Introducción al análisis de supervivencia avanzada

Juan Carlos Salazar Uribe
Ehidy Karime García Cruz
Carlos Gaviria Peña
Verónica Guarín Escudero

Universidad de San Buenaventura

e-ISBN: 978-958-8474-93-9

Introducción al análisis de supervivencia avanzada

UNIVERSIDAD DE SAN BUENAVENTURA

Colección Perfiles

Introducción al análisis de supervivencia avanzada

Juan Carlos Salazar Uribe
Ehidy Karime García Cruz
Carlos Gaviria Peña
Verónica Guarín Escudero

2020

Universidad de San Buenaventura
Introducción al análisis de supervivencia avanzada/Juan Carlos Salazar Uribe... [et al.]. –Medellín: Editorial Bonaventuriana, 2020
237 p. −−(Colección Perfiles)
Incluye referencias bibliográficas
e–ISBN: 978-958-8474-93-9
1. Análisis de supervivencia 2. Estadística 3. Martingalas (matemáticas) 4. Estimadores (estadística) 5. Modelos lineales (estadística) 6. Procesos estocásticos
519,546(CDD23)
U588

EDITORIAL BONAVENTURIANA

Colección Perfiles

Introducción al análisis de supervivencia avanzada

Facultad de Ingenierías
Universidad de San Buenaventura Medellín

Universidad de San Buenaventura Colombia

Universidad de San Buenaventura Medellín
Coordinación Editorial Medellín
Carrera 56c No. 51-110 (Medellín)
Calle 45 No 61 − 40 (Bello)
PBX: 57 (4) 5145600
editorial.bonaventuriana@usb.edu.co
www.usbmed.edu.co -
www.editorialbonaventuriana.usb.edu.co

Coordinador Editorial: Fraidy Alonso Alzate Pamplona
Asistente Editorial: Ezequiel Quintero Gallego
Corrección de estilo: Ezequiel Quintero Gallego
Diseño y diagramación: Carlos Gaviria Peña
Ilustración de carátula: María Gretel Álvarez Giraldo

e–ISBN: 978-958-8474-93-9
Cumplido el Depósito Legal (Ley 44 de 1993, Decreto 460 de 1995 y Decreto 358 de 2000).
Septiembre de 2020

Prefacio

Este libro es el producto de varios años de experiencia dictando material relacionado con análisis de supervivencia al nivel de pregrado y posgrado en la Universidad Nacional de Colombia, sede Medellín. Lo aquí expuesto surge a partir del material dictado en el curso de *Tópicos avanzados de análisis de supervivencia* que se dicta regularmente dentro del plan de estudios del doctorado en ciencias-estadística de la Universidad Nacional de Colombia, sede Medellín. Además, tiene la ventaja adicional de estar escrito en idioma español.

Cuando obtuve mi título doctoral en Estadística en el año 2004 en University of Kentucky, me di cuenta de la importancia del análisis de supervivencia en el manejo de datos relacionados con el tiempo que transcurre hasta la ocurrencia de un evento y su gran potencial como herramienta generadora de conocimiento basado en datos en nuestro medio. Por esta razón decidí formular primero un curso sobre el tema a nivel de pregrado y posteriormente un curso a nivel de posgrado. Este último fue mejorando gradual y sustancialmente, y con la valiosa retroalimentación de algunos estudiantes, fue progresando hasta alcanzar la forma que presentamos en este texto. En principio notamos que no había disponible mucho material relacionado con análisis de supervivencia avanzado en español y esto nos inspiró a emprender este proyecto con la convicción de que la forma en que está organizado el material puede ayudar a otros a comprender y a usar esta valiosa herramienta estadística.

Pensamos que nuestro libro es útil e importante ya que desarrolla muchas de las pruebas de resultados clásicos de una forma muy detallada y completa y además ilustra muchas de ellas con ejemplos basados en datos reales. Estos aspectos son particularmente necesarios para un estudiante que enfrenta este tipo de material por primera vez.

El análisis de supervivencia ha sido ampliamente estudiado en la literatura desde la propuestas de Kaplan y Meier, Nelson, Aalen y Cox (Kaplan y Meier, 1958; Nelson, 1969; Cox, 1972; Aalen, 1976). Existen excelentes libros sobre el tema que varían de complejidad, algunos con un enfoque práctico (Kleibaum, 1996; Collett, 2003; Allison, 2010) y otros con un enfoque puramente teórico (Andersen, Borgan, Gill, y Keiding, 1993; Fleming y Harrington, 1991). Este libro presenta una introducción al análisis de supervivencia avanzada y se recomienda como material de apoyo para los cursos de posgrado que se relacionen con estadística (como prerequisito se recomienda un curso de probabilidad y de procesos estocásticos, sin embargo, el texto incluye un repaso de algunos conceptos que se necesitan para el desarrollo de su contenido y que pretenden nivelar al lector a fin de que se familiarice con algunos conceptos necesarios para comprender el análisis de datos de supervivencia). Con este material se pretende discutir y exhibir la fusión entre el análisis clásico de supervivencia y los procesos estocásticos

(Andersen y cols., 1993; Therneau y Grambsch, 2000) que representa un avance muy importante en el área y hacen más flexible el manejo de los estimadores de las funciones de supervivencia y de las funciones de riesgo o *hazard* acumulado, a la vez que permite estudiar sus propiedades de una manera intuitiva, exhaustiva y apropiada.

El texto, a pesar de ser casi en su totalidad teórico, también incluye algunas ilustraciones con datos a fin de resaltar la utilidad del análisis de supervivencia; de hecho, en un apéndice el lector encontrará algunos programas en Python[1] y en R[2] para implementar el método de Kaplan-Meier, el método de Nelson-Aalen, el log-rank test y el modelo de riesgos proporcionales de Cox. Es importante anotar que estos métodos también se pueden implementar en SAS©, el cual es un software con licencia, aunque también existe una versión gratuita llamada SAS Studio© que se puede encontrar, por lo menos a la fecha que se publica este libro, en el link que aparece al pie de página[3]. Pensamos que al plasmar las notas del curso en este libro podemos ayudar a partir de nuestra experiencia a otros estudiantes que quieran profundizar en el análisis de datos de supervivencia.

Los autores agradecemos especialmente a la Universidad Nacional de Colombia, a la Universidad de San Buenaventura, a la Escuela de Estadística, a la Facultad de Ciencias y a todas aquellas personas que de forma directa o indirecta hicieron aportes importantes sin los cuales este proyecto no sería una realidad. Así también, un reconocimento muy especial a los coautores de este texto: Karime, Carlos y Verónica, pues sin su esfuerzo, aportes, dedicación, discusión y entusiasmo, este proyecto no hubiese sido posible.

[1] https://www.python.org/
[2] https://www.r-project.org/
[3] https://www.sas.com/en_us/software/university-edition.html

Tabla de contenido

Sobre los autores

Juan Carlos Salazar Uribe es profesor asociado de la Escuela de Estadística adscrita a la Facultad de Ciencias de la Universidad Nacional de Colombia, sede Medellín. Matemático de la Universidad Nacional de Colombia, sede Medellín y Magíster en Estadística de esta misma institución. En el 2004 obtuvo su título de Ph.D. en Estadística, otorgado por University of Kentucky. El profesor Salazar también cuenta con una amplia experiencia en docencia e investigación y ha sido autor de mútiples artículos académicos publicados en revistas de circulación nacional e internacional.

Ehidy Karime García Cruz es docente investigadora en la Universidad Pedagógica y Tecnológica de Colombia (UPTC), seccional Sogamoso. Licenciada en Matemáticas y Estadística de la UPTC, Magíster en Estadística de la Universidad Nacional de Colombia, sede Medellín y candiadta a doctora en Estadística de la Universidad Nacional de Colombia, sede Medellín. La profesora García cuenta con una amplia experiencia en docencia e investigación y ha sido autora de algunos artículos académicos publicados en revistas de circulación nacional.

Carlos Gaviria Peña es Licenciado en Matemáticas y Física de la Universidad de Antioquia (2006), Magíster en Matemáticas de la Universidad EAFIT (2010), Magíster en Ciencias-Estadística de la Universidad Nacional de Colombia, sede Medellín (2017), estudiante de Doctorado en Ciencias-Estadística de esta misma institución. Se ha desempeñado como docente desde el año 2006 en la Universidad de Antioquia, Universidad EAFIT, Universidad de Medellín e Instituto Tecnológico Metropolitano. Actualmente es docente-investigador asociado de la Universidad de San Buenaventura en el área de Ciencias Básicas. Ha participado en proyectos de investigación en la Universidad EAFIT y en la Universidad de San Buenaventura.

Julieth Verónica Guarín Escudero es estudiante de doctorado en Ciencias-Estadística de la Universidad Nacional de Colombia, sede Medellín. La estudiante es estadística y Magíster en Ciencias-Estadística de la misma institución y cuenta con experiencia como docente de cátedra en la Universidad de Antioquia en el departamento de Ingeniería de Sistemas y de Materiales. También es monitora de posgrado de la Universidad Nacional de Colombia, sede Medellín.

$\sum_j \int H_j dM_j(t)$

Capítulo 1

Introducción

El análisis de supervivencia ha sido una de las herramientas estadísticas más importantes para el usuario a nivel aplicado. Su capacidad de interpretación y disponibilidad de resultados computacionales, junto con los resultados teóricos, han hecho de este modelo una técnica muy popular. Para obtener estimaciones con este modelo, uno de los requerimientos fundamentales es la independencia de las observaciones muestrales.

1.1. ¿Qué es el análisis de supervivencia?

Es una rama de la Estadística que típicamente se enfoca en el estudio del tiempo hasta un evento de interés. En un sentido más general, el análisis de supervivencia (AS) consiste en un conjunto de técnicas para estudiar y modelar variables aleatorias positivas. Es una colección de métodos estadísticos para estudiar el tiempo que transcurre hasta la ocurrencia de un evento (también se conoce como análisis de tiempos de falla o análisis de sobrevida). El nombre de supervivencia se debe a que estos métodos son usados generalmente para estudiar tiempos hasta la muerte. Algunos campos de aplicación del AS incluyen: Ciencias de la Salud, Ingeniería (confiabilidad), Ciencias Sociales, Economía, Epidemiología, entre otros.

1.2. ¿Para qué es útil el análisis de supervivencia?

Se presentan a continuación algunos ejemplos de motivación donde el análisis de supervivencia puede ser útil: un investigador hace un seguimiento a un grupo de pacientes con leucemia durante algunas semanas y registra, entre otras variables, el tiempo que transcurre antes de que un paciente regrese a la clínica; un investigador hace un seguimiento a un grupo de pacientes durante algunos años para ver quiénes desarrollan una enfermedad del corazón; un ingeniero realiza una prueba con algunos motores de camión y registra el tiempo que transcurre antes de la primera falla; un investigador hace seguimiento a dos grupos de personas recién liberadas de la cárcel y registra el tiempo que transcurre antes de que alguno sea arrestado de nuevo; a partir de historias médicas, un investigador registra el tiempo que transcurrió hasta que un paciente desarrolló una cierta severidad de una enfermedad; un investigador registra el número de semestres que transcurren antes de que un estudiante universitario pierda la calidad de estudiante por primera vez. En los anteriores ejemplos es claro que no todos los eventos de interés se refieren a muertes, pero sí tienen en común un cambio de estado experimentado en las unidades.

En AS un evento se puede entender como un cambio cualitativo que se puede situar en el tiempo. Por cambio cualitativo se entiende el tránsito de un estado discreto a otro. Por ejemplo, de vivo a muerto, de soltero a casado, de sano a enfermo. Este tránsito se puede pensar —usando terminología de procesos estocásticos— como un modelo de dos estados con un estado absorbente: ESTADO 1: VIVO $->$ ESTADO 2: MUERTE (estado absorbente, no hay marcha atrás).

El análisis de supervivencia trata de responder preguntas tales como: ¿qué fracción de una población falla después de un tiempo determinado? ¿Por qué razón fallan? ¿Qué factores aceleran o no estas fallas?

1.3. Motivación

El análisis de supervivencia típicamente se enfoca en los datos de tiempo para eventos. En el sentido más general, consiste en técnicas para variables aleatorias de valor positivo. Para definir una variable aleatoria de tiempo de falla se requiere:

1. Un origen temporal no ambiguo (por ejemplo: asignación al azar a un ensayo clínico, compra de un automóvil).
2. Una escala de tiempo (por ejemplo: tiempo real —días o años—, kilometraje de un automóvil).
3. Definición clara del evento (por ejemplo: la muerte, la necesidad de cambiar la transmisión del automóvil).

Ejemplo 1 **Duración de la estancia en un hogar geriátrico (residencia de ancianos)**

El Centro Nacional de Investigación en Servicios de Salud estudió 36 hogares de ancianos con fines de lucro para evaluar los efectos de diferentes incentivos financieros en la duración de la estancia. Hogares "tratados" recibieron mayores bonificaciones y medicamentos para mejorar la salud de los pacientes y enviarlos a casa. El estudio incluyó 1601 pacientes ingresados entre el 1 de mayo de 1981 y el 30 de abril de 1982. Las variables incluyen (Lange, 1994):

LOS: duración de la estancia de un residente (en días).

EDAD: edad de un residente.

RX: asignación a un hogar de ancianos (1: con bonificaciones, 0: sin bonificaciones).

GÉNERO: género (1: masculino, 0: femenino).

CASADO: (1: casado, 0: no casado).

SALUD: estado de salud (2: segundo mejor, 5: peor).

CENSOR: indicador de censura (1: censurado, 0: dado de alta).

Primeras pocas líneas de datos:

```
LOS EDAD RX GÉNERO CASADO SALUD CENSOR
37  86   1  0      0      2     0
61  77   1  0      0      4     0
```

Ejemplo 2 **Fecundidad**

A mujeres que habían dado a luz recientemente se les pidió que recordaran cuánto tardaron en quedar embarazadas y si no fumaron durante ese tiempo. El resultado de interés es el tiempo hasta el embarazo (medido en ciclos menstruales). 19 sujetos no pudieron quedar embarazadas después de 12 meses. (Ciclo, Número de fumadoras, número de no fumadoras)

```
1 29 198, 2 16 107, 3 17 55, 4 4 38, 5 3 18, 6 9 22,
7 4 7, 8 5 9, 9 1 5, 10 1 3, 11 1 6, 12 3 6, 12+ 7 12
```

Ejemplo 3 **Ensayo clínico de prevención de MAC (Mycobacterium Avium Complex)**

El ACTG 196 (The Safety and Effectiveness of Clarithromycin and Rifabutin Used Alone or in Combination to Prevent Mycobacterium Avium Complex (MAC) or Disseminated MAC Disease in HIV-Infected Patients[4]) fue un ensayo clínico aleatorizado para estudiar los efectos de regímenes combinados para la prevención de MAC, una de las infecciones oportunistas más comunes en pacientes con SIDA.

Los 3 regímenes de tratamiento fueron: claritromicina (nueva), rifabutina (estándar), claritromicina más rifabutina. Otras características del ensayo: 1) Pacientes inscritos entre abril de 1993 y febrero de 1994, seguimiento finalizado en agosto de 1995. En febrero de 1994, la dosis de rifabutina se redujo de 3 píldoras / día (450 mg) a 2 píldoras / día (300 mg) debido a la preocupación sobre la uveítis (una afección ocular). El análisis principal de intención de tratar comparó los 3 tratamientos sin ajustar por este cambio de dosis.

Ejemplo 4 **Estudio HMO de supervivencia relacionada con el VIH**

Estos son datos hipotéticos utilizados por Hosmer y Lemeshow (1999) que contienen 100 observaciones sobre pacientes VIH positivos pertenecientes a una organización de mantenimiento

[4] https://clinicaltrials.gov/ct2/show/NCT00001030. To compare the efficacy and safety of clarithromycin alone versus rifabutin alone versus the two drugs in combination for the prevention or delay of Mycobacterium avium Complex (MAC) bacteremia or disseminated MAC disease. To compare other parameters such as survival, toxicity, and quality of life among the three treatment arms. To obtain information on the incidence and clinical grade of targeted gynecologic conditions.
Persons with advanced stages of HIV are considered to be at particular risk for developing disseminated MAC disease. The development of an effective regimen for the prevention of disseminated MAC disease may be of substantial benefit in altering the morbidity and possibly the mortality associated with this disease and its treatment.

de la salud (HMO: Health Maintenance Organization). La HMO quiere evaluar el tiempo de supervivencia de estas personas. En este conjunto de datos hipotéticos, los sujetos ingresaron al estudio desde el 1 de enero de 1989 hasta el 31 de diciembre de 1991. El seguimiento del estudio finalizó el 31 de diciembre de 1995. Las variables de interés:

ID: identificación del sujeto (1-100).

TIEMPO: tiempo de supervivencia en meses.

ENTDATE: fecha de entrada.

ENDDATE: fecha de finalización del seguimiento debido a muerte o censura.

CENSOR: indicador de muerte (1 = muerte, 0 = censura).

AGE: edad de la persona en años.

DROGAS: historial de uso de drogas por vía intravenosa (0 = no, 1 = sí).

Este conjunto de datos es utilizado por Hosmer y Lemeshow (1999) para motivar algunos conceptos en análisis de supervivencia en el capítulo 1 de su libro.

1.4. Algo de notación y definiciones

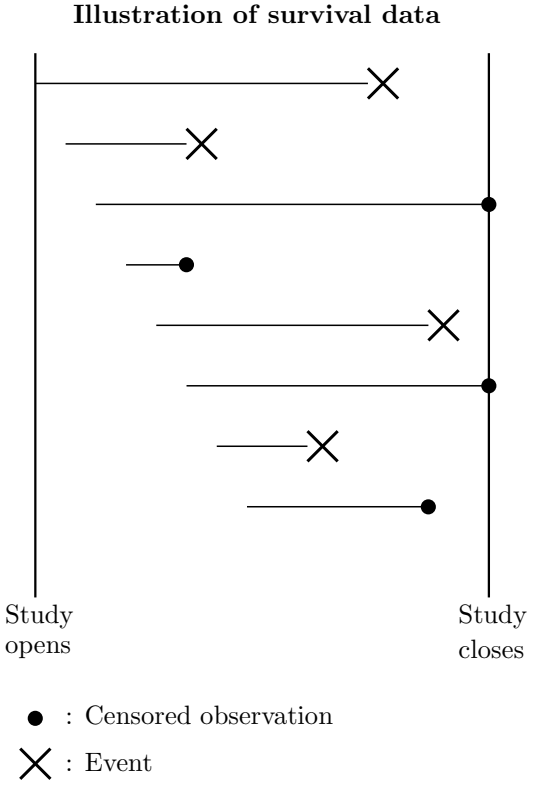
Las variables aleatorias de tiempo de falla son *siempre* no-negativas. Es decir, si se denota el tiempo de falla por T, entonces $T \geq 0$.

T puede ser *discreto* (toma un conjunto finito de valores), por ejemplo: $(a_1,\ a_2\ ,\ldots,\ a_n)$ o *continuo* (definido en $(0,\ \infty)$).

Una variable aleatoria X se denomina *variable aleatoria de tiempo de falla censurada* si $X = min(T,\ U)$, donde U es una variable de censura no negativa.

1.4.1 Tipos de censura

Varias características que normalmente se encuentran en el análisis de datos de supervivencia: todos los individuos no ingresan al estudio al mismo tiempo (*entrada tardía* o *entrada escalonada*). Cuando finaliza el estudio algunas personas aún no han tenido el evento de interés (censura); otros individuos abandonan o se pierden en medio del estudio y todo lo que sabemos de ellos es que la última vez que fueron vistos seguían "libres" del evento de interés (censura).



Censura a derecha: solo la variable aleatoria $X_i = min(T_i,\ U_i)$ se observa debido a la pérdida durante el seguimiento, la deserción o la finalización del estudio. A esto lo llamamos censura a derecha, porque los eventos no observados están a la derecha de nuestro tiempo de censura; es decir, lo que sí sabemos es que el evento no ha ocurrido al final del seguimiento. La censura a derecha es el tipo más común de censura y adoptaremos este supuesto en el texto.

Además de observar X_i también podemos observar el indicador de falla.

$$\delta_i = \begin{cases} 1 & \text{Si } T_i \leq U_i \\ 0 & \text{Si } T_i > U_i \end{cases}$$

Censura a izquierda: solo se puede observar $Y_i = max(T_i,\ U_i)$ y los indicadores de falla.

$$\delta_i = \begin{cases} 1 & \text{Si } U_i \leq T_i \\ 0 & \text{Si } U_i > T_i \end{cases}$$

Por ejemplo, en un estudio de la edad a la que algunos niños aprenden una tarea específica, es posible que algunos ya sepan hacerla (censurados a izquierda), otros aprenden durante el estudio (exacto), o algunos no la aprenden al final del estudio (censurados a derecha).

Censura de intervalo: en este caso se observa $(L_i,\ R_i)$ donde $T_i \in (L_i,\ R_i)$. Por ejemplo, tiempo hasta el cáncer de próstata, en este caso se observan mediciones longitudinales de PSA (antígeno prostático específico); o detectar recurrencia de cáncer de colon después de la cirugía, en este caso se hace un seguimiento a los pacientes cada 3 meses después de la extracción del tumor principal.

Censura independiente versus censura informativa: se dice que la censura es *independiente (no informativa)* si U_i es independiente de T_i.

Ejemplo 5

Si U_i es el final planificado del estudio, por ejemplo: 2 años después que el estudio empieza, generalmente, ese tiempo es independiente de los tiempos de los eventos.

Ejemplo 6

Si U_i es el momento en que un paciente abandona el estudio porque está empeorando su estado de salud y/o tuvo que dejar de tomar el tratamiento del estudio, posiblemente por efectos adversos severos, entonces U_i y T_i probablemente no sean independientes (censura informativa).

Una persona censurada en U debe ser representativa de todas las personas que sobreviven a U. Esto significa que la censura a U podría depender de las características de salud medidas al inicio del estudio, pero que entre todas las personas con las mismas características al inicio, la probabilidad de censura antes o al tiempo U debería ser la misma.

La censura se considera *informativa* si la distribución de U_i contiene información sobre los parámetros que caracterizan la distribución de T_i. *La censura informativa* se produce cuando los participantes se pierden durante el seguimiento debido a razones relacionadas con el estudio o el tratamiento administrado. Por ejemplo, en un estudio que compara la supervivencia libre de enfermedad después de dos tratamientos para el cáncer, el brazo de control puede ser ineficaz, lo que lleva a más recaídas y pacientes que empeoran su estado de salud durante el seguimiento. Por otro lado, los pacientes en el brazo de intervención podrían curarse completamente con un tratamiento eficaz y de esta manera dejar de sentir la necesidad o el deseo de seguir participando en el estudio. Si estos participantes son censurados rutinariamente, el verdadero efecto del tratamiento no se detectará y los resultados del estudio serán sesgados. Las tasas de supervivencia libres de enfermedad se basarían en los pacientes que continuaron con el seguimiento en el estudio, y estas podrían sobrestimarse para los controles y subestimarse para los pacientes tratados.

Tipos de tiempos de censura a derecha: suponga que se tiene una muestra de observaciones acerca de n personas:

$$X_1 = \text{mín}(T_1,\ U_1),\ X_2 = \text{mín}(T_2,\ U_2), \ldots, X_n = \text{mín}(T_n,\ U_n)$$

Censura tipo I: todos los U_i's son iguales. Por ejemplo, en estudios en animales, todos los animales son sacrificados después de 2 años.

Censura tipo II: en este caso $U_i = T_{(r)}$, el tiempo de la r-ésima falla. Por ejemplo, un estudio en animales se detiene cuando 4/6 tienen tumores (es decir, cuando el investigador observa un número específico de fracasos o fallas).

Censura tipo III: los U_i's son variables aleatorias y los δ_i's son indicadores de falla:

$$\delta_i = \begin{cases} 1 & \text{Si } T_i \leq U_i \\ 0 & \text{Si } T_i > U_i \end{cases}$$

Por ejemplo, un investigador está interesado en los factores relacionados con los divorcios y decide hacer un seguimiento a parejas de recién casados durante 10 años. El resultado de interés es el tiempo hasta el divorcio. Las parejas que aún están casadas después de 10 años se consideran datos censurados de tipo I. Pero en algunos casos, un miembro de la pareja podría haber muerto o haberse mudado a otra ciudad o país antes de 10 años o simplemente abandonar el estudio. Por lo tanto, los datos censurados de tipo III a menudo no están bajo el control del investigador (ocurren al azar).

Los tipos I y II se denominan *datos censurados simples*, mientras que el tipo III se denomina *censura aleatoria*.

Todos los ejemplos anteriores proporcionan evidencia del potencial de aplicación de los métodos de análisis de supervivencia. Hay que resaltar que estos ejemplos son solo de motivación y no necesariamente se trabajan en este libro, que tiene como objeto de estudio una exploración teórica a fondo de los estimadores más comunes del AS y sus formulaciones y conexiones con los métodos de conteo.

Capítulo 2

Esperanza condicional

2.1. Conceptos básicos de teoría de integración de Lebesgue y esperanza condicional

Se enuncian los siguientes resultados con la intención de tener una visión general de algunos conceptos de la teoría avanzada de probabilidad. Se sugiere hacer un estudio más profundo de estos conceptos, para tal propósito se recomienda estudiar los libros *Introducción a la teoría avanzada de la probabilidad* (2002) de Blanco y Muñoz, *Theory of Point Estimation* (1998) de Casella y Lehmann y *Elementos de teoría avanzada de probabilidad* (2019) de Salazar-Uribe y cols.

Definición 1 **Función de conjunto**

Sea Ω el conjunto formado por todos los posibles resultados de un experimento aleatorio. Una función conjunto es una función μ de $\mathcal{C}$ en $\overline{\mathbb{R}}$, donde $\mathcal{C}$ es una clase no vacía de subconjuntos de Ω y $\overline{\mathbb{R}}$ es el conjunto $\mathbb{R} \cup \{+\infty, \infty\}$.

Ahora, si μ es una función de conjunto, entonces a dicha función se le pueden asociar ciertas características teóricas importantes. Dichas características se dan en la siguiente definición.

Definición 2

Sean Ω el conjunto formado por todos los posibles resultados de un experimento aleatorio y μ una función de conjunto sobre $\mathcal{C}$.

1. μ se dice aditiva si para toda colección finita $A_1, A_2, \cdots, A_n$ disjuntos dos a dos de $\mathcal{C}$ tal que $\bigcup_{i=1}^{n} A_i \in \mathcal{C}$ se satisface:

$$\mu\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mu(A_i)$$

2. μ se dice σ-aditiva si para toda colección infinita $A_1, A_2, \cdots,$ disjuntos dos a dos de $\mathcal{C}$ tal que $\bigcup_{i=1}^{\infty} A_i \in \mathcal{C}$ se satisface:

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i)$$

Observación. Observe que en las definiciones 1 y 2 solo se exige que $\mathcal{C}$ sea una clase no vacía de subconjuntos de Ω, es decir, no se exige que $\mathcal{C}$ sea una σ-álgebra o que sea $\mathcal{P}(\Omega)$, por ejemplo. Ahora, entre todas las posibles funciones conjunto μ que se pueden considerar, existe una clase de estas que es de interés y esto tiene una relación directa con las σ-álgebras.

Definición 3 **Medida**

Sea $(\Omega, \mathfrak{F})$ un espacio medible. Una medida sobre $(\Omega, \mathfrak{F})$ es una función de conjunto μ definida sobre $\mathfrak{F}$ que satisface:

1. $\mu(A) \geq 0$ para todo $A \in \mathfrak{F}$.

2. $\mu(\emptyset) = 0$.

3. μ es σ-aditiva.

Para cada $A \in \mathfrak{F}$, $\mu(A)$ se denomina la medida de A y además la tripleta $(\Omega, \mathfrak{F}, \mu)$ se denomina espacio de medida. La medida μ es finita si $\mu(\Omega) < \infty$.

Observación. Sea $(\Omega, \mathfrak{F})$ un espacio medible. La medida P que además satisface $P(\Omega) = 1$ se llama medida de probabilidad. La tripleta $(\Omega, \mathfrak{F}, P)$ se llama espacio de probabilidad.

En estadística se estudian variables aleatorias de naturaleza discreta y continua y para cada una de estas se asocia una medida particular. En los dos siguientes ejemplos se mencionan estas importantes medidas.

Ejemplo 7

Sea Ω un conjunto contable y $\mathfrak{F} = \mathcal{P}(\Omega)$. Para cada $A \in \mathfrak{F}$ se define la función de conjunto μ como sigue:

$$\mu(A) = \begin{cases} |A| & \text{Si } A \text{ es finito} \\ \\ \infty & \text{Si } A \text{ es infinito} \end{cases}$$

Se puede verificar que μ es una medida y recibe el nombre medida contadora o de contar. El espacio $(\Omega, \mathfrak{F} = \mathcal{P}(\Omega), \mu)$ recibe el nombre de espacio de medida discreto.

Ejemplo 8

Sean Ω el espacio Euclidiano n-dimensional $\mathbb{R}^n$ y $\mathfrak{F}$ la σ-álgebra de Borel en $\mathbb{R}^n$. Para cada $A \in \mathfrak{F}$ se define la medida de conjunto μ como sigue:

$$\mu(A) = \prod_{i=1}^{n} (b_i - a_i)$$

donde $A = \{(x_1, x_2, \cdots, x_n) \in \mathbb{R}^n : a_i < x_i < b_i\}$, es decir, A es un rectángulo abierto en $\mathbb{R}^n$.

Observe que:

1. Si se considera $\mathbb{R}$, entonces μ es la longitud de el intervalo (a, b).

2. Si se considera $\mathbb{R}^2$, entonces μ es el área del rectángulo:

$$A = \{(x, y) \in \mathbb{R}^2 : a < x < b, c < y < d\}$$

3. Si se considera $\mathbb{R}^3$, entonces μ es el volumen de la caja:

$$A = \{(x, y, z) \in \mathbb{R}^3 : a < x < b, c < y < d, e < z < f\}$$

Se puede verificar que μ es una medida y recibe el nombre de medida de Lebesgue.

Observación. Los conjuntos a los cuales se les puede asignar la medida de Lebesgue se llaman conjuntos Lebesgue medibles o simplemente conjuntos medibles.

Sobre el conjunto de los números reales $\mathbb{R}$ se pueden identificar tres σ-álgebras bien definidas: $\mathcal{P}(\mathbb{R})$: partes de $\mathbb{R}$; $\mathcal{L}$: Conjunto de todos los subconjuntos de $\mathbb{R}$ que son Lebesgue medibles; y $\mathcal{B}$: Conjuntos Borel medibles. La relación entre estos conjuntos es la siguiente: $\mathcal{B} \subseteq \mathcal{L} \subseteq \mathcal{P}(\mathbb{R})$. Es por esta razón que los conjuntos borelianos[5] toman tal importancia en la estadística.

Además de definir y estudiar conceptos relacionados con σ-álgebras y medidas es necesario estudiar algunos conceptos generales de integral de una función de valor real f con respecto a una medida μ y sobre espacios abstractos.

Definición 4 Función Medible

Sean $(\Omega, \mathfrak{F})$ y $(\Omega', \mathfrak{F}')$ espacios medibles. Una función $f : (\Omega, \mathfrak{F}) \to (\Omega', \mathfrak{F}')$ se dice $\mathfrak{F} - \mathfrak{F}'$ medible si y solo si para todo $A \in \mathfrak{F}'$, se sigue $f^{-1}(A) \in \mathfrak{F}$.

Observación. Si en la definición 4 $(\Omega', \mathfrak{F}') = (\mathbb{R}, \mathcal{B})$, entonces la función f se dice función real medible.

De toda la gama existente de funciones medibles, las variables aleatorias son un caso particular de función medible.

[5]Sea Υ la colección de todos los conjuntos abiertos de $\mathbb{R}^n$. A $(\Upsilon, \mathbb{R}^n)$ se le llama espacio topológico euclidiano. La σ-álgebra generada por Υ se llama σ-álgebra de Borel de $\mathbb{R}^n$ y se denota por $\mathfrak{B}^n$. Sea $\Omega = \mathbb{R}$ y $\mathcal{G} = \{(a, b] : a < b\}$, donde a puede ser $-\infty$ y b puede ser ∞; la σ-álgebra generada por $\mathcal{G}$ se llama la σ-álgebra de conjuntos de Borel en $\mathbb{R}$. Un conjunto de Borel es un conjunto que pertenece a $\sigma(\mathcal{G})$. (Salazar-Uribe y cols., 2019)

Definición 5

Sea $(\Omega, \mathfrak{F}, P)$ un espacio de probabilidad. Una variable aleatoria es una función X tal que $X : (\mathcal{F}, (\Omega, \mathfrak{F}) \to (\mathbb{R}, \mathcal{B})$.

Para garantizar que una función de valor real X es una variable aleatoria, debe garantizarse que para todo $B \in \mathcal{B}$ se sigue $X^{-1}(B) \in \mathfrak{F}$, la cual es una tarea en general compleja. Por esta razón, a continuación se da un teorema que permite garantizar bajo qué condiciones una función de valor real X es una variable aleatoria.

Teorema 1

Sea $(\Omega, \mathfrak{F})$ un espacio medible y $f : (\mathcal{F}, (D \subseteq \Omega, \mathfrak{F}) \to (\overline{\mathbb{R}}, \overline{\mathcal{C}})$. Si $c \in \overline{\mathbb{R}}$, entonces las siguientes proposiciones son equivalentes:

1. f es $\mathfrak{F} - \overline{\mathcal{B}}$ medible.
2. $\{x \in \Omega : f(x) > c\} \in \mathfrak{F}$, esto es $f^{-1}((c, +\infty)) \in \mathfrak{F}$.
3. $\{x \in \Omega : f(x) \geq c\} \in \mathfrak{F}$, esto es $f^{-1}([c, +\infty)) \in \mathfrak{F}$.
4. $\{x \in \Omega : f(x) \leq c\} \in \mathfrak{F}$, esto es $f^{-1}((-\infty, c]) \in \mathfrak{F}$.
5. $\{x \in \Omega : f(x) < c\} \in \mathfrak{F}$, esto es $f^{-1}((-\infty, c)) \in \mathfrak{F}$.

Observación. A partir del teorema 1 se sigue que para mostrar que una función de valor real f es una función medible es suficiente probar alguna de las condiciones 2, 3, 4 o 5. Ahora, dado que la σ-álgebra de Borel es la σ-álgebra generada por los intervalos de la forma $(a, b]$ y cada uno de los conjuntos $(c, +\infty)$, $[c, +\infty)$, $(-\infty, c]$ y $(-\infty, c)$ pertenece a $\mathcal{B}$, entonces es suficiente mostrar que $f^{-1}(B) \in \mathfrak{F}$ para todo $B \in \mathcal{B}$. De manera directa, si se quiere probar que una función X es una variable aleatoria, entonces debe probarse que $X^{-1}(B) \in \mathfrak{F}$ para todo $B \in \mathcal{B}$.

La prueba del teorema 1 se puede encontrar en los libros *Introducción a la teoría avanzada de la probabilidad* (2002) de Blanco y Muñoz y *Elementos de teoría avanzada de probabilidad* (2019) de Salazar-Uribe y cols.

Teorema 2

Si $f : (\mathfrak{F}, (\Omega, \mathfrak{F})) \to (\Omega', \mathfrak{F}')$ es una función medible y μ es una medida sobre $(\Omega, \mathfrak{F})$, entonces la función $\mu_f : (\Omega', \mathfrak{F}') \to (\Omega, \mathfrak{F})$ dada por:

$$\mu_f(B) = \mu(f^{-1}(B)), B \in \mathfrak{F}'$$

define una medida sobre Ω'.

Observación. Si en el teorema 2 se toma una variable aleatoria $X : (\Omega, \mathfrak{F}) \to (\mathbb{R}, \mathcal{B})$, donde P es una medida de probabilidad sobre $(\Omega, \mathfrak{F})$, entonces:

$$P_X(B) = P(X^{-1}(B)), B \in \mathcal{B}$$

ahora, si $B = (-\infty, x]$, entonces:

$$\begin{aligned} P_X((-\infty, x]) &= P\left(X^{-1}\left((-\infty, x]\right)\right) \\ &= P(X \leq x) \\ &= F(x) \end{aligned}$$

es decir, la distribución de probabilidad F_X es una medida de probabilidad sobre $(\mathbb{R}, \mathcal{B})$.

Para definir el concepto de integral y otros conceptos importantes se utiliza la función indicadora, que se da en la siguiente definición.

Definición 6 **Función indicadora**

Sea $(\Omega, \mathfrak{F})$ un espacio medible y $A \in \mathfrak{F}$ fijo. La función indicadora de A, que se denota por I_A, es la función $I_A : \Omega \to \{0, 1\}$ tal que:

$$I_A(\omega) = \begin{cases} 1 & \text{Si } \omega \in A \\ 0 & \text{Si } \omega \notin A \end{cases}$$

Claramente la función indicadora es una función medible.

Como se mencionó arriba, el objetivo es definir algunos conceptos generales de la integral de una función de valor real f con respecto a una medida μ y sobre espacios abstractos. A continuación se consideran los casos de funciones integrables.

Caso 1. Funciones simples

Definición 7 **Función simple**

Sea $f : (\Omega, \mathfrak{F}) \to (\mathbb{R}, \mathcal{B})$ una función medible. La función f se dice simple si y solo si $Ran(f)$ es un conjunto finito formado por valores diferentes. Si $a_1 < a_2 < \cdots < a_n$, entonces f se deja expresar de la forma canónica:

$$f(\omega) = \sum_{i=1}^{n} a_i I_{A_i(w)}$$

donde $A_i = \{\omega : f(\omega) = a_i\} = f^{-1}(\{a_i\}) \in \Omega$, $A_i \cap A_j = \emptyset$ y $\bigcup_{i=1}^{n} A_i = \Omega$.

Una función simple es una función medible pues es una combinación lineal de funciones indicadoras, las cuales son medibles. De esta manera una función simple es una función integrable.