

Bioestadística Aplicada e Interactiva: Guía con R, Python y Aplicaciones Shiny

Deiver Suárez Gómez

2025-06-19

Tabla de contenidos

Preface	4
1 Introduction	6
2 Summary	8
2.1 Estructura Temática	8
2.2 Enfoque Pedagógico	9
2.3 Público Objetivo	9
3 Tipos y Naturaleza de los Datos	10
3.1 Introducción	10
3.2 ¿Qué son los datos?	10
3.3 Tipos de variables	11
3.3.1 Variables cualitativas	11
3.3.2 Variables cuantitativas	11
3.3.3 Tabla resumen	11
3.4 Cargar y explorar datos	12
3.4.1 En R	12
3.4.2 En Python	13
3.5 Conclusión	14
4 Estadística Descriptiva	15
4.1 1. Carga y descripción del conjunto de datos	15
4.2 2. Gráficos descriptivos	17
4.2.1 Boxplot de edad según presencia de ACV	17
4.2.2 Histograma de glucosa promedio	18
4.2.3 Frecuencias y proporciones	20
4.2.4 Aplicación interactiva	20
4.3 5. Recursos audiovisuales	21
4.3.1 Introducción a la estadística descriptiva	21
4.3.2 Visualización de datos en R (boxplots, histogramas)	21
4.3.3 Exploración con Python (Seaborn, pandas)	21
4.4 6. Conclusión	21
5 Probabilidades	22

6	Inferencia	23
7	Regresion	24
	References	25

Preface

La enseñanza de la bioestadística en contextos de salud pública, biología y ciencias biomédicas representa uno de los mayores retos pedagógicos de nuestro tiempo. Enfrentar datos reales, interpretar resultados y comunicar hallazgos de forma clara y rigurosa requiere no solo dominio conceptual, sino también habilidades prácticas y tecnológicas. Este libro nace precisamente de esa necesidad: ofrecer una guía moderna, aplicada e interactiva para aprender bioestadística desde la experiencia, con herramientas computacionales actuales como **R**, **Python** y **Shiny**.

Bioestadística Aplicada e Interactiva ha sido desarrollada a partir del trabajo docente realizado en los cursos **MPH 3102 – Métodos Estadísticos I** y **MPH 7202 – Inferential Statistics**, impartidos en la Universidad de Puerto Rico – Recinto de Mayagüez. A lo largo de múltiples sesiones, se han cubierto desde fundamentos básicos hasta técnicas avanzadas, siempre con una orientación aplicada e intuitiva.

Este libro tiene tres pilares fundamentales:

- **Aplicación práctica:** cada capítulo parte de ejemplos reales en salud pública, medicina, o investigación biológica. Los datos usados provienen de estudios auténticos, accesibles, y pertinentes para los desafíos actuales de investigación.
- **Interactividad:** se ha incorporado el desarrollo de aplicaciones **Shiny** y scripts reproducibles en **R** y **Python** que permiten al lector explorar los conceptos de manera dinámica. No se trata solo de leer, sino de *hacer* estadística.
- **Accesibilidad conceptual:** sin perder el rigor estadístico, se ha privilegiado un lenguaje claro, explicaciones paso a paso, y recursos visuales tomados de las presentaciones utilizadas en clase (transformadas para uso autónomo y progresivo del lector).

Los temas abordados incluyen:

- Estadística descriptiva y visualización de datos
- Probabilidades y distribuciones (Binomial, Poisson, Normal)
- Inferencia: estimaciones, intervalos de confianza, pruebas de hipótesis
- Comparación de grupos: t-tests, ANOVA, pruebas no paramétricas
- Modelos de regresión: lineal simple, múltiple, y logística
- Análisis de frecuencias: tablas de contingencia, chi-cuadrado, prueba exacta de Fisher
- Pruebas no paramétricas: Sign Test, Wilcoxon, Mann–Whitney, Kruskal–Wallis
- Análisis de supervivencia: estimación de curvas de Kaplan–Meier, prueba log-rank, modelo de riesgos proporcionales de Cox

Este libro también está diseñado para acompañarse de un repositorio de materiales interactivos, conjuntos de datos y aplicaciones, que pueden ser consultados y reutilizados por estudiantes e investigadores.

Finalmente, este esfuerzo busca integrar la enseñanza estadística con la capacidad de analizar críticamente datos biomédicos. Que esta guía sirva para formar no solo usuarios de herramientas estadísticas, sino también **pensadores críticos** capaces de transformar datos en decisiones informadas.

Dr. Deiver Suárez Gómez, PhD

Departamento de Biología

Universidad de Puerto Rico – Recinto de Mayagüez

1 Introduction

La bioestadística es una disciplina fundamental en las ciencias de la salud, la biología y la investigación biomédica. Su objetivo principal es proporcionar herramientas que permitan describir, analizar e interpretar datos cuantitativos provenientes de experimentos, estudios clínicos, encuestas y registros médicos. Comprender los principios de la estadística no solo es crucial para evaluar la validez de los hallazgos científicos, sino también para diseñar investigaciones rigurosas y tomar decisiones informadas basadas en evidencia.

Este libro ha sido estructurado con base en más de una docena de sesiones impartidas a estudiantes de maestría en salud pública y biología, organizadas en torno a los siguientes ejes temáticos:

- La **exploración inicial de datos** y la visualización descriptiva, abordando la importancia de las escalas de medición, la estructura de los conjuntos de datos, y las representaciones gráficas fundamentales.
- El uso de **herramientas computacionales modernas** como R y Python para aplicar conceptos estadísticos de forma práctica, reproducible e interactiva.
- La **probabilidad** como lenguaje para modelar la incertidumbre, incluyendo el enfoque clásico, empírico y bayesiano, y su relación con la toma de decisiones.
- El estudio de **distribuciones teóricas** fundamentales como la binomial, la de Poisson y la normal, esenciales para el desarrollo de la inferencia estadística.
- El desarrollo de **técnicas inferenciales**, como los intervalos de confianza y las pruebas de hipótesis, con énfasis en la interpretación correcta de los resultados.
- La comparación entre **modelos paramétricos y no paramétricos**, y la selección adecuada de pruebas según las características del diseño y los datos disponibles.
- La incorporación de **modelos de regresión** lineal y logística, así como el análisis de interacciones, efectos confusores y criterios de selección de variables.
- La enseñanza del **análisis de supervivencia**, incluyendo censura, curvas de Kaplan–Meier, prueba log-rank y modelo de riesgos proporcionales de Cox.

Este libro se diferencia de otros textos de bioestadística por su enfoque **altamente práctico e interactivo**. Cada capítulo incluye ejemplos basados en situaciones reales, ejercicios con datos reales, y aplicaciones **Shiny** que permiten explorar conceptos estadísticos en tiempo real.

Además, el libro ha sido concebido como un recurso integral para la docencia y el autoaprendizaje. No se requiere experiencia previa con programación: el lector será guiado

paso a paso en el uso de código en R y Python, con el objetivo de desarrollar competencia y autonomía en el análisis de datos.

En conjunto, este libro ofrece una experiencia de aprendizaje accesible, actualizada y centrada en la aplicación del conocimiento estadístico. Está dirigido a estudiantes de posgrado, investigadores, profesionales de la salud y docentes que deseen fortalecer su formación cuantitativa y aplicar la estadística de forma rigurosa y efectiva.

2 Summary

Bioestadística Aplicada e Interactiva: Guía con R, Python y Aplicaciones Shiny es un texto integral y didáctico diseñado para estudiantes y profesionales de la salud, biología, y ciencias afines que desean dominar la estadística aplicada en un contexto real y computacional. A partir de una docencia activa y más de una docena de sesiones desarrolladas en cursos como **MPH 3102** y **MPH 7202**, el libro articula teoría, práctica y tecnología para ofrecer un enfoque accesible, moderno e interactivo.

El contenido del libro abarca los fundamentos de la bioestadística, el análisis descriptivo, la teoría de la probabilidad y la inferencia estadística, hasta técnicas avanzadas como regresión múltiple, regresión logística y análisis de supervivencia. Todos los temas están acompañados por ejemplos reproducibles en R y Python, así como aplicaciones interactivas desarrolladas en Shiny que permiten explorar los conceptos de forma visual y práctica.

2.1 Estructura Temática

El libro se organiza en capítulos progresivos que abarcan:

- **Fundamentos de bioestadística:** variables, tipos de datos, escalas de medición y exploración inicial.
- **Visualización y estadística descriptiva:** gráficos, tablas, medidas de tendencia central y dispersión.
- **Teoría de la probabilidad:** enfoques clásico, empírico y bayesiano, eventos y reglas de probabilidad.
- **Distribuciones de probabilidad:** binomial, Poisson y normal, con aplicaciones biomédicas.
- **Inferencia estadística:** estimación de parámetros, intervalos de confianza y pruebas de hipótesis.
- **Comparaciones entre grupos:** t-student, ANOVA, pruebas no paramétricas (Wilcoxon, Kruskal–Wallis).
- **Modelos de regresión:**
 - Regresión lineal simple y múltiple
 - Regresión logística binaria
 - Inclusión de interacciones y análisis de confusión
 - Selección de variables y diagnóstico de modelos

- **Análisis de frecuencias:** tablas de contingencia, pruebas chi-cuadrado y prueba exacta de Fisher.
- **Análisis de supervivencia:** censura, curvas de Kaplan–Meier, log-rank test, modelo de Cox y supuestos.

2.2 Enfoque Pedagógico

Este libro ha sido diseñado no solo como material de consulta, sino como una herramienta **interactiva de aprendizaje autónomo**. Cada capítulo incluye:

- Explicaciones teóricas accesibles
- Casos reales y ejemplos contextualizados
- Código comentado en R y Python
- Ejercicios guiados y soluciones
- Aplicaciones **Shiny** interactivas para visualización y análisis

2.3 Público Objetivo

- Estudiantes de maestría y doctorado en salud pública, biología, epidemiología, psicología y áreas afines
- Profesionales que deseen fortalecer sus competencias en análisis de datos biomédicos
- Docentes que buscan recursos modernos y prácticos para sus cursos

Este libro busca transformar la forma en que se enseña y se aprende bioestadística: desde una práctica pasiva y memorística hacia una experiencia activa, exploratoria y fundamentada en datos reales.

3 Tipos y Naturaleza de los Datos

3.1 Introducción

Los análisis estadístico comienza con datos. Comprender su naturaleza, origen y estructura es fundamental para aplicar correctamente las técnicas estadísticas. Este capítulo introduce los conceptos básicos sobre tipos de datos, variables y su clasificación, ilustrados con ejemplos prácticos en **R** y **Python**.

Al finalizar este capítulo, podrás:

- Definir qué se entiende por datos en el contexto de salud pública
 - Clasificar variables según su naturaleza y escala de medición
 - Reconocer la diferencia entre variables cualitativas y cuantitativas
 - Crear y explorar conjuntos de datos simples en R y Python
-

3.2 ¿Qué son los datos?

En estadística, los **datos** representan observaciones o mediciones recolectadas sobre unidades de análisis: personas, comunidades, eventos u objetos.

En salud pública, los datos pueden provenir de:

- Encuestas poblacionales
- Registros clínicos o epidemiológicos
- Ensayos clínicos
- Estudios de vigilancia

Ejemplos comunes de variables recolectadas incluyen:

- Edad de los pacientes
- Presión arterial
- Nivel socioeconómico
- Diagnóstico de enfermedad

3.3 Tipos de variables

Las variables se pueden clasificar de acuerdo con:

- Su **naturaleza** (cualitativa o cuantitativa)
- Su **escala de medición** (nominal, ordinal, intervalo o razón)

3.3.1 Variables cualitativas

Representan categorías o atributos. No tienen significado numérico.

- **Nominales:** No poseen un orden inherente
Ejemplo: tipo de sangre (A, B, AB, O)
- **Ordinales:** Poseen un orden lógico
Ejemplo: nivel de dolor (leve, moderado, severo)

3.3.2 Variables cuantitativas

Representan cantidades numéricas.

- **Discretas:** Valores enteros contables
Ejemplo: número de hijos
- **Continuas:** Pueden tomar cualquier valor dentro de un intervalo
Ejemplo: estatura, peso corporal

3.3.3 Tabla resumen

Tipo	Subtipo	Ejemplo
Cualitativa	Nominal	Tipo de sangre
Cualitativa	Ordinal	Nivel de dolor (leve, moderado, severo)
Cuantitativa	Discreta	Número de visitas médicas
Cuantitativa	Continua	Índice de masa corporal

3.4 Cargar y explorar datos

A continuación, se presentan ejemplos prácticos para crear y explorar conjuntos de datos simples en **R** y **Python**, simulando variables típicas en salud pública.

3.4.1 En R

```
# Crear un DataFrame en R
datos <- data.frame(
  sexo = c("F", "M", "F", "M"),
  edad = c(23, 35, 29, 41),
  peso = c(55.2, 70.3, 60.1, 82.5)
)
# Ver datos
print(datos)
```

```
  sexo edad peso
1    F   23 55.2
2    M   35 70.3
3    F   29 60.1
4    M   41 82.5
```

```
# Ver estructura
str(datos)
```

```
'data.frame':  4 obs. of  3 variables:
 $ sexo: chr  "F" "M" "F" "M"
 $ edad: num  23 35 29 41
 $ peso: num  55.2 70.3 60.1 82.5
```

```
# Ver resumen
summary(datos)
```

sexo	edad	peso
Length:4	Min. :23.0	Min. :55.20
Class :character	1st Qu.:27.5	1st Qu.:58.88
Mode :character	Median :32.0	Median :65.20
	Mean :32.0	Mean :67.03
	3rd Qu.:36.5	3rd Qu.:73.35
	Max. :41.0	Max. :82.50

3.4.2 En Python

```
# Crear un DataFrame en Python
import pandas as pd

datos = pd.DataFrame({
    "sexo": ["F", "M", "F", "M"],
    "edad": [23, 35, 29, 41],
    "peso": [55.2, 70.3, 60.1, 82.5]
})

# Ver datos
datos
```

	sexo	edad	peso
0	F	23	55.2
1	M	35	70.3
2	F	29	60.1
3	M	41	82.5

```
# Ver estructura
datos.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4 entries, 0 to 3
Data columns (total 3 columns):
#   Column  Non-Null Count  Dtype
---  -
0   sexo    4 non-null         object
1   edad    4 non-null         int64
2   peso    4 non-null         float64
dtypes: float64(1), int64(1), object(1)
memory usage: 224.0+ bytes
```

```
# Ver resumen
datos.describe()
```

	edad	peso
count	4.000000	4.000000
mean	32.000000	67.025000

std	7.745967	12.082874
min	23.000000	55.200000
25%	27.500000	58.875000
50%	32.000000	65.200000
75%	36.500000	73.350000
max	41.000000	82.500000

3.5 Conclusión

Comprender los tipos de variables es el primer paso para aplicar correctamente métodos de análisis estadístico. Esta clasificación influye directamente en:

- La selección de pruebas estadísticas
- La representación gráfica
- El resumen numérico apropiado

En los siguientes capítulos, profundizaremos en la **estadística descriptiva** como base del análisis de datos en salud pública.

4 Estadística Descriptiva

Este capítulo utiliza un conjunto de datos real descargado de [Kaggle: Stroke Prediction Dataset](#), que contiene información demográfica, médica y conductual de pacientes. El objetivo es explorar los datos aplicando técnicas de estadística descriptiva, apoyados por gráficos, código en R y Python, una aplicación Shiny y material audiovisual.

4.1 1. Carga y descripción del conjunto de datos

```
datos <- read.csv("healthcare-dataset-stroke-data.csv")
summary(datos)
```

id	gender	age	hypertension
Min. : 67	Length:5110	Min. : 0.08	Min. :0.00000
1st Qu.:17741	Class :character	1st Qu.:25.00	1st Qu.:0.00000
Median :36932	Mode :character	Median :45.00	Median :0.00000
Mean :36518		Mean :43.23	Mean :0.09746
3rd Qu.:54682		3rd Qu.:61.00	3rd Qu.:0.00000
Max. :72940		Max. :82.00	Max. :1.00000
heart_disease	ever_married	work_type	Residence_type
Min. :0.00000	Length:5110	Length:5110	Length:5110
1st Qu.:0.00000	Class :character	Class :character	Class :character
Median :0.00000	Mode :character	Mode :character	Mode :character
Mean :0.05401			
3rd Qu.:0.00000			
Max. :1.00000			
avg_glucose_level	bmi	smoking_status	stroke
Min. : 55.12	Length:5110	Length:5110	Min. :0.00000
1st Qu.: 77.25	Class :character	Class :character	1st Qu.:0.00000
Median : 91.89	Mode :character	Mode :character	Median :0.00000
Mean :106.15			Mean :0.04873

3rd Qu.:114.09
Max. :271.74

3rd Qu.:0.00000
Max. :1.00000

```
head(datos)
```

	id	gender	age	hypertension	heart_disease	ever_married	work_type
1	9046	Male	67	0	1	Yes	Private
2	51676	Female	61	0	0	Yes	Self-employed
3	31112	Male	80	0	1	Yes	Private
4	60182	Female	49	0	0	Yes	Private
5	1665	Female	79	1	0	Yes	Self-employed
6	56669	Male	81	0	0	Yes	Private

	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
1	Urban	228.69	36.6	formerly smoked	1
2	Rural	202.21	N/A	never smoked	1
3	Rural	105.92	32.5	never smoked	1
4	Urban	171.23	34.4	smokes	1
5	Rural	174.12	24	never smoked	1
6	Urban	186.21	29	formerly smoked	1

```
import pandas as pd
datos = pd.read_csv("healthcare-dataset-stroke-data.csv")
datos.head()
```

	id	gender	age	...	bmi	smoking_status	stroke
0	9046	Male	67.0	...	36.6	formerly smoked	1
1	51676	Female	61.0	...	NaN	never smoked	1
2	31112	Male	80.0	...	32.5	never smoked	1
3	60182	Female	49.0	...	34.4	smokes	1
4	1665	Female	79.0	...	24.0	never smoked	1

[5 rows x 12 columns]

```
datos.describe(include='all')
```

	id	gender	...	smoking_status	stroke
count	5110.000000	5110	...	5110	5110.000000
unique	NaN	3	...	4	NaN
top	NaN	Female	...	never smoked	NaN
freq	NaN	2994	...	1892	NaN

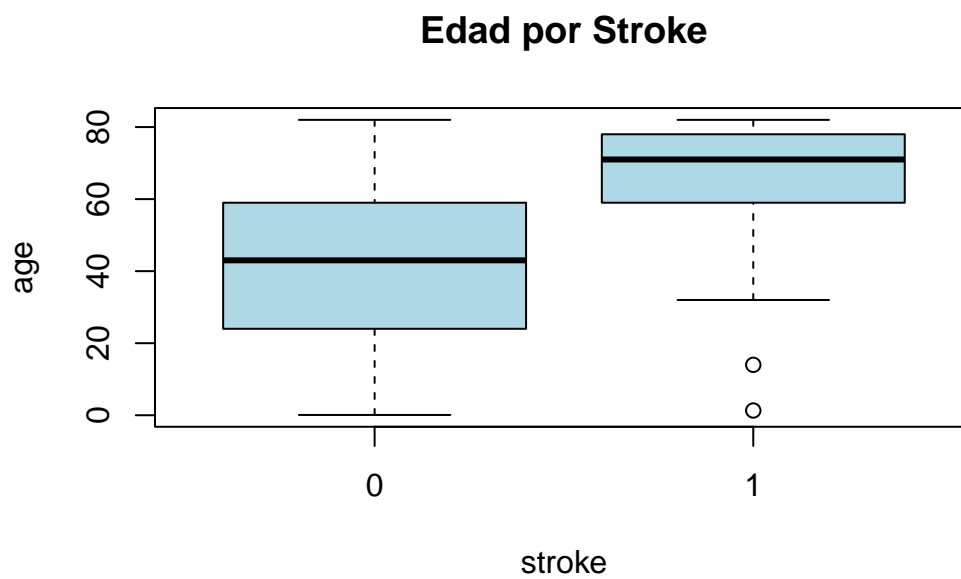
mean	36517.829354	NaN	...	NaN	0.048728
std	21161.721625	NaN	...	NaN	0.215320
min	67.000000	NaN	...	NaN	0.000000
25%	17741.250000	NaN	...	NaN	0.000000
50%	36932.000000	NaN	...	NaN	0.000000
75%	54682.000000	NaN	...	NaN	0.000000
max	72940.000000	NaN	...	NaN	1.000000

[11 rows x 12 columns]

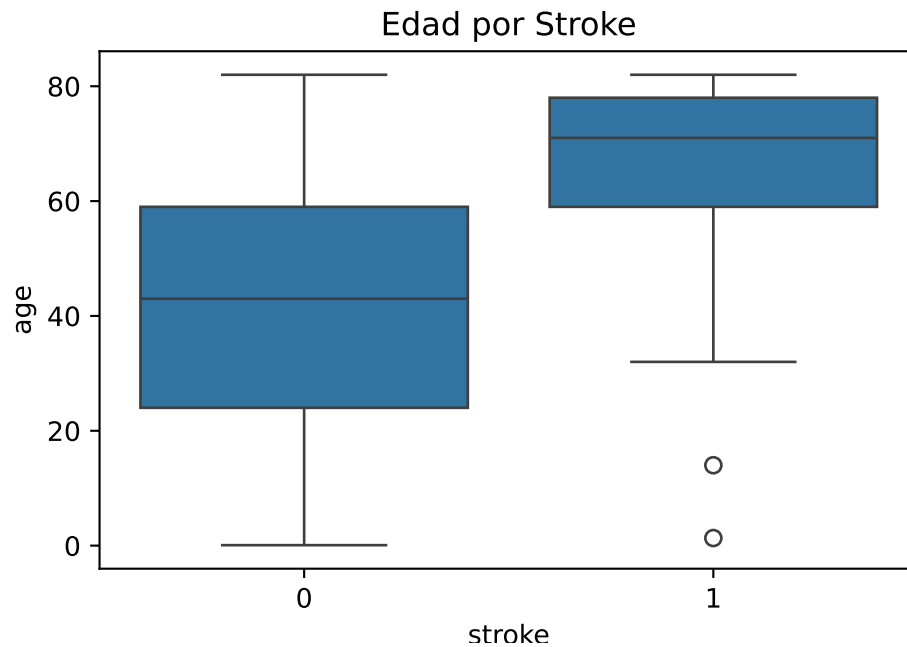
4.2 2. Gráficos descriptivos

4.2.1 Boxplot de edad según presencia de ACV

```
boxplot(age ~ stroke, data = datos, main = "Edad por Stroke", col = "lightblue")
```



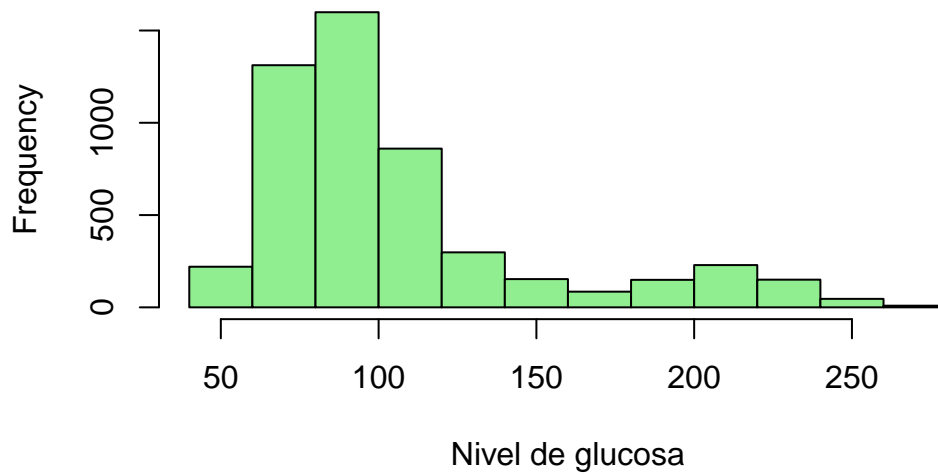
```
import seaborn as sns
import matplotlib.pyplot as plt
sns.boxplot(x="stroke", y="age", data=datos)
plt.title("Edad por Stroke")
plt.show()
```



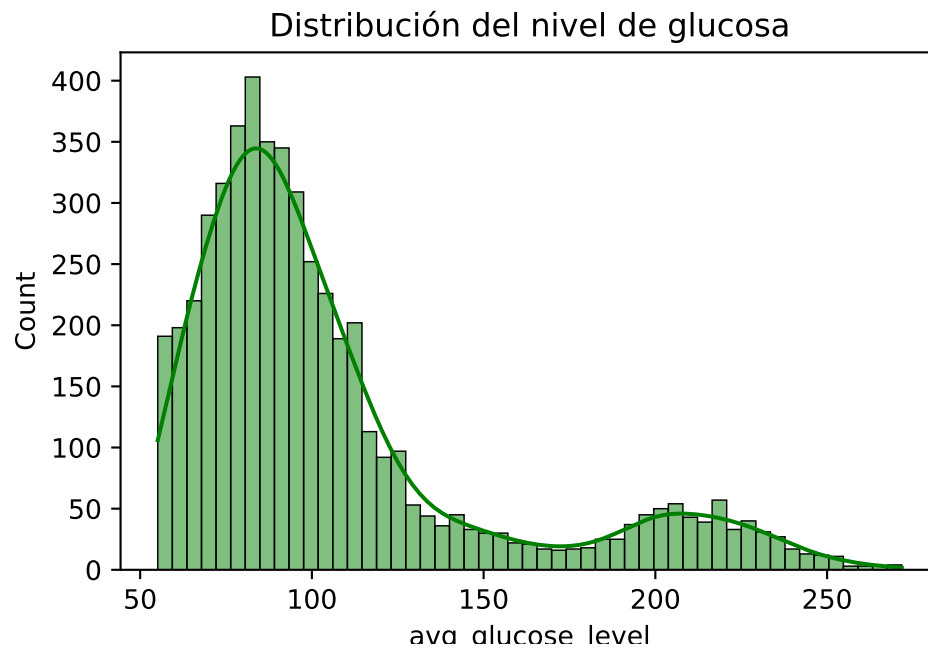
4.2.2 Histograma de glucosa promedio

```
hist(datos$avg_glucose_level, main = "Glucosa Promedio", xlab = "Nivel de glucosa", col = "1")
```

Glucosa Promedio



```
sns.histplot(data=datos, x="avg_glucose_level", kde=True, color="green")  
plt.title("Distribución del nivel de glucosa")  
plt.show()
```



4.2.3 Frecuencias y proporciones

```
table(datos$gender)
```

```
Female  Male  Other
  2994   2115     1
```

```
prop.table(table(datos$stroke))
```

```
      0      1
0.95127202 0.04872798
```

```
print(datos["gender"].value_counts())
```

```
gender
Female    2994
Male      2115
Other         1
Name: count, dtype: int64
```

```
print(datos["stroke"].value_counts(normalize=True))
```

```
stroke
0    0.951272
1    0.048728
Name: proportion, dtype: float64
```

4.2.4 Aplicación interactiva

Como complemento a este capítulo, se ha desarrollado una aplicación interactiva utilizando Shiny que permite explorar conceptos de estadística descriptiva y análisis exploratorio con visualizaciones dinámicas y opciones personalizables para el usuario.

Accede a la app aquí:

https://deiversg.shinyapps.io/app_statistical_Methods/

Video tutorial – ¿Cómo usar la app?

Video tutorial: [¿Cómo usar la app?](#)

4.3 5. Recursos audiovisuales

4.3.1 Introducción a la estadística descriptiva

Video: [Introducción a la estadística descriptiva](#)

4.3.2 Visualización de datos en R (boxplots, histogramas)

Video: [Visualización de datos en R](#)

4.3.3 Exploración con Python (Seaborn, pandas)

Video: [Exploración con Python](#)

4.4 6. Conclusión

El análisis exploratorio de datos con herramientas como R y Python permite obtener una comprensión inicial robusta de los patrones en datos biomédicos. Esto es esencial antes de aplicar modelos predictivos como regresión o clasificación. El uso de gráficos y resúmenes numéricos fortalece la interpretación clínica y estadística de los fenómenos observados.

5 Probabilidades

6 Inferencia

7 Regresion

References