Estadística descriptiva con datos cuantitativos

Ramon Ceballos

30/1/2021

1. Medidas de dispersión

Las **medidas de dispersión** evalúan lo dispersos que están los datos. Algunas de las más importantes son:

- El **rango** o **recorrido**, que es la diferencia entre el máximo y el mínimo de las observaciones.
- El **rango intercuartílico**, que es la diferencia entre el tercer y primer cuartil, $Q_{0.75} - Q_{0.25}$.
- La **varianza** (varianza poblacional), a la que denotaremos por s^2, es la media aritmética de las diferencias al cuadrado entre los datos x_i y la media aritmética de las observaciones, $\bar{x}$.

$$s^2 = \frac{\sum_{j=1}^{n}(x_j - \bar{x})^2}{n} = \frac{\sum_{j=1}^{k} n_j(X_j - \bar{x})^2}{n} = \sum_{j=1}^{k} f_j(X_j - \bar{x})^2$$

- La **desviación típica** (desviación típica poblacional) es la raíz cuadrada positiva de la varianza, $s = \sqrt{s^2}$. Permite volver a las unidades originales de las observaciones.
- La **varianza muestral** es la corrección de la varianza. La denotamos por $\tilde{s}^2$ y se corresponde con:

$$\tilde{s}^2 = \frac{n}{n-1}s^2 = \frac{\sum_{j=1}^{n}(x_i - \bar{x})^2}{n-1}$$

- La **desviación típica muestral**, que es la raíz cuadrada positiva de la varianza muestral, $\tilde{s} = \sqrt{\tilde{s}^2}$.

1.1 Propiedades de la varianza

- $s^2 \geq 0$. Esto se debe a que, por definición, es una suma de cuadrados de números reales.
- $s^2 = 0 \Longrightarrow x_j - \bar{x} = 0 \ \forall j = 1, \ldots, n$. En consecuencia, si $s^2 = 0$, entonces todos los datos son iguales.
- $s^2 = \frac{\sum_{j=1}^{n} x_j^2}{n} - \bar{x}^2$. Es decir, la varianza es la media de los cuadrados de los datos menos el cuadrado de la media aritmética de estos.

1.2 Varianza vs Varianza muestral

La diferencia entre ambas definiciones viene por la interrelación entre la estadística descriptiva y la inferencial.

Por un lado, es normal medir cómo varían los datos cuantitativos mediante su varianza definida como la media aritmética de las distancias al cuadrado de los datos a su valor medio. No obstante, por otro lado, el conjunto de nuestras observaciones, por lo normal, será una muestra de una población mucho mayor y nos interesará estimar entre otras muchas cosas su variabilidad.

La varianza de una muestra suele dar valores más pequeños que la varianza de la población, mientras que la varianza muestral tiende a dar valores alrededor de la varianza de la población.

Esta corrección, para el caso de una muestra grande no es notable. Dividir n entre $n-1$ en el caso de n ser grande no significa una gran diferencia y aún menos si tenemos en cuenta que lo que tratamos es de estimar la varianza de la población, no de calcularla de forma exacta.

En cambio, si la muestra es relativamente pequeña (digamos $n < 30$), entonces la varianza muestral de la muestra aproxima significativamente mejor la varianza de la población que la varianza.

La diferencia entre desviación típica y desviación típica muestral es análoga.

Con `R`, calcularemos la varianza y la desviación típica **muestrales**. Con lo cual, si queremos calcular las que no son muestrales, tendremos que multiplicarlas por $\frac{n-1}{n}$, donde n es el tamaño de la muestra. Lo veremos a continuación.

Nótese que tanto la varianza como la desviación típica dan una información equivalente. Entonces, es comprensible preguntarse por qué se definen ambas medidas si con una basta. Pues bien, las unidades de la varianza (metros, litros, años...), ya sea muestral o no, están al cuadrado, mientras que las de la desviación típica no.

2. Medidas de dispersión con R

La siguiente tabla recoge las instrucciones para obtener las principales medidas de dispersión en R.

Medida de dispersión	Instrucción
Valores mínimo y máximo	`range(x)`
Rango	`diff(range(x))`
Rango intercuartílico	`IQR(x, type = ...)`
Varianza muestral	`var(x)`
Desviación típica muestral	`sd(x)`
Varianza	`var(x)*(length(x)-1)/length(x)`
Desviación típica	`sd(x)*sqrt((length(x)-1)/length(x))`

Ejemplo 1

Cálculo de las medidas de dispersión para la variable cuantitativa de muestra dados2.

```
#Defino una semilla
set.seed(0)

#creo la variable dado2
dados2 = sample(1:6,15, replace = TRUE)
dados2
```

```
##  [1] 6 1 4 1 2 5 3 6 2 3 3 1 5 5 2
```

```
#anulo la semilla
set.seed(NULL)
```

```
#Esta es la variable de muestra
dados2
```

```
##  [1] 6 1 4 1 2 5 3 6 2 3 3 1 5 5 2
```

```
#Calculo el rango de la variable dados2
diff(range(dados2))
```

```
## [1] 5
```

```
#Calculo el rango intercuartílico de la variable dados2
IQR(dados2)
```

```
## [1] 3
```

```
#Calculo la varianza muestral de la variable dados2
var(dados2)
```

```
## [1] 3.209524
```

```
#Calculo la desviación típica muestral de la variable dados2
sd(dados2)
```

```
## [1] 1.791514
```

```
n = length(dados2)
#Calculo la varianza de la variable dados2
var(dados2)*(n-1)/n
```

```
## [1] 2.995556
```

```
#Calculo la desviación típica de la variable dados2
sd(dados2)*sqrt((n-1)/n)
```

```
## [1] 1.730767
```