos al principio no puede alcanzar 1 y si bien el resto de coeficientes se valoraba respecto a 1 como ocurria en el de Spearman y el de Pearson , este coeficiente c se compara con el coeficiente maximo

7.7 El coeficiente de correlacion biserial puntual (pbp)

Se utiliza cuando la naturalezza de las variables es distinta , asi se utiliza para dos variables, una de las cuales es de naturalezza continua y se presenta en intervalo y la otra es discreta y se presenta de forma dicotomica. Y es un desarrollo de coeficiente de correlacion producto momento de Pearson. El termino biserial se refiere a que esiste dos series de observaciones en x, las puntuaciones de Y que son 0 y 1 (Glass y Stanley). Es bastante usual en educacion permite relacionar variables como el sexo, la asistencia, niveles de escolaridad superacion de asignaturas…(con formato dicotomica) con otras variables recurrentes en educacion, CI, rendimento, rasgos personalidad, motivacion. Hay varias formas de hacerlo con dos formulas.

$$r_{bp}=\frac{\bar{X}_p-\bar{X}}{S_x}\sqrt{\frac{p}{q}}\;;\qquad r_{bp}=\frac{\bar{X}_p-\bar{X}_q}{S_x}\sqrt{p\cdot q}$$

P= proporcion de casos de una de las dos modalidades Y (Ej: hombre
Q= 1-p
$\bar{X}$p media de los casos que en la variable x psoseen la caracteristica p (media de nota de hombres por ejemplo)
$\bar{X}$q = media de los casos en la variable x posee la caracteristica q (ej: media nota mujeres)
Sx media de todos los datos
Este indice arroja en ocasiones unos resultados desconcertantes al dar resultados que van mas alla del +- 1 , lo cual no nos permitira esplicarlo, esto ocurre si se cumple uno de los supuestos exigidos para su utilizacion sobretodo si una variable no es continua(intervalo) . Tambien puede ser si la variable no se distribuye normalmente o que la muestra sea muy pequena y que la distribucion sea platicurtica o leptocurtica.

Considerando que en un aula universitaria los resultados obtenidos en una prueba de evaluación (variable X) y el sexo de los alumnos (variable Y), son los que aparecen recogidos en la tabla , determinar la correlación existente entre ambas variables. El sexo de los individuos se ha codificado como 1 cuando se trata de alumnos y 2 cuando se trata de alumnos.

x	18	12	14	16	14	9	20	16	17	14	12	10	15	16	13	12	19	20	15	16	14
y	1	1	2	2	1	1	2	2	2	1	1	1	2	2	1	1	2	2	1	1	1

Para determinar la correlación existente entre ambas variables, utilizaríamos el coeficiente de correlación biserial puntual. En primer lugar, calcularemos el valor de las proporciones de alumnos (p) y alumnas (q) teniendo en cuenta que en el grupo de 21 alumnos 12 son hombres (modalidad 1) y 9 mujeres (modalidad 2):

$$p = 12/21 = 0.57l4$$

$$q = 9/21 = 0.4285$$

A continuación calculamos los valores de la media de la variable X, la media de la variable X para los 12 sujetos de la modalidad 1 (en este caso los alumnos) y la desviación típica de X. Realizando los cálculos oportunos, que dejamos al lector, resulta:

$$\overline{X}=\frac{\Sigma X_i}{n}=14.86;\quad \overline{X}_p=\frac{\Sigma X_p}{n}=13.25;\quad S_x=\sqrt{\frac{\Sigma X_i^2}{n}-\overline{X}^2}=2.92$$

A partir de estos valores estamos en disposición de calcular el coeficiente de correlación biserial puntual. Aplicando una de las expresiones de cálculo de r_{bp}obtendremos:

$$r_{bp}=\frac{13.25-14.86}{2.92}\sqrt{\frac{0.57}{0.43}}=-0.635$$

Por tanto, el valor de la correlación entre ambas variables es -0.635. Al tratarse de un coeficiente de signo negativo, a puntuaciones altas en la variable X corresponde pertenecer a la categoría cuya proporción es q. Es decir, las puntuaciones altas en la prueba de evaluación se asocia a las alumnas; mientras que las puntuaciones bajas se asocian a los alumnos.

7.8 Otros coeficientes de correlacion

En educacion puede haber muchas situaciones diversas y por tanto muchos coeficientes diferentes que se adaptan mejor a cada situation.

7.8.1 El coeficiente phi φ

Busca la relacion entre dos variables dicotomicas y en algunos casos dicotomizadas es muy sencilla . Si asignamos los valores 0 y 1 a cada una de las dos modalidades de la variables dicotómicas X e Y, podremos construir una table, en la que quede reflejada la distribución conjunta de frecuencias para las dos variables.

		X	
		0	**1**
Y	**1**	a	b
	0	c	d

A partir de los valores a, b, c y d, que representan la frecuencia en cada una de las celdillas de la tabla, es posible calcular el coeficiente φ. Basta aplicar la siguiente fórmula:

$$\varphi = \frac{c \cdot b - a \cdot d}{\sqrt{(a+b) \cdot (c+d) \cdot (a+c) \cdot (b+d)}}$$

Como siempre todo se ve major con un ejemplo asi que aqui va

De un grupo de 200 estudiantes universitarios que han pasado una prueba objetiva, se sabe que 140 han acertado el ítem 34. Se sabe además que 30 varones han fallado, del grupo total de 80 varones. Determina el valor de la relación entre el sexo y el número de aciertos al ítem 34.

Consideraremos de una parte la variable sexo, con los valores 0 (hombre) y 1 (mujer), y de otra el resultado de la respuesta al ítem, con los valores 0 (error) y 1 (acierto). La tabla de contingencia con la que trabajamos puede completarse a partir de la información del enunciado .

		Ítem 34		
		0	**1**	
Sexo	**1**	30	90	120
	0	30	50	80
		60	140	200

Conociendo todos los valores de las celdas, podemos aplicar la fórmula del coeficiente φ:

$$\varphi = \frac{30 \cdot 90 - 30 \cdot 50}{\sqrt{(30+90) \cdot (30+50) \cdot (30+30) \cdot (90+50)}} = 0.13$$

El valor resultante no es muy elevado. El signo de la correlación indicaría que la tendencia observada es la asociación entre las modalidades 0 de cada variable y entre las modalidades 1. Así, acertar el ítem se asociaría a las mujeres y errarlo a los hombres

7.8.2 El coeficiente de correlacion tetracorico rt

El coeficiente de correlación tetracórica, expresado por r_t, se utiliza cuando las variables con las que trabajamos han sido dicotomizadas de manera artificial. Es más apropiado emplear el coeficiente φ cuando las variables son estrictamente dicotómicas, y recurrir a r_t cuando las variables, siendo original mente continuas, aparecen dicotomizadas.

El coeficiente r_t no es aplicación directa de r_{xy}, sino una estimación del valor de éste en el caso en que las dos variables no hubieran sido dicotomizadas y la relación entre ellas fuera lineal.

Se demuestra que el valor de r_t, viene dado por un complejo desarrollo en serie de potencias de r_t, que eludiremos presentar aquí. Sin embargo, como vía alternativa, el cálculo se ve enormemente facilitado por el uso de tablas que permiten encontrar el valor de r_t en función de las frecuencias alcanzadas para cada par de modalidades posibles.

Veamos cómo se procede al calcular el valor de este coeficiente. Si la distribución conjunta de frecuencias correspondiente a las variables X e Y es la que aparece en la tabla 9, obtenemos en primer lugar los productos ad y cb comparándolos entre sí, y construimos un cociente en el que el mayor de estos productos aparezca en el numerador:

si ad > cb, calculamos el cociente ad/cb.

si ad < cb, calculamos el cociente cb/ad.

		X	
		0	**1**
Y	**1**	a	b
	0	c	d

El valor obtenido para este cociente puede ser trasladado a la tabla 11 para determinar el coeficiente de correlación tctracórica que le corresponde. Conocido el valor de r_t, asignamos el signo con el siguiente criterio:

Si ad > cb, el coeficiente r_t es negativo.

Si ad < cb, el coeficiente r_t es positivo.

Las calificaciones obtenidas por los alumnos de un curso de Estadística (variable X) y su grado de cumplimiento con las tareas desarrolladas durante el curso (variable Y) han sido dicotomizadas del modo que muestra la tabla. Determinar la correlación existente entre ambas variables.

Datos correspondientes al ejemplo

		Desarrollo de tareas (Y)	
		Bajo(0)	**Alto(1)**
Calificaciones (X)	**Aprobado (1)**	2	10
	Suspenso (0)	7	4

Puesto que ad = 8 y cb = 70, se cumple que cb > ad, luego vamos a obtener un coeficiente de correlación positivo. Calculamos el cociente cb/ad:

$$\frac{c \cdot b}{a \cdot d} = \frac{7 \cdot 10}{2 \cdot 4} = \frac{70}{8} = 8.75$$

Consultando la tabla (*Coeficiente de correlación tetracórica en función de las frecuencias*, encontramos que 8.75 se encuentra comprendido en el intervalo [8.500, 8.910], por lo que el coeficiente de correlación tetracórica valdrá en este caso $r_t = 0.70$.

Un coeficiente positivo, como el que hemos obtenido, indica que a valores 1 de la variable X corresponden predominantemente valores 1 en la variable Y, y a valores 0 en X corresponden valores 0 en Y. Es decir, al aprobado corresponde un desarrollo alto de tareas, mientras que el suspenso se asocia a un nivel bajo en el desarrollo de tareas.

Tabla: Coeficiente de correlación tetracórica en función de las frecuencias

r_t	cb/ad o ad/cb	r_t	cb/ad o ad/cb	r_t	cb/ad o ad/cb
0'00	1'000	0'35	2'492-2'563	0'70	8'500-8'910
0'01	1'013-1'039	0'36	2'564-2'638	0'71	8'911-9'351
0'02	1'040-1'066	0'37	2'639-2'716	0'72	9'352-9'828
0'03	1'067-1'093	0'38	2'717-2'797	0'73	9'829-10'344
0'04	1'094-1'122	0'39	2'798-2'881	0'74	10'345-10'903
0'05	1'123-1'151	0'40	2'882-2'968	0'75	10'904-11'512
0'06	1'152-1'180	0'41	2'969-3'059	0'76	11'513-12'177
0'07	1'181-1'211	0'42	3'060-3'153	0'77	12'178-12'905
0'08	1'212-1'242	0'43	3'154-3'251	0'78	12'906-13'707
0'09	1'243-1'275	0'44	3'252-3'353	0'79	13'708-14'592
0'10	1'276-1'308	0'45	3'354-3'460	0'80	14'593-15'574
0'11	1'309-1'342	0'46	3'461-3'571	0'81	14'575-16'670
0'12	1'343-1'377	0'47	3'572-3'687	0'82	16'671-17'899
0'13	1'378-1'413	0'48	3'688-3'808	0'83	17'900-19'287
0'14	1'414-1'450	0'49	3'809-3'935	0'84	19'288-20'865
0'15	1'451-1'488	0'50	3'936-4'067	0'85	20'866-22'674
0'16	1'489-1'528	0'51	4'068-4'205	0'86	22'675-24'766
0'17	1'529-1'568	0'52	4'206-4'351	0'87	24'767-27'212
0'18	1'569-1'610	0'53	4'352-4'503	0'88	27'213-30'105
0'19	1'611-1'653	0'54	4'504-4'662	0'89	30'106-33'577
0'20	1'654-1'697	0'55	4'663-4'830	0'90	33'578-37'815
0'21	1'698-1'743	0'56	4'831-5'007	0'91	37'816-43'096
0'22	1'744-1'790	0'57	5'008-5'192	0'92	43'097-49'846
0'23	1'791-1'838	0'58	5'193-5'388	0'93	49'847-58'758
0'24	1'839-1'888	0'59	5'389-5'595	0'94	58'759-71'035
0'25	1'889-1'940	0'60	5'596-5'813	0'95	71'036- 88'964
0'26	1'941-1'993	0'61	5'814-6'043	0'96	88'965-117'479
0'27	1'994-2'048	0'62	6'044-6'288	0'97	117'480-169'503
0'28	2'049-2'105	0'63	6'289-6'547	0'98	169'504-292'864
0'29	2'106-2'164	0'64	6'548-6'822	0'99	292'865-923'687
0'30	2'165-2'225	0'65	6'823-7'115	1'00	923'688- ∞
0'31	2'226-2'288	0'66	7'116-7'428		
0'32	2'289-2'353	0'67	7'429-7'761		
0'33	2'354-2'421	0'68	7'762-8'117		
0'34	2'422-2'491	0'69	8'118-8'499		

7.8.3 Coeficiente de correlacion biserial (rb)

Se utiliza cuando se trata de ver relacines entre variables una quantitativa continua o discreta y otra dicotomizada . Las dos variables seran realmente cuantitativas pero una una se reducira a deos intervalos.

La ecuacion es

$$r_b = \frac{\bar{X}_p - \bar{X}_q}{\sigma_x} \bullet \frac{pq}{y}$$

$$r_b = \frac{\bar{X}_p - \bar{X}}{\sigma_x} \bullet \frac{p}{y}$$

Por eso necesitamos conocer

X es la variable continua

Y la dicotómica o dicotomizada.

p: proporción de casos de una de las dos modalidades de Y

q: 1 - p

y: ordenada de la curva normal que divide las áreas p y q. Ver Tabla 9.

$\bar{X}_p$ media de los casos que en la variable X poseen la característica p

$\bar{X}_q$ media de los casos que en la variable X poseen la característica q

$\bar{X}$ media de todos los casos en la variable X

σ_x desviación típica de la variable X

Un ejemplo típico de correlación biserial sería el siguiente:2

Hemos medido el peso (Y) y la estatura (X) a un grupo de individuos dividiéndolos según el peso en obesos (peso superior a la mediana) y delgados (peso inferior a la mediana) ¿hay alguna relación entre peso y altura así considerados?

X	Y(p)	Y(q)
176	1	
174	1	
172	1	
170	1	
169	1	
168		1
166		1
164		1
160		1
155		1

p=0,5, q=0,5, y=0,3989

$\overline{X}$ =167,4 σ x=6,40

$\overline{Xp}$ =172,20

$\overline{Xq}$=162,60
pq/y=0,6267
p/y=1,2530

$$r_b = \frac{\overline{X}_p - \overline{X}_q}{\sigma_x} \bullet \frac{pq}{y} = \frac{172{,}2 - 162{,}6}{6{,}4} \bullet \frac{0{,}5 \bullet 0{,}5}{0{,}3989} = 0{,}94$$

$$r_b = \frac{\overline{X}_p - \overline{X}}{\sigma_x} \bullet \frac{p}{y} = \frac{172{,}2 - 167{,}4}{6{,}4} = \frac{0{,}5}{0{,}3989} = 0{,}94$$

Por tanto, existe una alta relación entre tener valores altos en la variable altura y pertenecer al grupo de obesos y obtener valores bajos en altura y pertenecer al grupo de delgados.

7.9 La regresion lineal simple

A la hora de interpreter todos los coeficientes de correlacion anteriores se basan en cuanto se acerca a 0 o a +-1 como ya indicamos. Nos indica la variabilidad compartida la relacion existente entre el cambio de una variable en otra. Para su calculo se eleva el valor del coeficiente (Las formulas anteriores el resultado) al cuadrado obteniendose el coeficiente de determinacion, con el que podemos hallar los valores de una variable conociendo los de otra. Es lo que se llama regresion lineal simple (prediccion de una variable). Posible utilidad podemos predecir la nota que sacara un alumno segun las notas anteriores y si despues saca mas nota es que se ha esforzado. Francis Galton fue el primero en usarlo al tratar alturas de padres e hijos . El

procedimento de calculo de la cuantia de las prediciones recibe el nombre de regresion lineal. La formula es Y‘= ayx +byx ·xij
Ayx es una constante equivalente al valor de Y cuando x yes = a el
Byx nos indica la pendiente de esa recta Y sobre x o coeficiente x yes el valor de la variable predicatoria
Y ‘ij= a la puntuacion pronosticada