Estadística Aplicada

Terminología Básica

Clase 1

Ingeniería en ciberseguridad

La excelencia no se improvisa

1. INTRODUCCIÓN DE LA CLASE

En esta clase, iniciaremos con los fundamentos de la estadística y su propósito en la Ciencia de Datos, así como con la introducción del estudiante a las herramientas tecnológicas que se utilizarán para su implementación. Específicamente, en la primera parte, revisaremos el uso de R y RStudio, y su aplicación en distintas fases de la estadística.

En la segunda parte, abordaremos la terminología básica, los tipos de variables utilizados en el análisis de las medidas de tendencia central y las medidas de dispersión. También introduciremos gráficos que permitan entender de mejor manera los diversos estadísticos.

1)Terminología Básica

En primer lugar, empecemos indicando que es estadística: Existen múltiples definiciones de la estadística, posiblemente una por cada autor, a continuación, veremos algunas de las mejores definiciones:

- ✓ La estadística es la ciencia que estudia cómo debe emplearse la información y cómo dar una guía de acción en situaciones prácticas que entrañan incertidumbre., la cual implica su recolección, clasificación, síntesis, organización, análisis e interpretación, para la toma de decisiones frente a la incertidumbre. Gutiérrez 1998
- ✓ La estadística es la rama del conocimiento humano que tiene como objeto el estudio de ciertos métodos inductivos aplicables a fenómenos susceptibles de expresión cuantitativa. (López, 2020)

De estas definiciones podemos extraer los siguientes detalles:

De acuerdo con Gutierrez, (2007), debemos notar que hace referencia a la “incertidumbre”. Esto es correcto, aunque, contrario a lo que se piensa, la estadística no es una ciencia matemática cuyos resultados sean siempre exactos. En realidad, las operaciones estadísticas nos muestran valores dentro de un rango. No podemos asegurar cuál es el valor real, pero sabemos que estará dentro de un intervalo determinado.

Notaremos que la estadística abarca desde funciones que tiene que ver con la recolección de la información, su tratamiento y análisis de lo que nos está mostrando.

Por último, de acuerdo con López (2020), la estadística es una ciencia de carácter inductivo, lo que significa que se basa en los datos para tomar decisiones, a diferencia de los estudios deductivos, en los cuales se conoce de antemano lo que se intenta demostrar.

Existen tres ramas (o si se quiere tres etapas) de la estadística, cada una comprende una serie de técnicas y funciones que permiten contestar las siguientes preguntas:

¿Qué pasó? ¿Qué está pasando?

¿Qué pasará si seguimos de la misma manera en el futuro?

¿Qué podría pasar si se alteran las condiciones del negocio?

En esta clase, trataremos de responder la primera pregunta. El conjunto de métodos y procedimientos que nos ayudan a contestarla se denomina estadística descriptiva.

Estadística Descriptiva: Es la rama de las matemáticas que recolecta, presenta y caracteriza un conjunto de datos (por ejemplo, edad de una población, altura de los estudiantes de una escuela, temperatura en los meses de verano, etc.) con el fin de describir apropiadamente las diversas características de ese conjunto, Por lo tanto, la estadística descriptiva se apoya o trabaja en base a variables que describen un proceso, evento, negocio, una enfermedad etc.

Puede haber múltiples variables en un caso de estudio y estas variables pueden ser de distinta naturaleza. De la misma manera que el éxito de una aplicación es haber realizado un correcto análisis de las propiedades de los objetos y clases, el éxito de la estadística es haber recolectado todos los factores que afectan al proceso, evento, negocio etc.

Tipos de Datos:

Para iniciar con la estadística descriptiva, es importante mencionar que existen varios tipos de datos. Comúnmente, asociamos la palabra "dato" con una variable numérica, pero esto no siempre es el caso. Existen diversos tipos de datos dentro del análisis estadístico.

Empezaremos por catalogar dos grandes grupos de tipos de datos:

Variables cuantitativas: Las observaciones se expresan numéricamente. Dentro de esta categoría podemos ubicar algunas variantes:

Variables continuas: Son aquellas que puede tomar cualquier valor, con cualquier cantidad de decimales, en algunos casos podría estar acotado o dentro de un rango. Ejemplo pi = 3.14159

Variables discretas: Puede tomar determinados valores numéricos únicamente, por ejemplo, la edad de una persona. Un caso particular de esta es data binaria que solo puede tomar 2 valores: Verdadero/Falso.

> ***Series de Tiempo***: Es una subclase de cuantitativa representada por listas o vectores que representan el valor de la variable a través del tiempo, por ejemplo, temperatura de la ciudad durante el año

Ratio o Razón: Relación entre dos variables numéricas y generalmente se expresa en términos de porcentajes, por ejemplo, tasa de cambio de una moneda.

Intervalo: Variables que se hallan dentro de un rango. Esto es muy utilizado por ejemplo en secuencias en R o rangos en Python, ejemplo secuencia de conteo del 1 al 10, en R representado por la función seq(1::10)

Variables cualitativas: A este grupo pertenecen las variables que describen características que pueden presentarse en los elementos que conforman el conjunto de datos. En R también se les conoce como variables tipo factor. En esta categoría podemos ubicar los siguientes casos:

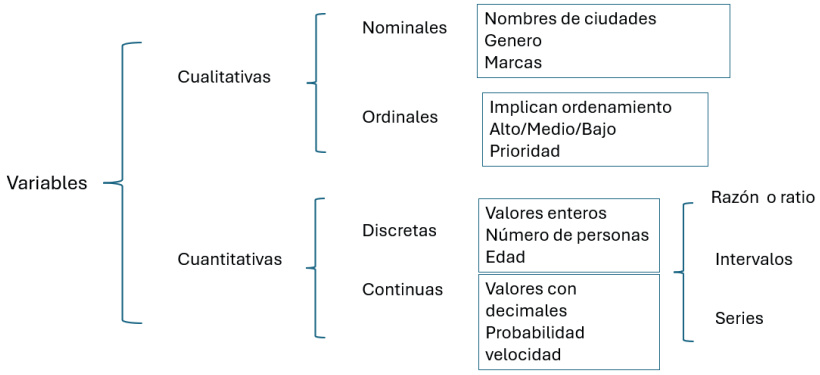
> **Variables categóricas:** Son variables que se consideran clasificadoras o calificadoras, ya que catalogan los elementos en grupos, conjuntos o categorías. Se conocen como dicotómicas aquellas que solo pueden adoptar dos valores diferentes, y como politómicas las que poseen un número de valores mayor a dos. A su vez, se subdividen en:
>
> > **Variables nominales:** Nombres o clasificaciones que se utilizan para datos en categorías, por ejemplo, país, genero, color que por su naturaleza no puede ser ordenada.
> >
> > Variables ordinales: Data que si bien no es numérica si nos da un sentido de prioridad. Por ejemplo, pequeño, mediano, grande, variables que pueden ser ordenadas.

La siguiente figura explica la clasificación:

Figura 1

Taxonomía de las variables

Creación autor: Alfonso Prado



1.1) **Medidas de tendencia central**

Como mencionamos anteriormente, el propósito de la estadística descriptiva es describir los datos. Existen múltiples herramientas para ello. A continuación, exploraremos algunas de las más comunes.

1.1.1 PROMEDIO

Promedio: Es el valor central de un conjunto de datos y nos indica cuál es el valor esperado para una variable aleatoria, considerando todas las posibilidades. En R, la función utilizada para calcularlo es mean(), cuya sintaxis y fórmula se describen en la siguiente figura.

Figura 2

Referencia: Función mean sintaxis y fórmula . www.rdocumentation.org

(a) Sintaxis de la función mean()

```
mean(x, …)

# S3 method for default
mean(x, trim = 0, na.rm = FALSE, …)
```

(b) Fórmula de la función

$$E[X] = \frac{\sum_{i=1}^{N} x_i}{N}$$

La función acepta los siguientes argumentos:

x representa el vector que contiene los datos.

trim: Es la fracción (0 a 0,5) de observaciones que se deben recortar de cada extremo de x antes de calcular la media. Los valores de recorte que se encuentren fuera de ese rango se toman como el punto final más cercano y es útil en caso de que el dataset tenga valores atípicos.

na.rm: Un valor lógico que indica si los valores NA deben eliminarse antes de continuar el cálculo. En R cuando un valor de una variable figura como NA indica que no se ha ingresado dicho valor y es diferente al caso de asumir que su valor es cero o nulo. La importancia de este argumento radica en que cualquier operación que se realiza sobre un vector que contiene al menos un valor NA dará como resultado NA.

Si analizamos lo fórmula de la función mean veremos que todos los valores tienen un peso igual. Esto podría no ser apropiado en algunos casos. Para esto utilizaremos la media ponderada:

Meda ponderada: La media aritmética ponderada es similar a una media aritmética, excepto que en lugar de que cada una de las observaciones contribuyan igualmente al promedio final, algunas contribuyen más que otras. Podemos obtener la media ponderada usando el paquete DescTools. La función y formula se muestran en la figura 3.

Figura 3

Media Ponderada sintaxis y fórmula

Referencia: Media Ponderada sintaxis y fórmula. www.rdocumentation.org

(a) Sintaxis	(b) Fórmula

`Mean(x, ...)`
 	
`# S3 method for Freq`
`Mean(x, breaks, ...)`
 	
`# S3 method for default`
`Mean(x, weights = NULL, trim = 0, na.rm = FALSE, ...)`
	$E[X] = \frac{\sum_{i=1}^{N} W_i X_i}{\sum_{i=1}^{N} W_i} = \sum_{i=1}^{N} P_i X_i$

Note que la función es Mean con mayúscula, en R mayúsculas no son igual que minúsculas.

Los argumentos son similares a la función mean, excepto que se introduce el argumento weights, que representa los pesos de cada una de la observaciones.

Media geométrica: Esta es un tipo de media que se usa generalmente en ambientes financieros debido a que su resultado es un poco más conservador. Su valor se obtiene mediante la raíz n de la multiplicación de los valores de las observaciones. Existen muchos paquetes que proveen esta funcionalidad, en nuestra clase utilizaremos la función gm() del paquete rob.compositions . Debido a que la multiplicación de los valores de las observaciones puede arrojar valores muy grandes alternativamente se puede calcular como la constante de Euler (e=2.7182) elevado a la potencia de la media de los logaritmos neperianos de la data, de esta forma garantizamos que no se producirá un “overflow” en su cálculo. Su función y fórmula se muestran en la figura 4.

Figura 4

Media geométrica, sintaxis y fórmula

(a)Sintaxis	(b) Fórmula
`exp(mean(log(data)))`	$\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}}$

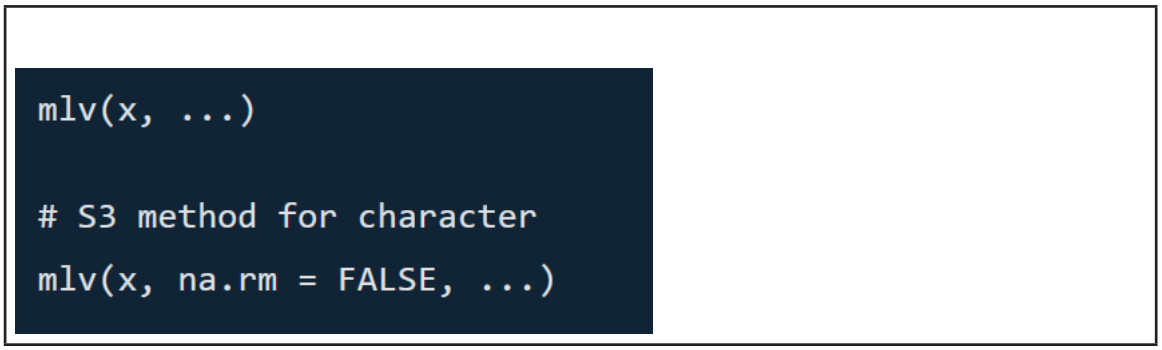
1.1.2 MODA

Moda: Esta función permite calcular el o los valor**(es)** más repetidos de un vector de datos. El paquete modeest contiene varias funciones para estimar la moda, incluyendo la función mlv que devuelve un vector numérico con los valores más repetidos. La sintaxis se muestra en la figura 5.

Figura 5

Sintaxis de la función mlv

Referencia: Moda sintaxis y fórmula . www.rdocumentation.org



mlv es una función genérica para estimar la moda de una distribución univariante. Se proporcionan diferentes métodos, el más común es "mfv" que significa el valor más frecuente.

1.2) Medidas de dispersión

Se llama dispersión de los datos a la variabilidad que existe entre ellos. Cuando tenemos un set de datos de una variable aleatoria, se podría pensar que estos podrían asumir cualquier valor. Sin embargo, en la realidad, cuando se analiza un fenómeno, los datos tienden a estar más o menos agrupados alrededor de la media. Existen varias funciones que nos permiten cuantificar esta dispersión.

1.2.1 PERCENTILES

Cuantiles: Los cuantiles nos dan una primera idea de la distribución de los valores. Un cuantil o percentil, le indica qué porcentaje de sus datos se encuentra por debajo de un cierto valor de la variable bajo análisis. Por ejemplo, el cuantil del 50 por ciento es lo mismo que la mediana. R tiene algunas funciones convenientes para ayudarlo a mirar los cuantiles.

La función quantile() puede proporcionarle cualquier cuantil que desee. Para eso, usa el argumento probs como un número fraccionario. Por ejemplo, para obtener el cuantil del 20%, utilizaremos 0,20 como valor de este argumento. Este argumento también toma un vector como valores, por lo que puede, por ejemplo, obtener los cuantiles de 5 y 95 por ciento de esta manera.

Es importante mencionar que existen distintos algoritmos para el caso de variables discretas y continuas y que no todas las herramientas informáticas utilizan el mismo algoritmo. Esto puede ser importante y hasta cierto punto desconcertante si se está trabajando con distintas herramientas a la vez, en cuyo caso es posible que dos herramientas provean resultados diferentes. Los algoritmos se numeran del 1 al 3 para variables discretas por ejemplo fechas o factores ordenados y del 4 al 9 para variables continuas. El algoritmo se define con el argumento "type" siendo el 7 el valor por defecto. La figura 6 muestra la sintaxis de la función quantile(). Algoritmos de la función quantile, describe las fórmulas aplicables a los distintos métodos. enlace

Figura 6

Sintaxis función quantile

Referencia: Quantile sintaxis. www.rdocumentation.org

```
quantile(x, …)

# S3 method for default
quantile(x, probs = seq(0, 1, 0.25), na.rm = FALSE,
         names = TRUE, type = 7, …)
```

1.2.2 VARIANZA

Varianza: Es una medida de dispersión que representa la variabilidad de una serie de datos respecto a su media. Una serie de datos podría tomar un número infinito de valores, pero en la práctica cuando se analiza un parámetro de interés tienden a estar dentro de un rango más o menos dispersos. La varianza indica la medida de esta dispersión.

Figura 7

Función y fórmula de la varianza

Referencia: Varianza sintaxis y fórmula . www.rdocumentation.org

(a) Función	(b) Fórmula
`var(x, y = NULL, na.rm = FALSE, use)`	$Var(x) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N - 1}$

1.2.3 DESVIACIÓN

La desviación estándar representa la raíz cuadrada de la varianza. El símbolo σ (sigma) se utiliza frecuentemente para representar la desviación estándar de una población, mientras que "s" se utiliza para representar la desviación estándar de una muestra.

Figura 8

Desviación estándar función y fórmula

Referencia: Desviación estándar función y fórmula, . www.rdocumentation.org

(a) Función	(b) Fórmula
`sd(x, na.rm = FALSE)`	$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$

Cómo visualizar la dispersión de los datos:

ggplot2 es un paquete de visualización de datos para el lenguaje R. Es un esquema general para la visualización de datos que divide los gráficos en componentes semánticos en una modalidad de capas. ggplot2 es un paquete mejorado de los gráficos base en R (plot).

Forma parte del paquete Tidyverse, un sistema para la manipulación, exploración y visualización de datos que comparten una filosofía de diseño común. Los paquetes de Tidyverse están destinados a ser usados por estadísticos y científicos de datos. Básicamente, tidyverse trata sobre las conexiones entre las herramientas que hacen posible el flujo de trabajo.

ggplot2 permite generar una gran cantidad de tipos de gráficos. Las gráficas de dispersión son más útiles para mostrar la relación entre dos variables continuas, o cuando una entidad está compuesta por 2 valores. Para visualizar la dispersión utilizaremos la función geom_point, cuya sintaxis se muestra en la figura 9. Información adicional sobre ggplot2, el siguiente enlace describe la operación básica del paquete enlace

Figura 9

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

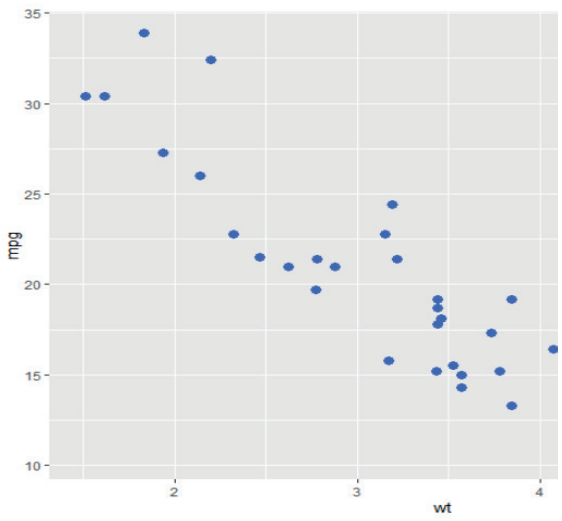
```
geom_point(
  mapping = NULL,
  data = NULL,
  stat = "identity",
  position = "identity",
  ...,
  na.rm = FALSE,
  show.legend = NA,
  inherit.aes = TRUE
)
```

Donde el “mapping” permite indicar que variables del dataset son asignadas a cada eje. La data, por supuesto, corresponde al dataframe que contiene los datos, “position” permite alterar la posición de las observaciones en situaciones de sobre-trazado para mejorar la visualización.

Figura 10

Código y visualización de la dispersión del dataset mtcars

Creación autor: Alfonso Prado

(a) Código	(b) Gráfico resultante
ggplot (data=mtcars, aes(x=wt,y=mpg)) + geom_point(color="BLUE)	

Referencias citadas en la Clase 1.

Gutierrez A. (2007) Estadística General Aplicada, Universidad Eafit

López W, (2020) , Estadística Práctica Aplicación y análisis para la toma de decisiones. Publicado Independiente

Definición de los términos citados en la Clase 1. *Glosario*

Nombre:	Definición
SERIES TEMPORALES	Conocida también como serie cronológica. Es la sucesión de observaciones cuantitativas ordenadas en el tiempo de un fenómeno. Los datos tienen un orden que no es posible variar. La información puede ser mensual, trimestral, anual o de cualquier otro intervalo temporal.
VARIABLE ALEATORIA	Conocida también como variable estocástica o probabilística. Es la característica considerada en un experimento aleatorio cuyo valor de ocurrencia solo puede saberse con exactitud una vez observado.

Profundización Clase 1.

Diagramas de dispersión

Muestra distintas técnicas para lograr este tipo de diagramas

enlace

PUCE Virtual

La excelencia no se improvisa

síguenos