Medidas de dispersion

Victor Lopez

2023-01-22

Medidas de dispersion

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, 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 es la raíz cuadrada positiva de la varianza, $s = \sqrt{s^2}$.

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, $\tilde{s} = \sqrt{\tilde{s}^2}$

Diferencias entre las medidas de dispersion muestrales y poblacionales

La varianza de una muestra suele dar valores más pequños 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 las dos es que la desviación estándar muestral es una estimación del valor real de la desviación estándar poblacional, y se suele utilizar cuando no se tiene información completa sobre la población.

Si te das cuenta, la medidas muestrales son muchos eficientes, y mas cuando se trata de estimar o aproximar la variabilidad, debido a que con una poblacion muy grande no hay mucha diferencia y al mismo tiempo con una muestra es muy eficiente.

Tambien debido, a que si estamos utilizando calculos con unidades de medida, la desviacion tipica es mucho mas eficiente, ya que no nos devuelve las unidades elevadas al cuadrado

Por eso en R, por defecto se calculan las medidas muestrales

Medidas de dispersion en R

```r
dados = sample(c(1:6), 100, replace = TRUE)

# Rango
diff(range(dados)) # range devuelve un vector con el min y max
```

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

```r
# Rango interquartilico
IQR(dados)
```

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

```r
# Varianza muestral
var(dados)
```

```
## [1] 2.764747
```

```r
# Varianza poblacional o verdadera
n = length(dados)
var(dados) * (n - 1) / n
```

```
## [1] 2.7371
```

```r
# Desviacion tipica muestral
sd(dados)
```

```
## [1] 1.662753
```

```r
# Desviacion tipica poblacional o verdadera
n = length(dados)
sd(dados) * sqrt((n - 1) / n)
```

```
## [1] 1.654418
```