Curso de Estadística básica para Data Scientists

Dae-Jin Lee < lee.daejin@gmail.com >

TEMA 4. Variables Aleatorias y Probabilidad

Índice

Regresar a la página principal

1. Concepto de Variable Aleatoria

En ocasiones, describir todos los posibles resultados de un experimento aleatorio no es suficiente.

- Lanzar una moneda 3 veces: $\{(CCC), (CCX), ...\}$
- Lanzar un dado dos veces: $\{(1, 1), (1, 2), (1, 3), ...\}$

A veces es útil asociar un número a cada resultado del experimento → Definir una variable

No conocemos el resultado del experimento antes de realizarlo

No conocemos el valor que va a tomar la variable antes del experimento

→ **Variable Aleatoria**

Ejemplo:

- Lanzar una moneda 3 veces: $\{(CCC), (CCX), ...\}$

  $X =$ "Nº de Caras en el 1er lanzamiento $X[(CCC)] = 1, X[(XCX) = 0, ...]$"

- Lanzar un dado dos veces: $\{(1, 1), (1, 2), (1, 3), ...\}$

$$Y = \text{“Suma de puntuaciones } Y[(1, 1)] = 2, Y[(1, 2)] = 3, ...\text{”}$$

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

Las variables aleatorias se representan por letras mayúsculas, normalmente empezando por el final del alfabeto: X,Y, Z, etc.

Los posibles valores que puede tomar la variable se representan por letras minúsculas, ej: $x = 1$ es un posible valor de la v.a. X.

Ejemplos:

- Número de unidades defectuosas en una muestra aleatoria de 5 unidades
- Número de defectos superficiales en un cm^2 de cierto material
- Tiempo de duración de una bombilla
- Resistencia a la compresión de un material de construcción

1.1. Variables aleatorias discretas y contínuas

El rango de una variable aleatoria es el conjunto de valores que puede tomar la variable.

Atendiendo al rango las variables se pueden clasificar como:

- Variables aleatorias discretas: Aquellas en las que el rango es finito infinito numerable.
- Variables aleatorias continuas: Aquellas en las que el rango es un intervalo de números reales.

1.1.1. Variables aleatorias discretas

Los valores de una variable aleatoria cambian de un experimento a otro al cambiar los resultados del experimento. Una variable aleatoria está definida por:

- los valores que toma
- la probabilidad de tomar cada uno esos valores

Función de probabilidad es una función que indica la probabilidad de cada valor

$$p(x_i) = P(X = x_i)$$

Ejemplo:

- Tiramos una moneda 3 veces. Representamos cara por c y cruz por z.

$$W = \{ccc, ccz, czc, zcc, czz, zcz, zzc, zzz\}$$

La probabilidad de cada suceso elemental es $1/8$*. Por ejemplo* $p(ccc) = 1/8$*, ya que la probabilidad de sacar cara en una tirada es* $1/2$ *según la definición clásica y las tiradas son independientes.*

Definimos la v.a. X=número de caras, que puede tomar los valores {0, 1, 2, 3}. Se buscan todos los puntos muestrales que dan lugar a cada valor de la variable y a ese valor se le asigna la probabilidad del suceso correspondiente.

X	Sucesos	P(X=x)
0	{zzz}	1/8
1	{czz, zcz, zzc}	3/8
2	{ccz, czc, zcc}	3/8
3	{ccc}	1/8

En ocasiones nos puede interesar la probabilidad de que una variable tome un valor menor o igual que una cantidad (**Función de probabilidad**)

$$F_n(X) = P(X \leq x_n)$$

En ejemplo anterior:

X	Sucesos	F(X)
0	{zzz}	1/8
1	{czz, zcz, zzc}	4/8
2	{ccz, czc, zcc}	7/8
3	{ccc}	1

1.1.2. Variables aleatorias continuas

Función de densidad $f(x)$ describe la distribución de probabilidad de una variable continua. Es una función continua que verifica:

- $f(x) \geq 0$
- $\int_{-\infty}^{\infty} f(x)dx = 1$
- $P(a \leq X \leq b) = \int_b^a f(x)dx$

Función de distribución

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)du \qquad -\infty < x < \infty$$

En el caso discreto la diferencia entre dos valores consecutivos de F(x) proporcionan la función de probabilidad. En el caso de variables continuas:

$$f(x) = \frac{dF(x)}{dx}$$

Propiedades:

- $a < b \rightarrow F(a) \leq F(b)$ es no decreciente
- $F(-\infty) = P(X \leq -\infty) = \int_{-\infty}^{\infty} = 0$; $F(+\infty) = P(X \leq +\infty) = \int_{-\infty}^{\infty} = 1$, es contínua

2. Medidas características de una variable aleatoria

La media μ o Esperanza matemática E

- $\mu = E(X) = \sum_i x_i p(x_i)$ (v.a. discreta)
- $\mu = E(X) = \int_{-\infty}^{\infty} x f(x) dx$ (v.a. contínua)

La **Varianza**

$$Var[X] = E\left[(X - E[X])^2\right] = E[X^2] - (E[X])^2$$

2.0.1. Desigualdad de Chebyshev

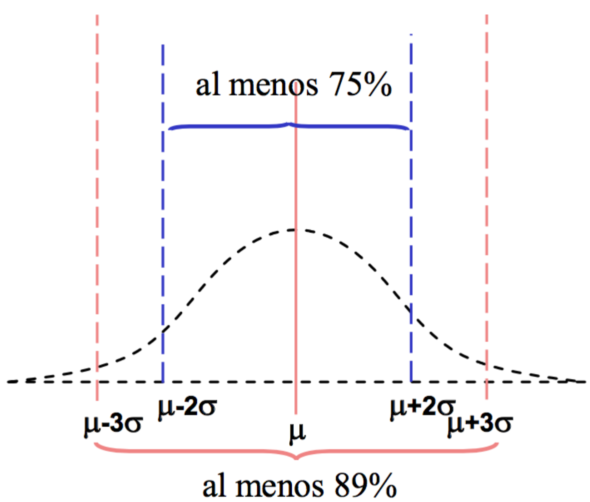
Si X es una variable aleatoria con:

$\mu = E[X]$ y $\sigma^2 = Var[X]$

Se puede demostrar que gran parte de la distribución está situada en un intervalo centrado en μ y que tiene una amplitud varias veces σ. En concreto:

$$\forall k > 0 \quad P(\mu - k\sigma \leq X \leq \mu + k\sigma) \geq 1 - \frac{1}{k^2}$$

Es decir, la probabilidad de realizar una observación de una variable y que esté en ese intervalo es mayor o igual que $1 - \frac{1}{k^2}$



Sea

$m_k = E[X^k]$ el momento de order k respecto al origen

$\mu_k = E[(X-\mu)^k]$ el momento de order k respecto a la media

Coeficiente de asimetría

$CA = \frac{\mu_3}{\sigma^3}$

Coeficiente de apuntamiento

$CA_p = \frac{\mu_4}{\sigma^4}$ ó $\frac{\mu_4}{\sigma^4} - 3$

2.0.2. Transformaciones lineales de variables aleatorias

Sea $Y = a + bX$

$E[Y] = a + bE[X]$ $Var[Y] = b^2 Var[X]$

3. Distribuciones de Probabilidad

3.1. Distribución Binomial $Bin(n, p)$

La variable aleatoria binomial, X, expresa el número de éxitos obtenidos en cada prueba del experimento.

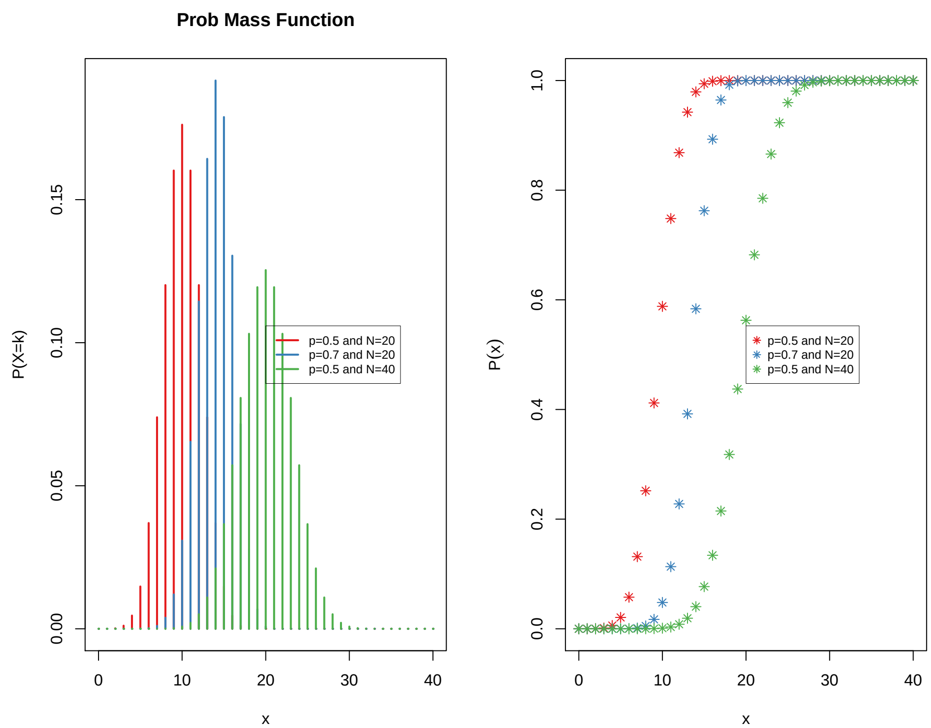
- En cada prueba del experimento sólo son posibles dos resultados:
- La probabilidad del suceso A es constante, es decir, que no varía de una prueba a otra. Se representa por p.
- El resultado obtenido en cada prueba es independiente de los resultados obtenidos anteriormente.

La distribución Binomial es una distribución de probabilidad discreta, que describe el resultado de un experimento de n pruebas independientes. Cada prueba se asume que tiene dos posibles valores. si la probabilidad de éxito es p, entonces la probabilidad de obtener k éxitos de un experimento con n pruebas independientes viene dado por la **función de masa de probabilidad**:

$$f(k, n, p) = \Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, 2, ..., n$$

La **función acumulativa de distribución** es:

$$F(k; n, p) = \Pr(X \leq k) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i}$$



con *media* np y *varianza* $np(1-p)$.

Pregunta:

Supongamos que en un examen de matemáticas hay 12 preguntas de respuesta múltiple. Cada una de las pregunras tiene 5 posibles respuestas de las cuáles tan sólo una de ellas es correcta.

- ¿Cuál es la probabilidad de obtener 4 o menos respuestas correctas si un estudiante contesta al azar?

- ¿Cuál es la probabilidad de que el alumno tenga 2 ó 3 respuestas correctas?

Pregunta:

Supongamos que la empresa A produce el producto B con una probabilidad de 0,005 de productos defectuosos. Supón que el prodcuto B se distribuye en lotes de 25 items. ¿Cuál es la probabilidad de que un lote seleccionado al azar tenga exactamente 1 producto defectuoso? ¿Cuál es la probabilidad de que un lote seleccionado al azar no tenga más de un item defectuoso?

Soluciones aquí

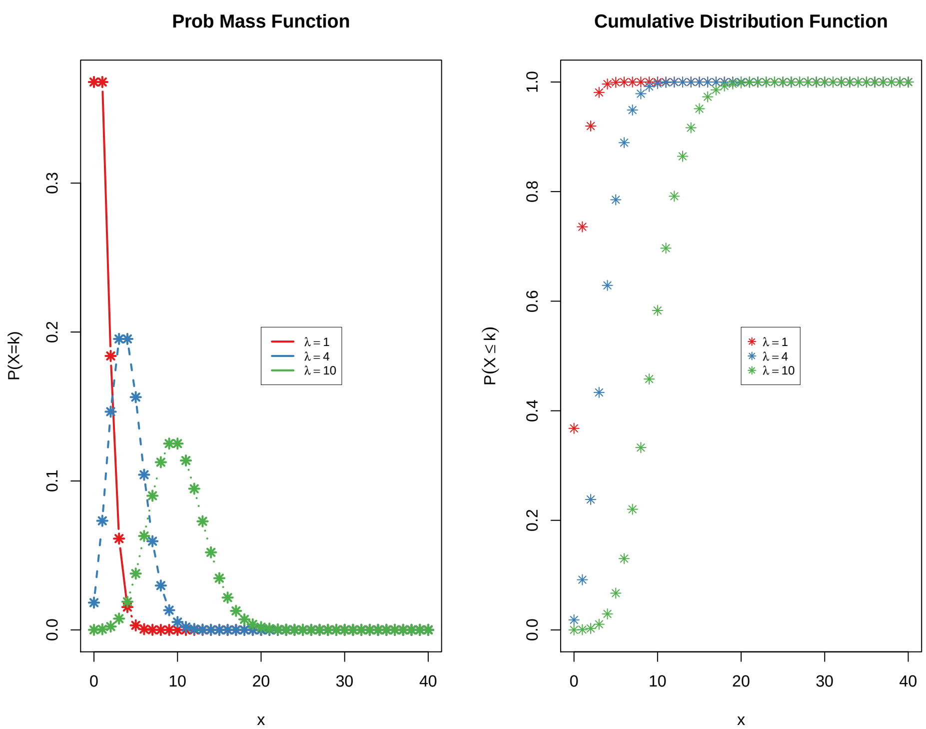
3.2. Distribución de Poisson $Pois(\lambda)$

La distribución de Poisson hace referencia a la modelización de situaciones en las que nos interesa determinar el número de hechos de cierto tipo que se pueden producir en un intervalo de tiempo o de espacio. Si el parámetro λ es la media de sucesos por intervalo, la probabilidad de tener k sucesos en el intervalo, se define como la función de probabilidad de masa:

$$\Pr(k \text{ sucesos en el intervalo}) = \frac{\lambda^k e^{-\lambda}}{k!}$$

La **función de densidad acumulada** es

$$P(X \leq x \mid \lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \quad \text{for } x = 0, 1, 2, ...$$



Pregunta:

Supongamos que el número de insectos en una plantación de un metro cuadrado, viene dada por una distribución de Poisson de media $\lambda = 10$. Calcula la probabilidad de encontrar exactamente 12 insectos en una parcela de 1 m^2.

```
## [1] 0.09478033
```

Pregunta: Si hay 12 coches cruzando un puente por minuto en media. Calcula la probabilidad de encontrar 17 ó más coches cruzando el puente en un minuto cualquiera.

Soluciones aquí

3.3. Aproximación de la Binomial y la Poisson

Ejemplo:

El 5 % de las bombillas de un árbol de navidad manufacturado por una compañia son defectuosas. El gerente del departamento de Control de calidad de la empresa está preocupado y ha seleccionado 100 bombillas de la cadena de montaje. Sea X el número de bombillas defectuosas. ¿Cuál es la probabilidad de que la muestra de 100 bombillas tenga a lo sumo 3 bombillas defectuosas?

```
p = 0.05
k = 3
n = 100
pbinom(k,size=n,prob=p)
```

```
## [1] 0.2578387
```

Se puede demostrar que la distribución Binomial se puede aproximar mediante la distribución de Poisson cuando n es grande. Mediante la distribución de Poisson, la media es $\lambda = np$

```
lambda <- n*p
sum(dpois(0:3,lambda))
```

```
## [1] 0.2650259
```

Esta aproximación es válida con n grande y p pequeño. En general, con $n \geq 20$ y $p \leq 0{,}05$, o bien cuando $n \geq 100$ y $p \leq 0{,}10$.

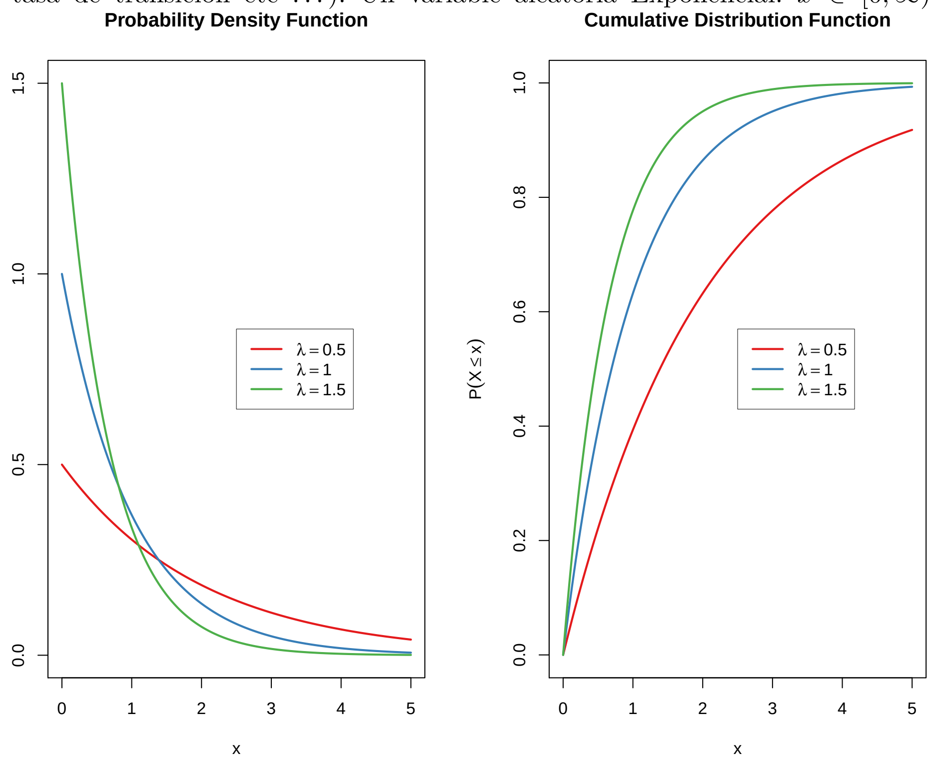
3.4. Distribucion exponencial $Exp(\lambda)$

La distribución Exponencial mide la probabilidad de que una variable aleatoria que describe el tiempo entre eventos de un proceso de Poisson, es decir un proceso que ocurre de manera contínua e independiente a una tasa constante. Es un caso particular de la distribución Gamma y el análogo contínua de la distribución geométrica y tiene la propiedad de falta de memoria.

La funcion de densidad de probabilidad es

$$f(x;\lambda) = \lambda \exp(-\lambda x)$$

donde $\lambda > 0$ es la tasa del evento (también conocido tasa de llegadas, tasa de transición etc ...). Un variable aleatoria Exponencial: $x \in [0, \infty)$



La función de distribución acumulada es

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

La media $\mathbb{E}(X) = 1/\lambda$, y $\mathbb{V}ar(X) = 1/\lambda^2$.

Pregunta: Supongamos que la cantidad de tiempo que los clientes de un banco pasan en él se distribuye como una exponencial con media 10 minutos, por tanto $\lambda = 1/10$.

- ¿Cuál es la probabilidad de que un cliente pase más de 15 minutos en el banco?
- ¿Cuál es la probabilidad de que un cliente pase más de 15 minutos en el banco dado que sigue en el banco después de 10 minutos?

Soluciones aquí

3.5. Distribución Normal $\mathcal{N}(\mu, \sigma^2)$

La función de densidad es

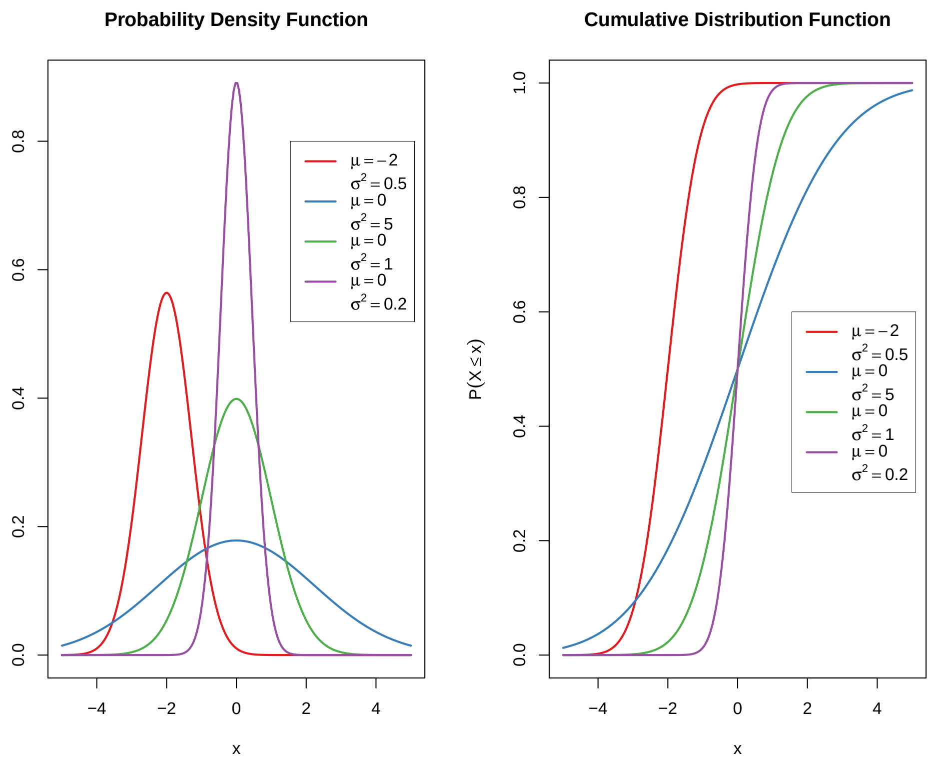
$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

donde

- μ es la media (tambien mediana y moda).
- σ es la desviación estándar ($\sigma > 0$).
- σ^2 la varianza.

El proceso de estandarización de la distribución Normal, consiste en transformar una variable $N(\mu, \sigma)$ en $N(0, 1)$, i.e.

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$



Pregunta:

Sea X una variable aleatoria distribuida como una Normal con media $\mu = 30$ y desviación estandard $\sigma = 4$. Calcula

a) $P(x < 40)$

b) $P(x > 21)$

c) $P(30 < x < 35)$

Pregunta:

El acceso a una Universdad viene dado por un examen a nivel Nacional. Las puntuaciones de este examen se distribuyen mediante una distribución Normal con media 500 y desviación estandard de 100. Tomás quiere ser admitido a esta universidad y sabe que debe obtener una nota media superior al 70 % de los estudiantes que hicieron el examen. Tomás sacó una nota de 585. ¿Será admitido?

Soluciones aquí