

Econometría I

Estructura de Datos



Carlos A. Yanes | Departamento de Economía | 2024-02-07





Preguntas de la sesión anterior?

Yo asumo que:

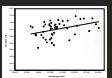
 Sabe consultar académicamente en youtube

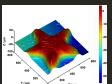
 tiene interés en R Markdown

 Sabe algo de estadística

Lo que aún no conoce:

 Interfaz del software o herramienta a usar

 Modelos de regresión

 Corregir problemas de estimación

Bases de datos



Bases de datos

Varias consideraciones pueden ser:

- Colección *específica* de datos.
- Formato "popular" de una tabla (matrices)
- Forma rectangular cuya organización aborda filas y columnas.
- Una **fila** tiene datos de una o varias variables para un mismo *individuo*.
- Una **columna** contiene valores de una *variable* para muchos individuos.

Bases de datos

Tipos de datos (Recordeis)

- **Índices:** Es la parte de nombres, números de identificación o cuestionario en una base de datos.
- **Binarios:** Variables que tienen sólo dos posibles respuestas. Ej: Si, no; Femenino, masculino, etc. Se codifican con *valores* de (0 y 1), y se les conoce como variable dummy.
- **De conteo:** Números enteros de no negación.
- **Continuos:** Aquellos que admiten decimales.
- **Nominales:** Respuestas no ordenadas y que amplían el espectro de las variables binarias, suelen ser datos categóricos.
- **Ordinales:** Admiten respuestas nominales pero en esencia *ordenadas* y son codificadas con números.

Estructuras de bases de datos

 Corte transversal

 Series de tiempo

 Panel longitudinal

Corte transversal



Datos de corte transversal

Individuos	Consumo	Ingreso	Carrera
Carlos	150	\$ 2'000.000	Economía
Fernando	160	\$ 3'500.000	Matemáticas
Luis	130	\$ 2'789.000	Economía
Mary	168	\$ 1'500.000	Economía
Paula	189	\$ 7'450.000	Psicología
...
...

Son *bases* donde la información obtenida varia por individuos (i) en un momento **determinado** de tiempo. Puede tratarse de un año o mes particular. Los datos se organizan por columnas asociando a cada uno de los individuos sus condiciones y/o características.

Por ejemplo: Recopilamos datos de varios **individuos** como $i \in I : \{Carlos, Fernando, \dots, Paula\}$ en un **periodo de tiempo** de un año, es decir, $T = 2023$. Regularmente se denotan las **variables** de forma (y_i, x_i) . La notación hace referencia a x_i como el valor de la variable x para la observación i , en un dataset con n observaciones, $i = 1, 2, 3, \dots, n$.

Datos de corte transversal

En  es posible hacerlo como:

```
# Si no importamos podemos entonces hacerlo a mano
individuos<- c("Carlos", "Fernando", "Luis", "Mary", "Paula")
Consumo<-c(150,160,130,168,189); Ingreso<-c(2000000,3500000,2789000,1500000,7450000); Carrera<-c("Economía", "Matemáticas")
datos_pr <- cbind(Consumo,Ingreso,Carrera)
rownames(datos_pr) <- individuos
datos_pr
```

```
#>           Consumo  Ingreso   Carrera
#> Carlos      "150"    "2e+06"  "Economía"
#> Fernando    "160"    "3500000" "Matemáticas"
#> Luis        "130"    "2789000" "Economía"
#> Mary         "168"    "1500000" "Economía"
#> Paula       "189"    "7450000" "Psicología"
```

Datos de corte transversal

Bases de datos con

Si le damos estructura

```
# Creamos un data frame y este se muestra:  
bd <- as.data.frame(datos_pr)  
bd$Consumo=as.numeric(bd$Consumo)  
bd$Ingreso=as.numeric(bd$Ingreso)  
bd
```

```
#>           Consumo Ingreso      Carrera  
#> Carlos       150 2000000    Economía  
#> Fernando     160 3500000 Matemáticas  
#> Luis          130 2789000    Economía  
#> Mary          168 1500000    Economía  
#> Paula         189 7450000 Psicología
```

Datos de corte transversal

Bases de datos con 

- Podemos crear rápidamente nuevas variables

```
bd$loging<-log(bd$Ingreso)
```

```
#>           Consumo Ingreso Carrera loging
#> Carlos      150 2000000 Economía 14.50866
#> Fernando    160 3500000 Matemáticas 15.06827
#> Luis        130 2789000 Economía 14.84119
#> Mary         168 1500000 Economía 14.22098
#> Paula       189 7450000 Psicología 15.82372
```

Datos de corte transversal

Bases de datos con

- Podemos crear rápidamente nuevas variables

```
bd$loging<-log(bd$Ingreso)
# Si no nos gusta el nombre
names(bd)[4]="Logaritmo_Ingreso"
```

	Consumo	Ingreso	Carrera	Logaritmo_Ingreso
#> Carlos	150	2000000	Economía	14.50866
#> Fernando	160	3500000	Matemáticas	15.06827
#> Luis	130	2789000	Economía	14.84119
#> Mary	168	1500000	Economía	14.22098
#> Paula	189	7450000	Psicología	15.82372

Series de Tiempo



Series de Tiempo

Fecha	PIB	I.P.C
2010	150 mill	101.2
2011	230 mill	104.6
2012	345 mill	110.4
2013	421 mill	115.3
2014	434 mill	118.2
...
...

Son *bases* donde la información viene por períodos de **tiempo** (t). Se usan mucho en economía, para análisis de tendencias y variaciones en el crecimiento

P.e: La notación hace uso del subíndice t en donde este hace referencia a la periodicidad ya sea en un día, mes, año, semestre, trimestre, específico. Las variables entonces se escriben de forma (y_t, x_t) . La parte de x_t es el valor de la variable x en el periodo t , en un dataset con T periodos, $t = 2018, 2019, 2020, \dots, T$.

Series de Tiempo

```
# Una serie de tiempo
Fecha<-c(2010,2011,2012,2013,2014); PIB<-c(150,230,345,421,434); I.P.C<-c(101.2,104.6,110.4,115.3,118.2)
datos_ts <- cbind(Fecha,PIB,I.P.C)
datos_ts

#>      Fecha PIB I.P.C
#> [1,] 2010 150 101.2
#> [2,] 2011 230 104.6
#> [3,] 2012 345 110.4
#> [4,] 2013 421 115.3
#> [5,] 2014 434 118.2
```

Series de Tiempo

- Si nos solicitan calcular inflación

```
datos_ts<- as.data.frame(datos_ts)
library(tidyverse)
datos_ts %>%
  mutate(inflacion=100*((I.P.C-lag(I.P.C))/lag(I.P.
```

```
#>   Fecha PIB I.P.C inflacion
#> 1 2010 150 101.2      NA
#> 2 2011 230 104.6  3.359684
#> 3 2012 345 110.4  5.544933
#> 4 2013 421 115.3  4.438406
#> 5 2014 434 118.2  2.515178
```

Panel de datos (longitudinal)



Panel de datos (longitudinal)

Individuo	Fecha	PIB	IPC
Colombia	2000	150 mill	101.2
Colombia	2001	230 mill	104.6
Ecuador	2001	347 mill	111.8
Ecuador	2001	347 mill	111.8
Ecuador	2002	358 mill	115.3
Perú	2000	434 mill	118.2
Perú	2001	452 mill	119.3
...
...

Son *bases* donde la información esta tanto por **individuos** (i) y por periodo de **tiempo** (t).
Son bases de datos mucho mas completas, eso si, un poco mas costosas, porque implica hacer seguimiento riguroso.

P.e: La notación hace uso tanto del subíndice t que sigue haciendo referencia a la periodicidad y se adiciona el subíndice i . Las variables entonces se escriben de forma (y_{it}, x_{it}) .

Modelos económicos y econométricos



Modelo Económico

Estos parten de una **función** como la siguiente:

$$\text{Salario} = f(\text{HorasT}, \text{Experiencia}, \text{CI}, \text{Edad})$$

Que significa que el salario *depende* (supuestamente) de variables como **horas de trabajo**, **Experiencia**, Coeficiente intelectual¹ (**CI**) y en efecto de la **Edad** de la persona.

[1] La psicología pone muchas veces esto en duda sin restarle -eso sí- importancia, aduce que hay otras variables mas allá de una prueba de inteligencia.

Modelos econométricos



Modelo Econométrico

Resuelven especificaciones y medidas de un modelo económico:

$$Salario_i = \beta_0 + \beta_1 HorasT_i + \beta_2 Experiencia_i + \beta_3 CI_i + \beta_4 Edad_i + \mu_i$$

Queremos estimar la relación entre las variables *explicativas* y la **dependiente**

$$Salario = f(\text{HorasT}, \text{Experiencia}, \text{CI}, \text{Edad})$$

Modelo Econométrico

Resuelven especificaciones y medidas de un modelo económico:

$$Salario_i = \beta_0 + \beta_1 HorasT_i + \beta_2 Experiencia_i + \beta_3 CI_i + \beta_4 Edad_i + \mu_i$$

Preguntas

- **P:** Como se interpreta β_1 ?
- **R:** Una hora adicional de trabajo correlacionado con el β_1 incrementa en una unidad monetaria el salario del individuo (controlando por Experiencia, coeficiente intelectual y Edad).
- **P:** Son los términos β_k parámetros poblacionales o estadísticos muestrales?
- **R:** Letras griegas significan **parámetros poblacionales**. Los estimados tienen gorro (*sombrerito*), p.e., $\hat{\beta}_k$

Modelo Econométrico

Resuelven especificaciones y medidas de un modelo económico:

$$Salario_i = \beta_0 + \beta_1 HorasT_i + \beta_2 Experiencia_i + \beta_3 CI_i + \beta_4 Edad_i + \mu_i$$

Preguntas

- **P:** Podemos interpretar el parámetro β_2 como causal?
- **R:** No, sobre todo si no hacemos hipótesis y/o supuestos sobre el proceso generador de datos.
- **P:** Qué es μ_i ?
- **R:** La desviación/perturbación aleatoria de un individuo con respecto a los parámetros de la población.

Modelo Econométrico

Resuelven especificaciones y medidas de un modelo económico:

$$Salario_i = \beta_0 + \beta_1 HorasT_i + \beta_2 Experiencia_i + \beta_3 CI_i + \beta_4 Edad_i + \mu_i$$

Preguntas

- **P:** ¿Qué supuestos imponemos al estimar con MCO?
- **R:**
 - La relación entre el salario y las variables explicativas es lineal en parámetros, y μ lo hace de forma aditiva.
 - Las variables explicativas son **exógenas**, p.e., $E[\mu|X] = 0$.
 - Tambien hay que asumir que :
 $E[\mu_i] = 0, E[\mu_i^2] = \sigma^2, E[\mu_i\mu_j] = 0$ para $i \neq j$.
 - Y (tal vez) μ_i se distribuye de forma normal.

Otro ejemplo

Curva de oferta de salario (Wooldridge, 2010). Suponga que la oferta del salario $wage^0$ esta expresado y dado por:

$$\log(wage^0) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 married + \epsilon$$

- Educ: Años de educación.
- Exper: Años de experiencia en el mercado laboral.
- Married: Variable de estado civil (marital)
- ϵ : Es el error aleatorio.
- β' s: Son los *Parámetros* del modelo.

Pregunta

La habilidad del trabajador sería un buen control en este modelo, pero ¿la observamos? ¿cómo se puede medir?

Hipótesis

Qué tan importante son?

Debe aprender lo **potente y flexible** que puede ser la regresión por mínimos cuadrados ordinarios (**MCO**).

Sin embargo, sus resultados requieren de **supuestos** y/o hipótesis

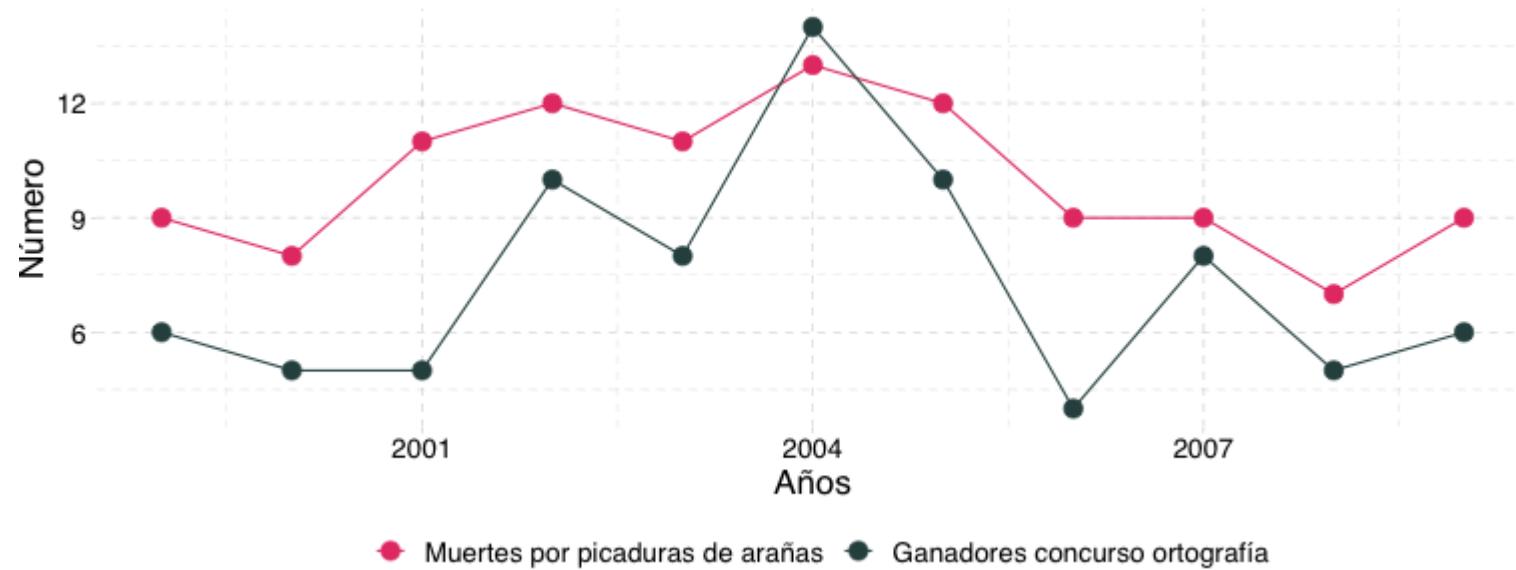
La vida real todo el tiempo viola estos supuestos.

Econometría I pregunta "**Qué ocurre cuando se violan los supuestos?**"

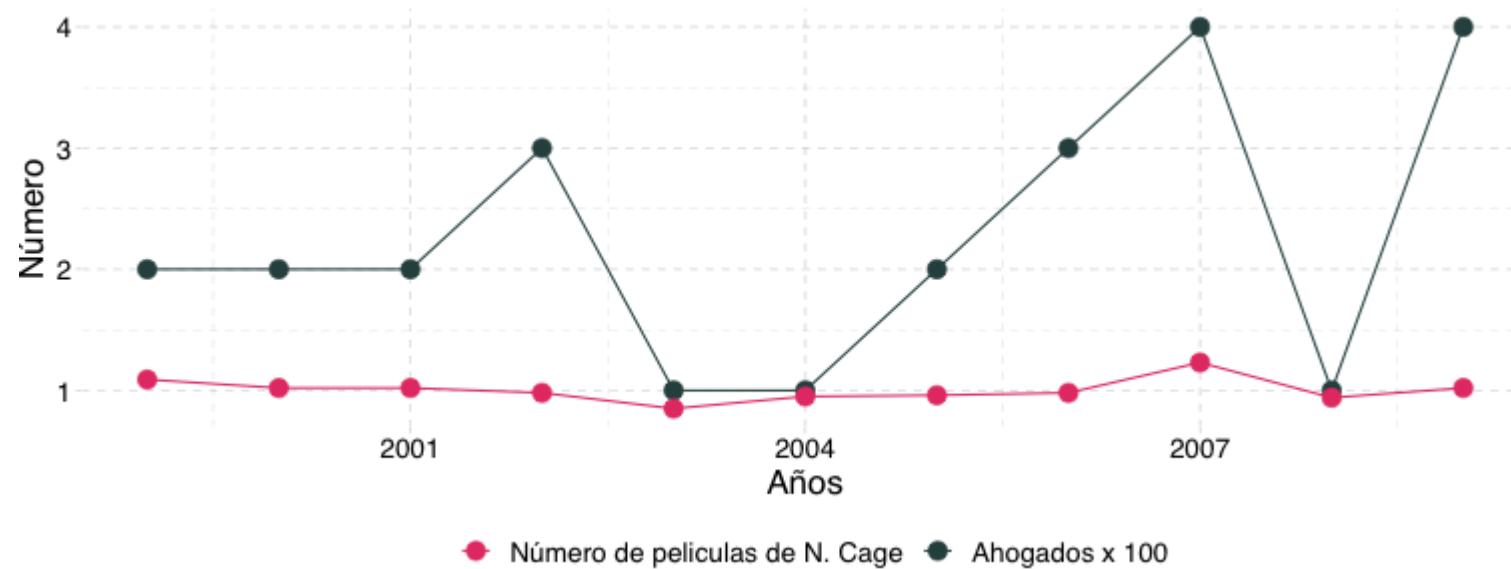
- Podemos arreglarlo? (Especialmente: cuando β es *causal*?)
- Qué ocurre si no los arreglamos o dejamos todo así?

El MCO sigue haciendo cosas increíbles, pero hay que saber cuándo hay que ser **prudente, confiado(a) o dudoso(a)**.

No todo es causal



Películas de Nicolas Cage



Econometría

Un econometrista aplicado [†] necesita un dominio sólido de (al menos) tres áreas:

1. La **teoría** subyacente a la econometría (supuestos, resultados, puntos fuertes, puntos débiles).
2. Cómo **aplicar los métodos teóricos** a los datos reales.
3. Métodos eficientes para **trabajar con datos**-limpiar, agregar, unir, visualizar.

Este curso tiene como objetivo profundizar en cada una de estas tres áreas.

- 1: Como antes.
- 2–3: R

[†]: *Econometrista aplicado* = Profesional de la econometría, e., analista, consultor, científico de datos.



Y que mas hay por aprender?

Básicos de R

- | | |
|-----------------------------------------------------|--------------------|
| 1. Todo es un objeto . | balon |
| 2. Todo objeto tiene nombre y valor . | balon <- 15 |
| 3. Puede usar funciones en esos objetos. | mean(balon) |
| 4. Estas estan en library (packages) | library(dplyr) |
| 5. R posee ayudas | ?dplyr |
| 6. R tiene avisos . | NA; error; warning |

Enfoque

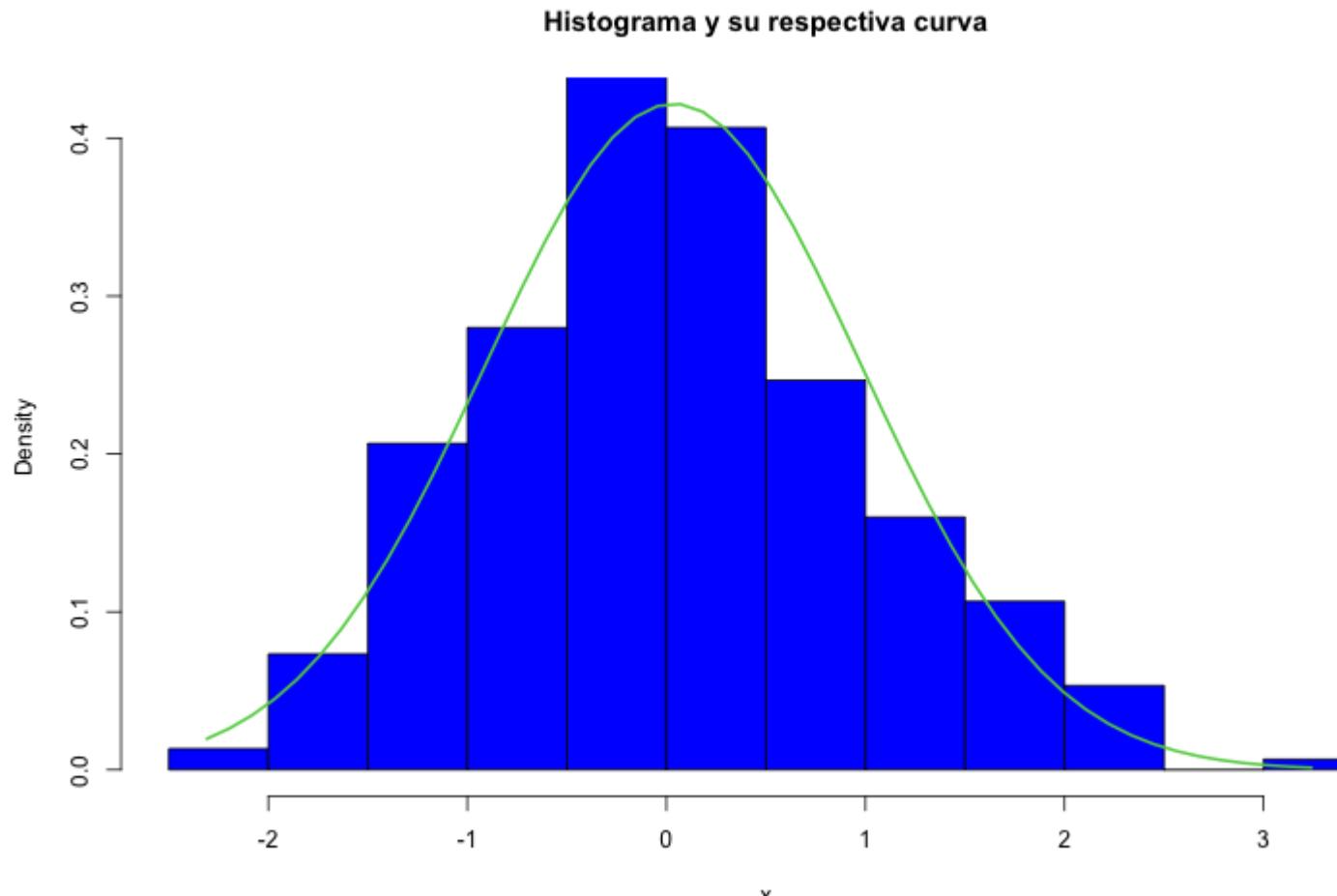


Enfoque

- Encontrar la mejor predicción de un valor real desconocido regularmente por un investigador a partir de datos recolectados (muestra) de una población.
- Medidas de tendencia central: **Media**, **varianza** y demás momentos de la *distribución*.
- Construcción de **intervalos de confianza** y proposición y/o planteamiento de **hipótesis**.

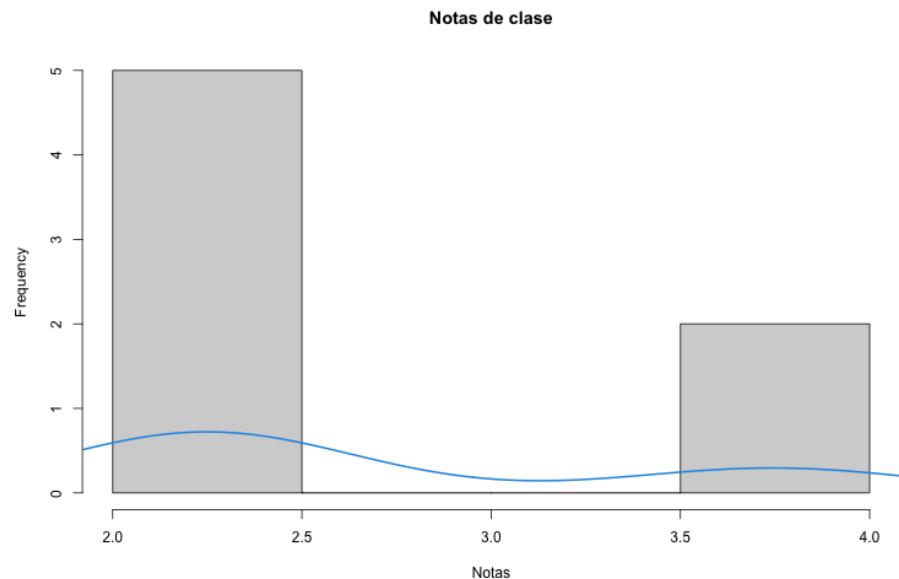
- | | |
|------------------------------------------------|------------------|
| 1. Media $\bar{x} = \sum x_i/n$ | mean(x) |
| 2. Varianza $\sigma^2 = \sum(x_i - \bar{x})^2$ | var(x) |
| 3. Paquete momentos | library(moments) |
| 4. Curtosis $K = 3.$ | kurtosis(x) |
| 5. Asimetría (+) o (-) | skewness(x) |

- Una función de distribución **normal**



```
# Instalar paquete
install.packages("moments")
library(moments)
Notas<-c(2.11,2.30,2.42,2.19,2.23,3.67,3.82)
mean(Notas)
sd(Notas)
kurtosis(Notas)
hist(Notas)
```

```
#> [1] 2.677143
#> [1] 0.7370146
#> [1] 1.912471
#> [1] 0.9053493
```



Bibliografía

- Gujarati, D. N., & Porter, D. C. (2011). *Econometria Básica*. Ed. Porto Alegre: AMGH..
- Stock, J. H., Watson, M. W., & Larrión, R. S. (2012). *Introducción a la Econometría*.
- Wooldridge, J. M. (2015). *Introductory econometrics: A modern approach*. Cengage learning.

Gracias por su atención!

Alguna pregunta adicional?

Carlos Andres Yanes Guerra

 cayanes@uninorte.edu.co

 keynes37