Informe del Proyecto Final de Estadística

Richard Alejandro Matos Arderí
Mauricio Sunde Jiménez

Grupo 311, Ciencia de la Computación.

Facultad de Matemática y Computación
Universidad de La Habana.

2024

Índice

1. Introducción

Este proyecto de análisis estadístico, realizado como parte del plan de estudios de Ciencia de la Computación, se centra en el conjunto de datos "Breast Cancer Wisconsin (Diagnostics)". Este dataset, ampliamente utilizado en la investigación de aprendizaje automático y análisis de datos biomédicos, contiene información crucial para la clasificación de tumores mamarios como benignos o malignos. Nuestro objetivo es ir más allá de una simple clasificación y profundizar en un análisis estadístico exhaustivo, explorando las características de los datos y sus relaciones intrínsecas.

En una primera etapa, emplearemos técnicas de estadística descriptiva para obtener una comprensión inicial del dataset. Esto incluirá el cálculo de medidas de tendencia central (media, mediana, moda) y medidas de dispersión (desviación estándar, varianza, rango intercuartílico) para cada variable, proporcionando una visión general de la distribución de los datos. Además, analizaremos la curtosis para determinar la forma de las distribuciones y la presencia de valores atípicos.

Posteriormente, nos adentraremos en el ámbito de la estadística inferencial. Comenzaremos con la estimación puntual y por intervalos de confianza de parámetros clave, como la media y la proporción, para inferir características de la población a partir de la muestra disponible. Realizaremos pruebas de normalidad (Shapiro-Wilk, Kolmogorov-Smirnov) para determinar si las distribuciones de las variables se ajustan a una distribución normal, un requisito para muchas pruebas paramétricas. Además, llevaremos a cabo pruebas de hipótesis sobre los parámetros de la población, examinando si existen diferencias significativas entre los grupos de tumores benignos y malignos.

Un aspecto fundamental de este proyecto será el análisis de las relaciones entre variables. Emplearemos técnicas de análisis de correlación (Pearson, Spearman) para identificar la fuerza y dirección de la asociación entre las características del tumor. También realizaremos pruebas de independencia de variables (chi-cuadrado) para evaluar si existe una relación estadísticamente significativa entre variables categóricas. Finalmente, exploraremos pruebas de homogeneidad para comparar la distribución de variables entre diferentes grupos, contribuyendo a una comprensión más profunda de las diferencias entre tumores benignos y malignos.

En resumen, este proyecto pretende ofrecer un análisis estadístico completo y riguroso del dataset "Breast Cancer Wisconsin (Diagnostics)", utilizando una variedad de técnicas descriptivas e inferenciales para extraer información relevante y contribuir a una mejor comprensión de las características y relaciones entre los atributos de los tumores mamarios. Los resultados obtenidos permitirán una mejor comprensión de los datos y podrán servir como base para futuros análisis y modelos predictivos.

2. Análisis Descriptivo de los datos

A continuación se muestra un cuadro con todas las variables presentes en el dataset, de conjunto con su clasificación estadística y su escala de medición.

Se brindará especial atención a las variables en rojo para el análisis. A continuación se expone una caracterización más detallada de las mismas:

- diagnosis:
- radius_mean:
- texture_mean:
- perimeter_mean:
- area_mean:
- smoothness_mean:
- compactness_mean:
- symmetry_mean:
- radius_worst:
- texture_worst:
- perimeter_worst:
- area_worst:
- smoothness_worst:
- compactness_worst:
- symmetry_worst:

3. Análisis de la distribución

3.1. Pruebas de Normalidad

3.2. Estimación de parámetros

3.2.1. Estimación Puntual

3.2.2. Estimación por Intervalos

3.3. Pruebas de Hipótesis

3.3.1. Pruebas de Hipótesis para una población

3.3.2. Pruebas de Hipótesis para dos poblaciones

4. Correlación e Independencia

Cuadro 1: Descripción de Variables del Dataset Breast Cancer Wisconsin (Diagnostics)

Variable	**Descripción**	**Clasificación Estadística**	**Escala de Medición**
ID	Identificador único del paciente	Cualitativa	Nominal
diagnosis	Diagnóstico del tumor (1 = maligno, 0 = benigno)	Cualitativa	Nominal
radius_mean	Radio medio del tumor en mm	Continua	Razón
texture_mean	Textura media del tumor	Continua	Razón
perimeter_mean	Perímetro medio del tumor en mm	Continua	Razón
area_mean	Área media del tumor en mm^2	Continua	Razón
smoothness_mean	Suavidad media del tumor	Continua	Razón
compactness_mean	Compacidad media del tumor	Continua	Razón
concavity_mean	Concavidad media del tumor	Continua	Razón
concave points_mean	Puntos cóncavos medios del tumor	Continua	Razón
symmetry_mean	Simetría media del tumor	Continua	Razón
fractal dimension_mean	Dimensión fractal media del tumor	Continua	Razón
radius_se	Desviación estándar del radio del tumor	Continua	Razón
texture_se	Desviación estándar de la textura del tumor	Continua	Razón
perimeter_se	Desviación estándar del perímetro del tumor	Continua	Razón
area_se	Desviación estándar del área del tumor	Continua	Razón
smoothness_se	Desviación estándar de la suavidad del tumor	Continua	Razón
compactness_se	Desviación estándar de la compacidad del tumor	Continua	Razón
concavity_se	Desviación estándar de la concavidad del tumor	Continua	Razón
concave points_se	Desviación estándar de los puntos cóncavos del tumor	Continua	Razón
symmetry_se	Desviación estándar de la simetría del tumor	Continua	Razón
fractal dimension_se	Desviación estándar de la dimensión fractal del tumor	Continua	Razón
radius_worst	Radio máximo del tumor en mm	Continua	Razón
texture_worst	Textura máxima del tumor	Continua	Razón
perimeter_worst	Perímetro máximo del tumor en mm	Continua	Razón
area_worst	Área máxima del tumor en mm^2	Continua	Razón
smoothness_worst	Suavidad máxima del tumor	Continua	Razón
compactness_worst	Compacidad máxima del tumor	Continua	Razón
concavity_worst	Concavidad máxima del tumor	Continua	Razón
concave points_worst	Puntos cóncavos máximos del tumor	Continua	Razón
symmetry_worst	Simetría máxima del tumor	Continua	Razón
fractal dimension_worst	Dimensión fractal máxima del tumor	Continua	Razón