Modelos probabilísticos y análisis estadístco / Métodos Estadísticos para Data Analytics

Elementos fundamentales de probabilidad

Rodrigo Gil Castañeda

rodrigo.gil@utadeo.edu.co

Departamento de Ciencias Básicas & Modelado
Facultad de Ciencias Naturales e Ingeniería
Universidad de Bogotá Jorge Tadeo Lozano

Tabla de contenidos

Introducción

Modelo estocástico

En un modelo estocástico, es aquel en el cual la información disponible no permite generar una relación o regla para determinar el resultado de un experimento. Muchos fenómenos (naturales o artificiales) son considerados **aleatorios** debido a que el resultado exacto no puede ser predicho; en este caso los resultados de mediciones repetidas sobre circunstancias similares generan un patrón.

* En un experimento aleatorio E, el conjunto de todos los posibles resultados de E se llama espacio muestral y se denota por la letra griega Ω. Ejemplo: en el experimento de lanzar una moneda al aire, Ω estaría compuesto por los posibles resultados: **cara** o **sello**; $\Omega = \{Ca, Se\}$. En la ejecución del experimento aletorio, aunque el resultado es impredecible, se sabe que debe corresponder a una de las opciones fijadas por el espacio muestral.

Medida numérica de la incertidumbre

```
install.packages("asbio")
library(asbio)
anm.coin(100, p.head = 0.3, interval=1)
```

Crea una gráfica animada que muestra los resultados del lanzamiento de monedas, y la convergencia resultante de P(Cara) a medida que el número de lanzamientos se incrementa.

Definiciones

Espacio Muestral (Ω): es el conjunto de todos los posibles resultados de un experimento aleatorio

Evento o suceso ($A : \ldots$): es cualquier subconjunto del espacio muestral

Eventos Excluyentes (Disyuntos): Sean A y B dos eventos de Ω si $A \cap B = \emptyset$, estos dos eventos se dicen excluyentes

Probabilidad clásica (teórica)

Si un experimento aleatorio resulta de n formas *igualmente posibles* y *mutuamente excluyentes* y si n_A de dichos resultados favorecen a la situación de interés A, la probabilidad de A, denotada $P(A)$, es:

$$P(A) = \frac{n_A}{n} = \frac{\text{Casos favorables}}{\text{Casos posibles}}$$

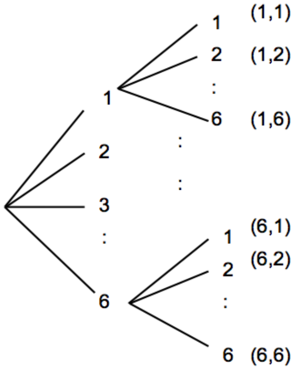
Probabilidad clásica (teórica)

Experimento aleatorio: lanzamiento de dos dados

Espacio muestral: $\Omega = \{(1, 1)\ (1, 2)\ (1, 3)\ (1, 4)\ (1, 5)\ (1, 6)\ (2, 1)\ (2, 2)\ (2, 3)\ (2, 4)\ (2, 5)\ (2, 6)\ (3, 1)\ (3, 2)\ (3, 3)\ (3, 4)\ (3, 5)\ (3, 6)\ (4, 1)\ (4, 2)\ (4, 3)\ (4, 4)\ (4, 5)\ (4, 6)\ (5, 1)\ (5, 2)\ (5, 3)\ (5, 4)\ (5, 5)\ (5, 6)\ (6, 1)\ (6, 2)\ (6, 3)\ (6, 4)\ (6, 5)\ (6, 6)\}$

Situación de interés: A = aparece el mismo número en ambos dados

Experimento aleatorio: lanzamiento de dos dados



- Diagrama de árbol para el experimento de lanzar dos dados
- Resultados igualmente posibles y mutuamente excluyentes

Casos favorables: (1, 1) (2, 2) (3, 3) (4, 4) (5, 5) (6, 6) $n_A = 6$

Probabilidad de ocurrencia del evento A: $P(A) = 6/36 = 0.167$

Probabilidad como frecuencia relativa

Si un experimento se repite n veces bajo las *mismas condiciones* y si n_B de dichas veces favorecen al atributo (situación) B entonces la probabilidad de B estará dada por:

$$P(B) = lim_{n\to\infty}\frac{n_B}{n}$$

Axiomas de probabilidad

Sea Ω un espacio muestral y A un evento cualquiera de Ω. Para todo A en Ω se asigna un *número* llamado probabilidad de A, notado $P(A)$, tal que:

i $P(A) \geq 0$

ii $P(\Omega) = 1$

iii Si $A = B_1 \cup B_2 \cup \ldots \cup B_k; B_i \cap B_j = \emptyset; \forall i \neq j; P(B_i) > 0$, entonces

$$P(A) = \sum_{i=1}^{k} P(B_i)$$

Algunos resultados importantes

Teorema 1: $P(\emptyset) = 0$

Teorema 2: $0 \leq P(A) \leq 1$

Teorema 3: $P(\bar{A}) = 1 - P(A)$

Ley aditiva de la probabilidad

Probabilidad para la *unión* de eventos

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - [P(AB) + P(AC) + P(BC)] + P(ABC)$$

Ejemplo

Las enfermedades I y II son comunes entre los individuos de cierta población. Se supone que el 10 % de la población contraerá la enfermedad I alguna vez durante su vida, 15 % contraerá eventualmente la enfermedad II, y el 3 % contraerá ambas. Encuentre la probabilidad de que un individuo elegido al azar de esta población, contraiga al menos una enfermedad

A: El individuo contrae la enfermedad I

B: El individuo contrae la enfermedad II

$P(A) = 0.1,\ P(B) = 0.15,\ P(AB) = 0.03$

$P(\text{contraiga al menos una enfermedad}) = P(A \cup B)$

$P(A \cup B) = P(A) + P(B) - P(AB) = 0.1 + 0.15 - 0.03 = 0.22$

Probabilidades conjunta, marginal y condicional

Caracterizar a un grupo de estudiantes según dos variables: género (H, M) y fumar (S, N)

	Fuma		
Género	Si (S)	No (N)	Total
Hombre (H)	20	40	60
Mujer (M)	10	30	40
Total	30	70	100

Probabilidad conjunta y marginal

	Fuma		
Género	Si (S)	No (N)	Total
Hombre (H)	20	40	60
Mujer (M)	10	30	40
Total	30	70	100

Probabilidad conjunta

- $P(HS) = 20/100 = 0.2$
- $P(HN) = 40/100 = 0.4$
- $P(MS) = 10/100 = 0.1$
- $P(MN) = 30/100 = 0.3$

Probabilidad marginal

- $P(H) = 60/100 = 0.6$
- $P(M) = 40/100 = 0.4$
- $P(S) = 30/100 = 0.3$
- $P(N) = 70/100 = 0.7$

Probabilidad condicional $p(A \mid B)$

	Fuma		
Género	Si (S)	No (N)	Total
Hombre (H)	20	40	60
Mujer (M)	10	30	40
Total	30	70	100

Condicionales por fila

- $P(S \mid H) = 20/60 = 0.33$
- $P(N \mid H) = 40/60 = 0.67$
- $P(S \mid M) = 10/40 = 0.25$
- $P(N \mid M) = 30/40 = 0.75$

Condicionales por columna

- $P(H \mid S) = 20/30 = 0.67$
- $P(M \mid S) = 10/30 = 0.33$
- $P(H \mid N) = 40/70 = 0.571$
- $P(M \mid N) = 30/70 = 0.429$

Probabilidad condicional $p(A \mid B)$

La relación entre las tres probabilidades viene dada por:

$$P(S \mid H) = \frac{20}{60} = \frac{\frac{20}{100}}{\frac{60}{100}} = \frac{P(SH)}{P(H)}$$

$$P(A \mid B) + P(\bar{A} \mid B) = 1$$

Ley multiplicativa de la probabilidad

Establece la forma de calcular la probabilidad de intersección de eventos

$$P(A \cap B) = P(A) \times P(B \mid A)$$

Ejemplo

En un estudio sobre alcohólicos se informa que el 40 % de los mismos tiene padre alcohólico, el 6 % madre alcohólica y el 42 % al menos uno de los padres alcohólicos. Obtener la probabilidad de que elegido uno al azar:

1. Tenga ambos padres alcohólicos.
   - $P(M_A \cap P_A) =?$
2. Tenga una madre alcohólica si lo es el padre.
   - $P(M_A \mid P_A) =?$

Independencia estadística

Dos eventos A y B son independientes si la ocurrencia de uno de ellos no afecta la probabilidad de ocurrencia del otro

$$P(A \mid B) = P(A)$$

$$P(AB) = P(A) \times P(B)$$

Ejemplo

Un contratista debe realizar tres trabajos similares. La experiencia indica que la probabilidad de que un determinado trabajo se demore mas de lo planeado es 0.4. Si el ritmo al cual trabaja es *independiente*,

a ¿cuál es la probabilidad de que los tres trabajos se terminen en el tiempo planeado?

b ¿cuál es la probabilidad de que los tres trabajos tengan retraso?

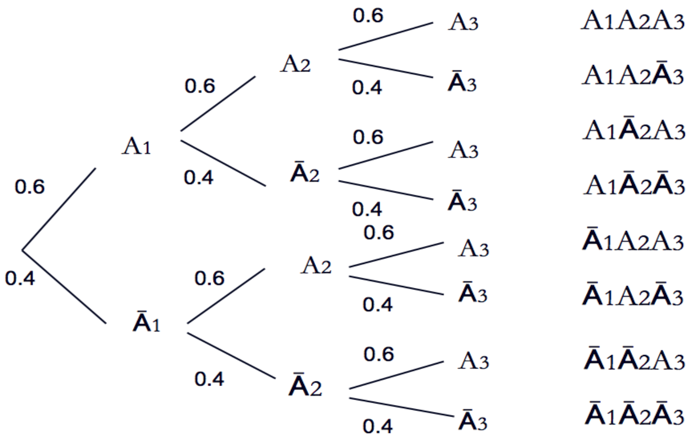
c ¿cuál es la probabilidad de que al menos uno de los trabajos tenga retraso?

Ejemplo

A_1: trabajo 1 se termina en el tiempo planeado $P(A_1) = 0.6$ $P(\bar{A}_1) = 0.4$

A_2: trabajo 2 se termina en el tiempo planeado $P(A_2) = 0.6$ $P(\bar{A}_2) = 0.4$

A_3: trabajo 3 se termina en el tiempo planeado $P(A_3) = 0.6$ $P(\bar{A}_3) = 0.4$



Ejemplo

a $P(A_1A_2A_3) = P(A_1) \times P(A_2) \times P(A_3) = 0.6 \times 0.6 \times 0.6 = 0.216$

La probabilidad de que los tres trabajos se terminen en el tiempo planeado es 0.126

b $P(\bar{A}_1\bar{A}_2\bar{A}_3) = P(\bar{A}_1) \times P(\bar{A}_2) \times P(\bar{A}_3) = 0.4 \times 0.4 \times 0.4 = 0.064$

La probabilidad de que los tres trabajos tengan retraso es 0.064

c $P(\bar{A}_1 \cup \bar{A}_2 \cup \bar{A}_3) = P(\bar{A}_1) + P(\bar{A}_2) + P(\bar{A}_3) - [P(\bar{A}_1\bar{A}_2) + P(\bar{A}_1\bar{A}_3) + P(\bar{A}_2\bar{A}_3)] + P(\bar{A}_1\bar{A}_2\bar{A}_3) = 3 \times 0.4 - 3 \times 0.16 + 0.064 = 0.784$

Otra forma: $P(\bar{A}_1 \cup \bar{A}_2 \cup \bar{A}_3) = 1 - P(A_1 \cap A_2 \cap A_3) = 1 - 0.216 = 0.784$

La probabilidad de que al menos uno de los trabajos tenga retraso es 0.784

Ley de probabilidad total

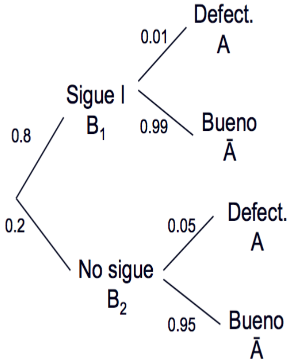
Sea Ω y $A \subset \Omega$. Si $\Omega = B_1 \cup B_2 \cup \ldots \cup B_k$ con $B_i \cap B_j = \emptyset \forall i \neq j$ y $P(B_i) > 0$, entonces:

$$P(A) = \sum_{i=1}^{k} P(B_i)P(A/B_i)$$

Ejemplo

La probabilidad que un artículo sea defectuoso es de 0.01 si el operario sigue con exactitud las instrucciones del protocolo de fabricación, y de 0.05 si el operario no las sigue. Si el operario sigue las instrucciones 80 % de las veces, ¿qué porcentaje de los productos fabricados por el operario serán defectuosos?

Ejemplo



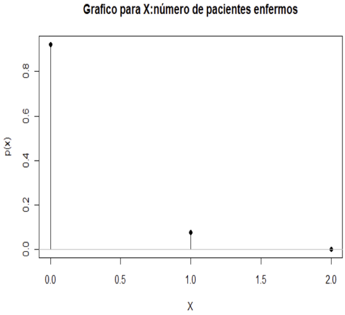
B_1: el operario sigue con exactitud el protocolo
B_2: el operario no sigue con exactitud el protocolo
A: el artículo es defectuoso

$P(B_1) = 0.8 \ P(A \mid B_1) = 0.01$
$P(B_2) = 0.2 \ P(A \mid B_2) = 0.05$

Variable aleatoria

Función que transforma los resultados del experimento (Ω) en puntos sobre la recta real $\Re$
X: número de pacientes enfermos

Evento	x	$p(x)$
EE	2	0.0016
ES, SE	1	0.0768
SS	0	0.9216



Clasificación

- **Discreta**: toma un conjunto *contable* de valores en un intervalo de los reales ($\Re$). En general, este tipo de variables procede de conteos
  - Número de personas que compran el producto A
  - Pacientes que presentan reacción adversa a un tratamiento
- **Continua**: aquellas variables que pueden tomar cualquier valor en un intervalo de los reales ($\Re$)
  - Tiempo de vida de un producto
  - Distancia que recorre un auto por galón de combustible
  - Contenido de agua en una fruta

Función de probabilidad

La función de probabilidad de una variable es una expresión matemática mediante la cual se pueden calcular las probabilidades asociadas al rango de variación de X (Ω_x)

La función de probabilidad de una variable *discreta* se le suele denominar *función de masa de probabilidad (fmp)*

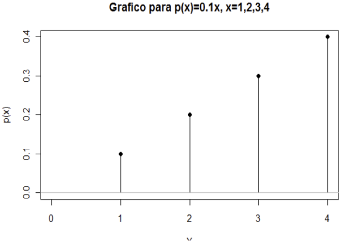
Si X es *continua* se llama *función de densidad de probabilidad* denotada con *fdb*

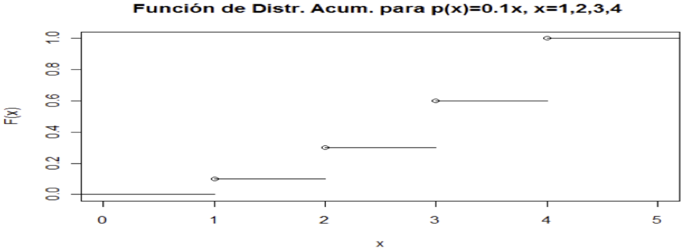
Función de probabilidad

- Discreta - *fmp*
  - $0 \leq p(x) \leq 1$
  - $\sum_{x \in \Omega} p(x) = 1$
- Continua - *fdp*
  - $f(x) \geq 0$
  - $\int_{-\infty}^{\infty} f(x)dx = 1$

X: número de días por semana que entrena un deportista aficionado
$f(x) = 0.1x, x = 1, 2, 3, 4$

x	$p(x)$	$F(x)$
1	0.1	0.1
2	0.2	0.3
3	0.3	0.6
4	0.4	1.0





Valor esperado y varianza

Discreta: $\mu = E(X) = \sum_{x \in \Omega} x \times p(x)$

Continua: $\mu = E(X) = \int_{-\infty}^{\infty} x \times f(x) dx$

$$\sigma^2 = V(X) = E(X - \mu)^2 = E(X^2) - \mu^2$$

Ejemplo

La vida máxima de la patente para un nuevo fármaco es 17 años. Si se resta el tiempo que requiere la FDA para probar y aprobar el fármaco se obtiene la vida real de la patente, es decir, el tiempo que la compañía farmacéutica tiene para recuperar los costos de la investigación y desarrollo y conseguir una utilidad. Considere la siguiente distribución del tiempo de vida de la patente para los nuevos fármacos

X (años)	3	4	5	6	7	8	9	10	11	12	13
$p(x)$	0.03	0.05	0.07	0.10	0.14	0.20	0.18	0.12	0.07	0.03	0.01

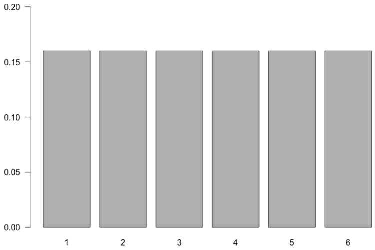
- Encuentre el número esperado de años de vida de la patente para un nuevo fármaco
- Determine la varianza y la desviación estándar de X

Distribución uniforme

Todos y cada uno de los resultados del experimento tienen la misma probabilidad de ocurrir.

Ejemplo

En el lanzamiento de un dado de 6 caras, es espacio muestral es $\Omega = \{1, 2, 3, 4, 5, 6\}$ y la ocurrencia de cada uno de estos resultados tiene una probabilidad de 0.1666667.



x	$P(x)$
1	0.16
2	0.16
3	0.16
4	0.16
5	0.16
6	0.16

Distribución Binomial

Distribución de probabilidad discreta que indica el número de éxitos al realizar n ensayos independientes entre sí, con un probabilidad individual fija de ocurrencia de éxito

- Consta de n ensayos (repeticiones)
- Cada ensayo puede resultar en un éxito (E) o fracaso (F)
- $P(E) = p(constante)$
- $P(F) = q = 1 - p$
- X: número de éxitos observados en los n ensayos $\Omega_x = 0, 1, 2, 3, ..., n$

Distribución Binomial

Función de masa de probabilidad
Si X es una variable que sigue las condiciones de un experimento binomial, entonces:

$$P(X = x) = P(x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x q^{n-x}, x = 0, 1, 2, \ldots, n$$

$$\left(\begin{array}{c} n \\ x \end{array} \right) = \frac{n!}{(n-x)!x!}$$

$$X \sim B(n; p)$$

1. n: número de veces que se repite el experimento
2. p: probabilidad individual de éxito

Distribución Binomial

Valor Esperado y Varianza
Si $X \sim B(n; p)$ entonces el valor esperado (promedio) y la varianza son:

$$\mu = E(X) = n \times p$$

$$\sigma^2 = V(X) = n \times p \times q$$

Distribución Poisson

Modelo para estudiar fenómenos que ocurren con baja frecuencia en alguna unidad dimensional (espacio, tiempo, volumen)
(X: número de eventos que ocurren por unidad)

- La probabilidad de que un evento ocurra en una unidad dimensional (intervalo) dada es la misma para todas la unidades
- La probabilidad de que un evento ocurra en un intervalo es proporcional a su longitud
- El número de eventos que ocurre en un intervalo es independiente de los que ocurren en los demás intervalos

Distribución Poisson

Función de masa de probabilidad
X: número de eventos que ocurren por unidad
μ: promedio de ocurrencias por unidad

$$p(x) = \frac{e^{-\mu}\mu^{x}}{x!}, x = 0, 1, 2, \ldots$$

$$X \sim P(\mu)$$

1. μ: media de la distribución

Distribución Poisson

Valor Esperado y Varianza
Si $X \sim P(\mu)$ entonces el valor esperado (promedio) y la varianza son:

$$E(X) = \mu$$

$$\sigma^2 = V(X) = \mu$$

Distribución Uniforme

$$X \sim U(a; b)$$

Función de densidad de probabilidad

$$f(x) = \frac{1}{b-a}; a \leq x \leq b$$

Función acumulada de probabilidad

$$F(x) = P(X \leq x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & x > b \end{cases}$$

1. a: valor mínimo de X
2. b: valor máximo de X

Distribución Uniforme

Valor Esperado y Varianza
Si $X \sim U(a; b)$ entonces el valor esperado (promedio) y la varianza son:

$$\mu = E(X) = \frac{b+a}{2}$$

$$\sigma^2 = V(X) = \frac{(b-a)^2}{12}$$

Distribución Exponencial

$$X \sim Exp(\lambda)$$

Función de densidad de probabilidad

$$f(x) = P(x) = \begin{cases} 0 & x < 0 \\ \lambda e^{-\lambda x} & x \geq 0 \end{cases}$$

Función acumulada de probabilidad

$$F(x) = P(X \leq x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-\lambda x} & x \geq 0 \end{cases}$$

Distribución Exponencial

Valor Esperado y Varianza
Si $X \sim Exp(\lambda)$ entonces el valor esperado (promedio) y la varianza son:

$$\mu = E(X) = \frac{1}{\lambda}$$

$$\sigma^2 = V(X) = \frac{1}{\lambda^2}$$

Distribución Normal

$$X \sim N(\mu, \sigma)$$

Función de densidad de probabilidad

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, -\infty < x < \infty, \sigma > 0$$

Función acumulada de probabilidad

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-x^2/2} dx$$

Distribución Normal Estándar

Si una variable normal tiene **promedio cero** y **varianza uno** se denomina normal estándar (típica)

$$Z \sim N(0; 1)$$

Estandarización:

$$Z = \frac{X - \mu}{\sigma}$$

$$T \sim t_v$$

Función de densidad de probabilidad

$$P(X = x) = \frac{\Gamma(\frac{v+1}{2})}{\sqrt{v\pi}\Gamma(\frac{v}{2})}(1 + \frac{x^2}{v})^{-\frac{v+1}{2}}$$

Función acumulada de probabilidad

$$P(X \leq x) = \frac{1}{2} + x\Gamma(\frac{v+1}{2}) \times \frac{{}_2F_1(\frac{1}{2}, \frac{v+1}{2}; \frac{3}{2}; -\frac{x^2}{v})}{\sqrt{v\pi}\Gamma(\frac{v}{2})}$$

donde ${}_2F_1$ es la función hipergeométrica

Distribución t-Student pt, qt

$$T \sim t_v$$

Promedio:

$$E(T_v) = 0, v > 1$$

Varianza:

$$V(T_v) = v/(v-2), v > 2$$

Teorema del Límite Central

Sea $X \sim ?(\mu; \sigma^2)$ y $X_1, X_2, \ldots, X_n$ una muestra aleatoria, con promedio $E(X_i) = \mu$ y varianza $V(X_i) = \sigma^2, i = 1, 2, \ldots, n$, entonces la distribución del promedio muestral tiende asintóticamente a una normal con promedio μ y varianza σ^2/n, a medida que n crece

$$\bar{X} \underset{n\to\infty}{\approx} N(\mu; \sigma/\sqrt{n}) \Rightarrow Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \underset{n\to\infty}{\approx} N(0; 1)$$

Distribución de muestreo para el promedio

Sea $X \sim N(\mu; \sigma^2)$ y $X_1, X_2, \ldots, X_n$ una muestra aleatoria, con promedio $E(X_i) = \mu$ y varianza $V(X_i) = \sigma^2, i = 1, 2, \ldots, n$ desconocida, entonces:

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

Intervalo de Confianza para el Promedio

Distribución muestral del promedio:

$$T_{n-1} = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

Intervalo de confianza:

$$\bar{x} \pm t_{\alpha/2;n-1} \frac{s}{\sqrt{n}}$$