eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Material de estudio

OCW 2019: *Curso práctico para el análisis e inferencia estadística con Mathematica*

Tema 2. Cálculo de estadísticos

Equipo docente del curso
Arrospide Zabala, Eneko
Martín Yagüe, Luis
Unzueta Inchaurbe, Aitziber
Soto Merino, Juan Carlos
Durana Apaolaza, Gaizka
Bikandi Irazabal, Iñaki

Departamento de Matemática Aplicada
Escuela de Ingeniería de Bilbao, Edificio II-I

OCW OpenCourseWare

TEMA 2. CÁLCULO DE ESTADÍSTICOS

Introducción

Definiciones

Se presentan una serie de magnitudes numéricas que permiten sintetizar los datos (valores de una variable estadística) contenidos en una serie estadística dada (muestra).

Estas magnitudes numéricas calculadas sobre los valores de una muestra se denominan estadísticos.

Las correspondientes magnitudes numéricas referidas a una población se denominan parámetros.

Un estimador es un estadístico utilizado para aproximar (estimar) un parámetro.

Tipos de estadísticos

Según las características de la muestra analizada, se tienen estadísticos:

- de tendencia central, que indican valores respecto a los cuales los datos parecen agruparse
- de dispersión, que señalan la concentración en conjunto de todos los datos de la distribución respecto de la medida o medidas de posición adoptadas
- de posición, aquellos valores de la variable que dividen a la distribución en partes que contienen el mismo número de individuos (para su cálculo los valores deben estar ordenados de menor a mayor)
- de forma, que estudian la simetría y el apuntamiento respecto de la distribución normal

Nota

En gran parte de los ejemplos de este tema se usan las siguientes listas, definidas en el *Tema 1*:

```
diam = {5, 5, 3, 4, 5, 6, 5, 7, 4, 5, 6};
```

```
long = {17, 17, 16, 15, 16, 16, 17, 18, 19, 19, 15};
```

Estadísticos de tendencia central

Definición

Son valores numéricos que resumen la serie. Representan un centro de la distribución en torno al cual se sitúa el conjunto de los datos.

Mediana

Es el valor de la variable que divide el conjunto de datos en dos partes iguales; es decir, es el valor que se encuentra en la mitad cuando los datos están ordenados de menor a mayor.

$$Me = \begin{cases} \text{término}\left(\frac{n+1}{2}\right) & \text{cuando } n \text{ es impar} \\ \frac{1}{2}\text{término}\left(\frac{n}{2}\right) + \text{término}\left[\left(\frac{n}{2}\right)+1\right] & \text{cuando } n \text{ es par} \end{cases}$$

- **Median[*list*]**. Devuelve la mediana de una *lista* indicada como argumento.

Ejemplo. Mediana de una lista con un número par de elementos.

```
listapar = Table[i, {i, 12}]
```

```
{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}
```

```
Median[listapar]
```

$$\frac{13}{2}$$

Ejemplo. Mediana de una lista con un número impar de elementos.

```
listaimpar = Table[i, {i, 13}]
```

```
{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13}
```

```
Median[listaimpar]
```

```
7
```

Ejemplo. Mediana de la lista **diam** de un ejemplo del tema 1.

```
diamord = Sort[diam] (* los valores deben estar ordenados de menor a mayor *)
```

```
{3, 4, 4, 5, 5, 5, 5, 5, 6, 6, 7}
```

```
medianadiam = Median[diamord]
```

```
5
```

Moda

Es el valor de la variable que más veces se repite, es decir, el valor que tiene mayor frecuencia absoluta en un conjunto de datos.

- **Commonest[*list*]**. Devuelve una lista con los elementos más comunes de la *lista* indicada como argumento.

Ejemplo. Moda de una lista. Serie bimodal.

```
lista1 = {a, b, c, a, b, c, a, b};
moda = Commonest[lista1]
```

```
{a, b}
```

Ejemplo. Moda de la lista **diam**.

```
Commonest[diam]
```

```
{5}
```

Media aritmética

Es el valor de la variable (estadístico) que se usa para indicar la tendencia central de un conjunto de datos calculada como la suma de los valores de la variable dividida por el número de observaciones:

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

- **Mean[*list*]**. Devuelve la media aritmética de una *lista* indicada como argumento.

Ejemplo. Media aritmética de una lista genérica.

```
lista2 = Table[xi, {i, 5}]
```

$\{x_1, x_2, x_3, x_4, x_5\}$

```
Mean[lista2]
```

$\frac{1}{5}(x_1 + x_2 + x_3 + x_4 + x_5)$

Ejemplo. Media aritmética de la lista **diam**.

```
Mean[diam]
```

5

Media geométrica

Es el valor que indica la tendencia central de un conjunto de datos usando el producto de sus valores:

$$\bar{x}_G = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

- **GeometricMean[*list*]**. Devuelve la media geométrica de una *lista* indicada como argumento.

Ejemplo. Media geométrica de una lista genérica.

```
lista2 = Table[xi, {i, 5}];
```

```
GeometricMean[lista2]
```

$(x_1\, x_2\, x_3\, x_4\, x_5)^{1/5}$

Ejemplo. Media geométrica de la lista **diam**.

```
GeometricMean[diam] // N
```

4.88498

Media armónica

Es el valor de la variable para la tendencia central de un conjunto de datos usando el inverso de la media aritmética de los inversos de dichos valores:

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

- **HarmonicMean[*list*]**. Devuelve la media armónica de una *lista* indicada como argumento.

Ejemplo. Media armónica de una lista genérica.

```
lista2 = Table[xi, {i, 5}];
```

```
HarmonicMean[lista2]
```

$$\frac{5}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \frac{1}{x_4} + \frac{1}{x_5}}$$

Ejemplo. Media armónica de la lista **diam**.

```
HarmonicMean[diam] // N
```

4.76289

Otras medias de interés

Media cuadrática.

- **RootMeanSquare[*list*]**. Devuelve la media cuadrática de una *lista* indicada como argumento.

Ejemplo. Media cuadrática de una lista genérica.

```
lista2 = Table[xi, {i, 5}];
```

```
RootMeanSquare[lista2]
```

$$\frac{\sqrt{x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2}}{\sqrt{5}}$$

Ejemplo. Media cuadrática de la lista **diam**.

```
RootMeanSquare[diam] // N
```

```
5.10793
```

Media contra-armónica.

- **ContraharmonicMean[*list*]**. Devuelve la media contra-armónica del argumento *lista*.

Ejemplo. Media contra-armónica de una lista genérica.

```
lista2 = Table[xi, {i, 5}];
```

```
ContraharmonicMean[lista2]
```

$$\frac{x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2}{x_1 + x_2 + x_3 + x_4 + x_5}$$

Ejemplo. Media contra-armónica de la lista **diam**.

```
ContraharmonicMean[diam] // N
```

```
5.21818
```

Estadísticos de dispersión

Definición

Son valores de la variable que indican la distancia de los datos respecto de un estadístico de posición central, generalmente, la media aritmética. Sirven como indicador de la variabilidad de los datos.

Rango

Se calcula como la diferencia entre la mayor observación, H, y la menor observación, L, del conjunto de datos. Indica la dispersión entre los valores extremos de una variable:

$$\boxed{R = \max(x_i) - \min(x_i) = H - L}$$

- **MinMax[*list*]**. Devuelve en una lista los valores mínimo (primer elemento) y máximo (segundo elemento) de la *lista* indicada como argumento.

Ejemplo. Rango de la lista **diam**.

```
r = MinMax[diam]; rango = r[[2]] - r[[1]]
```

```
4
```

Varianza y cuasivarianza

La varianza es un estadístico que indica el mayor o menor grado de dispersión de los valores de la muestra respecto de su media aritmética:

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

Como estimador de la varianza poblacional se utiliza la cuasivarianza muestral:

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

La relación entre ambas es:

$$S^2 = \frac{n}{n-1}\cdot s^2 \Rightarrow s^2 = \frac{n-1}{n}\cdot S^2$$

- **Variance[*list*]**. Devuelve la cuasivarianza de una *lista* indicada como argumento.

Ejemplo. Cuasivarianza de la lista **diam**.

```
Variance[diam] // N
```

```
1.2
```

Ejemplo. Varianza de la lista **diam**.

```
ndiam = Length[diam];
```

```
Variance[diam] * (ndiam - 1) / ndiam // N
```

```
1.09091
```

Desviación típica y cuasidesviación típica

La desviación típica es la raíz cuadrada de la varianza:

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

Como estimador de la desviación típica poblacional se utiliza la cuasidesviación típica muestral:

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

La relación entre ambas es:

$$S = \sqrt{\frac{n}{n-1}}\cdot s \Rightarrow s = \sqrt{\frac{n-1}{n}}\cdot S$$

- **StandardDeviation[*list*]**. Devuelve la cuasidesviación típica de una *lista* indicada como argumento.

Ejemplo. Cuasidesviación típica de la lista **diam**.

```
StandardDeviation[diam] // N
```

```
1.09545
```

```
Sqrt[Variance[diam]] // N
```

```
1.09545
```

Ejemplo. Desviación típica de la lista **diam**.

```
ndiam = Length[diam];
```

```
StandardDeviation[diam] * Sqrt[(ndiam - 1) / ndiam] // N
```

```
1.04447
```

Otras desviaciones de interés

Desviación media: media de las desviaciones absolutas de un conjunto de datos respecto de su media.

- **MeanDeviation[*list*]**. Devuelve la desviación media de los elementos de una *lista*.

$$D_{\bar{x}} = \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}|$$

Ejemplo. Desviación media de la lista **diam**.

```
MeanDeviation[diam] // N
```

```
0.727273
```

```
Total[Abs[diam - Mean[diam]]] / Length[diam] // N
```

```
0.727273
```

Desviación mediana: mediana de las desviaciones absolutas de un conjunto de datos respecto de su mediana (no queda afectado por los datos extremos).

- **MedianDeviation[*list*]**. Devuelve la desviación mediana de los elementos de una *lista*.

$$D_{Me} = Me\{|x_i - Me|\}$$

Ejemplo. Desviación mediana de la lista **diam**.

```
MedianDeviation[diam] // N
```

```
1.
```

```
Median[Sort[Abs[diam - medianadiam]]]
```

```
1
```

Coeficiente de variación de Pearson

Es el cociente entre la desviación típica y la media aritmética de un conjunto de datos:

$$CV = \frac{s_x}{\bar{x}}$$

Resulta interesante para comparar la dispersión de variables diferentes.

Ejemplo. Comparar la dispersión de los diámetros (lista **diam**) y las longitudes (lista **long**) de una muestra de tornillos.

```
sdiam = Sqrt[(Length[diam] - 1) / Length[diam]] StandardDeviation[diam] // N
```

```
1.04447
```

```
CVdiam = sdiam / Mean[diam]
```

```
0.208893
```

```
slong = Sqrt[(Length[long] - 1) / Length[long]] StandardDeviation[long] // N
```

```
1.33609
```

```
CVlong = slong / Mean[long]
```

```
0.0794429
```

```
(* resulta más dispersa la serie de los diámetros *)
```

Estadísticos de posición

Definición

Son valores numéricos que dividen a la distribución en un cierto número de partes que contienen el mismo porcentaje, p, de valores de la variable.

Cuantiles

Un cuantil de orden q $(0 < q < 1)$, C_q, es el valor de la variable por debajo del cual se encuentra una frecuencia relativa acumulada q; es decir, una proporción q de valores de la variable es menor que C_q.

- **Quantile[*list*,*q*,{{*a*,*b*},{*c*,*d*}}]**. Devuelve el q-ésimo cuantil de una *lista*. *Mathematica* presenta nueve métodos para el cálculo de los cuantiles para los cuales los parámetros a, b, c y d toman los siguientes valores:

{{0,0},{1,0}}	inverse empirical CDF (default)
{{0,0},{0,1}}	linear interpolation (California method)
{{1/2,0},{0,0}}	element numbered closest to q n
{{1/2,0},{0,1}}	linear interpolation (hydrologist method)
{{0,1},{0,1}}	mean-based estimate (Weibull method)
{{1,−1},{0,1}}	mode-based estimate
{{1/3,1/3},{0,1}}	median-based estimate
{{3/8,1/4},{0,1}}	normal distribution estimate

Para obtener el q-ésimo cuantil se puede proceder de la siguiente manera:

1) ordenar los datos de menor a mayor

2) siendo n el número de elementos se calcula el valor $n \cdot q$

2.1) si no es un número entero entonces el q-ésimo cuantil es el valor de la variables que se encuentra en la posición correspondiente al siguiente entero mayor que $n \cdot q$

2.2) si es un número entero el q-ésimo cuantil es el valor medio de los elementos ordenados situados en las posiciones $n \cdot q$ y $n \cdot q+1$

Para calcular con *Mathematica* el q-ésimo cuantil con el procedimiento indicado se deben combinar el primer y el cuarto método considerando, además, $q \in \{0, 1, 2, \cdots, 100\}$. Así, el código para el cálculo es:

```
If[IntegerQ[q / 100 * Length[lista]],
 Quantile[lista, q / 100, {{1 / 2, 0}, {0, 1}}],
 Quantile[lista, q / 100, {{0, 0}, {1, 0}}]]
```

Ejemplo. Cálculo de $C_{0.59}$ de una lista que contiene los cien primeros números enteros.

```
lista100 = Table[i, {i, 100}];

q = 59;

If[IntegerQ[q / 100 * Length[lista100]],
  Quantile[lista100, q / 100, {{1 / 2, 0}, {0, 1}}],
  Quantile[lista100, q / 100, {{0, 0}, {1, 0}}]] // N
```

```
59.5
```

Ejemplo. Cálculo de $C_{0.07}$ y $C_{0.77}$ de la lista **diam**.

```
Sort[diam]
```

```
{3, 4, 4, 5, 5, 5, 5, 5, 6, 6, 7}
```

```
q1 = 7; If[IntegerQ[q1 / 100 * Length[diam]],
          Quantile[diam, q1 / 100, {{1 / 2, 0}, {0, 1}}],
          Quantile[diam, q1 / 100, {{0, 0}, {1, 0}}]]
```

```
3
```

```
q2 = 77; If[IntegerQ[q2 / 100 * Length[diam]],
          Quantile[diam, q2 / 100, {{1 / 2, 0}, {0, 1}}],
          Quantile[diam, q2 / 100, {{0, 0}, {1, 0}}]]
```

```
6
```

Cuartiles

Los cuartiles son tres valores de la variable (Q_1, Q_2 y Q_3) que dividen el conjunto de datos en cuatro partes iguales (tres divisiones).

- **Quartiles[*list*]**. Devuelve una lista con los tres cuartiles de la *lista* indicada como argumento.

Ejemplo. Cálculo de los cuartiles de una lista que contiene los cien primeros números enteros.

```
lista100 = Table[i, {i, 100}];

Quartiles[lista100] // N
```

```
{25.5, 50.5, 75.5}
```

Ejemplo. Cálculo de los cuartiles de la lista **diam**.

```
Quartiles[diam] // N
```

```
{4.25, 5., 5.75}
```

Momentos

Definición

Los momentos de una variable estadística son los valores esperados de ciertas funciones definidas para la variable. Se pueden utilizar para determinar el modelo de distribución de probabilidad de la variable.

Los momentos muestrales son valores numéricos que caracterizan una muestra aleatoria de una variable. Se pueden utilizar para determinar el modelo de distribución de probabilidad de la variable.

Respecto al origen

Siendo n el número de elementos de una muestra $\{x_i\}$ de valores de una variable estadística, se puede definir el r-ésimo ($r \in \mathbb{N}$) momento de la variable respecto del origen como:

$$\alpha_r = \frac{1}{n} \sum_{i=1}^{n} (x_i)^r$$

- **Moment[*list*,*r*]**. Devuelve el r-ésimo momento respecto del origen de una *lista*.

El primer momento respecto del origen de los elementos de la lista coincide su media aritmética.

Ejemplo. Comprobación, para una lista genérica, de que la media aritmética es el primer momento respecto del origen.

```
lista2 = Table[x_i, {i, 5}];
```

```
Moment[lista2, 1] == Mean[lista2]
```

```
True
```

Ejemplo. Comprobación, para la lista **diam**, de que la media aritmética es el primer momento respecto del origen.

```
Moment[diam, 1] == Mean[diam]
```

```
True
```

Respecto a la media

Siendo n el número de elementos de una muestra $\{x_i\}$ de valores de una variable estadística, se puede definir el r-ésimo ($r \in \mathbb{N}$) momento de la variable respecto de la media (o momento central) como:

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

- **CentralMoment[*list*,*r*]**. Devuelve el r-ésimo momento respecto de la media de una *lista*.

El segundo momento central de los elementos de una serie coincide con su varianza.

Ejemplo. Comprobación, para la lista **diam**, de que la varianza es el segundo momento respecto de la media.

```
CentralMoment[diam, 2] == Variance[diam] * (Length[diam] - 1) / Length[diam]
```

```
True
```

Los momentos centrales tercero y cuarto de los elementos de una serie están relacionados con el estudio de la forma de la distribución: simetría y curtosis, respectivamente.

Estadísticos de forma

Definición

Son los valores numéricos que informan sobre la manera en que se distribuyen los datos en cuanto a la simetría y el apuntamiento.

Permiten identificar y describir la forma en que los valores de la serie tienden a agruparse de acuerdo con su frecuencia. Se compara la forma de la representación gráfica de una distribución con una distribución normal estándar en la que los datos se reparten en igual medida a ambos lados de la media.

Asimetría o sesgo

Una distribución de datos es simétrica cuando se reparten de igual forma a ambos lados de la media.

El coeficiente de asimetría de Fisher permite valorar la asimetría (sesgo) de una distribución de datos:

$$\gamma_1 = \frac{\mu_3}{s_x^3}$$

- **Skewness[*list*]**. Devuelve el coeficiente de asimetría de los elementos de la *lista*. Equivalente a:

$$\gamma_1 = \frac{\text{CentralMoment[list,3]}}{\text{CentralMoment[list,2]}^{3/2}}$$

En función de los valores del coeficiente, se tiene:

1) $\gamma_1 > 0$: asimetría a la derecha (las frecuencias descienden más lentamente por la derecha que por la izquierda)

2) $\gamma_1 = 0$: distribución simétrica

3) $\gamma_1 < 0$: asimetría a la izquierda

Ejemplo. Obtención del coeficiente de asimetría de la lista **diam**.

```
Skewness[diam] (* distribución simétrica *)
```

```
0
```

Ejemplo. Obtención del coeficiente de asimetría de la lista **long**.

```
Skewness[long] // N (* distribución asimétrica a la derecha *)
```

```
0.334537
```

Curtosis o apuntamiento

La curtosis es una medida para analizar el grado de concentración que presentan los datos de una distribución alrededor de la zona central.

Un coeficiente de curtosis es el cuarto momento con respecto a la media que se define como:

$$\beta_2 = \frac{\mu_4}{s_x^4}$$

El coeficiente de curtosis de Fisher permite valorar la simetría de una distribución de datos:

$$\gamma_2 = \frac{\mu_4}{s_x^4} - 3$$

Este coeficiente se aplica a distribuciones unimodales simétricas o con una asimetría moderada.

Se compara el apuntamiento de la distribución con el de la distribución normal estándar planteando la diferencia entre los momentos centrales de cuarto orden donde, para la distribución normal, $\beta_2 = 3$.

- **Kurtosis[*list*]**. Devuelve el coeficiente de curtosis de los elementos de la *lista*. Equivalente a:

$$\beta_2 = \frac{\text{CentralMoment[list,4]}}{\text{CentralMoment[list,2]}^2}$$

En función de los valores del coeficiente, se tiene:

1) $\beta_2 > 3$: distribución leptocúrtica (más apuntada que la normal estándar)

2) $\beta_2 = 3$: distribución mesocúrtica (igual apuntamiento que la normal estándar)

3) $\beta_2 < 3$: distribución platicúrtica (más aplanada que la normal estándar)

Ejemplo. Obtención del coeficiente de curtosis de la lista **diam**.

```
Kurtosis[diam] // N (* distribución platicúrtica *)
```

```
2.75
```

Ejemplo. Obtención del coeficiente de curtosis de la lista **long**.

```
Kurtosis[long] // N (* distribución platicúrtica *)
```

```
2.01055
```

Tipificación

Definición

La tipificación o estandarización es el proceso consistente en normalizar una variable estadística.

Consta de dos pasos:

1) centrar la variable, restando la media a cada uno de sus valores
2) reducir la variable, dividiendo cada uno de sus valores por la desviación típica

Una variable tipificada (centrada y reducida) tiene una media nula y su desviación típica es la unidad.

Unidades tipificadas

Sea una variable X cuyos valores $\{x_i\}$ proceden de una muestra (o una población) con media $\overline{x}$ (o μ) y desviación típica s (o σ).

El valor de la variable en unidades tipificadas (*z-score*), representado por z, se define como:

$$\boxed{z_i = \frac{x_i - \overline{x}}{s}} \qquad \boxed{z_i = \frac{x_i - \mu}{\sigma}}$$

- **Standardize[*list*]**. Desplaza y reescala los elementos de la *lista* para que su media sea cero y su cuasi-desviación típica sea 1.

El *z-score* indica el número de desviaciones típicas en que un valor dado, x_i, se sitúa por encima o por debajo de la media de su muestra ó población:

1) $z_i > 0$, la observación es mayor que la media
2) $z_i < 0$, la observación es menor que la media

Ejemplo. Tipificación de la lista **diam**.

```
{Mean[diam], StandardDeviation[diam]}
```

$$\left\{5, \sqrt{\frac{6}{5}}\right\}$$

```
zdiam = Standardize[diam]
```

$$\left\{0, 0, -\sqrt{\frac{10}{3}}, -\sqrt{\frac{5}{6}}, 0, \sqrt{\frac{5}{6}}, 0, \sqrt{\frac{10}{3}}, -\sqrt{\frac{5}{6}}, 0, \sqrt{\frac{5}{6}}\right\}$$

```
{Mean[zdiam], StandardDeviation[zdiam]}
```

```
{0, 1}
```

Utilidad

La tipificación permite:

- comparar valores de distribuciones distintas
- comparar valores de variables con unidades distintas
- detectar valores atípicos, aquellos tales que $z_i \notin [-3, 3]$, según la regla empírica