Unidad 1. Revisión de conceptos estadísticos básicos (Parte 1)

Erika R. Badillo
erika.badilloen@unaula.edu.co

Facultad de Economía
Universidad Autónoma Latinoamericana

En este tema

- Motivación
- Variables aleatorias y sus distribuciones de probabilidad
- Características de las distribuciones de probabilidad
- Distribución conjunta, condicionales e independencia

Lecturas

- Wooldridge, Jeffrey (2013). *Introducción a la econometría*. 5a edición, Cengage Learning. Apéndice B

- Gujarati, Damodar (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 reservaciones aceptar para un vuelo en el que hay 40 asientos disponibles. Existen diferentes situaciones que hacen el problema no trivial:
  - Si hay pocas personas reservando, todas serán aceptadas
  - Pero ¿qué pasa si hay muchas personas queriendo hacer una reservación?
    - si no existieran personas que a última hora cancelen su reserva, se aceptarían 40
    - pero si hay cancelaciones puede que algunos cupos no se llenen y esto ocasionaría pérdidas para la aerolínea
    - aceptar más de 40 esperando que hayan cancelaciones, pero corre el riesgo de tener que compensar a las personas que hicieron una reservación y no sean aceptadas para el vuelo

¿Se puede determinar la cantidad óptima de reservaciones que debe aceptar la compañia? Dada cierta información, puede utilizarse la probabilidad para encontrar una solución

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: número de personas que utilizan la reservación, es una variable aleatoria ya que, hasta que los datos son observados es incierto qué 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

- El número de personas que utilizan la reservación es una variable aleatoria ya que
  - cada dia el número de personas reservando es incierto, antes de un determinado vuelo no se sabe cuántos finalmente van utilizar la reserva
  - el número de personas que utilizan la reservación puede pensarse como resultado de un experimento, esto es, se puede repetir una cantidad infinita de veces (cada dia se presentan nuevas reservas) 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 el director de la aerolinea entienda que el número de reservaciones es una variable aleatoria, hace que el problema de estimar el número (óptimo) de reservaciones no sea trivial

Variables aleatorias y sus distribuciones de probabilidad

¿Se puede determinar la cantidad óptima de reservaciones que debe aceptar la compañía?

- Dada cierta información, por ejemplo la probabilidad con la que cada dia se acepta un número fijo de reservaciones (ej. 40), 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 personas reservando" con la información pasada de esta variable
- Este procedimiento permite con información ex-ante pronosticar el posible número de reservaciones aceptadas

Variables aleatorias y sus distribuciones de probabilidad

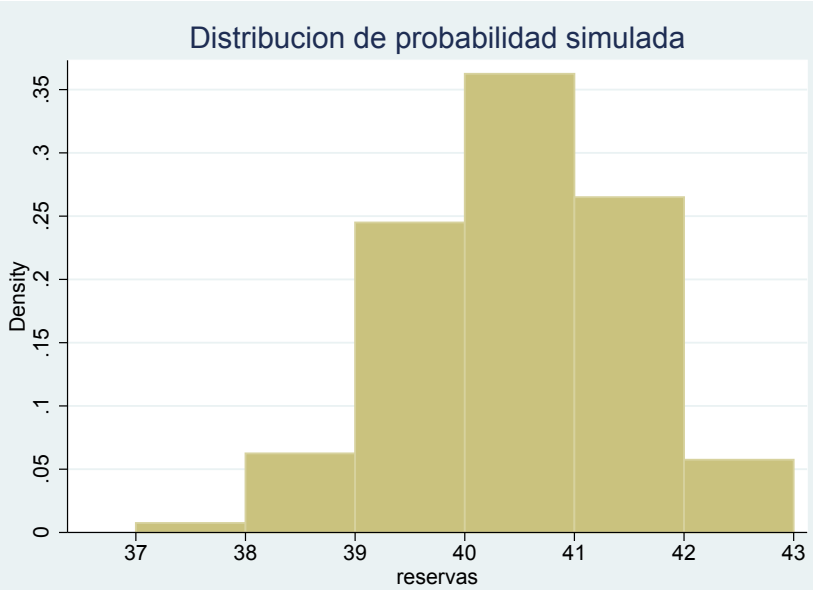
```
tab reservas
```

reservas	Freq.	Percent	Cum.
37	3	0.75	0.75
38	25	6.25	7.00
39	98	24.50	31.50
40	145	36.25	67.75
41	106	26.50	94.25
42	22	5.50	99.75
43	1	0.25	100.00
Total	400	100.00	

Simulando datos

```
set obs 400
gen x = rnormal(40)
gen reservas = round(x, 1)

hist reservas, w(1) ylabel(0 .05 .1 .15 .2 .25 .3 .35) ///
xlabel(37 38 39 40 41 42 43) ///
title("Distribución de probabilidad simulada")
```



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

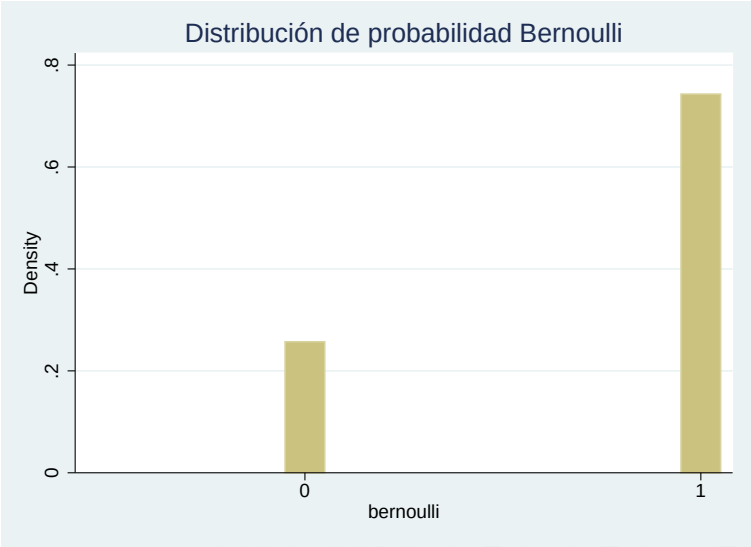
- Es una variable que sólo toma una cantidad finita de valores
- Ejemplos: variable aleatoria Bernoulli (o binaria, toma valores 0 o 1); estar desempleado o no; estar enfermo o no

```
. tab bernoulli

  bernoulli |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        250       25.00       25.00
          1 |        750       75.00      100.00
------------+-----------------------------------
      Total |      1,000      100.00
```

Simulando datos en Stata

```
set obs 1000
gen bernoulli=uniform()<=0.75
hist bernoulli, w(1) discrete xlabel(0 1) ///
barwidth(.1) title("Distribución de probabilidad Bernoulli")
```



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 reservaciones para un vuelo, el problema puede analizarse en el contexto de las variables aleatorias de Bernoulli de la siguiente forma: dado un cliente, tomado aleatoriamente, se define una variable aleatoria de Bernoulli como $X = 1$ si la persona utiliza su reservación y $X = 0$ si no la utiliza
- La probabilidad de que un determinado cliente utilice su reservación 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 variable aleatoria (VA) discreta queda completamente explicada cuando se determinan las probabilidades de cada valor que toma:

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

donde cada p_j está entre 0 y 1 y

$$p_1 + p_2 + ... + p_k = 1$$

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:

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

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

- 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:

$$f(0) = .20\ \ f(1) = .44\ \ f(2) = .36$$

Tipos de variables aleatorias

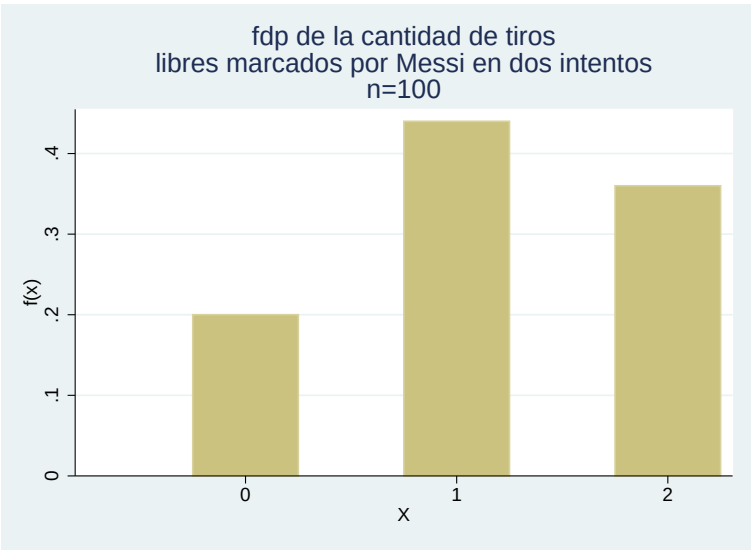
Discretas

Simulando datos

```
set obs 100
gen X=.
replace X = 0 if _n>=1 & _n<=20
replace X = 1 if _n>=21 & _n<=64
replace X = 2 if _n>=65 & _n<=100
hist X,  w(1) d xlabel(0 1 2) ytitle(f(x))  ///
barwidth(.5) title("fdp de la cantidad de ///
tiros" "libres marcados por Messi en dos intentos" "n=100")
```

```
. tab X
```

X	Freq.	Percent	Cum.
0	20	20.00	20.00
1	44	44.00	64.00
2	36	36.00	100.00
Total	100	100.00	



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

$$P(X \geq 1) = P(X = 1) + P(X = 2)$$
$$= .44 + .36$$
$$= .80$$

Tipos de variables aleatorias

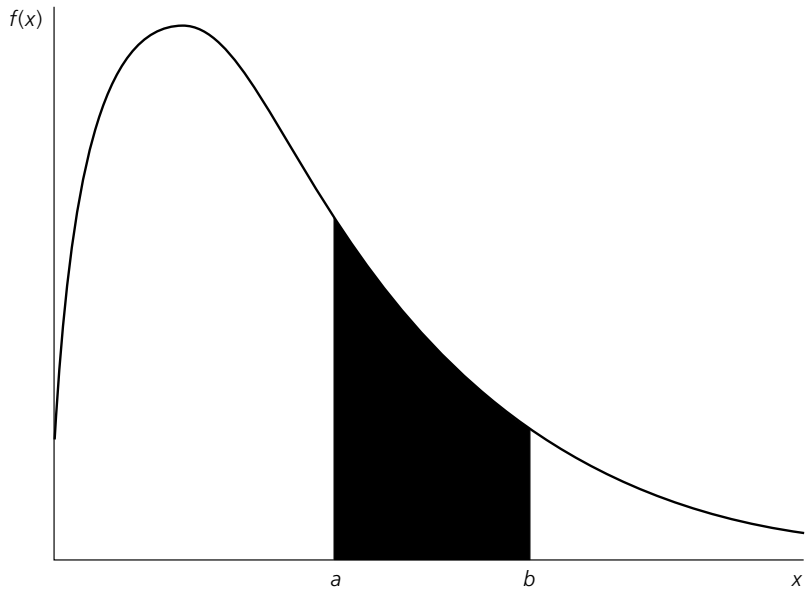
Continuas

- 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 VA son tantos que no es posible contarlos o hacerlos coincidir con los enteros positivos (es innumerable), por lo que la VA 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 VA
- Sin embargo, dado que no tiene sentido analizar la posibilidad de que una VA 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 puntos a y b



- Esta probabilidad es la integral de la función f entre los puntos 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 X 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

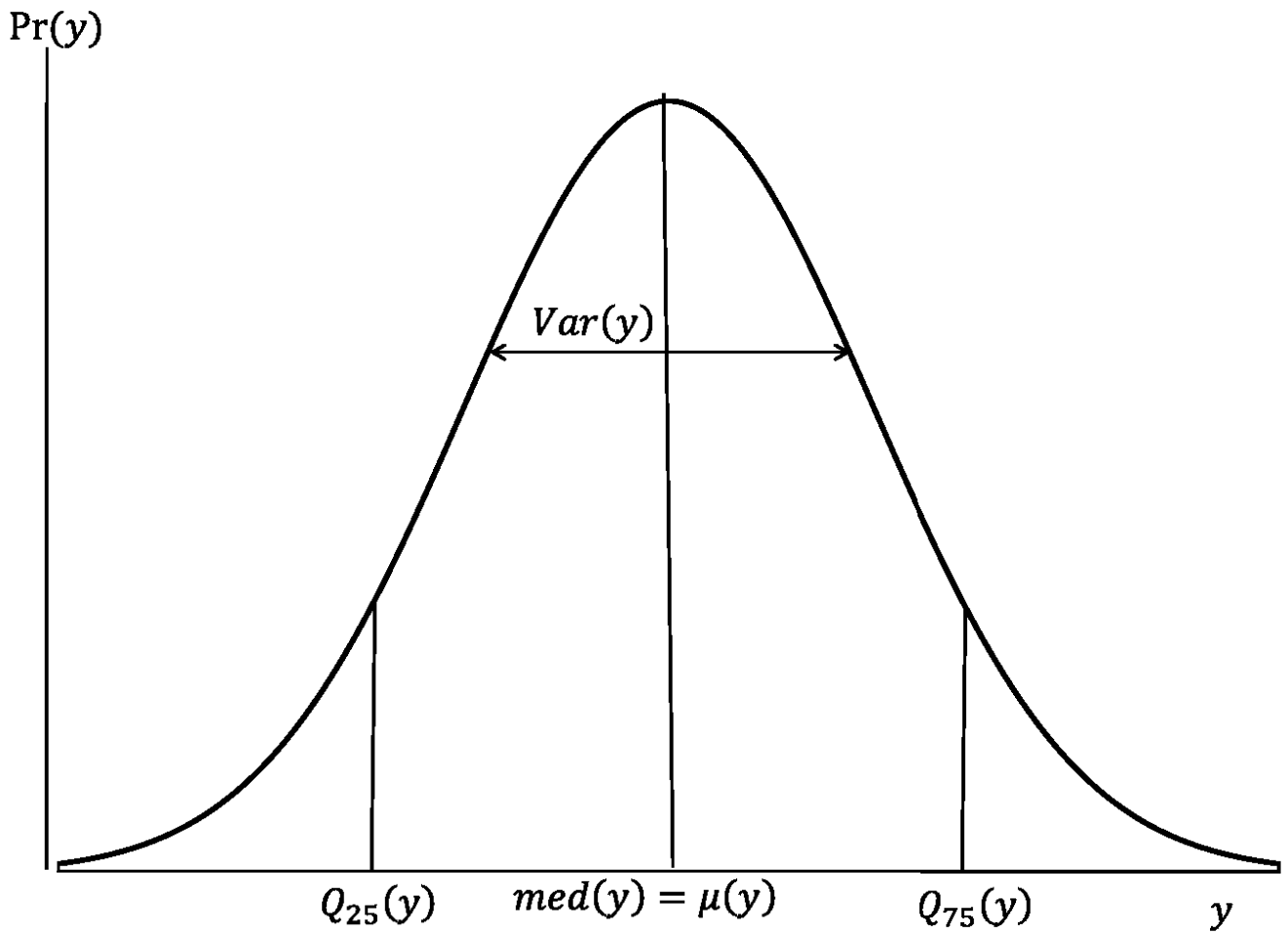
Propiedades

- 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

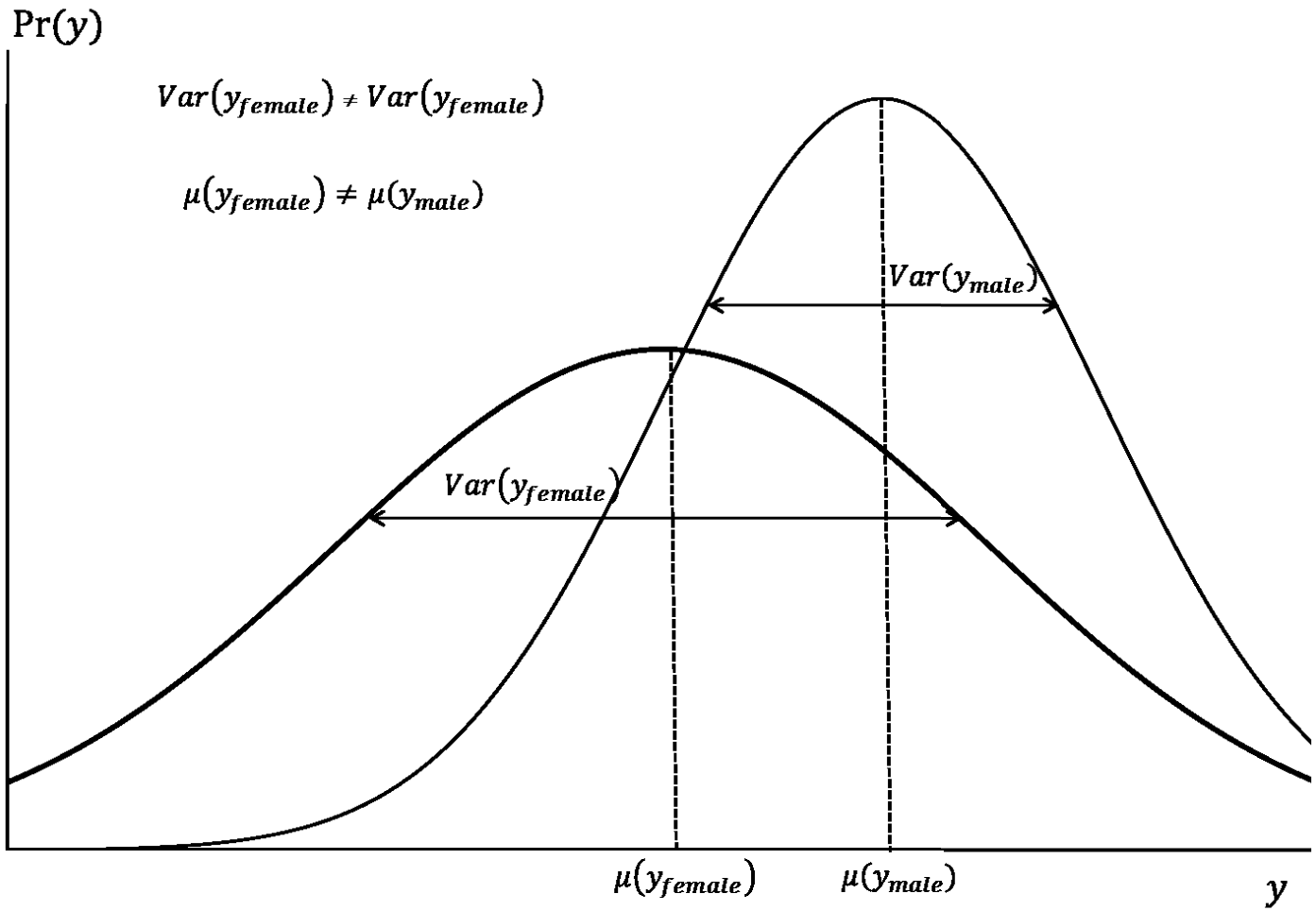
- Ejemplo: Función de densidad de los salarios y algunos estadísticos



Tipos de variables aleatorias

Continuas

- Ejemplo: Función de densidad de los salarios para hombres y mujeres



Características de las distribuciones de probabilidad

Las características de interés de las variables aleatorias se pueden agrupar en tres categorías: medidas de tendencia central, medidas de variabilidad o dispersión y medidas de la relación entre dos variables aleatorias

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:

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

- 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:

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

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:
$$E(a_1X_1 + a_2X_2 + ... + a_nX_n) = a_1E(X_1) + a_2E(X_2) + ... + a_nE(X_n) = \sum_{i=1}^{n} a_iE(X_i)$$

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

Medidas de variabilidad: varianza

- La varianza es una medida de qué tan lejos está 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 VA no modifica la varianza, pero multiplicar una VA 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

- $sd(c) = 0$
- $sd(aX + b) = |a| sd(X)$

Estandarización de una VA

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

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 VA: 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 VA 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

Distribución conjunta, condicionales e independencia

En economía, suele interesar la ocurrencia de eventos en los que participa más de una variable aleatoria. En particular, interesan:

- la probabilidad conjunta: probabilidad de que una persona que haga una reservación utilice su reservación y además sea un viajero de negocios
- la probabilidad condicional: dada la condición de que la persona es viajero de negocios, ¿qué probabilidad hay de que utilice su reservación?

Distribución de probabilidad conjunta

$$f_{X,Y}(x,y) = P(X = x, Y = y)$$

Independencia

Conocer el valor de X no modifica las probabilidades de los valores de Y (o viceversa). Las VA X y Y son independientes si y sólo si:

$$f_{X,Y}(x,y) = P(X = x, Y = y) = f_X(x) f_Y(y)$$

Donde f_X y f_Y son las funciones de densidad de probabilidad marginal

Distribución de probabilidad condicional

- En econometría, usualmente interesa saber cómo está relacionada una variable, con otra u otras variables
- Lo más que se puede saber acerca de cómo afecta X a Y está contenida en la distribución condicional de Y dada X. Esta información está resumida en la función de densidad de probabilidad condicional, definida por:

$$f_{Y|X}(y|x) = f_{X,Y}(x,y)/f_X(x)$$

- La interpretación de la anterior ecuación es más fácil cuando X y Y son dicretas:

$$f_{Y|X}(y|x) = P(Y = y|X = x)$$

- Una característica importante de las distribuciones condicionadas es que, si X y Y son variables aleatorias independientes, conocer el valor que toma X no dice nada acerca de la probabilidad de que Y tome diversos valores (y viceversa):

$$f_{Y|X}(y|x) = f_Y(y)$$
$$f_{X|Y}(x|y) = f_X(x)$$

Características de las distribuciones conjuntas y de las condicionales

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 $Cov(X,Y) = 0$
- $Cov(a_1X + b_1, a_2Y + b_2) = a_1a_2Cov(X,Y)$

Características de las distribuciones conjuntas y de las condicionales

Medidas de asociación

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

- La covarianza puede utilizarse para medir la relación entre dos variables, sin embargo, esta medida depende de las unidades de medición (se ve alterada)
- El coeficiente de correlación supera la deficiencia de la covarianza
- Es una medida de dependencia lineal, y su magnitud es más fácil de interpretar
- $-1 \leq Corr(X,Y) \leq 1$
- $Corr(a_1X + b_1, a_2Y + b_2) = Corr(X,Y)$, si $a_1 a_2 > 0$
- $Corr(a_1X + b_1, a_2Y + b_2) = -Corr(X,Y)$, si $a_1 a_2 < 0$

Características de las distribuciones conjuntas y de las condicionales

Esperanza condicional

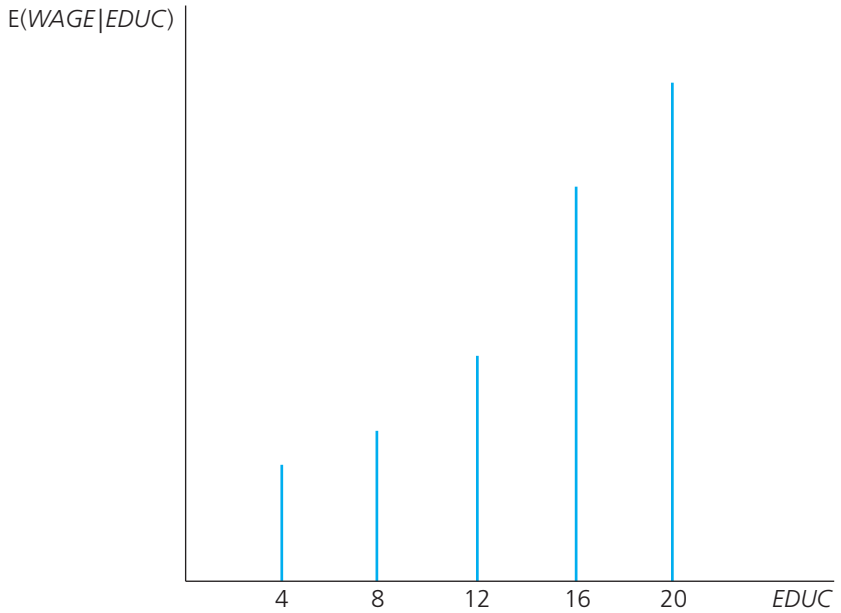
- La covarianza y la correlación miden la relación lineal entre dos variables aleatorias tratándolas simétricamente
- Sin embargo, en economía es frecuente que se desee explicar una variable Y, en términos de otra X. Por ejemplo, los salarios Y pueden ser explicados por los años de educación X
- Además, si Y se relaciona de forma no lineal con X, la covarianza y la correlación no capturarían este tipo de relaciones
- Se quiere ver cómo cambia la distribución de los salarios con base en el nivel de educación. Esta relación puede resumirse en la esperanza condicional de Y dado X, que también suele llamarse media condicional
- Este valor esperado se suele denotar por $E(Y|X = x) = E(Y|x)$ y la idea es calcular el valor esperado de Y dado que se conoce el valor de X
- Si Y es una variable aleatoria discreta, entonces:

$$E(Y|x) = \sum_{j=1}^{m} y_j f_{Y|X}(y_j|x)$$

Características de las distribuciones conjuntas y de las condicionales

Esperanza condicional

- Por ejemplo, sea (X, Y) la población de todos los individuos que trabajan, donde X es años de educación y Y es el salario por hora. Entonces
  - $E(Y|X = 12)$: es el salario promedio hora de la población que tiene 12 años de educación
  - $E(Y|X = 16)$: es el salario promedio hora de la población que tiene 16 años de educación



Características de las distribuciones conjuntas y de las condicionales

Esperanza condicional

En econometría, lo que se acostumbra hacer es dar funciones sencillas que expresan la relación entre las variables. Supongamos la siguiente relación lineal:

$$E(Salarios|Educ) = 1.05 + 0.45Educ$$

- El salario promedio de una persona que tenga 8 años de educación es $1.05 + 0.45(8) = \$4.65$
- El salario promedio de una persona que tenga 16 años de educación es $1.05 + 0.45(16) = \$8.25$
- El coeficiente de la variable $Educ$ implica que por cada año más de educación el salario esperado por hora aumenta en 0.45, es decir, 45 centavos