Fundamentos para al Análisis de Datos y la Investigación

Tema 5 - Modelos de distribuciones de datos y simulación

José María Sarabia

Máster Universitario en Ciencia de Datos
CUNEF Universidad

Contents

1 Introducción

- Veremos modelos de distribuciones de datos
- Estudiaremos modelos de una variable (discretos y continuos) y modelos multivariables
- La simulación es uno de los aspectos importantes, ya que nos permite recrear el modelo y obtener datos sintéticos
- Comentaremos la ley de los grandes números

Contents

2 Distribuciones de datos

- Los modelos de datos se obtienen por medio de las *variables aleatorias*, que es una variable cuyos valores son estocásticos, es decir, no se conocen con certidumbre
- Un modelo o una variable no aleatoria es determinista
- Existen dos tipos de modelos de datos: discretos (toman un número finito de valores) y continuos (pueden tomar infinitos valores)
- La variable Z se reserva para la distribución normal estándar
- Todo modelo tiene asociado tres funciones: la *función de densidad* (o función de cuantía en el caso discreto), la *función de distribución* y la *función de cuantiles*

2 Distribuciones de datos

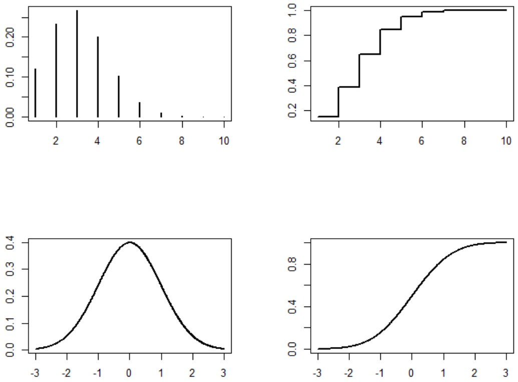

Figura: Funciones de densidad y de distribución de variables aleatorias discretas y continuas

Contents

3 Modelos de distribuciones univariantes

- Una *distribución de probabilidad* es un modelo paramétrico que representa un determinado fenómeno aleatorio por medio de una variable aleatoria.
- *Modelo paramétrico* significa, que viene especificado salvo un conjunto de parámetros que se estiman (o calibran) a partir de los datos.
- Una distribución de probabilidad puede ser: discreta (número morosos en una sucursal bancaria) o continua (precio de una acción).
- En la construcción de todo modelos estadístico tenemos tres etapas básicas: (i) especificación, (ii) estimación y (iii) validación.
- Para trabajar con una distribución en `R`, es suficiente con conocer la distribución de interés (p.e. distribución normal) y los valores de los parámetros correspondientes (en el caso normal, media y desviación típica).

3 Distribuciones de probabilidad en R

- Para cada distribución de probabilidad R proporciona los siguientes comandos:
  - `dxxx`: función de densidad de la distribución `xxx`
  - `pxxx`: función de distribución de la distribución `xxx`
  - `qxxx`: función cuantil de la distribución `xxx`
  - `rxxx`: generador de números aleatorios de la distribución `xxx`

3 Distribución Binomial

- Es la distribución de probabilidad que cuenta *el numero de éxitos en n ensayos con dos posibles resultados*
- La función de densidad, media y varianza son:

$$\begin{aligned} P(X = x) &= \binom{n}{x} p^x (1-p)^{n-x}, \; x = 0, 1, \ldots, n \\ E(X) &= np; \; var(X) = np(1-p) \end{aligned}$$

- Ejemplos:
  - Número de empresas del IBEX con rendimiento diario positivo un día fijado
  - Distribución del número de ases al lanzar 15 veces un dado
  - Distribución del número de victorias o empates de un equipo de fútbol en 10 partidos
  - Distribución del número de caras al lanzar 5 veces una moneda

3 Distribución Binomial

- Comandos:
  - `dbinom(x,nensayos,probexito)`
  - `pbinom(x,nensayos,probexito)`
  - `qbinom(prob,nensayos,probexito)`
  - `rbinom(nsample,nensayos,probexito)`

3 Distribución de Poisson

- Expresa la *distribución de ocurrencia de sucesos por unidad de tiempo* (a partir de la ocurrencia media por unidad).

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!},\ x = 0, 1, \ldots;\ E(X) = \lambda;\ var(X) = \lambda$$

- Ejemplos:
  - Número de páginas que acceden a una web por minuto.
  - El número de automóviles que pasan por un punto de una carretera en intervalos de un minuto.
  - El número de llamadas telefónicas en una central por minuto.
  - Número de estrellas en un determinado volumen de espacio
- Comandos:
  - `dpois(x,media)`
  - `ppois(x,media)`
  - `qpois(prob,media)`
  - `rpois(nsample,media)`

3 Distribución exponencial

- La distribución exponencial es la distribución del tiempo de espera entre dos sucesos, siempre que los sucesos ocurran según una distribución de Poisson.
- La función de densidad, media y varianza son:

$$f(x) = \frac{e^{-x/\mu}}{\mu},\ x > 0;\ E(X) = \mu;\ var(X) = \mu^2$$

- Ejemplos:
  - Tiempo de espera hasta que llegue el primer cliente a una sucursal bancaria.
  - El tiempo transcurrido en un call center hasta recibir la primera llamada del día.
  - El intervalo de tiempo entre huracanes de una determinada magnitud.

3 Distribución exponencial

- Comandos:
  - `dexp(x,rate=1/m)`
  - `pexp(x,rate=1/m)`
  - `qexp(prob,rate=1/m)`
  - `rexp(nsample,rate=1/m)`

3 Distribución uniforme continua

- La uniforme continua asigna la misma probabilidad a todos los valores de un intervalo $[a, b]$.
- La función de densidad, media y varianza son:

$$f(x) = \frac{1}{b-a}, \; a \leq x \leq b; \; E(X) = \frac{a+b}{2}; \; var(X) = \frac{(b-a)^2}{12}$$

- Comandos:
  - `dunif(x,min,max)`
  - `punif(x,min,max)`
  - `qunif(prob,min,max)`
  - `runif(nsampple,min,max)`

3 Otras distribuciones

- Otras distribuciones relevantes en data science incluyen:
  - *Distribución de Pareto o ley de potencias* (distribución de renta, sucesos extremos, etc.)
  - *Distribución gamma* (tiempos de espera, distribución renta, etc.)
  - *Distribución lognormal* (distribución de renta y riqueza, tamaños de empresas, precios, etc.)
  - *Distribución de Weibull* (análisis de la supervivencia, tiempo de fabricación, teoría de valores extremos, etc.)

Contents

4 Distribución multinomial

- **La distribución multinomial** es una generalización de la distribución binomial, cuando en cada experimento son posibles más de dos resultados.
- Supongamos un experimento donde son posibles k resultados con probabilidades $p_1, \dots, p_k$ de modo que $p_i \geq 0$, $i = 1, 2, \dots, k$ y $\sum_{i=1}^{k} p_i = 1$.
- *Ejemplo*. Distribución del tipo de inversor: conservador, moderado o arriesgado
- Se observan x_1 elementos del tipo 1, x_2 del tipo 2 y en general x_i del tipo i, con $i = 1, 2, \dots, k$ de modo que $\sum_{i=1}^{k} x_i = n$
- La función de densidad discreta de $(X_1, \dots, X_k)^\top$ viene entonces dada por,

$$P(X_1 = x_1, \dots, X_k = x_k) = \frac{n!}{x_1 \cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

donde $\sum_{i=1}^{k} x_i = n$ y $\sum_{i=1}^{k} p_i = 1$

4 Distribución multinomial

La distribución multinomial en R dispone de dos tipos de funciones:

- Función de de densidad:

```
dmultinom(x, size = NULL, prob, log = FALSE)
```

  donde:
  - `x`: vector de longitud k de enteros de 0:size.
  - `size`: entero, N número total de objetos en cada una de las k clases en un experimento multinomial (por defecto `sum(x)`).
  - `prob`: vector de probabilidades de longitud k.
  - `log` valor lógico; is es TRUE, se calculan log de las probabilidades.
- Simulación de muestras:

```
rmultinom(n, size, prob)
```

  siendo `n`: el número de vectores aleatorios a simular.

4 Distribución Normal Multivariante

- Se trata de la distribución multivariada más importante en Data Science
- Ejemplos: la altura y el peso; consumo y logaritmo de la renta, etc.
- La distribución normal multivariante viene determinada por el vector de medias y por la matriz de covarianzas. La estructura de dependencia es de tipo lineal
- En una distribución normal multivariante:
  - Las distribuciones marginales, condicionadas y cualquier combinación lineal de las marginales son nuevamente normales
  - Las regresiones son lineales y las varianzas condicionadas son homocedásticas (constantes).
  - Covarianza o correlación cero implica independencia. Esta propiedad no se extiende en general a otros tipos de distribuciones.

4 La Distribución Normal Bidimensional

- Veamos la *distribución normal bidimensional*
- La variable $(X_1, X_2)^\top$ sigue una distribución normal bidimensional si su función de densidad se puede escribir como:

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)} Q(x_1, x_2)\right\}$$

donde,

$$Q(x_1, x_2) = \left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2$$

donde $\mu_1, \mu_2 \in \mathbb{R}$, $\sigma_1, \sigma_2 \in \mathbb{R}^+$ y $\rho \in [-1, 1]$.

- La distribución depende de 5 parámetros

4 DLa Distribución Normal Bidimensional

- Representaremos

$$(X_1, X_2)^\top \sim \mathcal{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- El vector de medias es $\boldsymbol{\mu} = (\mu_1, \mu_2)^\top$.
- La matriz de covarianzas es

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

donde $\sigma_{12} = \rho\sigma_1\sigma_2$ es la covarianza.

4 La Distribución Normal Bidimensional

- Las distribuciones marginales son normales $\mathcal{N}(\mu_1, \sigma_1^2)$ y $\mathcal{N}(\mu_2, \sigma_2^2)$. Las combinaciones lineales son igualmente normales.
- Las distribuciones condicionadas son normales y vienen dadas por,

$$X_1|X_2 = x_2 \sim \mathcal{N}\left(\mu_1 + \frac{\rho\sigma_1}{\sigma_2}(x_2 - \mu_2); \sigma_1^2(1-\rho^2)\right),$$

y

$$X_2|X_1 = x_1 \sim \mathcal{N}\left(\mu_2 + \frac{\rho\sigma_2}{\sigma_1}(x_1 - \mu_1); \sigma_2^2(1-\rho^2)\right)$$

- Regresión de X_2 sobre X_1 es lineal

$$E(X_2|X_1 = x_1) = \mu_2 + \frac{\rho\sigma_2}{\sigma_1}(x_1 - \mu_1),$$

y

$$var(X_2|X_1 = x_1) = \sigma_2^2(1-\rho^2).$$

- Si $\rho = 0$, entonces las variables X_1 y X_2 son independientes. Los contornos de la distribución son *elipses*

4 La Distribución Normal Bidimensional

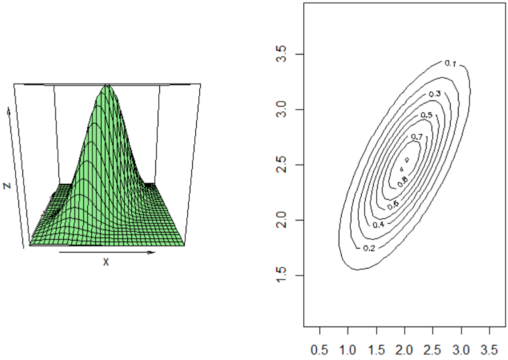

Figura: Función de densidad normal bidimensional y contornos

4 Distribución Normal Multivariante

- En este apartado presentamos la distribución normal n-dimensional, que es la extensión a n variables.
- Denotamos $\mathbf{X} = (X_1, \ldots, X_n)^\top \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
- La función de densidad viene dada por,

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \qquad (1)$$

- Nuevamente, tanto las distribuciones marginales como las condicionadas son de nuevo distribuciones normales.

4 Distribución Normal Multivariante en R

- Para el cálculo de probabilidades y simulación, podemos hacer uso de tres paquetes:
  - `MASS`: Incluye la normal bidimensional
  - `mvtnorm`: cálculo de probabilidades multivariadas normales y t de Student
  - `MVN`: permite estudiar bondad de ajuste, estimación y contraste

Contents

5 Método de la transformación inversa

- *El método de la transformación inversa* es el método básico de simular la distribución de una variable aleatoria
- Se X una variable aleatoria con función de distribución $F_X(x)$. Entonces, la nueva variable aleatoria

$$Y = F_X(X) \sim \mathcal{U}[0,1]$$

sigue una distribución uniforme en $[0,1]$.

- Si $F^{-1}(x) = \text{ínf}\{x : F(x) \geq u\}$, entonces

$$F_X(X) = U \Rightarrow X = F_X^{-1}(U)$$

- Para la simulación de datos de X hacemos:
  - Simular una muestra de tamaño n iid $u_1, \ldots, u_n \sim \mathcal{U}[0,1]$
  - La muestra simulada de la variable X es entonces

$$x_1 = F_X^{-1}(u_1), x_2 = F_X^{-1}(u_2), \ldots, x_1 = F_X^{-1}(u_n)$$

5 Simulación

Example

En una sucursal bancaria, el tiempo hasta que un cliente es atendido sigue una distribución exponencial de media 5 minutos. La función de distribución viene dada por,

$$F(x) = 1 - e^{-x/5}, \quad x \geq 0$$

Se trata de simular una muestra de tiempos de espera. Hallamos la inversa de la función de dfistribución:

$$y = 1 - e^{-x/5} \Rightarrow x = -5 \log(1 - y)$$

Por tanto si $U \sim U[0,1]$ entonces,

$$X = -5 \log(1 - U)$$

sigue una distribución exponencial de media 5, o también $X = -5 \log(U)$

5 Simulación de distribuciones en R

Distribución	Función para simular
Binomial $B(m, p)$	`rbinom(n,m,p)`
Poisson $P(\lambda)$	`rpois(n,lambda)`
Geométrica $G(p)$	`rgeom(n,p)`
Binomial negativa $BN(r, p)$	`rnbinom(n,r,p)`

Figura: Simulación de distribuciones discretas en R

5 Simulación de distribuciones en R

Distribución	Función para simular
Uniforme $U(a,b)$	`runif(n,a,b)`
Normal $N(\mu,\sigma)$	`rnorm(n,m,s)`
Exponencial	`rexp(n, rate = 1)`
Gamma	`rgamma(n, shape, rate = 1, scale = 1/rate)`
Weibull	`rweibull(n, shape, scale = 1)`
Lognormal	`rlnorm(n, meanlog = 0, sdlog = 1)`
t de Student	`rt(n, df, ncp)`

Figura: Simulación de distribuciones continuas en R

5 Simulación de variables aleatorias discretas

- Como ya hemos visto, el comando `sample` permite elegir muestras de tamaño `n` de un conjunto `x`, con o sin remplazamiento, y suponiendo equiprobabilidad o bien una determinada distribución de probabilidad. Por tanto, dicho comando permite la simulación de una distribución discreta.
- Viene definido por: `sample(x, size, replace = FALSE, prob = NULL)`

5 Simulación de distribuciones multivariantes

- Se trata de simular una distribución multivariante $(X_1, \ldots, X_p)^\top$
- Para ello, escribimos la función de distribución conjunta en términos de las *distribuciones condicionadas univariantes*

$$F_{12\ldots p}(x_1, \ldots, x_p) = F_1(x_1)F_{2|1}(x_2|x_1)\cdots F_{p|1\ldots p-1}(x_p|x_1, \ldots, x_{p-1})$$

- En el caso de $p = 2$

$$F_{12}(x_1, x_2) = F_1(x_1)F_{2|1}(x_2|x_1)$$

- Para simular una muestra bivariada de (X, Y) conociendo $F_1(x)$ y $F_{2|1}(y|x)$ hacemos
  - Generar
  $$x \sim F_1(x)$$
  - Con el valor de la x simulado previamente, se genera
  $$y \sim F_{2|1}(y|x)$$
  - El valor simulado es (x, y)

5 Simulación normal multivariante

- Existen diversas formas de siulación de la distribución normal multivariante
- La primera de ellas es haciendo uso del método anterior. Para el caso $p = 2$ (prácticas)

$$\begin{aligned} X &\sim \mathcal{N}(0,1) \\ Y|X = x &\sim \mathcal{N}(\rho x, \sqrt{1-\rho^2}) \end{aligned}$$

- Otra forma es haciendo uso de la *descomposición de Cholesky*
- La *transformación de Box-Muller* proporciona una pareja de variables $N(0,1)$ independientes. Si $U_1, U_2 \sim \mathcal{U}[0,1]$ iid, entonces

$$\begin{aligned} X_1 &= \sqrt{-2\log(U_1)}\cos(2\pi U_2), \\ X_2 &= \sqrt{-2\log(U_1)}\sin(2\pi U_2), \end{aligned}$$

son iid $N(0,1)$

5 Simulación normal multivariante

- Supongamos que queremos simular datos de una normal multivariada $\mathbf{X}$ de dimensión $p \times 1$ de vector de medias $\boldsymbol{\mu}$ y matriz de covarianzas $\boldsymbol{\Sigma}$
- En primer término usamos la descomposición de Cholesky $\Sigma = \mathbf{C}\mathbf{C}^\top$
- A contiuación generamos iid normales $(0, 1)$, $Z_1, \ldots, Z_p$, que representamos por $\mathbf{Z}$
- Finalmente $\mathbf{X} = \boldsymbol{\mu} + \mathbf{C}\mathbf{Z}$

Contents

6 Leyes de los Grandes Números

- Si $\{X_n\}$, $n = 1, 2, \ldots$ es una sucesión de v.a. independientes e igualmente distribuidas con $E(X_i) = \mu < \infty$ se verifica:

- Ley débil de los grandes números (convergencia en probabilidad): para todo $\epsilon > 0$:

$$\lim_{n\to\infty} P(|\overline{X}_n - \mu| > \epsilon) = 0$$

  o lo que es lo mismo

$$\overline{X}_n \xrightarrow{P} \mu$$

  si $n \to \infty$ donde $\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$.

- El resultado anterior significa que la media muestral se aproxima a la media de la población cuando aumenta el tamaño mustral. Se supone que dicha media poblacional existe.

- Ley fuerte de los grandes números (convergencia casi seguro):

$$P\left(\lim_{n\to\infty} \overline{X}_n = \mu\right) = 1$$

Fundamentos para al Análisis de Datos y la Investigación
Tema 5 - Modelos de distribuciones de datos y simulación

José María Sarabia

Máster Universitario en Ciencia de Datos
CUNEF Universidad