“Introducción a la Estadística, Probabilidad e Inferencia”

Maestría en Estadística Aplicada

Facultad de Ciencias Económicas y Estadística

UNR

Unidad 2 – Parte 2

- Análisis exploratorio de datos bivariados. Relaciones entre variables.
- Variables cuantitativas: diagramas de dispersión y medidas de asociación.
- Variables categóricas: tablas de contingencia y medidas de asociación.

La salud y el hábito de fumar…

En la actualidad, muchos estudios han concluido que el hábito del cigarrillo es perjudicial para la salud en muchos aspectos. Sin embargo, un estudio arrojó evidencia contradictoria.

En dicho estudio se les preguntó a 1314 mujeres británicas si fumaban o no. Luego de 20 años de seguimiento se observó si cada una de las mujeres había fallecido o continuaba con vida.

El objetivo perseguido por los investigadores era estudiar ***la existencia de relación*** entre el hábito de fumar y la sobrevida a un período de 20 años. Durante dicho período, fallecieron 24% de las fumadoras y 31% de las no fumadoras, lo cual sugeriría al hábito de fumar como un efecto protector…

Retomaremos este ejemplo más adelante.

Fuente: Ignoring a covariate: an example of Simpson's Paradox. *American Statistician, Vol 50, 340-341, 1996)*

Datos univariados y multivariados.

En las unidades anteriores se han presentado métodos para describir y analizar ***datos univariados***.

No obstante, es muy frecuente en la práctica encontrar situaciones como las del ejemplo, en las que el interés es analizar dos o incluso más variables sobre un mismo individuo u objeto de una población, y la forma en la que estas variables se relacionan entre sí.

Un conjunto de ***datos multivariados*** consiste de mediciones u observaciones de dos o más variables sobre un mismo individuo u objeto. Nos centraremos en el caso particular en el que se analizan dos variables, obteniéndose conjuntos de ***datos bivariados***.

Asociación: ***diremos que dos variables están asociadas cuando ciertos resultados de una de las variables son más probables de ocurrir con ciertos valores de la otra variable, que con otros.***

Análisis de datos bivariados.

En las unidades anteriores se ha hecho la distinción entre variables categóricas y variables cuantitativas, y también se ha hablado de la clasificación de las variables, según el rol que desempeñan en el análisis, en variable explicativa y variable respuesta.

Cuando se trabaja con datos bivariados, es importante reconocer tanto el tipo de variables que se tienen como (siempre que se pueda) el rol que cada una desempeña en el análisis para poder analizar adecuadamente la relación entre ellas.

*"…la **variable respuesta** es la variable de resultado sobre la cual se hacen las comparaciones y la **variable explicativa** es una variable que creemos que explica los resultados observados".*

Análisis de datos bivariados.

Cuando trabajamos con datos bivariados podemos encontrar tres situaciones diferentes:

1. **Una de las variables es categórica y la otra cuantitativa.** Por ejemplo, análisis del ingreso salarial y el género.
   - Cálculo de estadísticas descriptivas de la variable cuantitativa, para cada uno de los grupos definidos por la variable categórica.

2. **Las dos variables son categóricas.** Por ejemplo, hábito de fumar y sobrevida a los 20 años.
   - Tablas de contingencia y cálculo de proporciones condicionales

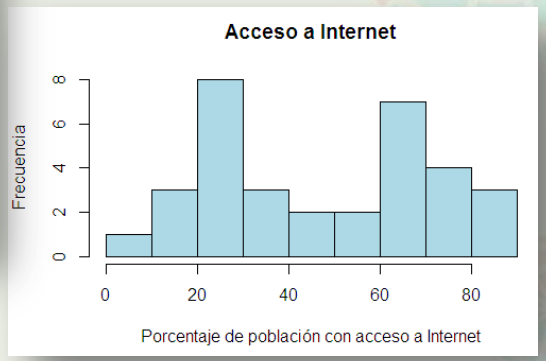
3. **Las dos variables son cuantitativas.** Por ejemplo, consumo diario de combustible para automóviles y nivel de contaminación en el aire.
   - Diagramas de dispersión y cálculo de coeficientes de correlación.

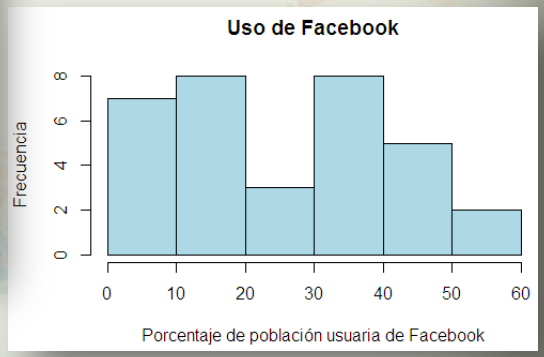
Asociación entre dos variables cuantitativas.

Ejemplo: Uso de internet y la red social Facebook

El número de usuarios de internet y de redes sociales como Facebook ha aumentado significativamente en la última década, aunque tal crecimiento no ha sido parejo alrededor del mundo.
Se tienen datos sobre el porcentaje de la población que tiene acceso a internet y el porcentaje de población que es usuaria de Facebook en 33 países alrededor del mundo *(Agresti, 2013, p.99)*.

```
##     Internet         Facebook
##  Min.   : 7.10   Min.   : 0.05
##  1st Qu.:24.90   1st Qu.:11.65
##  Median :49.40   Median :25.90
##  Mean   :46.89   Mean   :24.67
##  3rd Qu.:67.30   3rd Qu.:37.77
##  Max.   :82.90   Max.   :52.33
```



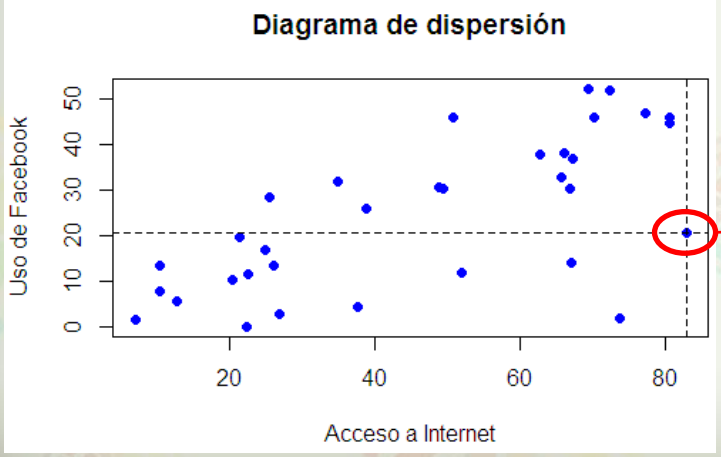


¿Cómo graficar ambas variables simultáneamente?

*Un **diagrama de dispersión** es una representación gráfica de dos variables cuantitativas simultáneamente, en la que se grafica la variable explicativa en el eje x y la variable respuesta en el eje y. Cada individuo u objeto queda representado en el diagrama mediante un punto ubicado en la combinación de los valores de ambas variables.*

Para los datos del ejemplo:

- ¿cuál es la variable explicativa y cuál la respuesta?
- ¿cómo describirían la relación entre ambas variables?

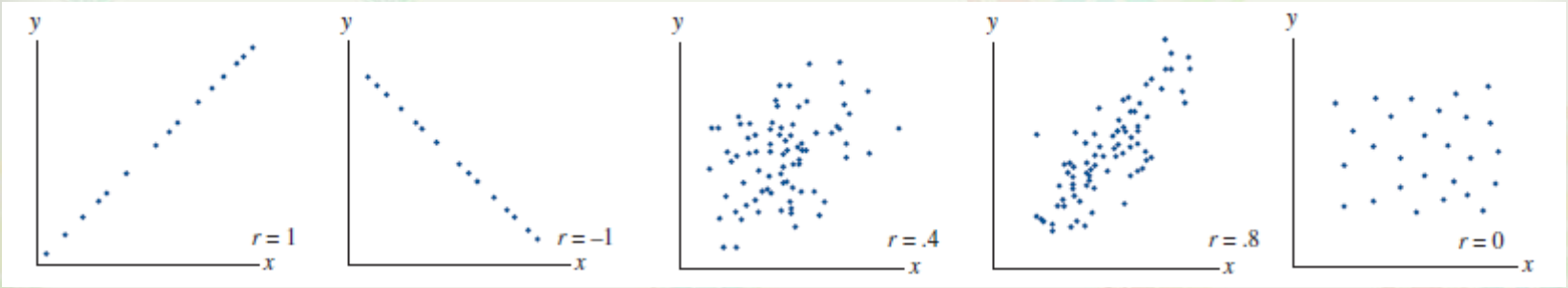


Punto correspondiente a Países Bajos:

- Acceso a internet: 82.90
- Uso de Facebbok: 20.54

¿Cómo resumir la relación entre las dos variables?

*El **coeficiente de correlación** es una medida de la intensidad y la dirección de la asociación lineal entre dos variables cuantitativas. Toma valores entre -1 y 1. Su signo indica la dirección de la asociación y su magnitud indica la fuerza o intensidad de la asociación.*



Fórmula para el coeficiente de correlación:

Existen varios tipos diferentes de coeficientes de correlación, siendo uno de los más utilizados el ***coeficiente de correlación de Pearson***:

- **Poblacional:**

$$\rho_{xy} = \frac{E\{[X - E(X)][Y - E(Y)]\}}{\sqrt{V(X)V(Y)}}$$

- **Muestral,** basado en una muestra de pares ordenados (x_i, y_i) con $i = 1, \ldots, n,$ de las variables X e Y:

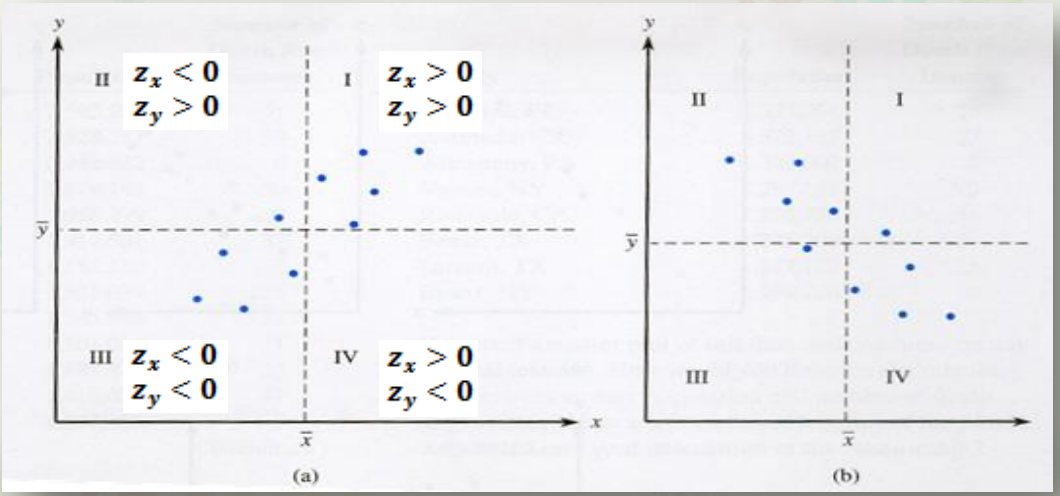
$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

Asociación entre dos variables cuantitativas.

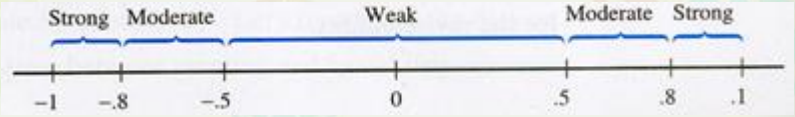
El coeficiente de correlación se basa en la suma de los productos de los valores estandarizados de x_i e y_i, en cada par.

Siendo $z_{x_i} = \frac{x_i - \bar{x}}{s_x}$ y $z_{y_i} = \frac{y_i - \bar{y}}{s_y}$, entonces $\boldsymbol{r = \frac{1}{n-1} \sum_{i=1}^{n} z_{x_i} z_{y_i}}$

Notar que los valores x_i más grandes que la media, tendrán un score Z_{x_i} positivo, mientras que los valores más chicos que la media tendrán un score negativo. Lo mismo sucede con los scores z_{y_i}.



Propiedades del coeficiente de correlación de Pearson:



- ***Varía entre -1 y 1***.
  Cuánto más cercano a 1 en valor absoluto sea el coeficiente, mayor es la fuerza de la asociación lineal. El valor cero indica ausencia de relación lineal entre las variables. Valores positivos del coeficiente indican asociación positiva (o directa) entre las variables, y valores negativos indican asociación negativa (o inversa).
- ***Es independiente de la unidad de medición de las variables***, puesto que para que su cálculo las variables se *estandarizan.*
- ***Es independiente de la elección del rol de las variables***, es decir, no importa cuál sea identificada como explicativa y cuál como respuesta.
- ***Sólo mide asociación lineal entre las variables.***

Para los datos del ejemplo resulta:

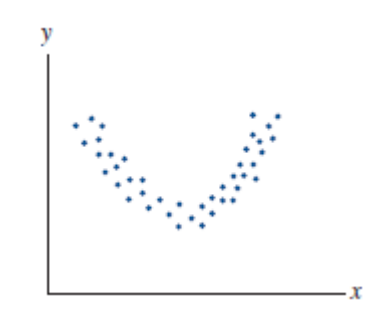
```
#Cálculo del coeficiente de correlación de Pearson
(cor(datos$Internet,datos$Facebook))
```

```
## [1] 0.681595
```

¿Cuándo es adecuado usar el coeficiente de correlación como medida de la asociación entre variables?

El coeficiente de correlación es una medida muy útil para resumir la información que los pares ordenados de valores proveen sobre la relación lineal entre las variables.
No obstante, debe recordarse que está diseñada para ***detectar sólo relaciones de tipo lineal***.

Precaución: el cálculo del coeficiente de correlación debe acompañarse siempre del diagrama de dispersión, de modo que pueda evaluarse si es apropiado el uso de este coeficiente para describir la relación.

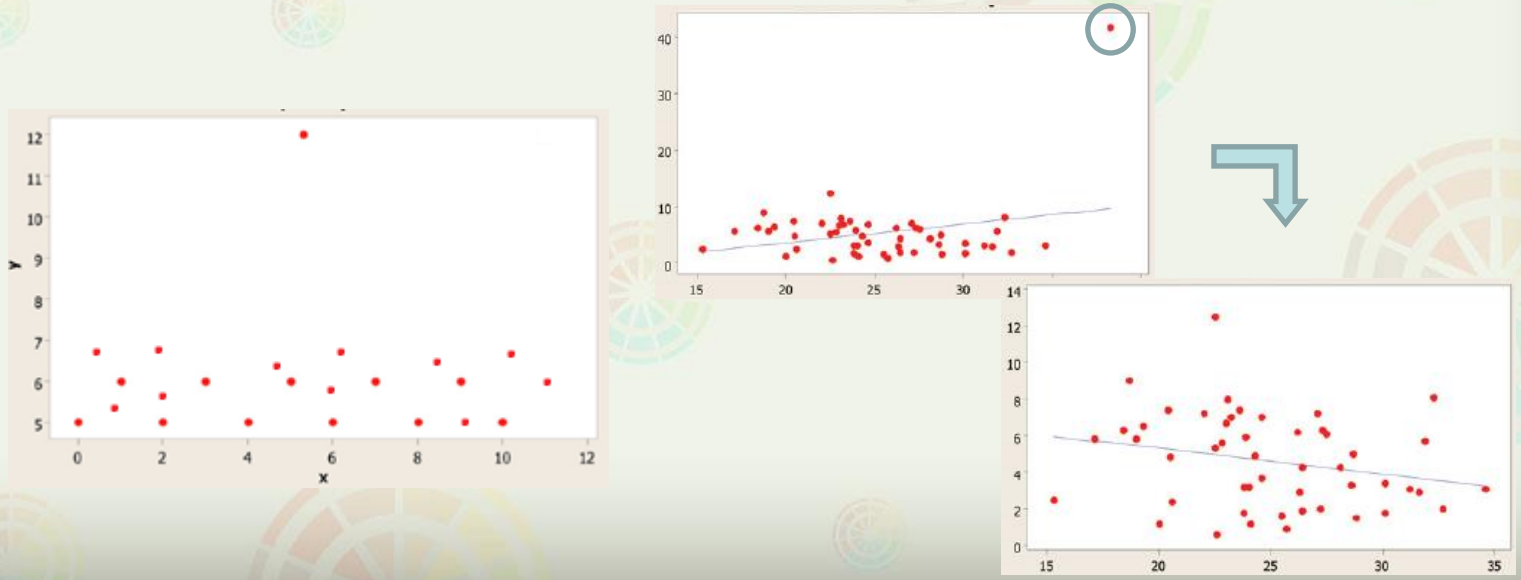


Los datos graficados arrojan un coeficiente de correlación igual a cero, aunque las variables están relacionadas entre sí. A medida que X crece, la variable Y primero tiende a decrecer y luego a crecer de nuevo, evidenciando una relación no lineal, por ende, no detectada por el coeficiente de correlación.

Precaución al analizar asociaciones

- **Extrapolación de resultados.** No debe concluirse sobre la relación entre las variables más allá del rango de valores de las variables que se haya estudiado. Nada asegura que la relación mantenga la misma forma por fuera del intervalo observado.

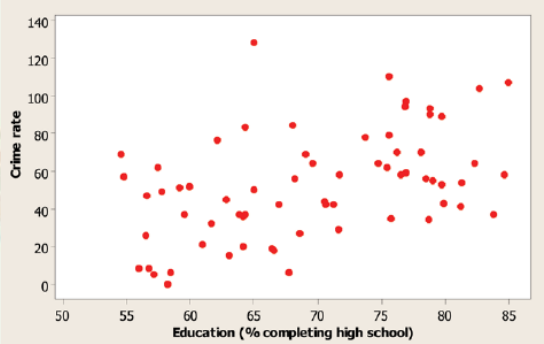
- **Presencia de valores atípicos influyentes.** Son valores que se alejan del patrón general del resto de los datos. Pueden llegar a tener una influencia grande en la determinación del coeficiente de correlación.

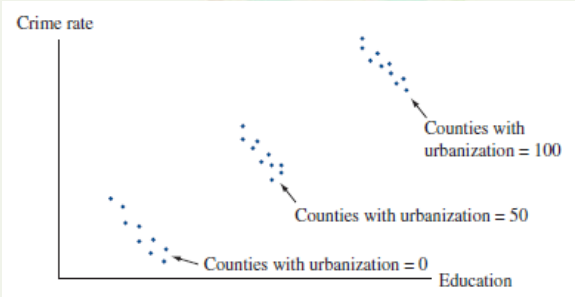


Precaución al analizar asociaciones

- **Correlación no implica causalidad.** Si bien existen situaciones en las que las correlaciones reflejan relaciones causales, esto no es necesariamente siempre así. La presencia de *variables latentes* y/o *variables de confusión* pueden generar que dos variables estén correlacionadas, aún cuando ello no tenga sentido.

Considérese el siguiente estudio en el que se midió para todos los condados de Florida el número de crímenes por cada 1000 habitantes en el último año (Y) y el porcentaje de residentes de al menos 25 años con al menos la preparatoria completa (X) (*Agresti, 2013, p.127*)





¿Pudo algo similar haber ocurrido en el ejemplo del hábito de fumar y la salud?

Otra medida de correlación lineal

Coeficiente de correlación de rangos de Spearman: Es una medida de la relación monótona entre dos variables, pero que, a diferencia del coeficiente de Pearson, no es sensible a la presencia de pares de valores extremos o inusuales.
Esto se logra trabajando sobre los ***rangos de las variables X e Y*** en lugar de los valores originales.

Para su cálculo se asignan rangos a cada valor de X y a cada valor de Y, y se trabaja luego sobre los pares de valores *rangueados*:

$$r_s = \frac{\sum_{i=1}^{n}\left(r_{x_i}\, r_{y_i}\right) - \frac{n(n+1)^2}{4}}{\frac{n(n-1)(n+1)}{12}}$$

Al igual que el coeficiente de Pearson, varía entre -1 y 1 y se interpreta de manera similar.

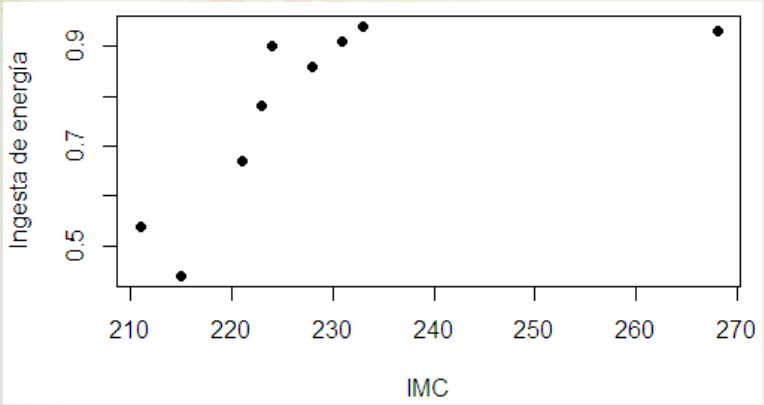
Coeficiente de correlación de rangos de Spearman

Ejemplo: se desea analizar la relación entre el índice de masa corporal (X) y la ingesta de energía en personas bajo dieta nutricional (Y).

Individuo	x	y	Rango x	Rango y	$R_x * R_y$
1	221	0.67	3	3	9
2	228	0.86	6	5	30
3	223	0.78	4	4	16
4	211	0.54	1	2	2
5	231	0.91	7	7	49
6	215	0.44	2	1	2
7	224	0.90	5	6	30
8	233	0.94	8	9	72
9	268	0.93	9	8	72

Coeficiente de correlación de rangos de Spearman

Ejemplo: se desea analizar la relación entre el índice de masa corporal (X) y la ingesta de energía en personas bajo dieta nutricional (Y).

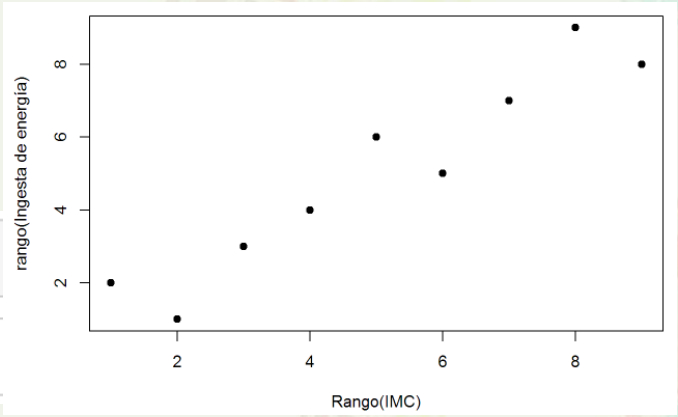



```
cor(x,y)
```

```
## [1] 0.6578901
```

```
cor(rank(x), rank(y))
```

```
## [1] 0.95
```



OBS: Existe una corrección que debe aplicarse en caso que existan valores “empatados ”.

Asociación entre dos variables categóricas.

Ejemplo: Pesticidas y comida orgánica

Una de las razones por las que los consumidores optan por la comida orgánica es que se asume que ésta no contiene pesticidas y que, por lo tanto, es mas saludable.
Dado el excesivo costo asociado a la comida orgánica, la Unión de Consumidores del estado de California (USA) llevó a cabo un estudio por muestreo para comparar la presencia de residuos de pesticidas en comida orgánica y en comida convencional.

¿Cuáles son las variables de interés? ¿De qué tipo son?
¿Puede identificarse una variable como explicativa y la otra como respuesta?
¿Cómo resumiría los datos del estudio?

Asociación entre dos variables categóricas.

Las variables categóricas pueden ser analizadas individualmente a través de sus distribuciones de frecuencias, pero si el interés es analizarlas en conjunto, es posible construir la ***distribución de frecuencias conjunta*** de ambas variables simultáneamente. Dicha representación recibe el nombre de *tabla de contingencia*.

*Una **tabla de contingencia** es una representación tabular de dos variables categóricas. En las filas se listan las categorías de una de las variables (la explicativa en general) y en las columnas las de la otra variable (la respuesta). Cada celda de la tabla contiene la frecuencia de unidades o individuos en la muestra que poseen esa combinación de categorías de ambas variables.*

El proceso de construcción de la tabla de contingencia a partir de los datos individuales se conoce como ***"tabulación cruzada"***.

La tabla de contingencia para los datos del ejemplo resulta:

	Residuos de Pesticida		Total
Tipo de comida	**Presentes**	**Ausentes**	
Orgánica	29	98	127
Convencional	19485	7086	26571
Total	19514	7184	26698

Distribución marginal de "Tipo de comida"

Distribución marginal de "Residuos de pesticida"

¿Cómo "*leer*" estos datos para analizar si están de acuerdo con la creencia de los consumidores o no?

Proporciones condicionales: $\frac{29}{127} = 0.2283$ **y** $\frac{19485}{26571} = 0.7333$

¿Qué representan estas proporciones?
¿Qué sugieren respecto de la creencia de los consumidores?

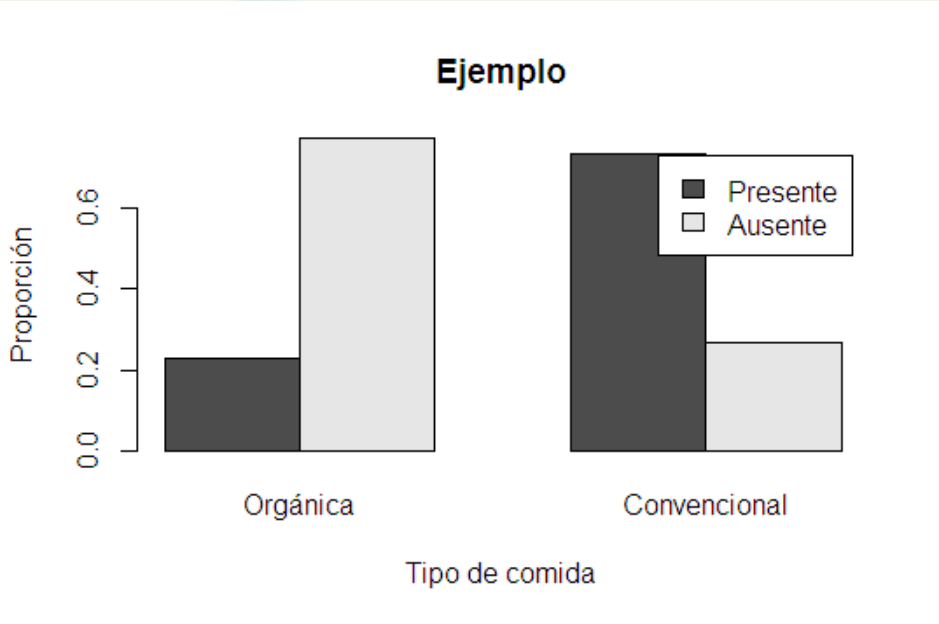
Estas proporciones se denominan *condicionales* porque brindan la distribución de una de las variables para niveles fijos o condicionado a los niveles de la otra variable.

Siempre que exista una distinción entre variable respuesta y explicativa, se acostumbra a calcular las proporciones condicionales de la variable respuesta para niveles fijos de la variable explicativa. En el ejemplo, interesa conocer la proporción de alimentos con y sin residuos de pesticida entre los alimentos orgánicos por un lado y entre los alimentos convencionales por otro.

Las proporciones calculadas a partir de los totales por fila o por columna se denominan ***proporciones marginales***, pues su cálculo representa la distribución de cada una de las variables, ignorando la información de la variable restante.

Asociación entre dos variables categóricas.

La información provista por una tabla de contingencia puede visualizarse también gráficamente.
Un ***gráfico de barras agrupadas*** permite representar las proporciones condicionales de interés en una tabla de contingencia y proporciona una herramienta visual para comparar tales proporciones.



¿Qué sugiere esta representación?

¿Cómo probar si las variables están asociadas?

Test de independencia

La pregunta que se intenta responder es si la ocurrencia de un determinado nivel de una de las variables *depende de* o está *asociado a* el nivel de la otra variable que se haya presentado.
En el ejemplo: ¿la presencia de residuos de pesticidas en los alimentos está asociado al tipo de alimento del que se trata?
Si las variables no están asociadas, se dice que son ***independientes***.

El ***test (o prueba) de independencia*** permite estudiar la presencia de asociación entre dos variables categóricas, mediante la comparación de las frecuencias observadas en la tabla de contingencia y *las esperadas bajo el supuesto de independencia*.

Existen distintos métodos para realizar el test de independencia, el más conocido y utilizado es el ***test χ^2 (Chi-cuadrado) de independencia***.

Test de independencia

$\boldsymbol{H_0}$) *Las variables son independientes*.
$\boldsymbol{H_1}$) *Las variables están asociadas*.

Estadístico de prueba:

$$X^2 = \sum \frac{(\boldsymbol{frec.observada} - \boldsymbol{frec.esperada})^2}{\boldsymbol{frec.esperada}}$$

Donde la suma se extiende a través de todas las celdas de la tabla de contingencia y las frecuencias esperadas se obtienen para cada celda (i, j), usando conceptos de probabilidad, mediante la fórmula:

$$\boldsymbol{Frec.esp._{ij}} = \frac{(\boldsymbol{total\ marginal\ fila\ i})(\boldsymbol{total\ marginal\ columna\ j})}{\boldsymbol{total\ general}}$$

Test de independencia

Las *frecuencias esperadas bajo independencia* son precisamente las frecuencias que se espera observar en la tabla de contingencia si realmente no existe asociación entre las variables, es decir, si ellas son independientes.

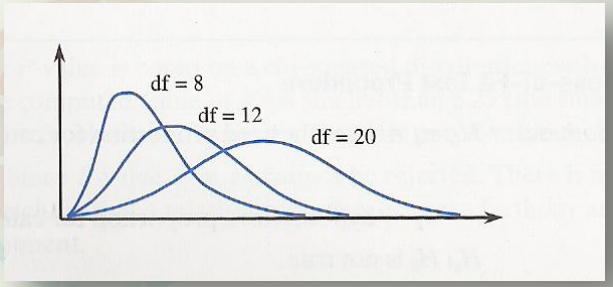
Por lo tanto, diferencias grandes entre lo realmente observado (frec.obs.) y lo esperado bajo independencia (frec.esp.) proveen evidencia a favor de la existencia de asociación entre las variables, mientras que diferencias pequeñas hablarán a favor de la hipótesis de independencia.

La estadística X^2 es una medida que cuantifica tales diferencias. A partir de su valor calculado en una situación particular, es posible calcular su ***probabilidad asociada p*** para compararla con el nivel de significación $\boldsymbol{\alpha}$ que se haya fijado y decidir si se rechaza o no la hipótesis de independencia.

Test de independencia

El cálculo de la probabilidad asociada se basa en la distribución bajo la hipótesis nula de la estadística de prueba, la cual puede aproximarse para tamaños de muestra grandes por la distribución χ^2 **(Chi-cuadrado)** con grados de libertad igual al producto entre el número de filas en la tabla menos 1 y el número de columnas en la tabla menos 1.

La probabilidad asociada a la estadística es el área a la derecha de su valor observado, bajo la curva de la distribución.



Precaución en el uso del test χ^2: Esta prueba es válida siempre que se cumpla una serie de condiciones:

- Los datos provienen de una muestra aleatoria, cuyos individuos son clasificados de acuerdo a dos variables categóricas.
- El tamaño muestral es suficientemente grande para asegurar que todas las frecuencias esperadas son mayores que cinco.

Test de independencia

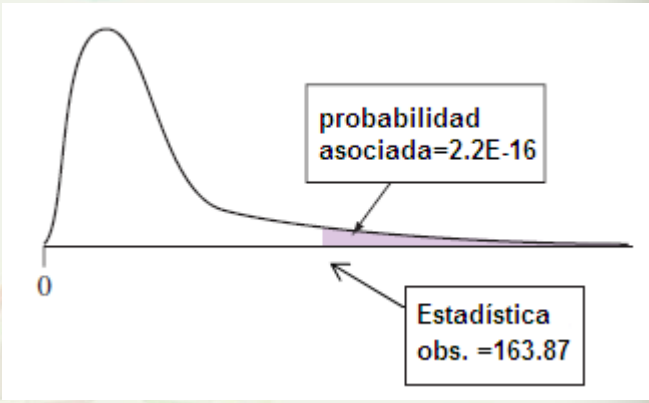
Para los datos del ejemplo:

```
# test de independencia chi-cuadrado
chisq.test(tabla1, correct=F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tabla1
## X-squared = 163.87, df = 1, p-value < 2.2e-16
```

Estadístico de prueba

gl=(2-1)(2-1)=1



Dado que la probabilidad asociada es menor que el nivel de significación $\alpha = 0.05$, se rechaza la hipótesis nula de independencia.

La presencia de residuos de pesticida en los alimentos está significativamente asociada al tipo de alimento.

Medidas de asociación

El test de asociación sólo provee información respecto de la presencia o ausencia de asociación entre las variables, pero en caso de concluir que las variables están asociadas, no ofrece ninguna información respecto de ***la fuerza*** ni ***el sentido*** de la asociación.

No es correcto interpretar valores grandes de la estadística X^2 (o de su probabilidad asociada) como evidencia de una asociación *"más fuerte"*. El test sólo brinda información acerca de cuán fuerte es la evidencia a favor de la asociación, pero nada dice de la fuerza o del grado de la asociación en sí misma.

De hecho, para un grado de asociación fijo entre dos variables, es posible aumentar tanto como se quiera el valor de la estadística del test, con sólo aumentar el tamaño de la muestra. Tamaños de muestra grandes, puede arrojar estadísticas del test con valores muy grandes, aún para asociaciones muy débiles.

Medidas de asociación

Una ***medida de asociación*** es una estadística que resume la fuerza de la dependencia entre dos variables.

Una de las medidas mas intuitivas es la ***diferencia de proporciones condicionales***, para una determinada categoría de respuesta. Por ejemplo, la proporción de alimentos con pesticida entre los orgánicos menos la proporción de alimentos con pesticida entre los convencionales.

Esta medida varía entre -1 y 1, siendo el cero el valor que indica independencia entre las variables.

Cuánto más cercana a 1 en valor absoluto sea la diferencia de proporciones, más fuerte será la asociación entre las variables.

Medidas de asociación

Otra de las medidas que se utiliza para reflejar la fuerza de una asociación, es el ***cociente de proporciones condicionales***. Este cociente es conocido como riesgo relativo en algunas aplicaciones, y resulta particularmente útil para evaluar la fuerza de una asociación, cuando las proporciones condicionales son cercanas a cero, en cuyo caso la diferencia de proporciones puede parecer insignificante.

Por ejemplo, supongamos que la proporción de muertes en accidentes de tránsito es de 0.00975 entre los conductores que no usan cinturón de seguridad y de 0.00124 entre los que sí lo usan. La diferencia de proporciones (0.00851) puede parecer despreciable.

Sin embargo, el cociente de proporciones: 0.00975/0.00124 = 7.86290, indica que ante un accidente de tránsito, la chance de morir de los conductores que no usan cinturón de seguridad es casi 8 veces mayor que la de los que sí usan cinturón de seguridad.

Medidas de asociación

El cociente de proporciones condicionales puede tomar cualquier valor real no negativo.

El valor unitario se obtiene cuando las proporciones condicionales son iguales, lo que implica que las variables son independientes. De este modo, valores cercanos a la unidad, representan asociaciones débiles.

Valores alejado de la unidad, en cualquiera de las dos direcciones, representan asociaciones fuertes. Notar que por tratarse de una medida relativa, dos valores representan la misma fuerza de asociación, cuando uno de los valores es el recíproco del otro.

Por ejemplo, un riesgo relativo de ¼, representa la misma fuerza de asociación que un riesgo relativo de 4: si $\frac{p_1}{p_2} = \frac{1}{4}$ entonces $p_1 = \frac{1}{4} p_2$ o bien $p_2 = 4\, p_1$.

Medidas de asociación

En el ejemplo de los pesticidas y la comida orgánica:

	Residuos de Pesticida		Total
Tipo de comida	Presentes	Ausentes	
Orgánica	29	98	127
Convencional	19485	7086	26571
Total	19514	7184	26698

$$\boldsymbol{Dif.de\ proporciones} = \frac{29}{127} - \frac{19485}{26571} = 0.2283 - 0.7333 = \mathbf{-0.505}$$

La proporción de alimentos con pesticida es menor entre los orgánicos que entre los convencionales.

$$\boldsymbol{RR} = \frac{29}{127} \Big/ \frac{19485}{26571} = \frac{0.2283}{0.7333} = \mathbf{0.311}$$

La chance de tener residuos de pesticida es casi un 70% menor entre los alimentos orgánicos que entre los convencionales (o la chance es 3.22 (1/0.311)veces mayor entre los convencionales que entre los orgánicos).

Frecuencias esperadas

Si el test de independencia se rechaza, entonces las frecuencias esperadas y observadas difieren. Conocer en cuáles celdas se producen esas diferencias puede ayudar a entender el sentido de la asociación.

Residuos estandarizados: son una medida estandarizada de la diferencia entre las frecuencias observadas y esperadas. Residuos de magnitud mayor a tres en valor absoluto indican celdas con una gran diferencia y que por lo tanto aportan al rechazo de la hipótesis de independencia. Por otra parte, el signo del residuo indica el sentido de la diferencia, teniendo en cuenta que el residuo se calcula en función de la frecuencia observada menos la esperada.

Ejemplo: En un estudio se estudió la asociación entre el sexo de las personas y la respuesta a la pregunta ¿cuán religioso se considera usted? Con opciones de respuesta: muy religioso, moderadamente religioso y poco religioso.

Frecuencias esperadas

Los valores entre paréntesis en la tabla de contingencia son los residuos estandarizados.

	Muy religioso	Moderad. religioso	Poco religioso	No religioso	Total
Femenino	241 (4.513)	499 (4.016)	208 (-4.606)	131 (-4.916)	1079
Masculino	133 (-4.513)	344 (-4.016)	258 (4.606)	186 (4.916)	921
Total	374	843	466	317	2000

En la primera fila, los residuos positivos (y mayores a tres) indican que se han observado más mujeres muy o moderadamente religiosas que las que se esperaban. Lo contrario ocurre entre los hombres, los residuos negativos indican que se observaron menos hombres muy o moderadamente religiosos que los esperados bajo independencia.

Puede interpretarse entonces que el hecho de ser religioso está asociado al sexo de las personas, siendo las mujeres las que tienden a ser mas religiosas.

Test exacto de Fisher

El test exacto de Fisher permite analizar si dos variables dicotómicas están asociadas (tablas de contingencia de dimensión 2x2). Dado que es un test exacto puede aplicarse para cualquier tamaño muestral, aunque su uso más habitual es como ***alternativa al test chi-cuadrado cuando el tamaño muestral no es suficientemente grande***.

Existen extensiones de este test para tablas de cualquier dimensión r x c, aunque no todos los softwares lo tienen implementado.

Ejemplo:

	Opción A	Opción B
Femenino	3	7
Masculino	10	5

```
fisher.test(tabla, alternative="two.sided")
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tabla
## p-value = 0.1107
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.02591934 1.55789518
## sample estimates:
## odds ratio
##   0.229258
```

```
test<-chisq.test(tabla, correct=F)
```

```
## Warning in chisq.test(tabla, correct = F): Chi-squared approximation may be
## incorrect
```

Precauciones

Asociación no implica causalidad. Al igual que en el caso cuantitativo la presencia de asociación entre dos variables categóricas no implica una relación de causa-efecto entre ellas.

La presencia de variables de confusión podría influenciar la asociación entre dos variables.

Un ejemplo de esto es el conocido caso de la denuncia por discriminación por género a la Universidad de California.

Los resultados globales de las admisiones del año 1973 señalaron una asociación significativa entre el género de los postulantes y el resultado de su solicitud de admisión.

	Admitidos	No admitidos	Total
Femenino	1512 (35%)	2809 (65%)	**4321**
Masculino	3714 (44%)	4728 (56%)	**8442**
Total	**5226**	**7537**	**12763**

```
tabla<-matrix(c(1512, 3714, 2809, 4728),2,2)
(test<-chisq.test(tabla, correct=F))
```

```
## data:  tabla
## X-squared = 95.793, df = 1, p-value < 2.2e-16
```

```
test$stdres
```

```
##           [,1]      [,2]
## [1,] -9.787375  9.787375
## [2,]  9.787375 -9.787375
```

Precauciones

Pero la Universidad está compuesta por diversos departamentos que administran sus solicitudes de admisión independientemente.

¿Qué sucede al considerar los datos discriminados por departamento?

Dpto. A	Admitidos	No admitidos	Total
Femenino	89 (82%)	19 (18%)	**108**
Masculino	511 (62%)	314 (38%)	**825**
Total	**600**	**333**	**933**

```
tabla<-matrix(c(511,89, 314, 19),2,2)
(test<-chisq.test(tabla, correct=F))
```

```
##  Pearson's Chi-squared test
##
## data:  tabla
## X-squared = 17.431, df = 1, p-value = 2.98e-05
```

```
test$stdres
```

```
##           [,1]      [,2]
## [1,]  4.175011 -4.175011
## [2,] -4.175011  4.175011
```

Dpto. B	Admitidos	No admitidos	Total
Femenino	17 (68%)	8 (32%)	**25**
Masculino	353 (63%)	207 (37%)	**560**
Total	**370**	**215**	**933**

```
tabla<-matrix(c(17, 353, 8, 207),2,2)
(test<-chisq.test(tabla, correct=F))
```

```
##  Pearson's Chi-squared test
##
## data:  tabla
## X-squared = 0.25372, df = 1, p-value = 0.6145
```