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 ¿Qué es la Bioestadística? **10**
3.0.1 Ejemplos: 10

4 Motivación para estudiar estadística **11**

5 Fuentes de datos **13**

6 Variables **14**
6.1 Tipos de variables 14

7 Variables aleatorias **15**
7.0.1 Ejemplo: 15

8 Escalas de medición **16**

9 Población y muestra **17**
9.0.1 Ejemplo: 17

10 Muestreo aleatorio simple **18**
10.0.1 Definición 18
10.0.2 Ventajas 18
10.0.3 Desventajas 18
10.0.4 En R 18
10.0.5 En Python 19

11 Conclusión **20**

12 Estadística Descriptiva **21**
12.1 1. Carga y descripción del conjunto de datos 21
12.2 2. Gráficos descriptivos 23
12.2.1 Boxplot de edad según presencia de ACV 23
12.2.2 Histograma de glucosa promedio 24
12.2.3 Frecuencias y proporciones 26
12.2.4 Aplicación interactiva 26
12.3 5. Recursos audiovisuales 27
12.3.1 Introducción a la estadística descriptiva 27
12.3.2 Visualización de datos en R (boxplots, histogramas) 27
12.3.3 Exploración con Python (Seaborn, pandas) 27
12.4 6. Conclusión 27

13 Probabilidades **28**

14 Inferencia **29**

15 Regresion **30**

References **31**

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 ¿Qué es la Bioestadística?

La **bioestadística** es una rama de la estadística esencial para entender la vida a través de los datos. Se aplica a información proveniente de organismos vivos, desde células individuales hasta poblaciones humanas enteras, y permite traducir esos datos en conocimiento útil para la salud pública, la medicina, la biología y otras ciencias de la vida.

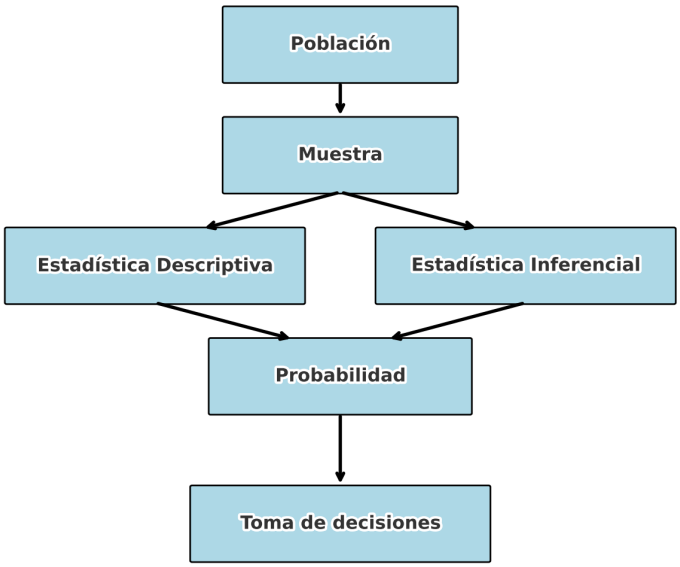
Su propósito es interpretar, analizar y contextualizar datos complejos para tomar decisiones informadas. Esto va desde evaluar si un tratamiento médico es efectivo, hasta identificar factores de riesgo en enfermedades crónicas, modelar la propagación de epidemias o entender cómo influyen los determinantes sociales en la salud.

3.0.1 Ejemplos:

- En una pandemia, la bioestadística permite estimar la velocidad de transmisión de un virus y proyectar su impacto.
- En un hospital, ayuda a decidir si un nuevo medicamento mejora significativamente la supervivencia de pacientes con cáncer.
- En biología molecular, permite identificar genes cuya expresión cambia en respuesta a una enfermedad.

La bioestadística nos invita a pensar críticamente y a ver más allá de los números: a encontrar patrones, formular hipótesis y validar teorías con base en evidencia. En última instancia, es una disciplina que busca mejorar vidas a través de decisiones basadas en datos confiables.

4 Motivación para estudiar estadística



La estadística es esencial en la investigación científica porque actúa como el lenguaje con el que interpretamos la variabilidad inherente al mundo natural y social. Nos permite responder preguntas complejas con evidencia numérica, y transforma la incertidumbre en conocimiento útil y accionable.

Gracias a la estadística, podemos:

- **Extraer conclusiones significativas** a partir de datos, por ejemplo, determinar si un tratamiento realmente mejora la salud de los pacientes o si una intervención de salud pública reduce la incidencia de una enfermedad.
- **Medir la incertidumbre**, lo cual es fundamental cuando se trabaja con muestras en lugar de poblaciones completas. Esto permite comunicar el grado de confianza en los resultados obtenidos, algo crucial al tomar decisiones clínicas o políticas.

- **Tomar decisiones fundamentadas** basadas en análisis objetivos. Desde la asignación de recursos hospitalarios hasta la aprobación de nuevas vacunas, la estadística provee la base para justificar esas decisiones.

En esencia, la estadística no solo ayuda a responder "¿qué está ocurriendo?", sino también "¿por qué?", "¿cuánto?" y "¿con qué grado de certeza?". Es una brújula intelectual que guía la ciencia hacia conclusiones robustas y reproducibles.

5 Fuentes de datos

Los datos en ciencias de la salud provienen de múltiples fuentes:

- Registros clínicos
- Encuestas poblacionales
- Ensayos clínicos
- Observaciones experimentales

6 Variables

Una **variable** es una característica que puede tomar diferentes valores entre individuos, objetos o situaciones. Ejemplos: estatura, peso, edad.

6.1 Tipos de variables

Tipo	Subtipo	Ejemplos
Cualitativas	Nominales	Sexo, diagnóstico médico, etnicidad
	Ordinales	Nivel educativo, escala de dolor
Cuantitativas	Discretas	Número de hijos, admisiones hospitalarias
	Continuas	Edad, IMC, presión arterial

Ejercicio

Clasifique las siguientes variables como nominal, ordinal, discreta o continua:

- Estado civil:
- Nivel educativo:
- Medallas de los Juegos Panamericanos:
- Calificaciones en un curso (A, B, C, D, F):
- Puntos anotados en un juego de baloncesto:
- Número de habitantes de un municipio:
- Ingreso anual de una familia en Puerto Rico:

7 Variables aleatorias

Una **variable aleatoria** es aquella cuyo valor depende del azar. En salud, por ejemplo, la estatura, el peso o la edad son variables aleatorias, ya que varían entre personas.

7.0.1 Ejemplo:

Si lanzamos una moneda:

- Variable aleatoria (X):
  - Resultado: cara (1) o cruz (0)

Ejemplos en salud

- Número de hijos por madre
- Nivel de glucosa en sangre
- Respuesta al tratamiento

8 Escalas de medición

Las escalas de medición determinan cómo se interpretan los valores observados:

Escala	Características	Ejemplos
Nominal	No orden	Sexo, tipo de sangre
Ordinal	Orden, sin distancias fijas	Grado de dolor
Intervalo	Orden y distancia, sin cero absoluto	Temperatura en °C
Razón	Orden, distancia y cero absoluto	Peso, edad, ingresos

9 Población y muestra

- **Población:** Conjunto total de elementos que se desea estudiar (personas, células, plantas, etc.). Puede ser finita o infinita.
- **Muestra:** Subconjunto de la población seleccionado para su análisis.

9.0.1 Ejemplo:

- Población: Todos los estudiantes de una universidad
- Muestra: 100 estudiantes seleccionados aleatoriamente

10 Muestreo aleatorio simple

10.0.1 Definición

Una muestra de tamaño (n) se selecciona de una población de tamaño (N), de forma que cada elemento tiene la misma probabilidad de ser elegido.

10.0.2 Ventajas

- Todos los individuos tienen igual oportunidad.
- Resultados más generalizables si el tamaño es adecuado.

10.0.3 Desventajas

- Se necesita una lista completa de la población.
- Muestras pequeñas pueden no ser representativas.

10.0.4 En R

```
poblacion <- 1:100
set.seed(123)
# Sin reemplazo
muestra1 <- sample(poblacion, size = 10, replace = FALSE)
muestra1
```

```
 [1] 31 79 51 14 67 42 50 43 97 25
```

```
# Con reemplazo
muestra2 <- sample(poblacion, size = 10, replace = TRUE)
muestra2
```

```
 [1] 90 91 69 91 57 92  9 93 99 72
```

10.0.5 En Python

```
import random
poblacion = list(range(1, 101))
random.seed(123)
# Sin reemplazo
muestra1 = random.sample(poblacion, 10)
print("Muestra sin reemplazo:", muestra1)
```

```
Muestra sin reemplazo: [7, 35, 12, 99, 53, 14, 5, 49, 69, 72]
```

```
# Con reemplazo
muestra2 = [random.choice(poblacion) for _ in range(10)]
print("Muestra con reemplazo:", muestra2)
```

```
Muestra con reemplazo: [43, 44, 7, 21, 18, 44, 72, 43, 90, 32]
```

11 Conclusión

Comprender el tipo de datos, las variables involucradas y cómo se seleccionan las muestras es fundamental para aplicar correctamente los métodos estadísticos en salud. Estos conceptos son la base para avanzar hacia la estadística descriptiva e inferencial.

12 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.

12.1 1. Carga y descripción del conjunto de datos

```r
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                                          3rd Qu.:0.00000
 Max.   :271.74                                          Max.   :1.00000
```

```r
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
```

```python
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]
```

```python
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
```

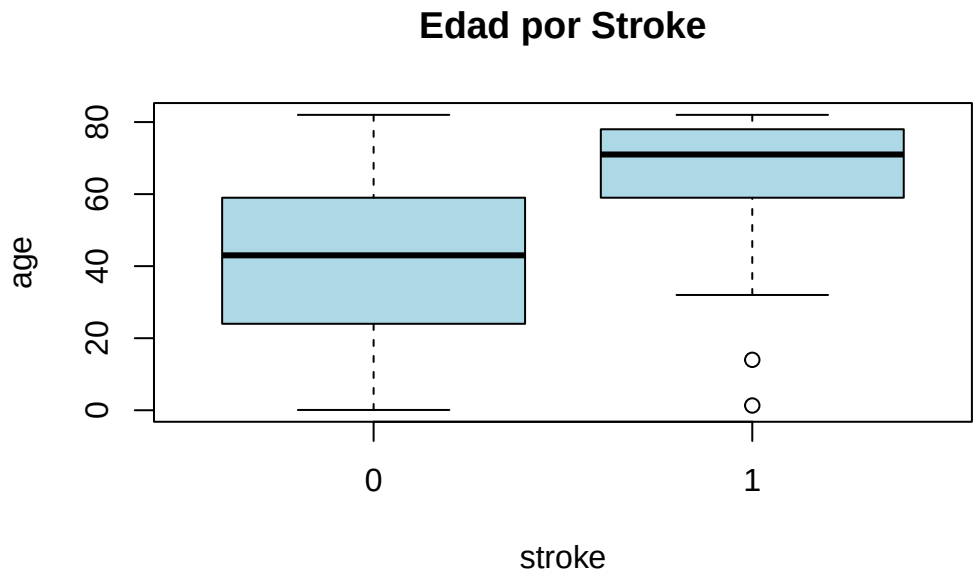
```
mean    36517.829354     NaN  ...             NaN     0.048728
std     21161.721625     NaN  ...             NaN     0.215320
min        67.000000     NaN  ...             NaN     0.000000
25%     17741.250000     NaN  ...             NaN     0.000000
50%     36932.000000     NaN  ...           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

[11 rows x 12 columns]
```

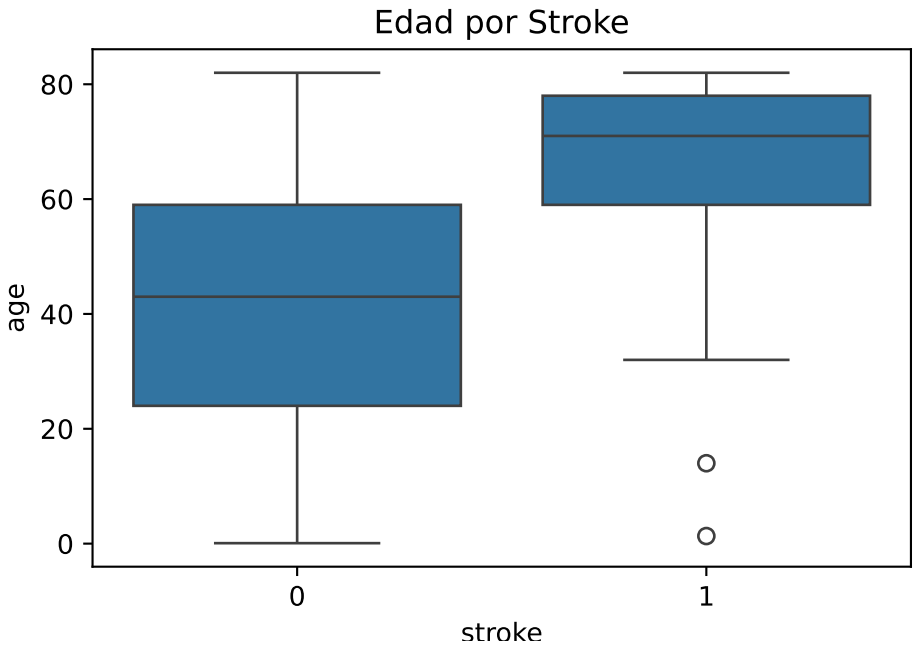
12.2 2. Gráficos descriptivos

12.2.1 Boxplot de edad según presencia de ACV

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

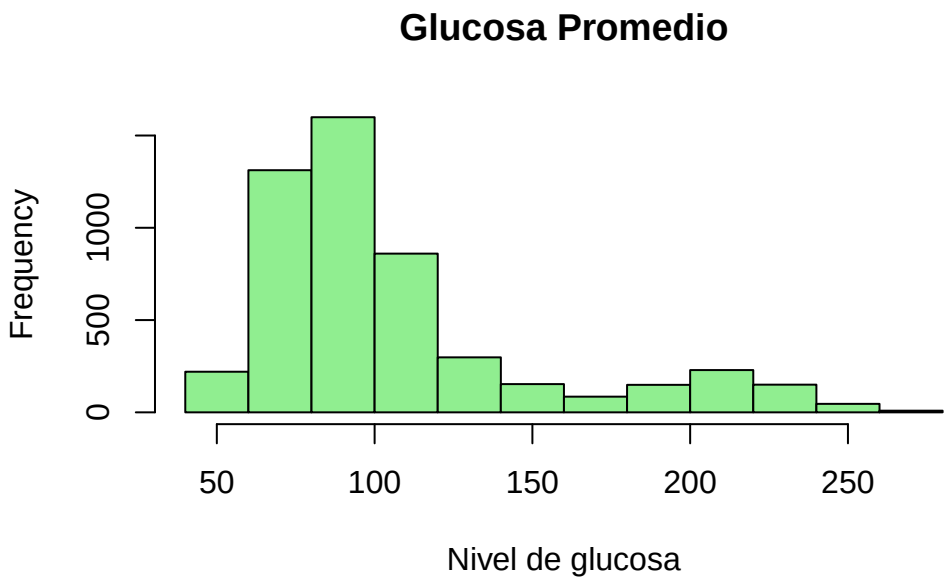


```
import seaborn as sns
import matplotlib.pyplot as plt
sns.boxplot(x="stroke", y="age", data=datos)
plt.title("Edad por Stroke")
plt.show()
```

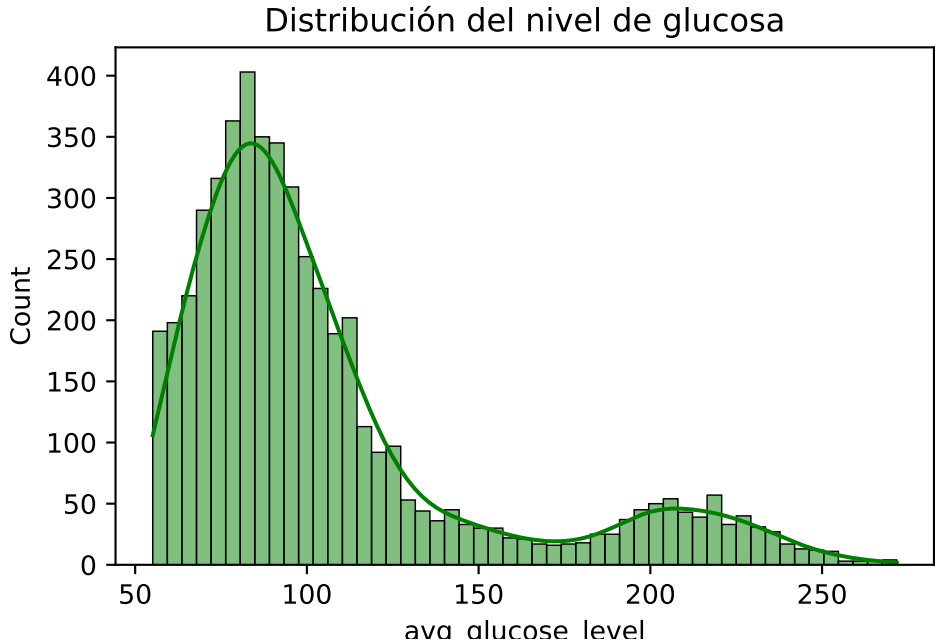


12.2.2 Histograma de glucosa promedio

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




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



12.2.3 Frecuencias y proporciones

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

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

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

```
         0          1
0.95127202 0.04872798
```

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

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

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

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

12.2.4 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?

12.3 5. Recursos audiovisuales

12.3.1 Introducción a la estadística descriptiva

Video: Introducción a la estadística descriptiva

12.3.2 Visualización de datos en R (boxplots, histogramas)

Video: Visualización de datos en R

12.3.3 Exploración con Python (Seaborn, pandas)

Video: Exploración con Python

12.4 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.

13 Probabilidades

14 Inferencia

15 Regresion

References