Modulo II

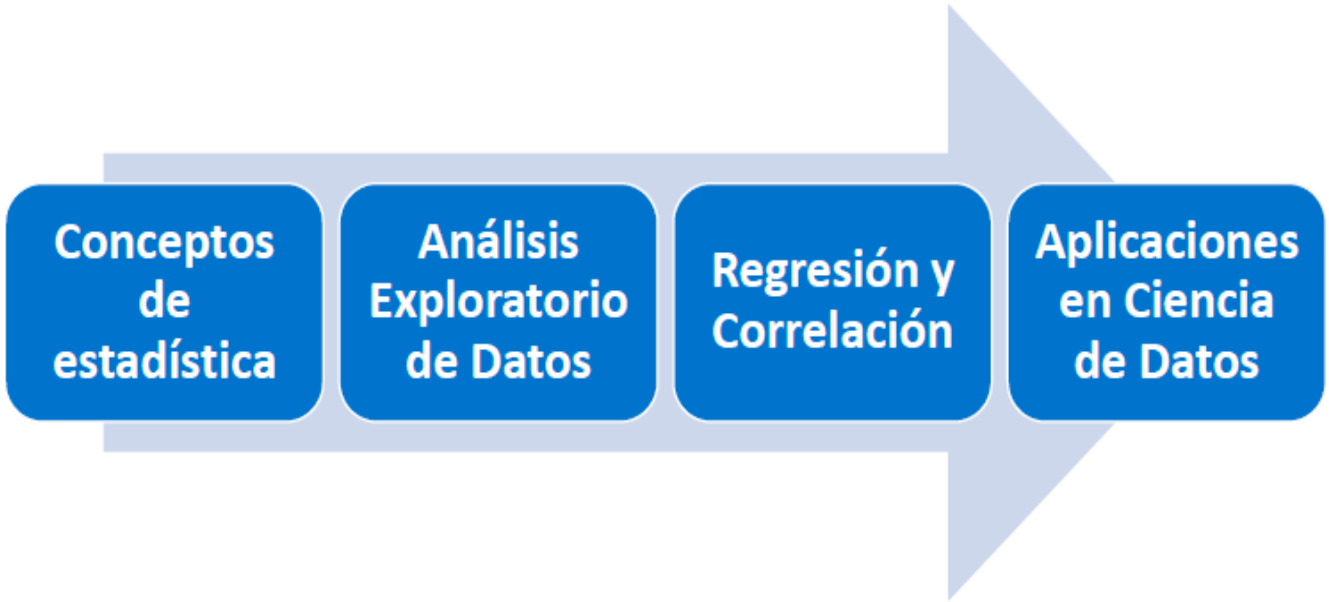
Introducción a la Estadística para Ciencia de Datos

Dr. Edgar Ramírez Galeano

Propósitos específicos del Módulo

Enseñar los fundamentos prácticos de la estadística necesarios para la realización de proyectos, donde los estudiantes deberán utilizar técnicas estadísticas para analizar conjuntos de datos reales, empleando herramientas como R o Python.

2. Introducción a la Estadística para Ciencia de Datos



Bibliografía

- Freedman, D., Pisani, R., & Purves, R. (2007). Estadística. Madrid: Pearson Prentice Hall.
- Freund, J., & Simon, G. (2011). Estadística elemental. México: Pearson Educación.
- Triola, M. F. (2018). Introducción a la estadística. Pearson Educación.
- Agresti, A., & Finlay, B. (2009). Métodos estadísticos para las ciencias sociales. México: Pearson Prentice Hall.
- Infante, S. & Zarate, G., (2012) Métodos Estadísticos. México:Colegio de Postgraduados.

¿Qué es la Ciencia de Datos?

Definición

La ciencia de datos es un campo interdisciplinario que utiliza métodos, procesos, algoritmos y sistemas científicos para extraer conocimiento y perspectivas de datos en diversas formas, estructuradas o no estructuradas.

Interdisciplinariedad en Ciencia de Datos

Definición

La ciencia de datos es interdisciplinaria porque combina conocimientos, técnicas y herramientas de diversas áreas del conocimiento para abordar problemas complejos relacionados con la extracción de conocimiento y perspectivas a partir de datos.

- Matemáticas y Estadística.
- Informática y Tecnología.
- Negocios y Economía.
- Ciencias Sociales y Humanidades.
- Ciencias Naturales y Salud.

Razones por las cuales la Ciencia de Datos es Interdisciplinaria

- **Matemáticas y Estadística:** La ciencia de datos se basa en fundamentos matemáticos y estadísticos para desarrollar modelos, realizar inferencias y analizar datos. Conceptos como álgebra lineal, cálculo, probabilidad y estadística son fundamentales en este campo.
- **Informática y Tecnología:** La ciencia de datos utiliza técnicas de programación, desarrollo de software y tecnología de la información para manipular, almacenar y procesar grandes volúmenes de datos de manera eficiente.
- **Negocios y Economía:** En el contexto empresarial, la ciencia de datos se utiliza para analizar datos financieros, comportamiento del consumidor, optimización de procesos empresariales, pronósticos de ventas, entre otros.

Importancia de la Ciencia de Datos

- La ciencia de datos permite tomar decisiones basadas en evidencia.
- Ayuda a identificar patrones y tendencias en grandes conjuntos de datos.
- Facilita la automatización de tareas y procesos.
- Es fundamental para el desarrollo de inteligencia artificial y aprendizaje automático.
- Tiene aplicaciones en una amplia gama de campos, incluyendo negocios, medicina, finanzas, entretenimiento, etc.

Tareas en Ciencia de Datos

1. Recopilación de datos.
2. Limpieza y preprocesamiento de datos.
3. Análisis exploratorio de datos.
4. Modelado y predicción.
5. Comunicación de resultados.

Herramientas en Ciencia de Datos

- Lenguajes de programación: Python, R, SQL.
- Bibliotecas y frameworks: Pandas, NumPy, Scikit-learn, TensorFlow, Tidyverse.
- Herramientas de visualización: Matplotlib, Seaborn, shiny, ggplot2.
- Entornos de desarrollo integrado (IDE): Spyder, Jupyter Notebook, RStudio.
- Bases de datos: MySQL, PostgreSQL, MongoDB.

¿Qué es la Estadística?

Definición

La estadística es una disciplina que se encarga de recolectar, organizar, analizar, interpretar y presentar datos.

La estadística es esencial en ciencia de datos para obtener descubrir patrones, tendencias, relaciones o características importantes que pueden no ser evidentes a simple vista, pero que proporcionan información valiosa para comprender un fenómeno o problema. y tomar decisiones basadas en datos.

Estadística Descriptiva

Definición

La estadística descriptiva se enfoca en resumir y describir características importantes de un conjunto de datos.

- Medidas de tendencia central: media, mediana, moda.
- Medidas de dispersión: desviación estándar, rango intercuartil.
- Gráficos: histogramas, gráficos de barras, diagramas de caja.

Estadística Inferencial

Definición

La estadística inferencial se utiliza para sacar conclusiones sobre una población basadas en una muestra de datos.

- Pruebas de hipótesis.
- Intervalos de confianza.
- Análisis de regresión.
- Análisis de varianza (ANOVA).

Visualización de Datos

Definición

La visualización de datos utiliza gráficos y diagramas para representar información de manera visual y comprensible.

- Histogramas.
- Gráficos de dispersión.
- Diagramas de caja.
- Gráficos de barras.

Aprendizaje Estadístico

Definición

El aprendizaje estadístico utiliza técnicas estadísticas para construir modelos predictivos a partir de datos.

- Regresión lineal.
- Regresión logística.
- Clasificación y agrupamiento.

Estadística e Incertidumbre

La incertidumbre en estadística se refiere a la falta de certeza o precisión en las estimaciones y resultados obtenidos a partir de datos muestrales. Esta incertidumbre puede surgir de diversas fuentes, como variabilidad inherente en los datos, errores de medición, muestreo y modelos estadísticos.

Ejemplo

Queremos saber la producción de maíz por ha en una región, pero nuestros recursos económicos nos permiten entrevistar sólo a 100 de los 1000 agricultores de la zona

Cualquier conclusión contendrá elementos de incertidumbre, ya que no es imposible que entrevistemos a los 100 peores (o mejores) agricultores y por tanto nuestras afirmaciones, basadas en los datos, subestimará (o sobrestimará) la producción verdadera de maíz.

Estadística e Incertidumbre

- El papel de la Estadística en el caso analizado es es cuantificar la incertidumbre que es inseparable de las conclusiones obtenidas.
- La cuantificación se logra mediante el uso de los conceptos y técnicas de la Probabilidad.
- Con el conocimiento de la probabilidad, podremos desarrollar los métodos estadísticos de naturaleza inductiva, que se conocen como: Inferencia Estadística o Estadística Inductiva.

Conceptos Básicos

En muchas ocasiones, para llevar a cabo una investigación se hacen encuestas, las cuales son dirigidas a una **muestra representativa** de la población.

- **Población:** Es un conjunto de personas, eventos o cosas de las cuales se desea hacer un estudio, y tienen una característica en común

- **Muestra:** Es un subconjunto cualquiera de la población; es importante escoger la muestra en forma aleatoria (al azar), pues así se logra que sea representativa y se puedan obtener conclusiones más afines acerca de las características de la población.

Población y Muestra

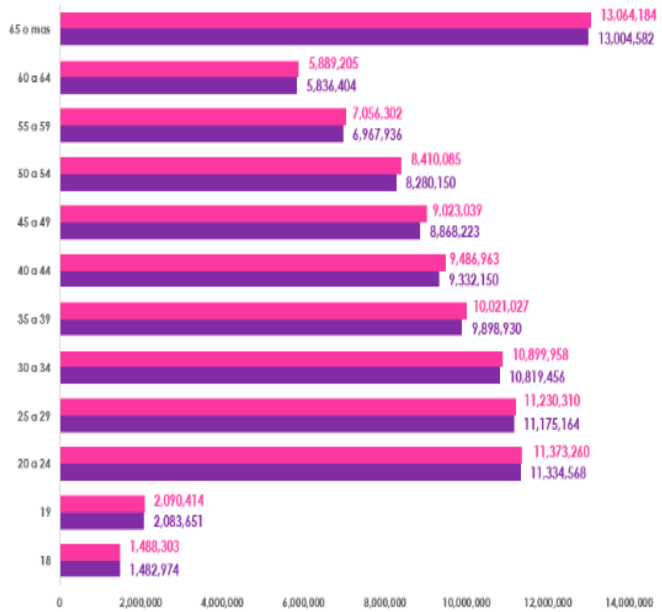
Frente a la dificultad de hacer un censo (estudio de toda la población), se examina una muestra estadística que representará a la totalidad de los sujetos. Con los resultado obtenidos mediante la muestra, se intentará inferir las propiedades de todos los elementos, mediante la estadística inferencial.

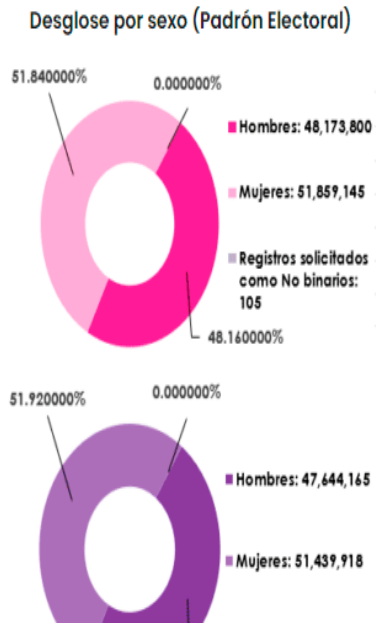
Sheinbaum al frente de la encuesta del Reforma

Total de las y los ciudadanos registrados en Territorio Nacional y Extranjero

Corte al 27 de marzo de 2024

Padrón Electoral	Lista Nominal del Electorado
100,033,050	99,084,188





VARIABLES ESTADÍSTICAS

Es el conjunto de valores que puede tomar cierta característica de la población sobre la que se realiza el estudio estadístico y sobre la que es posible su medición.

- **Variables cualitativas** son aquellas que se refieren a categorías o atributos de los elementos (individuos) estudiados.

- **Variables cuantitativas** son aquellas cuyos datos son de tipo numérico.

TIPOS DE VARIABLES CUALITATIVAS

- **Dicotómicas:** Sólo hay dos categoría, que son excluyentes una de la otra.
  - Ejemplo: planta enferma-sana, se rego-no se rego.
- **Nominal:** Tiene mas de dos categorías y no hay orden entre ellas.
  - Ejemplo: marca de tractores, tipos de sembradora (neumática, mecánica).
- **Ordinal:** Tiene varias categorías y hay orden entre ellas.
  - Ejemplo: grado de salinidad, calidad de la fumigación realizada.

TIPOS DE VARIABLES CUANTITATIVAS

- **Continuas:** La variable puede adquirir cualquier valor dentro de un intervalo de valores determinado.
  - Ejemplo: Consumo de combustible, Potencia del motor.

- **Discretas:** La variable solo puede tomar valores en número determinado de valores. Se asocia con el concepto de conteo.
  - Número de tractores, Número de cosechadoras por estado

Hay ocasiones en las que las medidas cuantitativas continuas son transformadas en ordinales mediante la utilización de uno o varios puntos de corte.
Ejemplo: La variable conductividad hidráulica del suelo es codificada en varias categorías y se utiliza en términos como: permeabilidad alta, media o baja.

Datos HEART

Los datos HEART contienen una variable "HD" de 303 pacientes que se presentaron con dolor en el pecho.

Un valor de resultado de "Yes" indica la presencia de enfermedad del corazón basado en una prueba angiográfica, mientras que "No" significa que no hay enfermedades del corazón.

Descripción de las variables:

- AHD: Diagnóstico de enfermedades del corazón (Yes,No).
- Age: Edad en años.
- Sex: Sexo del paciente (0 Femenino 1 Masculino).
- ChestPain: Tipo de dolor en el pecho (asymptomatic, nonanginal,nontypical, typical).
- RestBP: Presión Arterial en Reposo.
- Chol: Colesterol en sangre en mg/dl (25 a 200 mg/dL normal).
- Fbs: Azúcar en sangre (fasting blood sugar ¿ 120 mg/dl ;1 = SI; 0 = NO)

Datos HEART

Descripción de las variables:

- RestECG: Resultados Electrocardiográficos en Reposo:
  - 0: Normal
  - 1: Tener anomalía en la onda ST-T
  - 2: Mostrar hipertrofia ventricular izquierda
- MaxHR: Frecuencia cardíaca máxima alcanzada.
- ExAng: Angina de pecho inducida por el ejercicio (1=SI 0=NO).
- Oldpeak: Depresión sanguínea inducida por el ejercicio en relación con el reposo.
- Slope: Pendiente del segmento ST(Segmento del electrocardiograma entre la onda S y la T) de ejercicio máximo.
  - 1: Ascendente.
  - 2: Plano.
  - 3: Pendiente descendente.
- Ca: Número de vasos principales (0-3) coloreados por fluoroscopia.
- Thal: 3 = Normal, 6 = Defecto fijo, 7 = Defecto reversible.