



1 — Introducción

Los retos y alcances que la ingeniería moderna necesita superar requieren el apoyo de herramientas científicas que le permitan fortalecer tanto su investigación como las aplicaciones que surjan de su ejercicio y así afrontar la solución de problemas. De tales herramientas, la estadística brinda el apoyo necesario a los ingenieros, quienes muchas veces se ven enfrentados a problemas que tienen varias variables, y por tanto necesitan conocer algunas técnicas del ámbito univariado y multivariado. Este texto no pretende convertirse en un libro de análisis multivariado, ni en un libro de estadística para ingenieros ya que no contiene la totalidad de los temas que clásicamente se acostumbra en estos textos; tan solo se han seleccionado algunos conceptos, técnicas de estadística y métodos en una y más variables que el ingeniero puede revisar para ampliar su abanico de herramientas.

El criterio de selección de temas para ser incluidos en este texto obedece a las siguientes razones. La primera es la maravillosa oportunidad de que los estudiantes de pregrado y posgrado que no han tenido la conveniencia de conocer estas teorías incursionen en ellas; la segunda es el importante poder aplicativo que tienen los temas seleccionados; la tercera es que, a pesar de que algunos de estos conceptos no se tratan en el pregrado, es factible su aprendizaje porque hay convencimiento de la inmensa capacidad de los alumnos, que se potencializa con una explicación adecuada; la cuarta es que algunos de los temas tratados en este texto son temas principales de una asignatura optativa en Ingeniería Industrial en la Universidad del Magdalena, y la quinta es que con el contenido del presente texto se pueden realizar trabajos investigativos en solución de problemas, pronósticos, simulaciones o entendimiento de un fenómeno en particular en ingeniería. De esta forma, siendo estos temas no menos o más importantes que otros en estadística, se podría contribuir a que el ingeniero mejore su potencial de aplicaciones y complemente sus conocimientos científicos.

Aunque se explicó la forma de seleccionar las temáticas tratadas en este libro, los capítulos 2 y 3 merecen una mención aparte en la medida en que se incluyeron como contexto para brindar los conocimientos básicos y comunes a todos los ingenieros y que resultan necesarios para el estudio de los demás capítulos. Casi que cualquier herramienta estadística usada, sea de inferencia o de modelación, necesita de las bases tratadas en dicho apartado, si bien ciertos temas como estadística descriptiva y probabilidad básica, reglas de probabilidad, eventos y técnicas de conteo no se incluyen, por lo que para el estudio de este libro el lector debe tener clara la teoría al respecto. En todo caso, aunque el lector este enfrentado a temas nuevos a lo largo del texto, gracias al capítulo 2 siempre encontrará la forma de conectar estos conceptos con los temas básicos de estadística a fin de facilitar su comprensión.

En el libro se encuentra además una serie de ejemplos que le ayudarán al lector a entender cada tema tratado. El desarrollo de algunos de estos ejemplos se presenta mediante códigos originales del autor, escritos en R. De este modo el texto constituye un material complementario para la educación básica en ingeniería y puede servir a su vez como fuente de consulta y referencia en temas particulares. Por ejemplo, dentro de temas como distribuciones de probabilidad o series de tiempo, este texto puede desempeñar el rol de material de consulta, y al mismo tiempo puede incentivar la exploración de nuevos conocimientos, y quizás alguno o algunos de los capítulos sean de utilidad a la hora proponer tesis de grado o investigaciones.

Por la forma como se explican algunos de los temas con la intención de clarificar, este texto no está dirigido a personas con formación en estadística; está dirigido a ingenieros que deseen fortalecer algunos conceptos básicos de estadística y luego usarlos para introducirse en alguno o algunos de los temas ofrecidos por el texto en el resto de capítulos. Estos son: (1) series de tiempo, (2) hipótesis multivariadas, (3) análisis de componentes principales, (4) análisis de correspondencias, (5) regresión multivariada, (6) mixturas para series de tiempo normales multivariadas y sus estimaciones, y (7) modelos de eficiencia basados en el análisis envolvente de datos.

1.1 Acerca de capítulos 2 y 3. Estadística univariada

Estos capítulos pretenden brindar herramientas conceptuales de estadística como preparación para los capítulos posteriores. Hay un equilibrio entre la explicación contextual y la explicación matemática, en el sentido de que no se persigue que este capítulo sea solo de demostraciones matemáticas ni tampoco que sea demasiado intuitivo. La estrategia en estos apartados incluye explicaciones desde contextos reales y la utilización de ecuaciones y conceptos matemáticos cuando es necesario. Los conceptos de esta parte de la publicación están enlazados en uno o en varios de los capítulos posteriores, de tal forma que en estos últimos se hace alusión a lo visto en estos capítulos.

Los temas básicos de esta sección son (1) esperanza matemática y varianza con sus propiedades, (2) distribuciones de probabilidad más importantes, (3) distribución de probabilidad conjunta, marginal, condicional, (4) conceptos de correlación y covarianza, (5) transformación de variables, función generadora de momentos, (6) método de máxima verosimilitud y (7) conceptos bayesianos.

En ese orden de ideas, la esperanza matemática y la varianza se presentan en el capítulo 2, con el fin de poder abordar su extensión y su utilización en los capítulos 5, 6, 7, 8 y 9. Luego se hace un repaso de las distribuciones de probabilidad. La familiarización y el estudio de estas permitirán estudiar capítulos posteriores con mayor claridad. Por ejemplo, en este repaso se aborda la distribución normal, cuya extensión hacia la normal multivariada se estudia posteriormente en el capítulo 5.

Asimismo, los capítulos 2 y 3 plantean ciertas ecuaciones y definiciones de estadística univariada que permiten comprender más adelante, en el capítulo 9, los conceptos de probabilidad conjunta, distribuciones marginales, independencia estadística y correlación. De igual forma, el concepto de correlación entre dos variables desarrollado en estas instancias del libro se extiende en el capítulo 4 a la autocorrelación.

Cabe mencionar, en particular, el método de máxima verosimilitud, necesario para comprender algunos conceptos importantes en el capítulo 9. En este caso, si bien la función de verosimilitud es un poco más compleja por ser multivariada, se hace la respectiva cita específica de la ecuación en univariado y la explicación para facilitarle al lector que haga la extensión del concepto univariado al multivariado.

Estos dos capítulos no están inconexos con el resto del texto y buscan brindar un apoyo y el material de consulta necesario en el abordaje de los subsiguientes capítulos. Sin embargo, un ingeniero podría encontrar una oportunidad en aplicaciones directas cuando se estudian temas como distribuciones de probabilidad, sin necesidad de avanzar en el resto del libro.

1.2 Acerca del capítulo 4. Series de tiempo

Este capítulo ofrece las bases teóricas iniciales para entender conceptos básicos e importantes de las series de tiempo. Tiene algunas demostraciones matemáticas de los modelos teóricos básicos al calcular media, varianza, autocovarianza y autocorrelación de modelo ruido blanco, promedios móviles de orden uno (MA(1)) y autorregresivo de orden uno (AR(1)). Luego, a través de ejemplos, se va enseñando cómo una serie de datos puede ser modelada mediante estos modelos básicos o la combinación de ellos, y se da una introducción con ejemplos resueltos con la metodología ARIMA. Con este capítulo se pretende ayudar al ingeniero a incursionar en las series de tiempo con los elementos teóricos básicos para que adquiera habilidades que lo conduzcan a la investigación del tema o para que realice aplicaciones con los conocimientos adquiridos en este texto. El capítulo

también presenta una serie de ejemplos resueltos en el software libre R, que le permitirá al estudiante realizar aplicaciones prácticas con las respectivas interpretaciones.

1.3 Acerca del capítulo 5. Algunas hipótesis multivariadas

Este apartado introduce la inferencia estadística multivariada e incluye tres tipos de pruebas de hipótesis: la primera es sobre la igualdad de una media multivariada con un vector de constantes; la segunda corresponde a una comparación de dos medias multivariadas pertenecientes a dos grupos, y la tercera se trata de una comparación de varios grupos donde la variable de respuesta es un vector multivariado. Los primeros dos tipos están basados en la distribución T^2 de Hotelling, que tiene relación con la distribución t Student univariada. El tercer tipo se trabaja con la técnica de análisis de varianza multivariado conocida como MANOVA. También se explican los supuestos sobre los cuales están construidas estas técnicas y se dan ejemplos resueltos para correr en software libre como R. Estas técnicas están basadas en el supuesto de multinormalidad.

Asimismo, se ofrecen unos conceptos de preparación que sirven para abordar la estadística multivariada de este capítulo y del siguiente (6). De esta forma se introducen conceptos como los valores y vectores propios, con algunos ejemplos que ayuden a clarificar su uso y aplicación dentro de temas multivariados. Además se aborda, por primera vez en el texto, un tema vital como el de matriz de varianzas y covarianzas, explicando la diferencia entre correlación y covarianza y su relación.

Este capítulo también introduce la distribución normal multivariada, de manera que el ingeniero pueda relacionarla con cada elemento de la normal univariada y pueda encontrar así la forma de ampliar sus conocimientos en el ámbito multivariado con mayor claridad.

Como se explica al inicio de este capítulo introductorio, estas explicaciones de empalme univariado al multivariado no son necesarias para un profesional con formación en estadística porque podrían resultarle muy obvias, aunque serían de utilidad para un ingeniero que carezca de experiencia en aspectos multivariados. Al lograr claridad, se puede esperar que el ingeniero mejore sus aplicaciones, adquiera capacidad de investigar temas nuevos para él, o simplemente amplíe su panorama de conocimientos adquiridos.

1.4 Acerca del capítulo 6. Métodos multivariados

Existen varias técnicas clásicas de proyección de puntos u observaciones que pertenecen a muchas dimensiones hacia un sistema con una cantidad reducida de dimensiones para lograr mejor entendimiento y fortalecer el análisis sobre el problema en estudio. Esta proyección de observaciones en R^n a un sistema en R^p , donde $p \ll n$, requiere de un tratamiento estadístico interesante. Este capítulo se dedicará a dos de esas técnicas de proyección, conocidas como análisis de componentes principales y análisis de correspondencias. Ambas

son similares, pero trabajan sobre diferentes características de las variables de estudio: la primera está diseñada para variables cuantitativas; la segunda, para variables cualitativas. Este capítulo contiene algunas explicaciones matemáticas, ejemplos resueltos y el uso del software libre R para que el estudiante practique lo aprendido.

1.5 Acerca del capítulo 7. Regresión multivariada

Otro aporte que se pretende hacerle al ingeniero con conocimientos básicos de estadística es ampliar la idea de la regresión lineal simple y regresión múltiple a un sistema de regresión multivariada, donde la variable de respuesta es un vector multivariado. Para este fin se realiza un empalme de los conceptos de regresión como pruebas de hipótesis y estimación de parámetros al caso multivariado. Dicho empalme también se ve cuando se asume el valor esperado de los errores, la varianza de los errores y el supuesto de normalidad, los cuales también son extendidos al caso multivariado con su respectiva explicación. En este capítulo se usan algunos estadísticos vistos en el capítulo 5 y que pueden ser más entendibles en este apartado.

1.6 Acerca del capítulo 8. Mixturas multivariadas en series

Es uno de los aportes novedosos del texto, ya que los libros clásicos de análisis multivariado poco lo tratan. Se refiere a las mixturas normales multivariadas aplicadas a las series de tiempo, un tema que puede ser bastante aplicable y útil en ingeniería. Si bien este texto no se enfoca en el análisis multivariado, las mixturas normales multivariadas han sido seleccionadas por ser poco conocidas por estudiantes de pregrado y posiblemente también por algunos ingenieros con estudios de posgrado. Así, siguiendo el interés y la motivación de llevarles a estas personas nuevas herramientas que alimenten su creatividad para solucionar problemas o investigar, es posible que estas técnicas puedan potencializar su desempeño profesional. Finalmente, este capítulo trata conceptos relacionados con las series de tiempo como las mixturas normales multivariadas y la asignación de clúster a modelos de mixtura, de manera que sus principales bases se encuentran en el capítulo 4, y asimismo se presentaría la teoría básica para estimar los parámetros de estas mixturas, la cual se complementaría luego, en el capítulo 9.

1.7 Acerca del capítulo 9. Estimación en series de mixtura

La introducción en el capítulo 8 sobre la mixtura de modelos normales multivariados será ampliada en este capítulo al explicar matemáticamente cómo se estiman los parámetros de estos modelos. Existen unos algoritmos basados en el método de máxima verosimilitud llamados algoritmos EM (*Expectation Maximization*) y AECM (*Alternating Expectation Conditional Maximization*). Este capítulo brinda la fundamentación teórica de cómo estos

algoritmos logran estimar parámetros por máxima verosimilitud, lo que puede constituir un aporte científico a estudiantes de pregrado y posgrado.

La estrategia para explicar este capítulo consiste en desarrollar ejemplos que ayuden a contextualizar las ecuaciones y permitan enlazar conceptos previos con los de este capítulo. Con el fin de agilizar la comprensión de este apartado, se establecen relaciones con conceptos con los que un ingeniero puede estar más familiarizado, aunque de todos modos el capítulo 2 presenta elementos que permitirán entender esta parte del texto a cualquier lector. Es común que en la literatura de estimación de modelos de mixtura multivariados los autores asuman que el lector tiene los conocimientos para entender el equivalente al caso más sencillo o univariado, pero este no es el enfoque de este texto y asume que el lector está viendo estos temas por primera vez. En este sentido, este libro es más didáctico al hacer referencias al concepto que se necesita para abordar las nuevas temáticas.

El lector puede tomar dos caminos con referencia a este capítulo: el primero es tener en cuenta solo los resultados e intentar realizar aplicaciones y programaciones de las ecuaciones que maximizan el logaritmo de la función de verosimilitud, para así estimar sus parámetros; el segundo camino es ser un poco más exhaustivo para entender las demostraciones y aprovechar las ayudas que se brindan en el texto, puesto que estas por lo general no se encuentran en las referencias bibliográficas que tratan con dichas funciones de verosimilitud y su maximización. Asumir este último camino, aunque requiere un poco de más tiempo de estudio, puede ofrecer muchas ventajas en la comprensión. Mediante la interiorización de estos conceptos, se puede dar el paso para programar, en algún software como R o Matlab, la estimación de parámetros en una aplicación de interés para el usuario.

1.8 Acerca del capítulo 10. Modelos de eficiencia

Otro aporte novedoso de este libro es la inclusión del tema de modelos de eficiencia, que aunque no está considerado dentro de la categoría análisis multivariado, es una técnica no paramétrica que considera diversas variables en la entrada y en la salida de un sistema. Luego, al comparar varios sistemas, cada uno de ellos con diferentes niveles en sus variables de entrada y salida, se puede establecer, por medio de los modelos vistos en este capítulo, la eficiencia a la que cada sistema trabaja. Es un tema bastante aplicable y que puede contribuir mucho a la investigación de estudiantes de pregrado y posgrado.



2 — Estadística univariada, parte I

Una vez presentadas brevemente algunas definiciones como variable aleatoria, esperanza matemática y varianza con sus respectivas propiedades, en este capítulo se dará una mirada a algunas distribuciones de probabilidad discretas y continuas. En ciertos casos se presentará una situación cotidiana que permita ver el trasfondo del modelo matemático de la distribución. Luego de dicho planteamiento de situación cotidiana se llegará a la función de probabilidad para el caso de variable discreta o a la función de densidad de probabilidad para el caso de variable continua. Se hablará de algunas características de las distribuciones como sus parámetros, media y varianza, y todo esto servirá de preparación para entender los siguientes capítulos.

Abordadas las anteriores funciones de probabilidad, se presenta el concepto de distribución de probabilidad conjunta, lo que hace imprescindible hablar de distribución marginal y probabilidad condicional. Estas distribuciones se definen tanto para variable aleatoria discreta como continua. Entender este tipo de funciones es importante para un espacio que no es unidimensional sino multidimensional, y es aquí donde se empieza a pensar en un caso diferente al univariado. Estos conceptos se complementan con los del siguiente capítulo.

2.1 Definiciones sobre variable aleatoria

El concepto de variable aleatoria nace en el de espacio muestral.

Definición 2.1

El **espacio muestral** Ω es el conjunto de todos los posibles resultados de un experimento.

Un ejemplo siempre muy común en cursos de estadística es definir como experimento el lanzamiento de un dado: el espacio muestral son los números $1, \dots, 6$, mientras que como eventos se tienen los hechos de que caiga un resultado par, o un resultado impar, o un número menor a 3. A los eventos se les puede asignar una probabilidad teniendo en cuenta el espacio muestral. Ahora, si bien el número de resultados en el lanzamiento de un dado es finito, hay casos en los que esta cantidad posible es infinita. Este caso se da cuando se tiene una variable aleatoria continua y el hecho de que ocurra un número particular entre infinitas posibilidades tiene probabilidad cero y por eso $P(X = x) = 0$. Pero ¿qué es una variable aleatoria?

Definición 2.2

Una **variable aleatoria** es una función que asocia un número real con cada elemento del espacio muestral.

A las variables aleatorias se les puede calcular su esperanza matemática o su media y su varianza entre varias medidas. En la siguiente sección se presentan la esperanza y la varianza de una variable aleatoria y sus propiedades.

2.2 Esperanza matemática

Definición 2.3

Sea X una variable aleatoria con distribución de probabilidad $f(x)$. La media o **valor esperado** de X es

$$\mu = E(X) = \sum_x x f(x) \quad (2.1)$$

si X es discreta, y

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x) dx \quad (2.2)$$

si X es continua.

2.2.1 Propiedades esperanza y varianza

Linealidad de la esperanza matemática

1. $E(X + Y) = E(X) + E(Y)$
2. $E(kX) = kE(X)$ para todo número real k .
3. $E(k) = k$ para todo número real k .
4. $E(aX + b) = aE(X) + b$ para todo par de números reales a y b .

Esperanza de un producto

1. $E(X \cdot Y) = E(X) \cdot E(Y)$ únicamente para el caso de que X e Y sean variables independientes.

Propiedades de la varianza

La varianza de una variable aleatoria se define a continuación:

Definición 2.4

Sea X una variable aleatoria con función de masa de probabilidad $f(x)$. La varianza de X es

$$V(X) = \sum_x (x - \mu)^2 f(x) \quad (2.3)$$

si X es discreta, y

$$V(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx \quad (2.4)$$

si X es continua.

1. $V(X) \geq 0$
2. $V(aX + b) = a^2 V(X) + V(b) = a^2 V(X)$, siendo a y b números reales cualesquiera.
3. $V(X - Y) = V(X) + V(Y) - 2Cov(X, Y)$, donde $Cov(X, Y)$ es la covarianza de X con Y . Si estas dos variables son independientes, la covarianza es cero.
4. $V(X + Y) = V(X) + V(Y) + 2Cov(X, Y)$ donde $Cov(X, Y)$ es la covarianza de X con Y . Si estas dos variables son independientes, la covarianza es cero.

Nótese que en la segunda propiedad de la varianza, la varianza de una constante es cero $V(b) = 0$. También es importante aclarar que, de acuerdo con las propiedades 3 y 4 de la varianza, si dos variables son independientes, entonces la covarianza es cero. Sin embargo, si al explorar la covarianza de dos variables esta es cero, no significa que estas son independientes; significa que o bien son independientes o bien no están relacionadas linealmente.

Cuando se habla de valor esperado o media como las de las ecuaciones (2.1) o (2.2), se refiere a la media de un modelo teórico o de una variable aleatoria que en últimas es un modelo matemático teórico. No obstante, cuando se tiene una muestra, la media muestral se calcula con

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}. \quad (2.5)$$

De la misma forma, cuando se habla de varianza como las de las ecuaciones (2.3) o (2.4), se refiere a la varianza de un modelo teórico o de una variable aleatoria, pero cuando se tiene una muestra la media varianza muestral se calcula o bien con el estimador máximo verosímil dado por la ecuación

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}, \quad (2.6)$$

o bien con el estimador insesgado dado por la ecuación

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}. \quad (2.7)$$

2.3 Distribuciones de probabilidad más importantes

En la presente sección se da una mirada breve a algunas distribuciones discretas y continuas, al menos las más conocidas. Se verá en esta sección que unas distribuciones de probabilidad contribuyeron al planteamiento de unas nuevas. Se hace énfasis en características importantes como el valor esperado y la varianza y se verá cómo estas medidas siempre son dependientes de los parámetros de la distribución. Toda distribución tiene uno o varios parámetros que son realmente importantes a la hora de relacionar la muestra aleatoria con el modelo matemático estadístico.

2.3.1 Algunas distribuciones discretas

Distribución Bernoulli

La variable aleatoria (en adelante v.a.) Bernoulli es la más básica de todas, ya que solo puede tener como valores 1 o un 0, $x = 1$ o $x = 0$. Permite modelar situaciones donde se presenten dos alternativas disyuntas y exhaustivas; por ejemplo, en una muestra de personas, el sexo biológico tiene dos alternativas: hombre o mujer; en una muestra de productos fabricados en una empresa, estos pueden estar defectuosos o en buen estado, o en una muestra de estudiantes se puede evaluar si aprobaron el semestre o no.

Definición 2.5

Sea p un valor numérico tal que $0 \leq p \leq 1$ y sea X una variable aleatoria. Se dice que X tiene distribución Bernoulli de parámetro p si su función de densidad está dada por

$$f(x) = f(x; p) = \begin{cases} p^x(1-p)^{1-x}, & \text{para } x = 0, 1. \\ 0, & \text{de otra forma.} \end{cases}$$

Ejercicio 2.1 Reemplace en la definición 2.5 el valor de $x = 0$ y observe el resultado de la probabilidad $f(x)$. También reemplace el valor de $x = 1$ y observe el resultado de la probabilidad $f(x)$. ■

Siguiendo la definición 2.3, el valor esperado de la distribución Bernoulli es

$$E(X) = 0 \times (1 - p) + 1 \times p = p.$$

$$E(X) = p. \quad (2.8)$$