Análisis e Interpretación de Datos

MÁSTER UNIVERSITARIO EN ANÁLISIS Y VISUALIZACIÓN DE DATOS MASIVOS / VISUAL ANALYTICS AND BIG DATA

Miller Janny Ariza Garzón

Tema 5. Probabilidad Condicional y Variables Aleatorias II

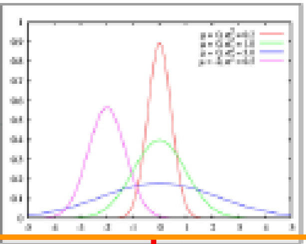
unir LA UNIVERSIDAD EN INTERNET

Universidad Internacional de La Rioja

Tabla de contenido

Contenido

Probabilidad condicional y variables aleatorias

- **Probabilidad condicional e independencia**
  - $P(A \text{ sabiendo que ha ocurrido } B) = P(A|B)$
  - $P(A|B) = \dfrac{P(A \cap B)}{P(B)}$
  - sucesos independientes P(A|B)=P(A)
- **Distribución Binomial**
  - $X \sim Bi(n, p)$
  - $P(X = k) = \binom{n}{k} p^k q^{n-k}$
- **Distribución Normal**
  - $X \sim N(\mu, \sigma)$
  - 
- **Variables aleatorias**
  - **Discretas** (un dado, una moneda etc.)
    - $P(X = x_i)$
    - $F(x_i) = P(X \leq x_i)$
    - $E(X) = \sum x_i p_i$
    - $\sigma_x^2 = \sum (x_i - E(X))^2 p_i$
  - **Continuas** (peso, tiempo etc.)
    - $P(a < x < b) = \int_a^b f(x)dx$
    - $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$
    - $E(X) = \int x f(x)dx$
    - $V(X) = \int (x-\mu)^2 f(x)dx$

2.

1.

Variable aleatoria

Se llama variable aleatoria a toda función que asocia a cada elemento del espacio muestral, un número real.

$$X: \Omega \rightarrow R$$

Se utilizan letras mayúsculas para designar las variables aleatoria: *X*, *Y*, *Z*; y sus respectivas letras minúsculas para los valores concretos de las mismas: *x*, *y*, *z*.

Variable aleatoria discreta Es la que solo puede tomar un numerable o infinito numerable de valores.

$$x \in \{x_1, x_2, \ldots, x_n\} \; o \; x \in \{x_1, x_2, \ldots\}$$

Variable aleatoria continua Es aquella que puede tomar infinitos valores dentro de un intervalo de números reales

$$x \in \{x: a < x < b;\ a, b \in R\}$$

Variable aleatoria

Ej. Exp: Lanzamiento de 2 monedas
X(s): Número de CARAS

$$\Omega = \{CC, CS, SC, SS\}$$

$$X : \Omega \rightarrow R$$

$$s \in \Omega \rightarrow X(s)$$

$$CC \rightarrow 2$$

$$CS \rightarrow 1$$

$$SC \rightarrow 1$$

$$SS \rightarrow 0$$

Función de probabilidad

$$f(x): R \rightarrow [0,1]$$

Para v.a. discretas

Función másica de probabilidad

$$f(x_i) = P(X = x_i);\ i = 1,2\ 3, \ldots$$

i. $P(X = x) \geq 0$

ii. $\sum_x P(X = x) = 1$

Probabilidades se hallan mediante:

$$\sum$$

Para v.a. continuas

Función de densidad

$$f(x)$$

i. $f(x) \geq 0$

ii. $\int_{-\infty}^{\infty} f(x)dx = 1$

Probabilidades se hallan mediante:

$$\int$$

Función de probabilidad

Para v.a. discretas

Función másica de probabilidad

$$f(x_i) = P(X = x_i);\ i = 1,2\ 3, \ldots$$

$$\left.\begin{array}{l} P(a < X < b) \\ P(a \leq X \leq b) \\ P(a \leq X < b) \\ P(a < X \leq b) \end{array}\right\} \text{Pueden ser diferentes}$$

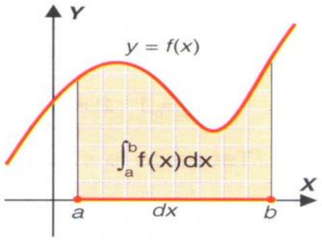
- $P(X = a)\ \epsilon\ [0,1]$

Para v.a. continuas

Función de densidad

$$f(x)$$

$$P(a < X < b) = \int_a^b f(x)dx$$



$$P(a < X < b) = P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b)$$

- $P(X = a) = 0$

Función de probabilidad

Ej, v.a. discretas

X = "el número de mujeres (M) que una pareja puede tener si tienen dos hijos"

$$\Omega = \{HH, HM, MH, MM\}$$

X	0	1	2
$P(X = x)$	¼	½	¼

- $P(X \leq 1) = ½ + ¼ = 0.75$
- $P(X < 1) = ¼$
- $P(X = 1) = ½$

Ej. v.a. continuas

X = Tiempo de espera antes de atención en una llamada a sanidad. (en minutos, máximo de 2 minutos)

$$f(x) = \begin{cases} \frac{x}{2} \; si \; 0 \leq x \leq 2 \\ 0 \; e.o.c. \end{cases}$$

$$\int_{-\infty}^{\infty} \frac{x}{2} dx = \int_0^2 \frac{x}{2} dx = \frac{x^2}{4}\bigg|_0^2 = 1$$

- $P(X < 1) = P(X \leq 1) = \int_0^1 f(x)dx = \frac{1}{4}$
- $P(X = 1) = 0$

Función acumulada de probabilidad

Se define para todos los reales $-\infty < x < \infty$.

$$F(x) = P(X \leq x)$$

Toda probabilidad se puede hallar en función de $F(x)$

Ej, v.a. discretas

$$F(a) = P(X \leq a) = \sum_{x \leq a} P(X = x)$$

Ej. v.a. continuas

$$F(a) = P(X \leq a) = \int_{-\infty}^{a} f(x)dx$$

A la Función acumulada de probabilidad también se le llama Función de distribución

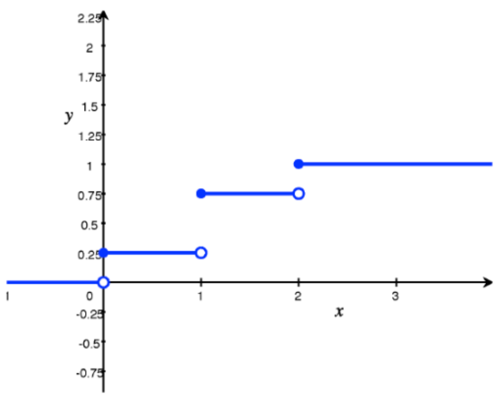
Función acumulada de probabilidad

Ej, v.a. discretas

X = "el número de mujeres (M) que una pareja puede tener si tienen dos hijos"

X	0	1	2
$P(X=x)$	¼	½	¼

$$F(x) = \begin{cases} 0, & \text{for } x < 0 \\ 0.25 & \text{for } 0 \le x < 1 \\ 0.75 & \text{for } 1 \le x < 2 \\ 1 & \text{for } x \ge 2. \end{cases}$$

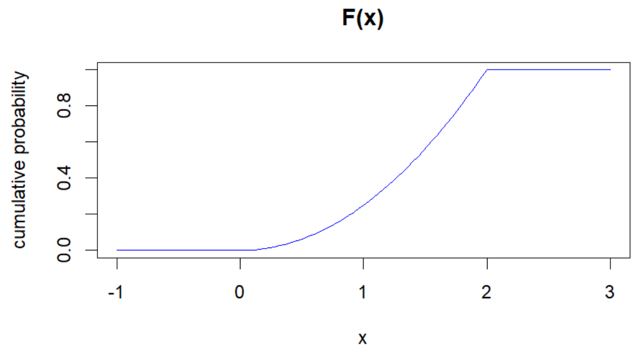


Ej. v.a. continuas

X = Tiempo de espera antes de atención en una llamada a sanidad. (en minutos, máximo de 2 minutos)

$$f(x) = \begin{cases} \frac{x}{2} \; si \; 0 \le x \le 2 \\ 0 \; e.o.c. \end{cases}$$

$$F(x) = \begin{cases} 0 \; si \; x < 0 \\ \frac{x^2}{4} \; si \; 0 \le x \le 2 \\ 1 \; si \; x > 2 \end{cases}$$



Función acumulada de probabilidad

Ej, v.a. discretas

X = "el número de mujeres (M) que una pareja puede tener si tienen dos hijos"

X	0	1	2
$P(X = x)$	¼	½	¼

$$F(x) = \begin{cases} 0, & \text{for } x < 0 \\ 0.25 & \text{for } 0 \leq x < 1 \\ 0.75 & \text{for } 1 \leq x < 2 \\ 1 & \text{for } x \geq 2. \end{cases}$$

- $P(X \leq 1) = F(1) = 0.75$
- $P(X < 1) = F(0) = 0.25$
- $P(X = 1) = F(1) - F(0) = 0.75 - 0.25 = 0.5$

Ej. v.a. continuas

X = Tiempo de espera antes de atención en una llamada a sanidad. (en minutos, máximo de 2 minutos)

$$f(x) = \begin{cases} \frac{x}{2} \; si \; 0 \leq x \leq 2 \\ 0 \; e.o.c. \end{cases}$$

$$F(x) = \begin{cases} 0 \; si \; x < 0 \\ \frac{x^2}{4} \; si \; 0 \leq x \leq 2 \\ 1 \; si \; x > 2 \end{cases}$$

- $P(X < 1) = F(1) = \frac{1}{4} = 0.25$
- $P(0.5 < X < 1.5) = F(1.5) - F(0.5) = \frac{1.5^2}{4} - \frac{0.5^2}{4} = 0.5$

$$\boxed{\boldsymbol{P(a \leq X \leq b) = F(b) - F(a)}}$$

Momentos de una variable aleatoria

$E(X) = \mu_X$ → Esperanza o valor esperado. Valor medio

$V(X) = \sigma_X^2 = E(X - \mu_X)^2$ → Varianza

Ej, v.a. discretas

$$\boldsymbol{E(X) = \sum_x x * P(X = x)}$$

$$\boldsymbol{V(X) = \sum_x (x - \mu_X)^2 * P(X = x)}$$

Ej. v.a. continuas

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

$$\boldsymbol{V(X) = \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x) dx}$$

Momentos de una variable aleatoria

Ej, v.a. discretas

X = “el número de mujeres (M) que una pareja puede tener si tienen dos hijos”

X	0	1	2
$P(X = x)$	¼	½	¼

$\boldsymbol{E(X) = 0} * ¼+\mathbf{1}* ½ \mathbf{+2} * ¼ = 1$

$V(X) = (0-1)^2 * ¼+(1-1)^2 *½ +(2-1)^2 * ¼=0.5$

$\sigma = 0.707$

Ej. v.a. continuas

X = Tiempo de espera antes de atención en una llamada a sanidad. (en minutos, máximo de 2 minutos)

$$f(x) = \begin{cases} \frac{x}{2} \; si \; 0 \leq x \leq 2 \\ 0 \; e.o.c. \end{cases}$$

$$E(X) = \int_0^2 x\left(\frac{x}{2}\right) dx = \left[\frac{x^3}{6}\right]_0^2 = \frac{8}{6} = 4/3$$

$$V(X) = \int_0^2 (x-\frac{4}{3})^2 \left(\frac{x}{2}\right) dx = \int_0^2 x^2 \left(\frac{x}{2}\right) dx - \left(\frac{4}{3}\right)^2 = 2 - \frac{16}{9} = 2/9$$

$\sigma = 0.471$

Momentos de una variable aleatoria

Propiedades:

- $E(k) = k$
- $E(k + x) = k + E(x)$
- $E(kx) = kE(x)$
- $E(ax + b) = aE(x) + b$

- $V(k) = 0$
- $V(k + x) = V(x)$
- $V(kx) = k^2V(x)$
- $V(ax + b) = a^2V(x)$

Si X y Y son dos variables aleatorias:

- $E(X + Y) = E(X) + E(Y)$
- $V(X \pm Y) = V(X) + V(Y) \pm 2Cov(X, Y) = V(X) + V(Y) \pm 2\rho_{XY}\sigma_X\sigma_Y$
- Si X y Y son independientes: $V(X \pm Y) = V(X) + V(Y)$

Correlación de X y Y

Momentos de una variable aleatoria

Ejemplo: Si quieres estudiar en EEUU (e incluso a nivel laboral) te pueden exigir realizar la prueba conocida como SAT, la cual a su vez consta de varias partes.

Aptitud matemática	Aptitud lingüística	
$E(X) = 419$	$E(Y) = 407$	
σ_X=105	σ_Y=91	ρ_{XY}=0.81

¿Cuál es el valor esperado y la varianza para las dos pruebas en conjunto?

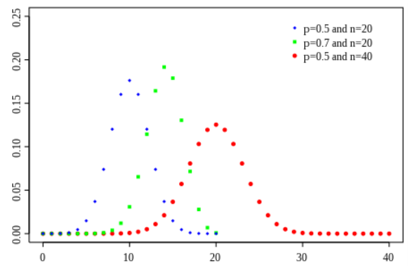
$$E(X + Y) = 419 + 407 = 826$$
$$V(X + Y) = 105^2 + 91^2 + 2 * 0.81 * 105 * 91 = 34781.1$$

Ejemplos distribuciones de probabilidad

Ej, v.a. discretas

Distribución binomial

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



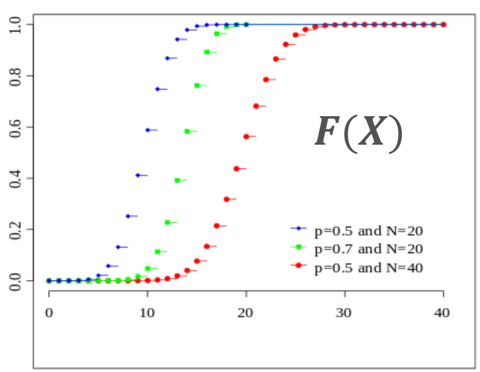
$\boldsymbol{X}$: **Número de éxitos** en n ensayos independientes (o una muestra de tamaño n).

$$X = 0,1, \dots, n$$

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

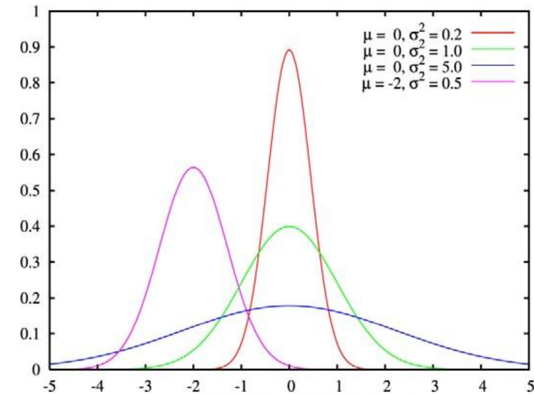
$$\mu = np$$

$$\sigma^2 = np(1-p)$$



Ej. v.a. continuas

Distribución normal



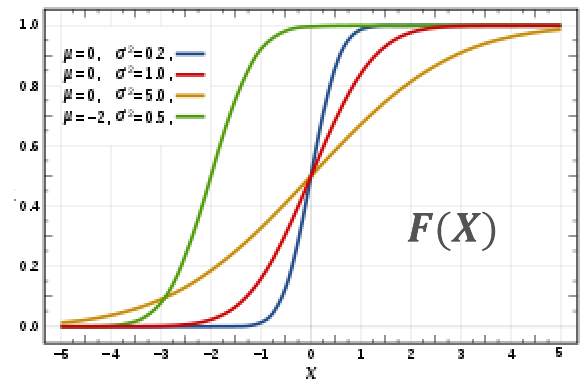
$$X \sim N(\mu, \sigma)$$

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

– La forma acampanada de la función de densidad se llama campana de Gauss.

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

Z es el número de desviaciones estándar en que se aleja un valor x de su respectiva media.



Ejemplos distribuciones de probabilidad

Ej, v.a. discretas

Distribución binomial

X =Número de caras al lanzar una moneda en 6 lanzamientos

$$X \sim Bi(6, \tfrac{1}{2})$$

$$P(X=2) = \binom{6}{2} \times 0{,}5^2 \times 0{,}5^4 = 15 \times 0{,}25 \times 0{,}0625 = 0{,}234$$

$$E(X) = 2 \times \tfrac{1}{2} = 1 \; cara \quad \text{y} \quad V(X) = 2 \times \tfrac{1}{2} \times \tfrac{1}{2} = 1/2$$

σ = 0.70

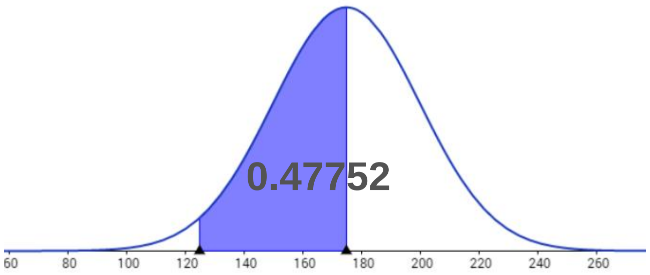
Ej. v.a. continuas

Distribución normal

Una empresa de telecomunicaciones ha determinado, que la transmisión de datos **promedio es de 175 segundos** con una **desviación estándar de 25 segundos** y una **distribución normal** definida. Con el fin de analizar la demanda de servicios y maximizar la oferta de los mismos se desea obtener la probabilidad de que los tiempos de comunicación oscilen entre 125 y 175 segundos

x = tiempo de transmisión en segundos
µ = 175 seg. (media)
σ = 20 seg. (desviación estándar)

$$P(125 < X < 175) = F(175) - F(125)$$



Ejemplos distribuciones de probabilidad

Ej, v.a. discretas

Distribución binomial en R

- dbinom(x, size, prob) Probabilidad puntual
- pbinom(x, size, prob) Probabilidad acumulada
- qbinom(p, size, prob) Quantile
- rbinom(n, size, prob) Números aleatorios

•**x** is a vector of numbers.
•**p** is a vector of probabilities.
•**n** is number of observations.
•**size** is the number of trials.
•**prob** is the probability of success of each trial.

Ej. v.a. continuas

Distribución normal en R

- dnorm(x, mean, sd) Valor en la función de densidad
- pnorm(x, mean, sd) Probabilidad acumulada
- qnorm(p, mean, sd) Quantile
- rnorm(n, mean, sd) Números aleatorios

•**x** is a vector of numbers.
•**p** is a vector of probabilities.
•**n** is number of observations(sample size).
•**mean** is the mean value of the sample data. It's default value is zero.
•**sd** is the standard deviation. It's default value is 1

Próxima sesión

Tema 6: Distribución en el muestreo

- Distribución en el muestreo del conteo y la proporción muestral
- Central del Límite y distribución de la media muestral
- Aplicabilidad del Teorema Central del Límite en ámbitos Big Data
- Estimación puntual vs estimación por intervalos
- Propiedades de los estimadores

unir

LA UNIVERSIDAD
EN INTERNET

www.unir.net