Tema 2. Probabilidad y distribución

Gustavo A. García

ggarci24@eafit.edu.co

Econometría para la Toma de Decisiones

Maestría en Economía Aplicada

Escuela de Finanzas, Economía y Gobierno

Universidad EAFIT

Link slides en formato html

Link slides en formato PDF

En este tema

- Motivación
- Variables aleatorias y sus distribuciones de probabilidad
- Tipos de variables aleatorias
- Características de las distribuciones de probabilidad
- Medidas de asociación

Lecturas

- Wooldridge, Jeffrey (2013). *Introducción a la econometría*. 5a edición, Cengage Learning. Apéndice B
- Gujarati, D. y Porter, D. (2010). *Econometría*. 5a edición, Mc Graw Hill. Apéndice A

Motivación

El análisis de probabilidad de una variable aleatoria ayuda a encontrar soluciones a diferentes situaciones

Por ejemplo, el caso de una aerolínea que tiene que decidir cuántas reservas aceptar para un vuelo en el que hay 100 asientos disponibles. Existen diferentes situaciones que hacen el problema no trivial:

- Si hay pocas reservas, todas serán aceptadas
- Pero ¿qué pasa si hay muchas personas reservando, más de 100?
  - aceptar máximo 100. Pero si hay personas que cancelan se corre el riesgo de no llenar el avión
  - aceptar más de 100. Habrá un exceso de reservas y tendrá que compensar a los pasajeros

¿Se puede determinar la cantidad óptima de reservar que deba hacer la aerolínea?

Variables aleatorias y sus distribuciones de probabilidad

- Entender el concepto de variable aleatoria es muy importante en economía y en general en las ciencias sociales $\Longrightarrow$ todo el tiempo los economistas estamos trabajando con variables aleatorias

- La variable X: el número de reservas que recibe la aerolínea, es una variable aleatoria ya que, hasta que los datos son observados, es incierto que valores tomará X

- Una variable aleatoria es aquella que toma un valor numérico que será determinado por un experimento

- Un experimento, en general, es un procedimiento que puede repetirse una cantidad infinita de veces y que tiene un conjunto bien definido de resultados

Variables aleatorias y sus distribuciones de probabilidad

- Por ejemplo, el número de reservas que recibe la aerolínea es una variable aleatoria y es importante ya que es crucial determinar el óptimo de reservas que debe aceptar la aerolínea

- El número de reservas que recibe la aerolínea es una variable aleatoria ya que
  - en cada vuelo el número de reservas es incierto, no se sabe cuantas finalmente se van hacer
  - el número de reservas puede pensarse como un resultado de un experimento, esto es, se puede repetir una cantidad infinita de veces (cada vez que haya un vuelo) y tiene un conjunto bien definido de resultados (de cero reservas al óptimo definido por la aerolínea)

- Problema de información incompleta: el hecho de que la aerolínea entienda que el número de reservas es una variable aleatoria hace que el problema de estimar el número (óptimo) de reservas no sea trivial

Variables aleatorias y sus distribuciones de probabilidad

¿Se puede determinar la cantidad óptima de reservas que debe aceptar la aerolínea?

- Dada cierta información, por ejemplo la probabilidad con la que en cada vuelo se aceptan un número fijo de reservas (ej. 100), puede emplearse la probabilidad básica para encontrar la solución a este caso

- En otras palabras, se puede construir la distribución de probabilidad de la variable aleatoria número de reservas con la información pasada de esta variable

- Este procedimiento permite con información ex-ante pronosticar el posible número de reservas

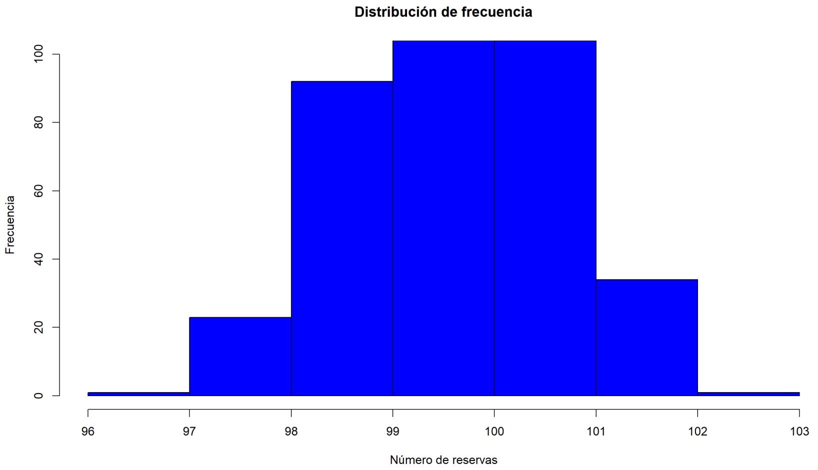
Variables aleatorias y sus distribuciones de probabilidad

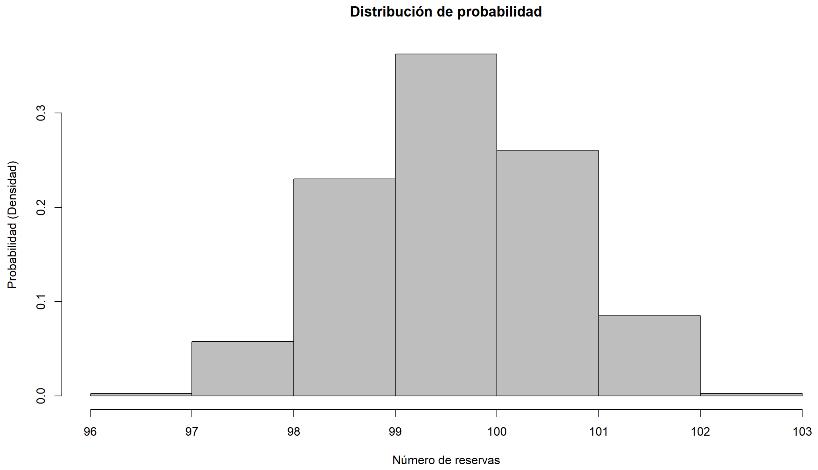
```r
library(summarytools)
x <- rnorm(400,100)
x <- round(x, digits = 0)
freq(x, headings=F)
```

```
              Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
----------- ------ --------- -------------- --------- --------------
         97      1      0.25           0.25      0.25           0.25
         98     23      5.75           6.00      5.75           6.00
         99     92     23.00          29.00     23.00          29.00
        100    145     36.25          65.25     36.25          65.25
        101    104     26.00          91.25     26.00          91.25
        102     34      8.50          99.75      8.50          99.75
        103      1      0.25         100.00      0.25         100.00
       <NA>      0                               0.00         100.00
      Total    400    100.00         100.00    100.00         100.00
```

```r
hist(x, breaks = c(96,97,98,99,100,101,102,103), xlab="Número de reservas"
     ylab="Frecuencia", ylim=c(0,100), main="Distribución de frecuencia",c
```

```r
hist(x, freq=F, breaks = c(96,97,98,99,100,101,102,103), xlab="Número de r
     ylab="Probabilidad (Densidad)", main="Distribución de probabilidad",c
```





Variables aleatorias y sus distribuciones de probabilidad

Otras variables aleatorias importantes

- Hogar: ingreso anual del hogar, tamaño del hogar, tasa de participación laboral del hogar
- País: número de pensionados, número de individuos bajo la línea de pobreza, precio del petróleo, precio del dólar, tasa de homicidio, tasa de desempleo
- Firma: costo de producción, insumos utilizados, número de empleados
- Otras menos importantes: cantidad de caras en 10 lanzamientos, la cantidad de veces que cae el 6 al tirar un dado, número de penaltis acertados cobrados por Messi

Tipos de variables aleatorias

Discretas

- Es una variable que sólo toma una cantidad finita de valores
- Ejemplos: variable aleatoria Bernoulli (o binaria, toma valores 0 o 1); estar desempleado o no; estar enfermo o no

```
x <- rbinom(100, 1, 0.75)
freq(x, h=F, cumul=F)
```

```
              Freq   % Valid   % Total
----------- ------ --------- ---------
          0     23     23.00     23.00
          1     77     77.00     77.00
       <NA>      0                0.00
      Total    100    100.00    100.00
```

```
hist(x, freq=F, xlim=c(0,1), ylim=c(0,8), xlab="Bernoulli", ylab="Probabilidad", main="", col="skyblue")
```

Tipos de variables aleatorias

Discretas

- Para describir por completo el comportamiento de una variable aleatoria discreta es necesario determinar la probabilidad que toma cada valor de la variable

- En el ejemplo de número de reservas óptimo por la aerolínea, el problema puede analizarse en el contexto de las variables aleatorias de Bernoulli de la siguiente forma: dada una reserva realizada por una persona, tomado aleatoriamente, se define una variable aleatoria de Bernoulli como $X = 1$ si la aerolínea ha tomado la reserva y $X = 0$ si no la ha tomado

- La probabilidad de que una reserva haya sido tomada por la aerolínea puede ser cualquier numero entre cero y uno. Si llamamos a esta probabilidad θ, se tienen las siguientes probabilidades:

$$P(X = 1) = \theta$$

$$P(X = 0) = 1 - \theta$$

- De manera general, cualquier va aleatoria discreta queda completamente explicada cuando se determinan las probabilidades de cada valor que toma:

$$p_j = P(X = x_j), j = 1, 2, \ldots, k$$

donde cada p_j está entre 0 y 1 y

$$p_1 + p_2 + \ldots + p_k = 1$$

Tipos de variables aleatorias

Discretas

- La función de densidad de probabilidad (fdp) de X resume la información concerniente a los valores que puede tomar X y a sus correspondientes probabilidades:

$$f(x_j) = p_j, j = 1, 2, \ldots, k$$

- Dada la fdp de cualquier variable aleatoria discreta, es fácil calcular la probabilidad de cualquier evento relacionado con esa variable aleatoria

- Por ejemplo, suponga que X es la cantidad de tiros libres anotados por Messi en dos intentos, entonces X puede tomar los valores $\{0, 1, 2\}$. Suponga que la fdp de X está definida por:

$$f(0) = .17; f(1) = .50; f(2) = .33$$

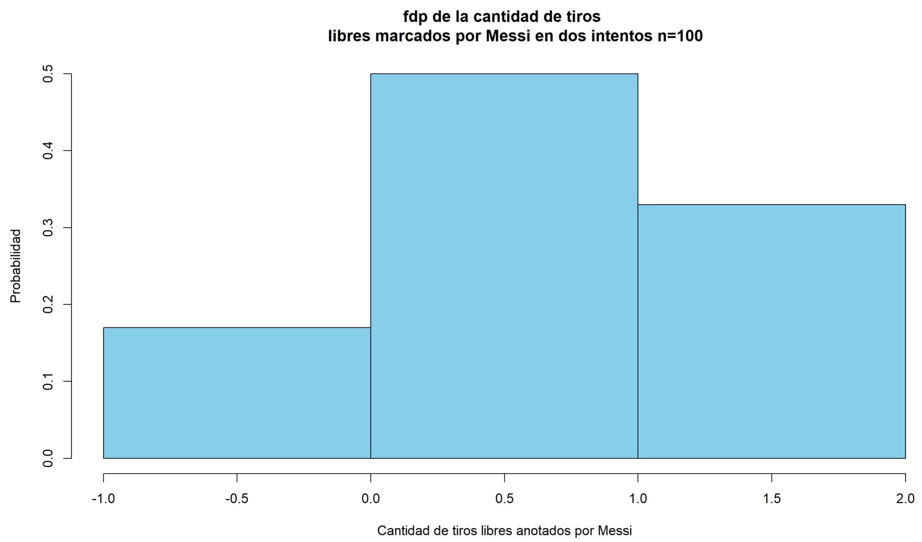
Tipos de variables aleatorias

Discretas

```
set.seed(123456789)
x <- rbinom(100, 2, 0.6)
freq(x, h=F, cumul=F)
```

```
hist(x, freq=F, breaks=c(-1,0,1,2), xlab="Cantidad de tiros libres anotado
    libres marcados por Messi en dos intentos n=100", col="skyblue")
```

```
              Freq   % Valid   % Total
----------- ------ --------- ---------
          0     17     17.00     17.00
          1     50     50.00     50.00
          2     33     33.00     33.00
       <NA>      0                0.00
      Total    100    100.00    100.00
```



- Estas tres probabilidades suman 1
- La probabilidad de que Messi anote por lo menos un tiro libre:

$$P(X \geq 1) = P(X = 1) + P(X = 2) = .50 + .33 = .83$$

Tipos de variables aleatorias

Continuas

- Es una variable en la cual la probabilidad de que esta tome un valor cualquiera es cero. Es decir, los valores que puede tomar la variable aleatoria son tantos que no es posible contarlos o hacerlos coincidir con los enteros positivos (es innumerable), por lo que la variable aleatoria puede tomar cada uno de estos valores con probabilidad cero

- Ejemplos: salarios, riqueza, tasa de cambio, precio del dólar, ingresos operacionales, millas recorridas al final de la carrera futbolística de Messi

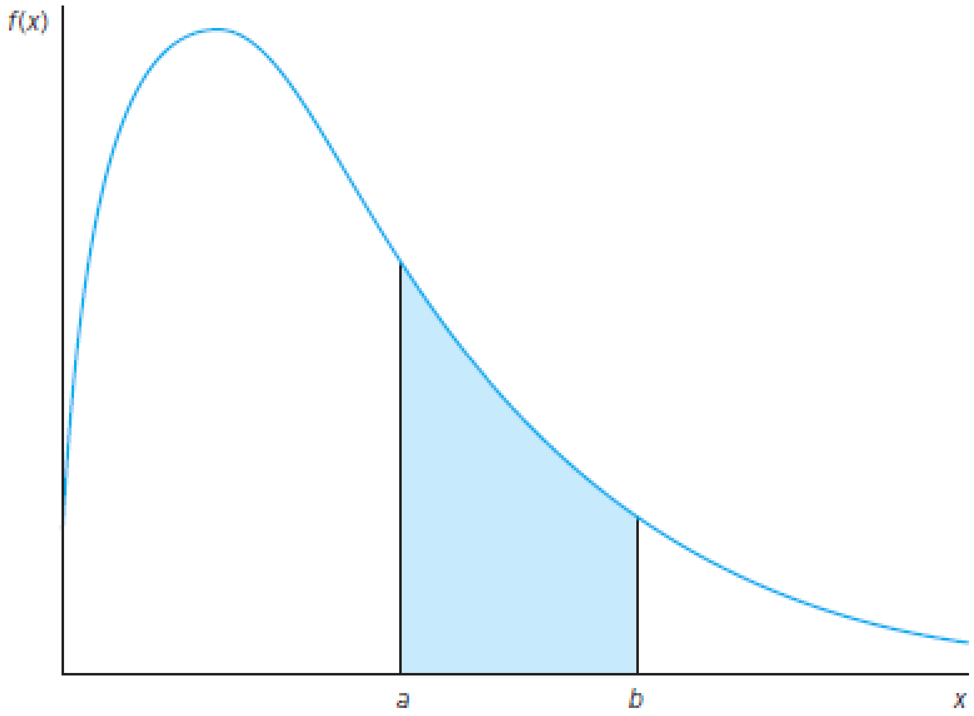
- También puede definirse una fdp y proporciona información sobre los posibles valores de esta variable aleatoria

- Sin embargo, dado que no tiene sentido analizar la posibilidad de que una variable aleatoria continua tome un determinado valor, la fdp de una variable de este tipo sólo se usa para calcular eventos que comprenden un rango de valores

Tipos de variables aleatorias

Continuas

Por ejemplo, si a y b son constantes y $a < b$, la probabilidad de que X se encuentre entre a y b, $P(a \leq X \leq b)$, es el área bajo la fdp entre los punto a y b



Esta probabilidad es la integral de la función f entre los punto a y b. Todo el área bajo la fdp siempre debe ser igual a 1

Tipos de variables aleatorias

Continuas

- Para calcular probabilidades de va continuas, es más fácil emplear la función de distribución acumulada (fda). La fda de se define como:

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

- En el caso de va discretas, la fda se obtiene sumando las fdp de todos los valores x_j tales que $x_j \leq x$

- En el caso de va continuas, $F(x)$ es el área bajo la fdp, f, a la izquierda del punto x

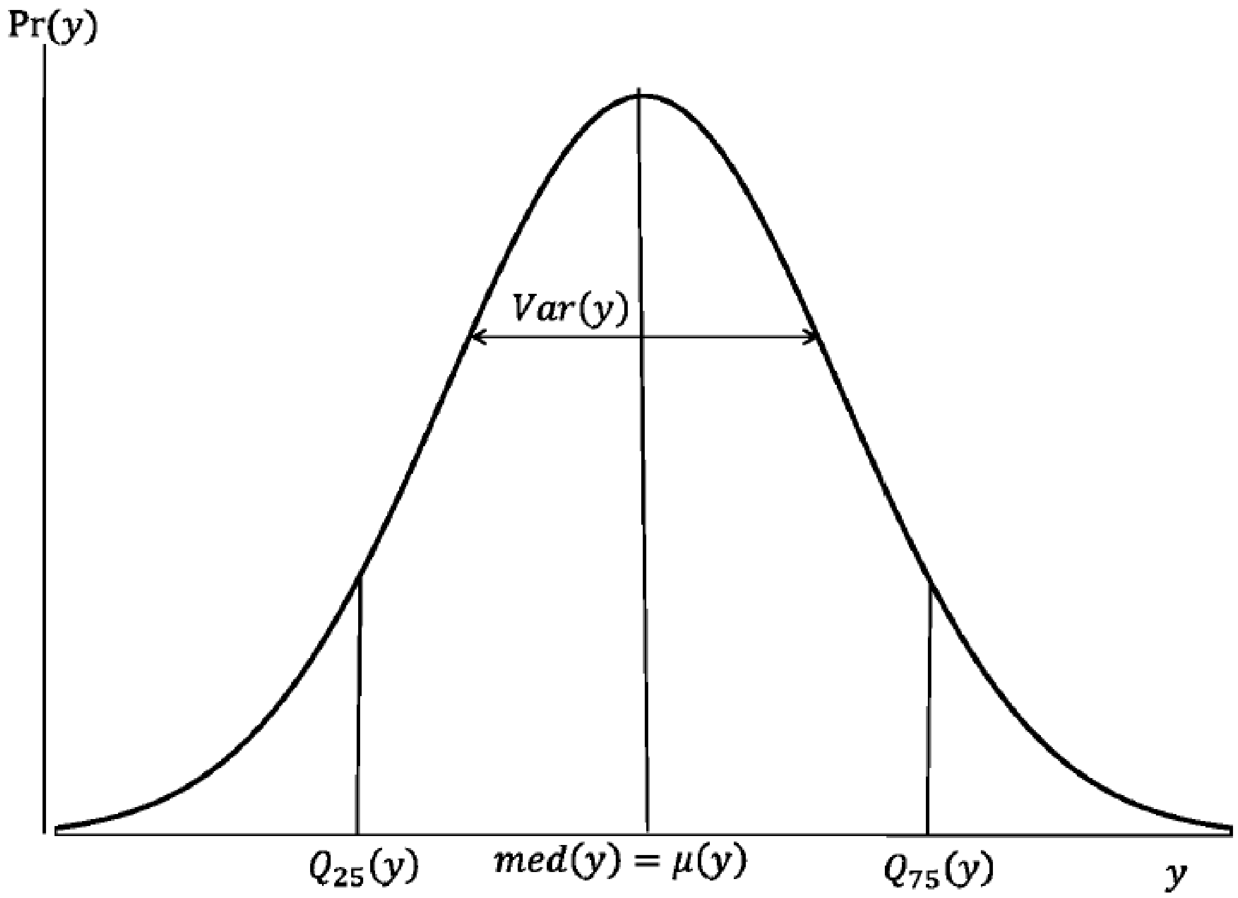
Propiedades

- Como $F(x)$ es una probabilidad, su valor estará siempre entre 0 y 1
- Si $x_1 < x_2$, entonces $P(X \leq x_1) \leq P(X \leq x_2)$, es decir que $F(x_1) \leq F(x_2) \Longrightarrow$ la fda es una función creciente o por lo menos no decreciente de x
- $P(X > x) = 1 - F(x)$
- $P(a < X \leq b) = F(b) - F(a)$, para todo par de números $a < b$

Tipos de variables aleatorias

Continuas

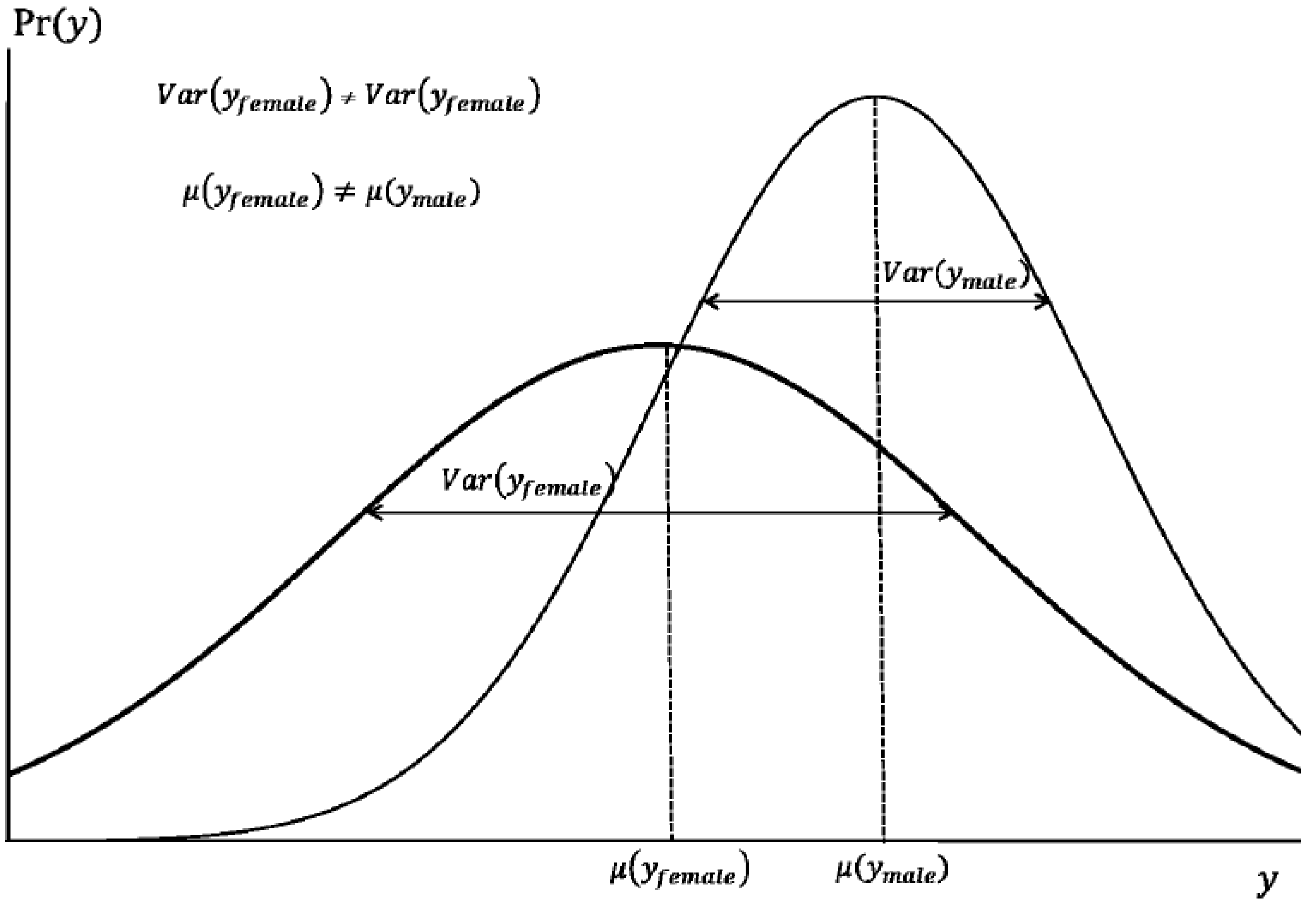
Ejemplo: función de densidad de los salarios y algunos estadísticos



Tipos de variables aleatorias

Continuas

Ejemplo: función de densidad de los salarios para hombres y mujeres



Características de las distribuciones de probabilidad

Medidas de tendencia central: el valor esperado

- El valor esperado o la esperanza de X (o media poblacional), que se denota por $E(X)$ o μ_X o simplemente μ, es un promedio ponderado de todos los posibles valores de X
- Valor esperado para va discretas:

$$E(X) = x_1 f(x_1) + x_2 f(x_2) + \ldots + x_k f(x_k) = \sum_{j=1}^{k} x_j f(x_j)$$

- Si X es una va continua, entonces $E(X)$ está definida como una integral:

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

- Si $g()$ es una función, entonces $g(X)$ es una variable aleatoria y su valor esperado será:

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

Características de las distribuciones de probabilidad

Medidas de tendencia central: el valor esperado

Propiedades

- Para toda constante c, $E(c) = c$

- Sea a y b dos constantes, $E(aX + b) = aE(X) + b$

- El valor esperado de una suma es la suma de valores esperados:

$$\begin{aligned} E(a_1X_1 + a_2X_2 + \ldots + a_nX_n) &= a_1E(X_1) + a_2E(X_2) + \ldots + a_nE(X_n) \\ &= \textstyle\sum_{i=1}^{n} a_iE(X_i) \end{aligned}$$

Medidas de tendencia central: la mediana

- La mediana de X, llámese m, es el valor tal que una mitad del área bajo la curva de la fdp queda a la izquierda de m y la otra mitad del área queda a la derecha de m

- Si X es una va discreta, la mediana se obtiene ordenando todos los posibles valores de X y seleccionando después el valor medio

Características de las distribuciones de probabilidad

Medidas de variabilidad: varianza

- La varianza es que tan lejos esta X de su valor esperado
- $Var(X) = \sigma^2 = E[(X - E(X))^2]$

Propiedades

- $Var(c) = 0$
- $Var(aX + b) = a^2 Var(X)$

Esto significa que sumar una constante a una variable aleatoria no modifica la varianza, pero multiplicar una variable aleatoria por una constante aumenta la varianza en un factor igual al cuadrado de la constante

Medidas de variabilidad: desviación estándar

- $sd(X) = \sqrt{Var(X)} = \sqrt{E[(X - E(X))^2]} = \sqrt{E(X^2) - E(X)^2}$

Propiedades

- $sd(c) = 0$

Características de las distribuciones de probabilidad

Estandarización de una variable aleatoria

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

Entonces

$$E(Z) = \frac{1}{\sigma}E(X) - \frac{\mu}{\sigma} = \frac{1}{\sigma}\mu - \frac{\mu}{\sigma} = 0$$

$$Var(Z) = \frac{1}{\sigma^2}Var(X) = \frac{1}{\sigma^2}\sigma^2 = 1$$

Otras características de la distribución de una variable aleatoria: el sesgo y la curtosis

- **El sesgo**: el tercer momento de la va Z

$$E(Z^3) = E[\frac{(X - \mu)^3}{\sigma^3}]$$

  El sesgo sirve para determinar si la distribución es simétrica

- **La curtosis**: el cuarto momento de la variable aleatoria Z

$$E(Z^4) = E[\frac{(X - \mu)^4}{\sigma^4}]$$

Valores mayores de la curtosis significan que las colas de la distribución de X son más gruesas

Medidas de asociación

Covarianza

$$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = E(XY) - \mu_X\mu_Y$$

- Es una medida de dependencia lineal entre dos variables aleatorias
- Si la covarianza es positiva indica que las dos variables se mueven en la misma dirección, mientras que si es negativa indica que las dos variables se mueven en dirección opuesta
- La interpretación de la magnitud es difícil
- Si X y Y son independientes entonces $Cov(X,Y) = 0$
- Limitación: la covarianza puede utilizarse para medir la relación entre dos variables, sin embargo esta medida depende de las unidades de medición, esto es, la covarianza se verá alterada si las variables están medidas en diferentes unidades de medición

Coeficiente de correlación

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X\sigma_Y} = \rho_{XY}$$

- El coeficiente de correlación supera la deficiencia de la covarianza
- La interpretación de la magnitud es más fácil
- $-1 \leq Corr(X,Y) \leq 1$