Contents

Módulo 3, Clase 1: Introducción al Modelado de Datos de Entrada y Distribuciones Comunes

1. Objetivos Específicos de la Clase:

- Comprender la importancia del modelado de datos de entrada en simulaciones.
- Identificar las distribuciones de probabilidad más comunes utilizadas para modelar datos de entrada.
- Conocer métodos gráficos exploratorios para la selección inicial de una distribución.
- Comprender el concepto de ajuste de una distribución a un conjunto de datos.
- Familiarizarse con el software R y su entorno para el análisis estadístico.

2. Contenido Teórico Detallado:

2.1. Introducción al Modelado de Datos de Entrada:

En la simulación, los datos de entrada representan los elementos del mundo real que impulsan el modelo. La calidad de estos datos impacta directamente la validez y confiabilidad de los resultados de la simulación. El modelado de datos de entrada implica:

- **Recolección de datos:** Obtención de datos relevantes del sistema real.
- **Análisis de datos:** Exploración y resumen de los datos recolectados.
- **Selección de distribución:** Identificación de una distribución de probabilidad que represente adecuadamente los datos.
- **Estimación de parámetros:** Cálculo de los parámetros de la distribución seleccionada.
- **Validación:** Verificar que la distribución ajustada represente adecuadamente los datos.

Un modelo de datos de entrada mal especificado puede llevar a conclusiones erróneas sobre el sistema real.

2.2. Distribuciones de Probabilidad Comunes:

Existen diversas distribuciones de probabilidad que son útiles para modelar datos de entrada. Algunas de las más comunes son:

- **Distribución Uniforme:** Adecuada cuando todos los valores dentro de un rango tienen la misma probabilidad de ocurrencia. Útil cuando se tiene poca información sobre la distribución subyacente. (Ejemplo: Tiempo entre llegadas cuando se sabe que es entre 5 y 10 minutos, pero no se tiene más información).
- **Distribución Exponencial:** Se utiliza comúnmente para modelar el tiempo entre eventos independientes, como el tiempo entre llegadas en una cola. Caracterizada por su propiedad de "falta de memoria". (Ejemplo: Tiempo entre fallas de un componente).
- **Distribución Normal (Gaussiana):** Aparece frecuentemente en fenómenos naturales y es útil para modelar variables que son la suma de muchos efectos independientes. (Ejemplo: Altura de personas en una población).
- **Distribución Triangular:** Útil cuando solo se conoce el valor mínimo, máximo y más probable de una variable. (Ejemplo: Duración de una tarea en un proyecto cuando se tienen estimaciones optimistas, pesimistas y más probables).
- **Distribución Gamma:** Flexible y puede modelar una variedad de formas de datos. (Ejemplo: Tiempo de servicio en una cola o tiempo hasta la falla de un sistema).
- **Distribución de Poisson:** Modela el número de eventos que ocurren en un intervalo de tiempo o espacio fijo. (Ejemplo: Número de clientes que llegan a una tienda en una hora).

2.3. Métodos Gráficos Exploratorios:

Antes de ajustar formalmente una distribución, es útil explorar visualmente los datos. Algunos métodos comunes son:

- **Histogramas:** Muestran la frecuencia de los valores de los datos en intervalos. Permiten visualizar la forma de la distribución y detectar posibles asimetrías o multimodalidad.
- **Diagramas de dispersión:** Útiles para identificar relaciones entre dos variables. Aunque menos comunes para la selección inicial de distribuciones univariadas, pueden revelar patrones importantes si se tienen datos multivariados.
- **Boxplots:** Representan la mediana, los cuartiles y los valores atípicos de los datos. Útiles para comparar distribuciones y detectar asimetrías.
- **Gráficos Q-Q (Quantile-Quantile):** Comparan los cuantiles de los datos con los cuantiles de una distribución teórica. Si los datos se ajustan bien a la distribución teórica, los puntos del gráfico Q-Q estarán cerca de una línea recta.

2.4. Ajuste de una Distribución:

El proceso de ajuste de una distribución implica encontrar los parámetros de la distribución que mejor se ajusten a los datos observados. Esto se puede hacer utilizando diferentes métodos de estimación, como:

- **Método de Máxima Verosimilitud (MLE):** Encuentra los valores de los parámetros que maximizan la probabilidad de observar los datos dados.
- **Método de Momentos:** Estima los parámetros igualando los momentos muestrales (media, varianza, etc.) a los momentos teóricos de la distribución.

El paquete `fitdistrplus` en R proporciona funciones para ajustar distribuciones utilizando varios métodos de estimación.

2.5. Introducción a R:

R es un lenguaje de programación y un entorno de software para análisis estadístico y gráficos. Es ampliamente utilizado en la simulación y el modelado de datos.

- **Instalación de R y RStudio:** R se puede descargar e instalar desde el sitio web oficial de R Project. RStudio es un entorno de desarrollo integrado (IDE) para R que facilita la escritura y ejecución de código R.
- **Paquetes:** R utiliza paquetes para extender su funcionalidad. `fitdistrplus` es un paquete importante para el ajuste de distribuciones. Los paquetes se instalan utilizando la función `install.packages()`.
- **Funciones básicas:** R proporciona una amplia gama de funciones para análisis estadístico, manipulación de datos y gráficos.

3. Ejemplos o Casos de Estudio:

Caso de Estudio 1: Tiempo entre llegadas de clientes a un banco.

Supongamos que se han recolectado datos sobre el tiempo entre llegadas de clientes a un banco durante una hora. Los datos (en minutos) son: 1.2, 2.5, 0.8, 3.1, 1.9, 0.5, 2.8, 1.5, 2.2, 1.0.

- **Análisis exploratorio:** Se crea un histograma de los datos para visualizar la distribución. Se observa que los datos tienen una forma asimétrica hacia la derecha, lo que sugiere una posible distribución exponencial.
- **Ajuste de distribución:** Se ajusta una distribución exponencial a los datos utilizando el paquete `fitdistrplus` en R.
- **Interpretación:** Se obtienen los parámetros estimados de la distribución exponencial (tasa). Se puede usar esta distribución para simular futuras llegadas de clientes al banco.

4. Problemas Prácticos o Ejercicios con Soluciones:

Ejercicio 1:

Se han recolectado datos sobre el tiempo de reparación de una máquina (en horas): 5, 7, 3, 8, 6, 4, 9, 5, 7, 6.

1. Crea un histograma de los datos utilizando R.
2. ¿Qué tipo de distribución crees que podría ser apropiada para modelar estos datos? Justifica tu respuesta.

Solución:

1. Código R:

“‘R

Datos del tiempo de reparación

tiempo_reparacion <- c(5, 7, 3, 8, 6, 4, 9, 5, 7, 6)

Crear un histograma

hist(tiempo_reparacion, main=”Histograma del Tiempo de Reparación”, xlab=”Tiempo (horas)”, ylab=”Frecuencia”, col=”lightblue”, border=”black”) “‘

1. Dado que los datos parecen tener una forma unimodal y relativamente simétrica, una distribución Normal o Gamma podría ser apropiada. Se necesitaría un análisis más detallado y pruebas de bondad de ajuste para determinar cuál de las dos se ajusta mejor.

Ejercicio 2:

¿Cuál es la diferencia clave entre una distribución uniforme y una distribución exponencial en términos de sus aplicaciones en el modelado de datos de entrada?

Solución:

La distribución uniforme se utiliza cuando se conoce el rango de valores posibles, pero no se tiene información sobre la probabilidad relativa de cada valor dentro de ese rango. La distribución exponencial se utiliza para modelar el tiempo entre eventos independientes, como llegadas en un sistema de colas, y se caracteriza por su propiedad de ”falta de memoria”.

5. Materiales Complementarios Recomendados:

- **Libro:** ”Simulation Modeling and Analysis” de Averill M. Law.
- **Artículo:** ”Input Data Analysis” de Barry L. Nelson.
- **Tutorial:** Tutoriales online sobre el uso del paquete `fitdistrplus` en R. Buscar en Google ”fitdistrplus tutorial”.
- **Documentación:** Documentación oficial del paquete `fitdistrplus` en CRAN (Comprehensive R Archive Network).

Esta clase proporciona una base sólida para comprender el modelado de datos de entrada y el uso de distribuciones de probabilidad comunes. En las siguientes clases, se profundizará en las técnicas de ajuste, la validación de modelos y el uso de software estadístico.