How can the sample and population variance be calculated in R?

Authored by

stats writer

April 22, 2024

In R, the sample and population variance can be calculated using the "var" function. This function takes a vector or a data frame as input and calculates the variance for the given data. The sample variance is calculated using the formula (n-1)/n * sum((x-mean)^2), where n is the sample size and x is the data values. On the other hand, the population variance is calculated using the formula 1/n * sum((x-mean)^2), where n is the population size. The "var" function in R automatically adjusts the formula based on the input data, giving the appropriate result for either sample or population variance. This allows for efficient and accurate calculation of variance in R, making it a useful tool for statistical analysis.

Calculate Sample & Population Variance in R

The variance is a way to measure how spread out data values are around the mean.

The formula to find the variance of a population is:

$\sigma 2 = \Sigma (xi - \mu)2 / N$

where μ is the population mean, xi is the ith element from the population, N is the population size, and Σ is just a fancy symbol that means "sum."

The formula to find the variance of a sample is:

$s2 = \Sigma (xi - x)2 / (n-1)$

where x is the sample mean, xi is the ith element in the sample, and n is the sample size.

Example: Calculate Sample & Population Variance in R

Suppose we have the following dataset in R:

```
#define dataset
data <- c(2, 4, 4, 7, 8, 12, 14, 15, 19, 22)
```

We can calculate the sample variance by using the var() function in R:

```
#calculate sample variance
var(data)
```

```
46.01111
```

And we can calculate the population variance by simply multiplying the sample variance by (n-1)/n as follows:

```
#determine length of data
n <- length(data)

#calculate population variance
var(data) * (n-1)/n
```

```
41.41
```

Note that the population variance will always be smaller than the sample variance.

In practice, we typically calculate sample variances for datasets since it's unusual to collect data for an entire population.

Example: Calculate Sample Variance of Multiple Columns

```
#create data frame
data <- data.frame(a=c(1, 3, 4, 4, 6, 7, 8, 12),
b=c(2, 4, 4, 5, 5, 6, 7, 16),
c=c(6, 6, 7, 8, 8, 9, 9, 12))

#view data frame
data

a b c
1 1 2 6
2 3 4 6
3 4 4 7
4 4 5 8
5 6 5 8
6 7 6 9
7 8 7 9
```

8 12 16 12

We can use the sapply() function to calculate the sample variance of each column in the data frame:

```
#find sample variance of each column
sapply(data, var)

a b c
11.696429 18.125000 3.839286
```

And we can use the following code to calculate the sample standard deviation of each column, which is simply the square root of the sample variance:

```
#find sample standard deviation of each column
sapply(data, sd)

a b c
3.420004 4.257347 1.959410
```

You can find more R tutorials here.