PUCE Virtual
EDUCACIÓN FLEXIBLE

Estadística 1 para administración

Introducción y terminología básica

Clase 1

La excelencia no se improvisa

INTRODUCCIÓN DE LA CLASE

En esta clase iniciaremos con fundamentos de la estadística y su propósito en el Ciencia de Datos, así como introduciremos al estudiante en las herramientas tecnológicas que se usarán para su implementación. Específicamente, en la primera parte revisaremos terminología básica, tipos de variables que se utilizan en el análisis de medidas de tendencia central, y teoría del muestreo .

En la segunda parte veremos la representación visual de los datos, esta sección nos permite ver la distribución visual tanto de los datos de variables de variables cuantitativas, simples y agrupadas y los mismo para variables cualitativas

RDA 1: Analizar información de un conjunto de datos, mediante tablas, gráficos y medidas descriptivas.

CLASE 1: Introducción y terminología básica

Definiciones básicas

Población: La población estadística es el total de individuos o conjunto de ellos que presentan o podrían presentar el rasgo característico que se desea estudiar(economipedia). Algunos autores llaman a este concepto el universo. Veamos algunos ejemplos:

Si el gerente de una empresa desea investigar los problemas de producción, la empresa tiene 5 plantas o fábricas, dado que los resultados son específicos para esta empresa entonces podemos decir que la población será todas las 5 plantas, observaciones población o universo.

Si los ingresos de los 121 millones de asalariados de los Estados Unidos son de interés para un economista que asesora al Congreso en la formulación del plan nacional tributario, entonces los 121 millones de personas constituyen la población.

Si se desea estudiar la preferencia electoral se entrevistan a 10.000 personas, la población en este caso sería los ciudadanos con derecho a voto.

Muestra: Es una parte representativa de la población que se selecciona para ser estudiada. Hay varias razones por las cuales un trabajo de investigación no es posible hacerlo con toda la población: Por ejemplo:

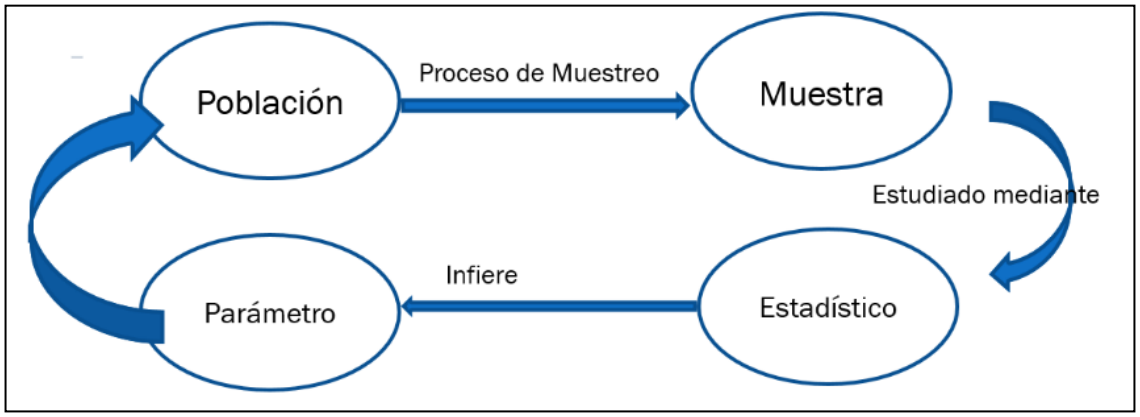
- ✓ Ejecutar prueba sobre el 100% de la población es muy costoso, como probar una vacuna en toda la población para medir el efecto de la misma.
- ✓ La población está geográficamente dispersa, por lo que sería muy costoso movilizarse a muchos lugares.
- ✓ El análisis sobre cada elemento de la población termina dañando a la población. Por ejemplo, queremos medir la resistencia de un ladrillo a la presión, en este caso el experimento termina con las destrucciones del ladrillo. Si lo hacemos sobre el 100% de la producción de ladrillos de una fábrica, esta terminará sin inventario de ladrillos.

Parámetro y Estadístico:

Cuando nos referimos a la estadística nos referimos al conjunto de conocimientos que estamos estudiando. Cuando nos referimos al estadístico nos referimos a una métrica de define una característica del conjunto. Adicionalmente, cuando nos referimos a un parámetro nos referimos a una característica de toda la población. Los dos conceptos son similares, pero aplican a distintos grupos. Se define como parámetro a una medida descriptiva de la población total, mientras que el estadístico representa lo mismo, pero es el valor obtenido de la muestra. Por ejemplo: siguiendo con el ejemplo del gerente de la empresa que tiene 5 plantas de producción, el gerente ha seleccionado 2 de ellas para analizar la producción diaria, este análisis sería mediante la medición de la producción diaria, a esto lo llamamos el estadístico, en base al cual podemos extrapolar la producción de las 5 plantas, que es el parámetro. La figura 1 muestra la relación entre estos dos conceptos:

Figura 1: Relación entre estadístico y parámetro

Creación del autor Alfonso Prado



Error de Muestreo:

El error de muestreo se define como la diferencia entre el parámetro desconocido de la población y el estadístico de la muestra utilizado para calcular el parámetro. La exactitud de toda estimación depende en gran parte del muestreo y que este sea representativo, sin embargo, como no podemos garantizar que la muestra sea 100% representativa, siempre habrá un error de muestreo. Otra forma de entender este error sería la siguiente: si midiéramos un estadístico una y otra vez sobre varias muestras obtendríamos una diferencia que sería el error de muestreo.

Sesgo Muestral:

También conocido como "bias" es la tendencia a favorecer la selección de ciertos elementos de muestra en lugar de otros. Esto es cuando hay una situación (intencional o no intencional) a seleccionar determinados elementos de la muestra en lugar de otros. Por ejemplo, estimar la preferencia de voto en ciertas localidades porque no hay presupuesto para muestrear todas las localidades.

Tipos de muestreo:

Una muestra puede ser obtenida de dos tipos: probabilística y no probabilística. Las técnicas de muestreo probabilísticas, permiten conocer la probabilidad que cada individuo a estudio tiene de ser incluido en la muestra a través de una selección al azar. En cambio, en las técnicas de muestreo de tipo no probabilísticas, la selección de los sujetos a estudio dependerá de ciertas características, criterios, etc. que él (los) investigador (es) considere (n) en ese momento.

Tipos de Muestreos
Contiene una descripción pros y contras de cada tipo
Enlace: https://scielo.conicyt.cl/pdf/ijmorphol/v35n1/art37.pdf

Medidas de tendencia central

Para iniciar con el análisis estadístico es importante mencionar que existen varios tipos de datos. Comúnmente asociamos la palabra dato con una variable de tipo numérico, pero esto no es el caso, de hecho, existen muchos tipos de datos que pueden estar incluidos dentro de la función estadística.

Empezaremos por catalogar dos grandes grupos de tipos de datos:

Variables cuantitativas: Si las observaciones pueden expresarse 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 cantidad de personas. 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.

Adicionalmente a la clasificación anterior, podemos hacer una clasificación en base al concepto de causa-efecto.

Predictoras (anteriormente llamadas variables independientes) son variables que representa un argumento para obtener una salida

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

Por ejemplo, el número de personas atendiendo en servicio al cliente es una variable predictora de una respuesta del tiempo de espera del cliente.

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.

Figura 2
Representación de una función Y =f(x)
Creación del autor Alfonso Prado

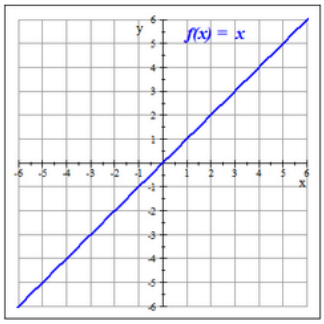
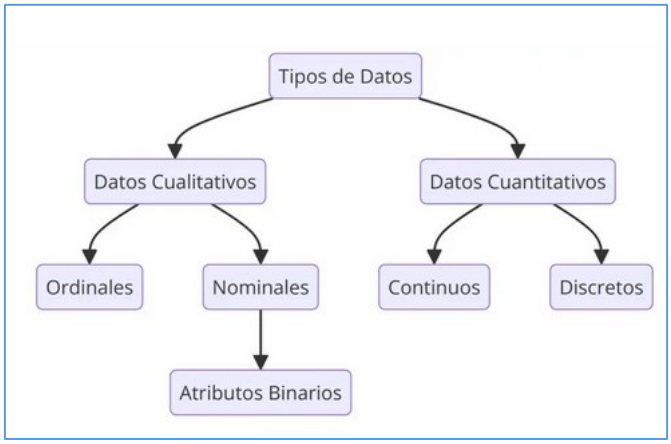


Figura 3: Tipos de datos
Creación del autor Alfonso Prado



Como mencionamos anteriormente, el propósito de la estadística descriptiva es describir los datos. En estadística existen múltiples funciones que nos ayudan a este concepto, a continuación, presentaremos un conjunto inicial de funciones comúnmente utilizadas.

Promedio: Es el promedio aritmético de los valores de la serie o un vector y nos indica cual es el valor esperado para de un evento aleatorio, considerando todas las posibilidades.

Figura 4

Función matemática y programática de la media simple Prado A. (2025)

Adaptación de la fórmula para la resolución de medias de Webster (2000) capitulo 3

$E[X] = \frac{\sum_{i=1}^{N} X_i}{N}$	`mean(x, …)`
	
	`# S3 method for default`
	`mean(x, trim = 0, na.rm = FALSE, …)`

Media 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

Figura 5

Función matemática y programática de la media ponderada Prado A. (2025)

Adaptación de la fórmula para la resolución de medias de Webster (2000) capitulo 3

Creación del autor Alfonso Prado

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

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 mg() del paquete rob.compositions (note mg en minúsculas).

Figura 6

Función matemática y programática de la media geométrica Prado A. (2025)

Adaptación de la fórmula para la resolución de medias de Webster (2000) capitulo 3

$MG = \sqrt[n]{\prod_{i=1}^{n} x_i}$	exp(mean(log(data)))

Moda: Esta función permite calcular el/los valor(es) más repetidos de un vector de datos. El paquete modeest contiene varias funciones para estimar la moda que incluye la función mlv que devuelve un vector numérico con los valores más repetidos. Es una forma de calcular una media para variables cualitativas.

Figura 7

Función programática de la moda , donde mfv representa el valor más repetido Prado A. (2025)

Adaptación de la fórmula para la resolución de la moda Rdocumentation (s.f.), https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/mlv

```
mlv(x, bw = NULL, method="mfv",na.rm = FALSE)
```

Suma, máximo , mínimo y rango:

Estos estadísticos no requieren mayor explicación: La función sum() corresponde a la suma de los valores de un vector de observaciones. La función max () , min() y range() entregan el valor máximo, mínimo y el rango de valores encontrados dentro de un vector.

Funciones de conteo:

Comúnmente, nos interesa contar la cantidad de elementos que contiene un objeto. En R existen varias funciones pero aplican a distintos tipos de objetos (vector, data frame o matriz).

Tabla 1: Funciones para conteo

Creación del autor Alfonso Prado

count(x)	Del paquete dplyr ,aplicado sobre una variable de tipo carácter, Del paquete dplyr permite contar las observaciones "únicas" es decir no repetidas en un vector y la cantidad de veces que se repiten
count(data_frame)	Cuando la función se aplica a un data frame contará la cantidad de observaciones (filas) que contiene el data frame.
length()	A diferencia de count() esta función cuenta estrictamente la cantidad de elementos de un vector.
NROW(data frame)	Esta función cuenta la cantidad de observaciones (filas) que contiene un objeto de la clase data frame
nrow(arreglo)	Esta función es similar a la anterior, pero se aplica a objetos de tipo arreglo o matrices y cuenta la cantidad de filas del objeto
ncol(arreglo)	Cuenta la cantidad de columnas que contiene un objeto del tipo arreglo o matriz

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.

Es importante mencionar que existen distintos algoritmos para el cálculo de cuantiles 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.

Algoritmos aplicables en cuantiles

Describe el uso de diversos algoritmos

Enlace: algoritmos

Introducción a la graficación

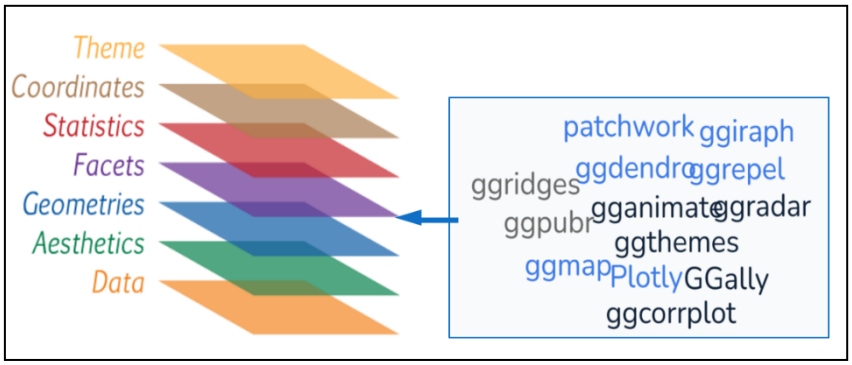
En el presente curso vamos a trabajar sobre el paquete R ggplot2, como su nombre lo indica es la segunda generación de paquete de graficación, previamente el paquete plot era utilizado en la creación de gráficos, pero estos resultaban estáticos y con muy pocas

capacidades de personalización. Ggplot2 fue creado justamente con la idea de que a un gráfico básico se le pueda añadir componentes y efectos visuales a voluntad.

Figura 8

Arquitectura básica de Ggplot2

Fuente: https://ggplot2/tidiverserse.org



A continuación, detallamos el uso de cada capa

Data: Especifica el dataset que se usará, solo tiene un argumento llamado "data". Es importante mencionar que ggplot solo trabaja con data frames, otras estructuras como vectores y matrices deben convertirse previamente. Capas inferiores pueden definir datasts adicionales a fin de combinarlos en un mismo gráfico.

Aesthetics: Especifica el mapeo de los datos a las características gráficas y su función básica es aes().

Geometrics: Especifica tipo de gráfico y combina: datos, mapeo estético, funciones estadísticas (stats) y ajuste de posición. Todas las funciones de esta capa empiezan con "geom_"

Facets: Es una capa que permite desagregar relaciones complejas en pequeños múltiples, muy útil cuando existen múltiples factores en la data. Todas sus funciones empiezan "facet_"

Statistics: Cada geom está asociado a una función estadística, pero puede ser cambiado mediante esta capa para especificar exactamente lo que se desea ver. Todas sus funciones empiezan con "stat_"

Coodinates: El sistema de coordenadas, determina cómo se combinan valores con las coordenadas x/y para colocar elementos en el gráfico. Todas sus funciones empiezan con "coord_"

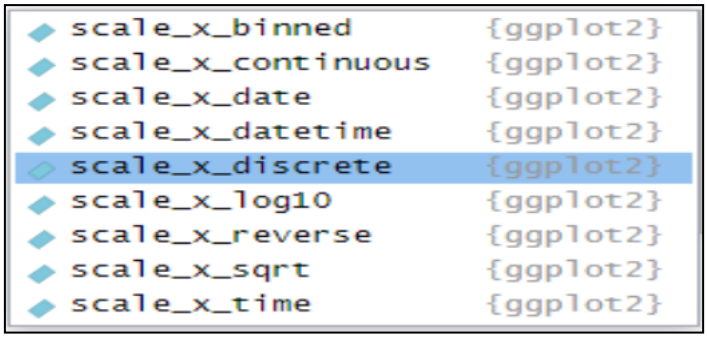
Themes: Los temas controlan los elementos visuales asociados a los datos como colores, líneas, tamaño letra etc. Existen temas preestablecidos y todos empiezan con "theme_"

Además de los mencionado anteriormente es importante entender el concepto de escalas: Las escalas no son nada más que la forma de indicar a ggplot que se debe hacer una transformación. Las transformaciones pueden ser de posición, por ejemplo, cuando se desea cambiar la posición de un valor numérico por su logaritmo, porque de esa forma obtenemos una mejor visualización. Estas pueden aplicarse tanto al eje X como Y .

Tabla 2

Escalas de mapeo de posición en R

Creación del autor Alfonso Prado



De acuerdo a Wickham (2020) "Las escalas de posición se utilizan para controlar las ubicaciones de las entidades visuales en un gráfico y cómo se asignan esas ubicaciones a los valores de los datos. Cada gráfico tiene dos escalas de posición, que corresponden a las funciones x e y. En la mayoría de los casos, esto está claro en la especificación del gráfico, porque el usuario especifica explícitamente las variables asignadas a x e y."

Las escalas de posición continuas más comunes son las funciones predeterminadas scale_x_continuous() y scale_y_continuous(). En el caso más simple, se asignan linealmente desde el valor de los datos a una ubicación en el gráfico. Existen otras escalas de posición para variables continuas (scale_x_log10(), scale_x_sqrt()), la mayoría de las

cuales son funciones de conveniencia que se utilizan para proporcionar un acceso fácil a las transformaciones cuando se desea escalas logarítmicas o raíz cuadrada.

Por último, los mapeos de formas y tamaños son muy simples y básicamente se realiza asignando la variable de dataset a la característica visual.

BIBLIOGRAFIA

Pareja C. (2025) Estadística descriptiva , (https://economipedia.com/definiciones/estadistica-descriptiva.html)

Webster A. ,(2000), Estadística aplicada a los negocios y la economía, Irwin McGraw Hill

Wickham H. (2016), R for Data Science , O´Reiley Media

PUCE Virtual

La excelencia no se improvisa

síguenos