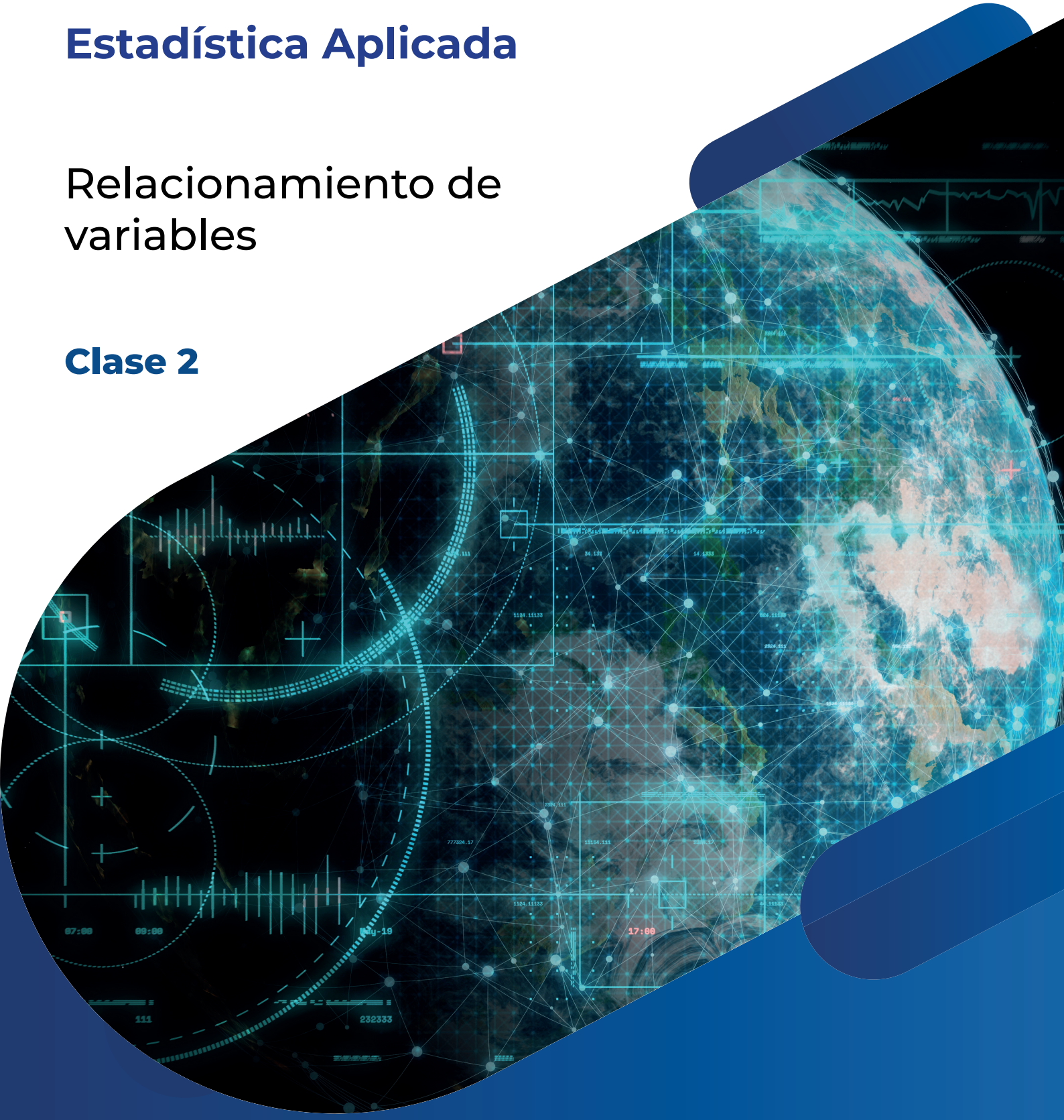
Estadística Aplicada

Relacionamiento de variables

Clase 2



Ingeniería en ciberseguridad

La excelencia no se improvisa

1. INTRODUCCIÓN DE LA CLASE

En esta clase, iniciaremos con el estudio de las relaciones entre variables; específicamente, trataremos sobre distintas funciones y algoritmos para medir la correlación y covarianza entre dos o más variables, así como técnicas para visualizar la correlación.

Un tema importante en la estadística es tratar de entender cómo se relacionan las variables entre sí. Para ello, introduciremos los temas de correlaciones, covarianza y causación. Lo anterior estará apoyado por el conocimiento del marco general del lenguaje R y la interfase RStudio, los mismos que se usarán en los laboratorios propuestos y que reforzarán el aprendizaje incorporando algunos enlaces externos. Un tema particular en este caso es la determinación de la existencia de relaciones espurias.

2) Relacionamiento de variables

En la clase anterior, revisamos la clasificación de las variables desde el punto de vista de los tipos de datos; es así que mencionamos la distinción entre variables continuas y discretas, o variables nominales y ordinales. Ahora presentaremos otra clasificación basada en la relación que se observa entre estas variables, en base al concepto de causa-efecto. La idea es entender si una variable es la causante de que otra variable cambie de valor.

VARIABLES PREDICTORAS Y DE RESPUESTA

Predictoras (anteriormente llamadas variables independientes) son variables que representa un argumento para obtener un efecto sobre otra variable.

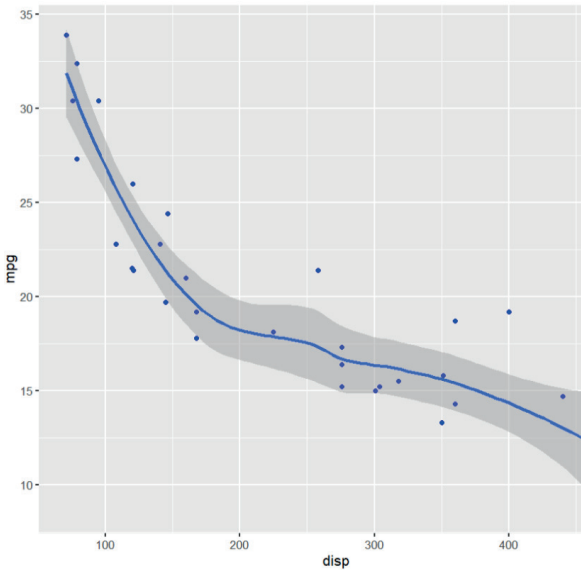
Variables de Respuesta (anteriormente llamadas dependientes) son variables que obtienen su valor en base a variables predictoras.

Para entender mejor el tema imagine la representación de una función matemática en un plano cartesiano Y=f(x), donde la variable Y tomará un valor derivado de la variable X y no al revés.

Por ejemplo, en R tenemos un dataset llamado *mtcars*, el cual contiene una decena de variables relacionadas con características de vehículos. Dos de estas características son: disp (que representa el volumen de los cilindros de un vehículo, establecido en centímetros cúbicos) y mpg (el rendimiento o millas por galón), y queremos entender qué tipo de relación tiene este par de variables.

Figura 1

Relación disp vs mpg , Creación de autor: Alfonso Prado



De la figura 1 notamos que a medida que los cc. del motor aumenta el rendimiento de consumo de gasolina disminuye.

En estadística, una relación monótona entre dos variables se refiere a un escenario en el que un cambio en una variable generalmente se asocia con un cambio en una dirección específica en otra variable.

Hay dos tipos de relaciones monótonas:

Monotónico positivo: Cuando el valor de una variable aumenta, el valor de la otra variable tiende a aumentar también.

Monotónica negativa: Cuando el valor de una variable aumenta, el valor de la otra variable tiende a disminuir

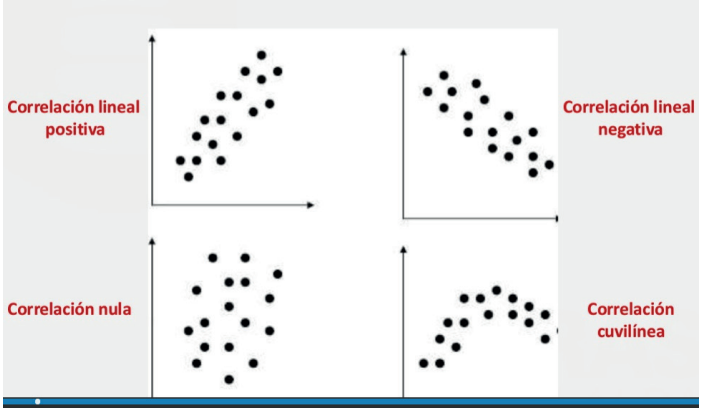
No monotónica: Si dos variables generalmente no cambian en la misma dirección, entonces se dice que tienen una relación no monótona.

Cuando analizamos la relación entre dos variables, también es importante distinguir entre un comportamiento lineal y no lineal. En algunos casos, podemos observar que dicha relación es monotónica positiva solo en cierto rango y luego pasa a ser monotónica negativa, o simplemente se estabiliza en cierto valor.

En base a estos dos conceptos podemos establecer las siguientes posibilidades:

Figura 2

Distintos tipos de correlaciones



2.1) Medición de la correlación y covarianza

Una vez que hemos entendido la relación entre dos variables, nos interesa cuantificar qué tan fuerte es dicha relación; para esto, existen distintos métodos para cuantificarla. Estos métodos podemos distinguirlos en dos clases.

PRUEBAS PARAMÉTRICAS Y NO PARAMÉTRICAS

La diferencia principal entre estos dos tipos es su base matemática para el cálculo. Las pruebas paramétricas se basan en el entendimiento de la distribución de las variables. Este concepto lo veremos más adelante en el curso. Por el momento, diremos que las pruebas paramétricas asumen una cierta distribución.

Ventajas de las Pruebas Paramétricas

- ✓ Tienen más poder de eficiencia
- ✓ Más sensibles a los rasgos de los datos recolectados
- ✓ Menos posibilidad de errores

- ✓ Dan estimaciones probabilísticas bastante exactas

2.1.1 Correlación

Para el cálculo de la correlación tenemos las siguientes pruebas:

Correlación de Pearson, que mide una dependencia lineal entre dos variables (X, Y) que es una prueba de tipo paramétrica. Se puede usar solo cuando X e Y son de distribución normal (esto lo veremos más adelante).

Figura 3

Fórmula de la Correlación de Pearson

$$r_{xy} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

Donde:

$\bar{x}$ $\bar{y}$ representan la media del vector X y Y respectivamente

S_x y S_y representan las desviaciones estándar de X y Y respectivamente

n es la cantidad de observaciones

Correlación de Kendall y Spearman, son coeficientes de correlación basados en rangos (no paramétricos): Las pruebas no paramétricas son aquellas que se encargan de analizar datos que no tienen una distribución particular y se basan en hipótesis, pero los datos no están organizados de forma normal. Aunque tienen algunas limitaciones, cuentan con resultados estadísticos ordenados que facilita su comprensión.

Para el cálculo de la correlación usaremos la función cor() cuya sintaxis se muestra en la siguiente figura.

Figura 4

Sintaxis de la función cor

Referencia: Sintaxis de la función, www.rdocumentation.org

```
cor(x, y = NULL, use = "everything",
    method = c("pearson", "kendall", "spearman"))
```

Retorna el coeficiente de correlación

Es importante mencionar que cor() puede tomar 2 vectores (X y Y) o un dataframe completo que puede tener múltiples columnas, por lo tanto, si la data es muy extensa su resultado será una matriz de correlaciones entre todos los vectores del data.frame.

Esta matriz tiene las siguientes características:

- ✓ La matriz es simétrica, la correlación entre a y b es la misma que entre b y a
- ✓ Los valores de la correlación varían entre -1 y 1, el signo indica si la relación es monotónica negativa o positiva. Y el valor absoluto indica que tan fuerte es la relación.
- ✓ La diagonal es siempre 1 (una variable siempre se correlaciona consigo mismo)

Es recomendable limitar las columnas que intervienen mediante "subsetting", es decir un subconjunto de columnas de interés.

Tabla 1

Matriz de correlación en el dataset mtcars

Creación de autor: Alfonso Prado

```
> data(mtcars)
> names(mtcars)
 [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"
[10] "gear" "carb"
> class(mtcars)
[1] "data.frame"
> cor(mtcars[,1:4])
            mpg        cyl       disp         hp
mpg   1.0000000 -0.8521620 -0.8475514 -0.7761684
cyl  -0.8521620  1.0000000  0.9020329  0.8324475
disp -0.8475514  0.9020329  1.0000000  0.7909486
hp   -0.7761684  0.8324475  0.7909486  1.0000000
```

Cuando la matriz es pequeña como en la tabla anterior, analizar las relaciones es simple, pero cuando se tiene decenas o centenas de variables se torna difícil, en este caso recurrimos a la visualización.

Por ejemplo, utilizando el datset economics, este dataset contiene las siguientes columnas:

Pce: Gastos de consumo personal,

Pop: Población total, en miles,

Psavert: Tasa de ahorro personal,

Uempmed: Duración media del desempleo, en semanas

Unemploy: Número de desempleados en miles

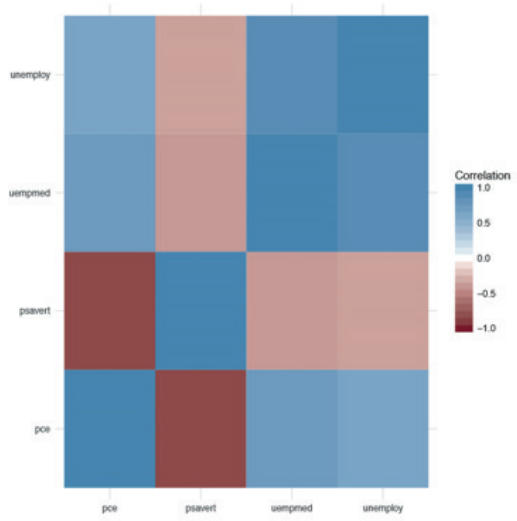
Utilizando ggplot, podemos crear un mosaico, cuyos colores representan la fuerza de la correlación. Colores fuertes representan una correlación alta, color neutro como el blanco indican una correlación nula. Estos gráficos son conocidos como correlogramas.

Existen muchas formas de crear estos correlogramas, el siguiente enlace cubre algunas posibilidades.

Figura 5

Diagrama calor de la correlación del dataset economics

Creación de autor: Alfonso Prado



2.1.2 Covarianza:

Hemos visto que la correlación indica la relación que existe entre dos variables; es decir, si una de ellas (la variable predictora) cambia, la variable de respuesta también cambiará, ya sea hacia arriba o hacia abajo.

Por otro lado, la covarianza es un valor que indica el grado de variación conjunta de dos variables aleatorias respecto a sus medias. Es el dato básico para determinar si existe una dependencia entre ambas variables y es esencial para estimar otros parámetros fundamentales, como el coeficiente de correlación lineal o la recta de regresión. También ayuda a entender si estas dos variables están relacionadas de alguna forma.

Retorna un valor que puede ser positivo (si se mueven juntas), negativa (se mueven en direcciones opuestas), o nula (no se mueven juntas de manera apreciable). Existen varios métodos (algoritmos) para su cálculo.

En otras palabras, la covarianza trata de encontrar la relación entre dos variables predictoras. Se asume que hay una variable predictora que provoca cambios entre las variables de respuesta. Queremos encontrar la relación entre estas dos variables predictoras

La función para determinar la covarianza y su fórmula de cálculo se puede ver en la siguiente gráfica:

Figura 6

Sintaxis de la función cov (a) y su fórmula de cálculo (b)

Referencia: Sintaxis de la función, www.rdocumentation.org

(a)

```
cov(x, y = NULL, use = "everything",
    method = c("pearson", "kendall", "spearman"))
```

(b)

$$Cov(X,Y) = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{n}$$

2.1.3 CORRELACIÓN CON VARIABLES CATEGÓRICAS

Hasta el momento, hemos visto cómo cuantificar la relación entre variables cuantitativas numéricas, pero, como mencionamos antes, estas no son el único tipo de variables que la estadística puede procesar. Otro caso importante es el de las variables nominales. Para este tipo de datos, utilizamos tablas de contingencia.

De acuerdo con Agresti A. (2013), el análisis de datos categóricos se basa típicamente en tablas de contingencia de dos o más dimensiones, tabulando la frecuencia de ocurrencia de niveles de datos nominales y/o ordinales.

Una tabla de contingencia es una herramienta que permite organizar datos categóricos en filas y columnas, mostrando la frecuencia de ocurrencia de cada combinación de valores. Esta tabla nos permite medir la interacción entre dos variables y obtener información valiosa para comprender mejor los resultados de una investigación.

Por ejemplo, trabajando con el datase Oncho, este dataset describe la cantidad de personas encuestadas para saber si padecen de esta enfermedad (onchocercosis). Tenemos una variable que indica si está infectada la persona, pero también contiene otras variables como el área (Savannah o Rainforest) , sex (el género masculino o femenino) y grupo de edad (0-20, 20-39,40+). Todas estas variables son tipo categórico.

En R, estas variables pueden ser convertidas al tipo factor. Un factor no es nada más que una variable categórica que permite agrupación y las distintas opciones de cada variable toman el nombre de niveles del factor. Por ejemplo, para el factor "área" hay 2 niveles Savannah o Rainforest.

Con estos datos podemos crear una tabla de contingencia que nos permita medir la relación entre estas variables en términos de frecuencia de ocurrencia en el dataset, por ejemplo, no infectados en la Savannah tenemos 267 personas mientras que infectadas en la misma área 281.

Figura 7

Dataset (a) y tabla de contingencia (b)

(a)

id ID	mf Infection	area Residence	agegrp Age group (years)	sex Sex
1	Infected	Savannah	20-39	Female
2	Infected	Rainforest	40+	Male
3	Infected	Savannah	40+	Female

(b)

	Residence Savannah, N = 548	Rainforest, N = 754	Overall, N = 1,302
Infection			
Not-infected	267	213	480
Infected	281	541	822

Debemos notar que la cantidad de observaciones de un nivel nominal no va a ser necesariamente la misma de otro nivel. Por ejemplo, la cantidad de encuestados hombres (Male) no es igual a las mujeres (Female).

En algunos campos de investigación como la bioestadística se estila crear la tabla con nombres genéricos como Exposure y Outcome, donde las filas (Exp+ y Exp -) indican haber estado sujeto a un "tratamiento" y las columnas conocidas como Out +, Out - haber desarrollado un efecto o no. En nuestro dataset Oncho, el área sería el "tratamiento" y mf el "efecto".

Los datos de frecuencia de la tabla por si solos no indican mayor cosa, por lo tanto, la tabla debe ser procesada a fin de obtener algunos estadísticos que nos indiquen si existe una relación entre un par de variables. El proceso de análisis va a ser diferente si la tabla es de 2X2 (dos filas 2 columnas) o de más filas, como por ejemplo si analizamos la variable "agegrp".

Para el caso de variable dicotómicas, (esto es variables que solo tienen dos posibles niveles) la función epi.2by2() nos ayuda a obtener los estadísticos necesarios, pero para tablas más grandes la función chisq.test() sería la apropiada. Analizaremos estas funciones más adelante en el curso.

Tratamiento de tablas de contingencia

Describe distintas funciones utilizadas en el análisis

enlace

2.2) Causación

Las medidas de correlación y covarianza nos indican como cambia una variable cuando la otra variable cambia. Es decir, solo nos indican como se mueven las variables en el dataset, pero no necesariamente nos indica que una variable causa a la otra variable.

Ejemplo: Usted dispone de un dataset que contiene población de una ciudad, el consumo de carne y el consumo de pescado en la ciudad. Existe una correlación directa entre las 2 últimas variables y la población. Esto es correcto porque se asume que mientras más grande es la población de una ciudad hay más consumo

Pero, si se efectúa una correlación entre consumo de pescado y consumo de carne también encontrará correlación directa. ¿Significa esto que el consumo de carne hace que se consuma más pescado? Por supuesto que no. En este caso existe correlación, pero no existe causación, es decir, el consumo de carne no produce que se consuma más pescado. Por lo tanto, diremos que la causación implica correlación, pero la correlación no implica causación.

Análisis de Correlación

Describe varias funciones para análisis de correlación y causación

enlace

2.2.1 Diseño de Experimentos

De acuerdo con Mendiburu F, Yaseen M. (2020), el propósito del diseño de experimentos (DoE) es poder obtener en forma segura la relación causa-efecto entre dos variables. Por ejemplo: saber si un cierto medicamento ayudó (o no) en la recuperación del paciente.

Para aseverar que lo anterior es verdadero ciertas consideraciones deben cumplirse.

Asociación

El primer criterio para establecer un efecto causal es una asociación observada entre las variable predictora y de respuesta. Generalmente esto se logra mediante un análisis de correlación. Por otro lado, debemos garantizar que existe la relación tiene un ordenamiento en el tiempo. Se debe asegurar que la variación de la variable predictora se produjo antes del tiempo de la variación de la variable de respuesta.

Detección de asociaciones espurias

El segundo criterio es que esta relación no sea espuria, definimos como espuria a una relación que es ficticia o fraudulenta, muchas veces nos topamos con la existencia de una tercera variable que en realidad afecta a las dos primeras. Este tipo de variable se las conoce como “confounding” (que produce confusión).

La siguiente figura ilustra la relación entre la edad del paciente y el riesgo de enfermedad coronaria. En principio parecería que la edad es un factor, pero la variable confounding sería el ejercicio.

Figura 8

Correlación vs Causación en enfermedad coronaria. Creación de autor: Alfonso Prado

Para garantizar que la relación no sea espuria debemos cumplir las siguientes condiciones:

- ✓ Dos grupos de comparación (en el caso más simple, un grupo experimental y un grupo de control), para establecer asociación
- ✓ Variación en la variable independiente se produce antes del cambio en la variable dependiente, para establecer el orden temporal
- ✓ Asignación aleatoria a los dos (o más) grupos de comparación, para establecer que la relación no es espuria.

REFERENCIAS CITADAS

Webster A., (2000), Estadística aplicada a los negocios, Irwin Professional

Publishing

Agresti A. (2013), Categorical Data Analysis, John Wiley & sons publication

Mendiburu F, Yaseen M. (2020), Experimental Designs with agricolae, https://myaseen208.com/agricolae/articles/ExperimentalDesign.html

Definición de los términos citados en la Clase 1.

Nombre:	Definición
CORRELOGRAMA	Es un gráfico que permite apreciar las autocorrelaciones $r_1, r_2, \ldots\ldots, r_k$ mediante el cual se identifican si los datos de una serie de tiempo tienen las siguientes características: estacionalidad, aleatoriedad, tendencia y estacionariedad.

PREUBA CHI CUADRADO	De acuerdo con Ramirez-Alan ,(2016) "El test Xi2 considera la hipótesis nula (H0) de que las variables son independientes, si esto es verdad la frecuencia de ocurrencia debería estar dada por la cantidad de casos totales multiplicada por la probabilidad esperada". En el ejemplo anterior, si la probabilidad de compra es igual para los 3 niveles socioeconómicos (πij = 33%) entonces si se han vendido 1000 ítems estos deberían estar distribuidos en las 3 categorías (n*πij =1000*.33) . Este valor conocido como μij . Pero si Ha está en lo correcto va a existir una diferencia entre frecuencia observada (ηij) y la esperada (μij) indicando que existe algún fenómeno por detrás que influencia la frecuencia observada. La prueba Xi2 calcula esta diferencia de la siguiente forma $$X^2 = \sum \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}}$$

Profundización Clase 1.

Diagramas de mosaicos

Muestra distintas técnicas para lograr este tipo de diagramas

enlace

PUCE Virtual

La excelencia no se improvisa

síguenos