Medidas de Tendencia Central y Dispersión

Dr. Delfino Vargas Chanes

Facultad de Economía
Universidad Nacional Autónoma de México

15 de febrero de 2024

Índice

Media Muestral

Definición 1.1 (Media Muestral)

La media de una muestra de n respuestas medidas $x_1, x_2, ..., x_n$ es dada por

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

La media de la población correspondiente se denota como μ

- La media muestral es el promedio de las observaciones en la muestra
- No es común poder medir la media de una población, μ es una constante desconocida que estimamos con muestras de la población.
- La media muestral se ve afectada por valores extremos
- La media es la suma de todas las observaciones dividida entre el numero total de observaciones que participan en la suma.

Ejemplos de la Media

- Se tienen diez consultorios médicos y se registra el número de pacientes por consultorio. La media es:

$$\overline{x} = \frac{7+23+4+8+2+12+6+13+9+4}{10} = \frac{88}{10} = 0.88$$

- Excluyendo amistosos, Lionel Messi ha jugado 1,047 partidos, en los que ha metido 821 goles. La media de gol por partido de Messi es:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{821}{1,047} = 0.78$$

- Excluyendo amistosos, Cristiano Ronaldo ha jugado 1,204 partidos, en los que ha metido 873 goles. La media de gol por partido de CR7 es:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{873}{1,204} = 0.73$$

Media Ponderada

Definición 1.2 (Media Ponderada)

Sea una tupla de datos finita $(x_1, ..., x_n)$ y no vacía, con pesos asociados $(w_1, ..., w_n)$ y no negativos, definimos a la media ponderada como:

$$\overline{x} = \frac{\displaystyle\sum_{i=1}^{n} w_i x_i}{\displaystyle\sum_{i=1}^{n} w_i}$$

- La media ponderada es aquella que tiene asociados pesos.
- La media muestral es un caso particular de la media ponderada.

Ejemplos Media Ponderada

- Se desea estimar el número de productos vendidos en tres tipos de tiendas, repartidas en tres estratos. El peso probabilístico de cada entrato se indica por w_i Obtener la media y la media ponderada.

Tipo	Número de productos	w_i	Producto
Hipermercado	140	0.16	22.4
Supermercado	200	0.18	36.0
Mercado	163	0.21	34.2
		0.55	92.63

- **Media Muestral:** 167.67
- **Media Ponderada:** 168.42

Varianza

Definición 1.3 (Varianza)

La varianza de una muestra de resultados $x_1, \ldots, x_n$ es la suma del cuadrado de las diferencias entre los resultados y su media, dividido entre $n-1$. Es decir,

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

La varianza poblacional asociada se denota por σ^2

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n}{x_i}^2 - N\mu^2 = \frac{\sum_{i=1}^{n}{x_i}^2}{N} - \mu$$

- La varianza mide la magnitud de la dispersión con respecto a una medida de tendencia central, por ejemplo, la media.
- Solo sirve para datos cuantitativos en escala intervalar o de razón.

Desviación Estandar

Definición 1.4 (Desviación Estándar)

La desviación estándar de una muestra de respuestas es la raíz cuadrada positiva de la varianza, es decir

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}}$$

La desviación estándar poblacional asociada se denota como $\sigma = \sqrt{\sigma^2}$

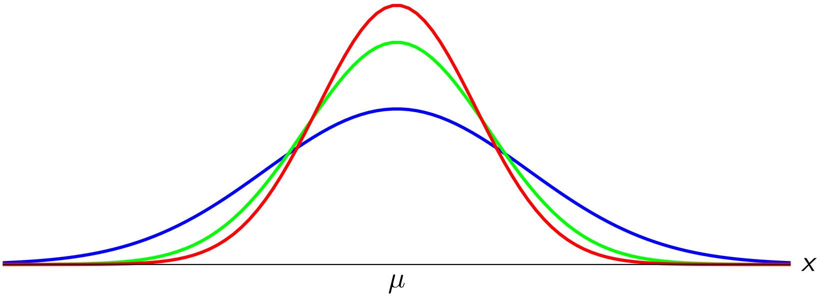
Ejemplo Varianza y Desviación Estándar en R

- Estaremos usando la libreria **Applied Econometrics with R**
- Cargamos una base de datos utilizada en un paper publicado en el JPE sobre la infidelidad (Fair 1978).
- Variable numérica affairs $\Rightarrow$ ¿Con qué frecuencia participó en relaciones sexuales extramatrimoniales durante el último año? $0 =$ ninguna, $1 =$ una vez, $2 =$ dos veces, $3 =$ 3 veces, $7 =$ de 4 a 10 veces, $12 =$ mensual, $12 =$ semanal, $12 =$ diariamente.
- ¿Que promedio y varianza tiene esta variable?

Ejemplo Varianza y Desviación Estándar en R

```
install.packages("AER")
library(AER)

data("Affairs")
range(Affairs$affairs)

mean(Affairs$affairs)
[1] 1.455907

var(Affairs$affairs)
[1] 10.8818

sd(Affairs$affairs)
[1] 3.298758
```

La Dispersión



- La figura muestra la comparación tres distribuciones simétricas con diferente grado de dispersión. Tenemos $X \sim \mathcal{N}(3, 1)$, $X \sim \mathcal{N}(3, 0.7)$, y $X \sim \mathcal{N}(3, 0.6)$, donde definimos a $X \sim \mathcal{N}(\mu, \sigma^2)$.
- La desviación estándar y la varianza son dos medidas que miden la dispersión de una distribución de datos.

Datos de Ejecutivos

- Ejemplo se tienen datos de 81 ejecutivos que se encuentran en el archivo "ejecutivos.dta", a los cuales se les miden la siguientes variables: Cuestionario, Caso, Planeación.
- Se desea obtener las Medidas de Tendencia central (media), mínimo, máximo. Medias de Dispersión: varianza, desviación estándar. Medias de Forma: Asímetría y Curtosis. (**TAREA: Reproducir en R**)

	Mín	Máx	Media	Desv.	Skew.	Curtosis
Cuest.	0	10	9.09	1.832	-2.396	7.019
Caso	1	10	6.99	2.519	-.329	-.781
Plan.	1	5	3.16	.798	.004	.776

Asimetría, Curtosis

Medidas de Forma

Buscan describir la función de distribución de frecuencias, indicando la simetría y curtosis.

Medidas de asimetría

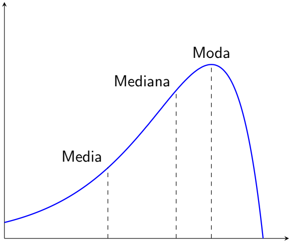
Miden la simetría de la distribución de datos.

Medidas de curtosis

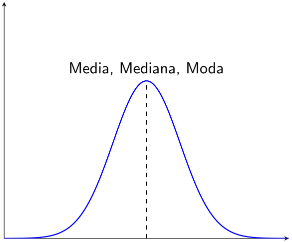
Miden la picudez de la distribución de datos.

- Si los datos tienen una distribución normal entonces cumplen que son simétricos con curtosis moderada (mesocurticos).

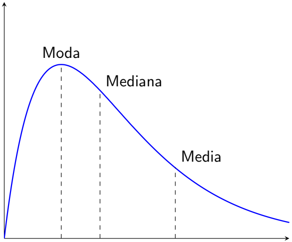
Asimetría (Skewness)



$S_k < 0$



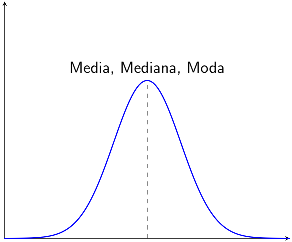
$S_k \approx 0$



$S_k > 0$

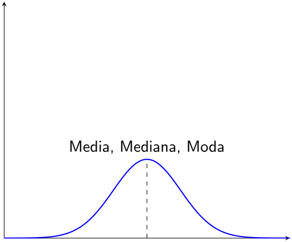
Curtosis

$K > 3$



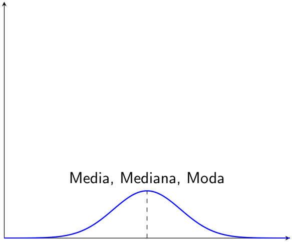
Distribución leptocúrtica

$K = 1$

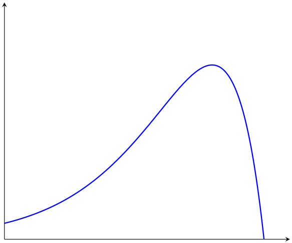


Distribución mesocúrtica

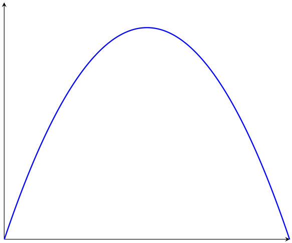
$K < 3$



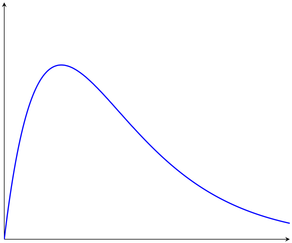
Distribución platocúrtica



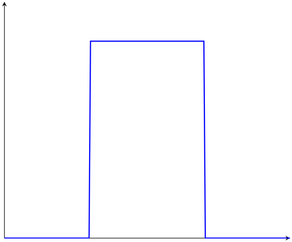
Sesgada a la izquierda



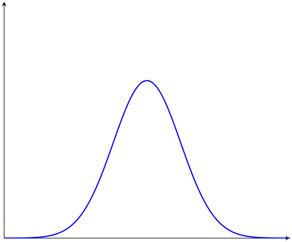
Simétrica



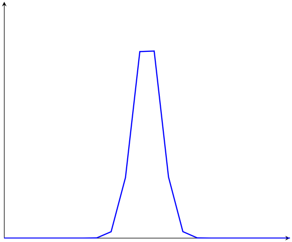
Sesgada a la derecha



Uniforme



Acampanada



Picuda

Sesgo y Curtosis

- La anterior gráfica nos enseña diferentes formas de curvas de densidad de probabilidad.
- En la primera fila, diferentes grados de simetría, y en la segunda, de picudez.
- Una curtosis alta es mayor que 1.0 o menor que 1.0
- Una asimetría alta es mayor que 1.0 o menor que -1.0
- La asimetría alta (en valor absoluto) y la curtosis alta (en valor absoluto), indican que las mediciones no tienen una distribución simétrica

Definición de Momentos

Definición 3.1

El k-ésimo momento central o centrado de una variable aleatoria x se define como

$$m_r = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^r$$

El primer momento central es 0, el segundo es la varianza (σ^2) donde σ es la desviación estándar. Podemos encontrar el tercer y cuarto momento en la definición de asimetría y de curtosis.

Curtosis:: *psych* en R

- Para R la paquetería psych devuelve la curtosis.

  Tipo 1: Definición típica utilizada en libros de texto antiguos:

$$g_2 = \frac{m_4}{m_2^2 - 3}$$

  Tipo 2: Utilizada en SAS, SPSS, STATA y EXCEL:

$$G_2 = \frac{[(n-1)g_2 + 6](n-1)}{(n-2)(n-3)}$$

  Tipo 3: Utilizado en MINITAB y BMDP:

$$b_2 = \frac{m_4}{s^4} - 3$$

Curtosis y Asimetría:: *moments* en R

- Para la paquetería *moments* solo hay una manera de calcular la asimetría, que es igual al **Tipo 1** para *psych*, en la que se usa la definición típica en los libros de texto, la formula es dada por:

$$g_2 = \frac{m_4}{m_2^2 - 3}$$

- La Curtosis para la paquetería *moments* calcula el estimador de la medida de curtosis de Pearson más 3. Quiere decir que nos da el valor exacto de la curtosis y no el exceso.

$$K_p = \frac{m_4}{m_4^2} + 3$$

¿Por qué existen diferencias entre *psych* y *moments*?

- A lo largo de los años, se han propuesto varias medidas de asimetría y curtosis para muestras, en donde la mayoría difieren en su error cuadrático medio y en el sesgo.
- Diversos autores han sugerido que la asimetría y la curtosis deben verse como "conceptos vagos" que pueden formalizarse de muchas maneras (Groeneveld 1998). Se han sugerido definiciones distintas.
- Entonces, de acuerdo con los autores de la paquetería psych (Joanes y Gill 1998) para muestras grandes, hay muy poco que elegir entre las distintas medidas ya que la diferencia entre los valores se vuelve insignificante. El problema radica cuando nos enfrentamos a muestras pequeñas en donde las diferencias pueden ser bastante sorprendentes.

Cálculo de Asimetría y Curtosis en R

```r
install.packages("psych")
install.packages("moments")
library(moments)
library(psych)

x<- c(6.3, 7.3, 7.3, 7.5, 7.8, 8, 8.1, 8.5, 8.6, 10)

#Usando la libreria psych

psych::kurtosi(x,na.rm = TRUE, type = 1)
psych::kurtosi(x,na.rm = TRUE, type = 2)
psych::kurtosi(x,na.rm = TRUE, type = 3)
describe(x,type=1)
describe(x,type=2)
describe(x,type=3)
```

Cálculo de Asimetría y Curtosis en R

```
install.packages("psych")
install.packages("moments")
library(moments)
library(psych)

x<- c(6.3, 7.3, 7.3, 7.5, 7.8, 8, 8.1, 8.5, 8.6, 10)

#Usando la libreria moments

moments::skewness(x)
moments::kurtosis(x)
```

Resultados (Comprobar con R)

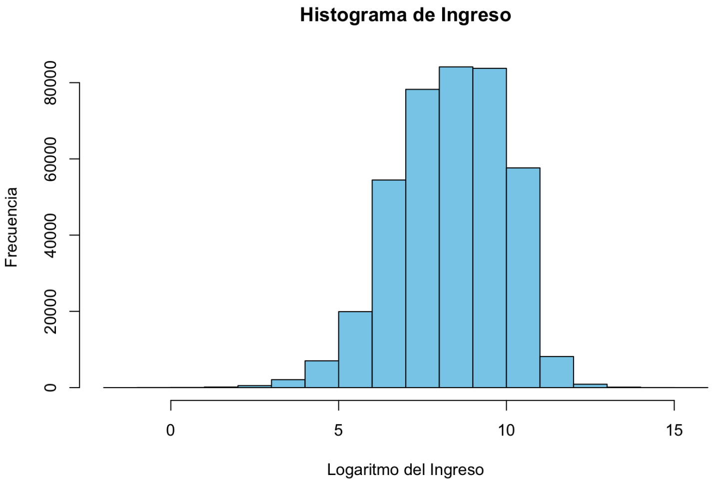
Tipo	Skewness	Curtosis
1	0.51	0.3880583
2	0.61	1.650317
3	0.44	-0.2556727
Moments	0.5119674	3.388058

- Estos resultados los podemos encontrar con otros programas estadísticos como Excel, STATA, SPSS, pero el enfoque práctico de este curso es el uso de RStudio.

Algunas consideraciones

- La mediana puede inclusive coincidir con los cuartiles o con los límites de los bigotes.
- Esto sucede cuando se concentran muchos datos en un mismo punto, en este caso, muchas personas tienen los mismos puntajes. Pudiera ser este un caso particular de una distribución sesgada o el caso de una distribución muy homogénea.
- En seguida se muestra el histograma de frecuencias correspondiente de los mismos datos que el diagrama de caja de la presentación pasada.

Histograma en R



Código en R

```r
library(ggplot2)
theme_set(theme_bw())

hist(log_ingreso,
     main = "Histograma de Ingreso",
     ylab = "Frecuencia",
     xlab = "Logaritmo del Ingreso",
     col = "skyblue",
     border = "black"
)

summary(log_ingreso)
quantile(log_ingreso)
describe(log_ingreso)
```

Bibliografía

Bhattacharyya, G. K., y Johnson, R. A. (1977). *Statistical concepts and methods*. Wiley.

Bickel, P. J., y Doksum, K. A. (2015). *Mathematical statistics: basic ideas and selected topics, volumes i-ii package*. CRC Press.

Fair, R. C. (1978). A theory of extramarital affairs. *Journal of political economy*, *86*(1), 45–61.

Groeneveld, R. A. (1998). A class of quantile measures for kurtosis. *The American Statistician*, *52*(4), 325–329.

Joanes, D. N., y Gill, C. A. (1998). Comparing measures of sample skewness and kurtosis. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *47*(1), 183–189.

Wackerly, D., Mendenhall, W., y Scheaffer, R. L. (2014). *Mathematical statistics with applications*. Cengage Learning.

¡Gracias por su atención!