Análisis Estadístico para Ciencia de Datos

Este módulo es para estudiantes y profesionales de ciencia de datos que quieren mejorar su análisis estadístico. Veremos desde la teoría hasta la práctica en Python, para que tomes decisiones basadas en datos.

Aprenderás sobre probabilidad, pruebas de hipótesis, simulación Monte Carlo y aplicaciones en ciencia de datos.

¡Iniciemos este aprendizaje!

Dr Mauricio Rosales Rivera

Universidad Autónoma del Estado de Morelos

Centro de Investigación en Ciencias

Fundamentos del Análisis Estadístico

Exploraremos los conceptos básicos del análisis estadístico de una manera más visual:

Tipos de datos

Datos cuantitativos y cualitativos.

Medidas de tendencia central

Media, mediana y moda.

Medidas de dispersión

Rango, varianza y desviación estándar.

Cada tipo de dato influye en las técnicas estadísticas que podemos aplicar, así como las medidas que nos indicarán qué tan dispersos están nuestros datos alrededor de la media.

Probabilidad y Distribuciones

Exploraremos la probabilidad y las distribuciones, con un enfoque en cómo se aplican en diversos escenarios en ciencia de datos. Aprenderemos a generar y visualizar estas distribuciones en Python.

Distribuciones comunes

Exploraremos las distribuciones de probabilidad más comunes y sus características:

- Normal
- Binomial
- Poisson

Generación y visualización en Python

Aprenderemos a generar distribuciones usando:

- NumPy
- SciPy

Y visualizarlas con:

- Matplotlib
- Seaborn

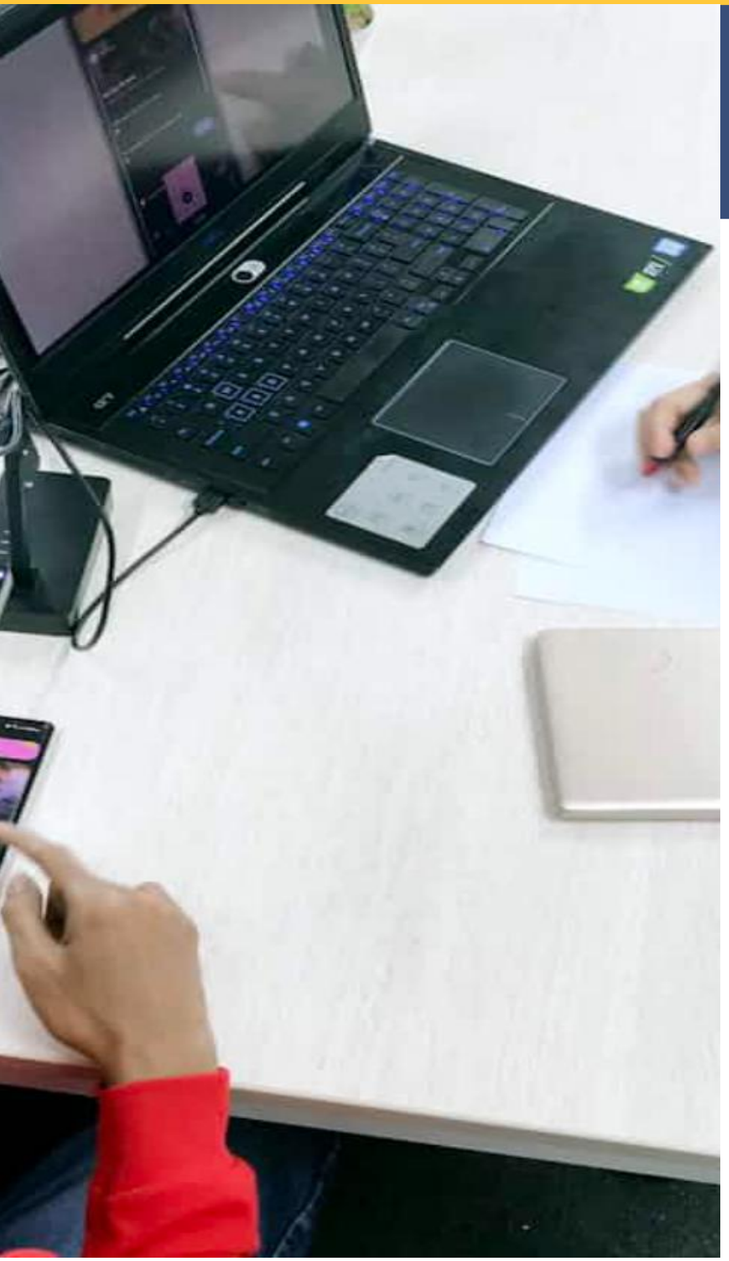
Pruebas de Hipótesis

Abordaremos las **pruebas de hipótesis**, un componente esencial del análisis estadístico. Aprenderemos sobre los **errores** Tipo I (falso positivo) y Tipo II (falso negativo), y cómo afectan nuestras conclusiones.

Entenderemos el concepto del valor **p** y el nivel de significancia **α**, y cómo se utilizan para tomar decisiones sobre si rechazar o no la hipótesis nula. Examinaremos las pruebas paramétricas, que asumen una distribución específica de los datos, y las pruebas no paramétricas, que son más robustas y no requieren tales suposiciones.

1. Errores Tipo I y II
2. Valor p y α
3. Pruebas paramétricas y no paramétricas

Centro de Investigación en Ciencias



A/B Testing en Ciencia de Datos

A/B testing es una técnica poderosa para comparar dos versiones de una variable y determinar cuál funciona mejor. A continuación, los pasos clave en el proceso:

Diseño del experimento

Define la hipótesis y objetivos.

Selección de métricas

Identifica las métricas clave.

Pruebas de hipótesis

Pruebas estadísticas para resultados significativos.

El A/B testing se utiliza en la optimización de sitios web, aplicaciones móviles, campañas de marketing, comparativa de modelos, etc.

Simulación Monte Carlo

Exploraremos la simulación Monte Carlo, una técnica que usa números aleatorios para simular procesos y obtener resultados probabilísticos. Esta técnica tiene aplicaciones en finanzas, ingeniería, ciencia de datos, etc.

Aprenderemos a implementarlo en Python con NumPy y SciPy. Usaremos estas simulaciones para hacer predicciones y tomar decisiones informadas. Ejemplos incluyen la estimación de riesgos, la optimización de portafolios y la simulación de sistemas complejos.

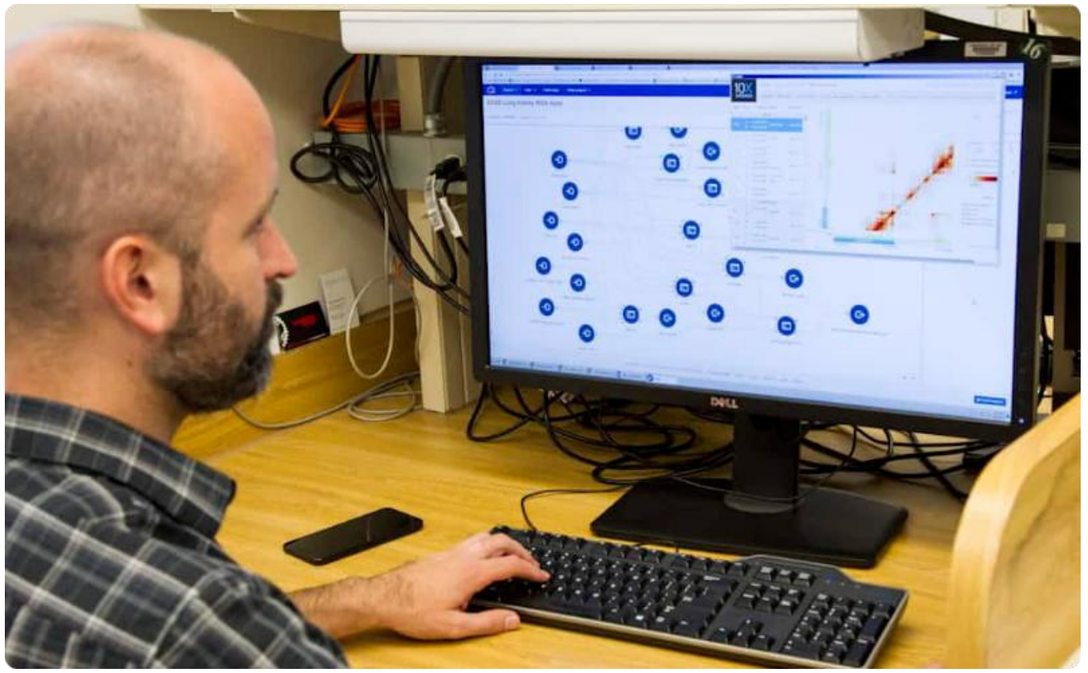
- Números aleatorios
- Predicciones
- Implementación en Python

Aplicaciones en Ciencia de Datos

Exploraremos aplicaciones prácticas de la estadística en ciencia de datos, resaltando web scraping y análisis exploratorio.

Web Scraping

Extracción automatizada de datos de sitios web con Beautiful Soup y Scrapy.



Análisis Exploratorio (EDA)

Resumen y visualización de datos para identificar patrones y anomalías.

Estadística en Machine Learning y Deep Learning

La estadística es clave en Machine Learning y Deep Learning. Nos ayuda a entender y evaluar modelos, elegir las características importantes y evitar el sobreajuste.

Usaremos estadística para seleccionar características con pruebas como chi-cuadrado y ANOVA. Veremos técnicas contra el sobreajuste, como la regularización y la validación cruzada. También, métricas estadísticas para evaluar modelos, como precisión, recall y F1-score.

Selección de características

Evitar el sobreajuste

Métricas estadísticas

Evaluación de Modelos con Métricas Estadísticas

Evaluaremos modelos usando métricas estadísticas, un paso crítico para determinar el rendimiento y las mejoras necesarias.

1 Métricas Clave de Evaluación

Precisión, Recall, F1-Score, AUC-ROC: métricas esenciales para entender el rendimiento del modelo.

2 Técnicas de Validación Cruzada

Utilización de la validación cruzada para obtener estimaciones precisas y confiables del rendimiento del modelo.

3 Importancia de la Interpretación

Aprender a elegir y interpretar las métricas apropiadas para cada tipo de problema de ciencia de datos.