Lab1

Anastasia Piadi, Nazli Bilgic

2023-11-06

Question 2:

a) Write your own R function, myvar, to estimate the variance in this way.

```
## myvar function with the example data(c(1,2,3,4)): 1.666667
```

```
## var function with the example data(c(1,2,3,4)): 1.666667
```

b) Generate a vector x = (x1, . . . , x10000) with 10000 random numbers with mean 108 and variance 1.

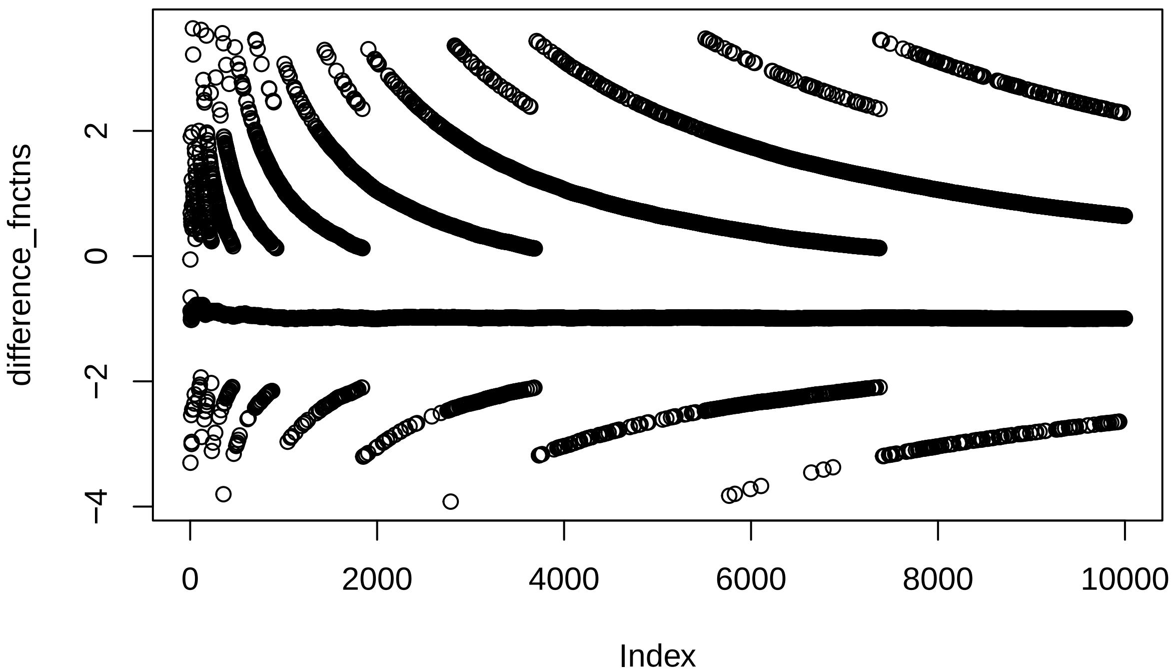
```
## myvar function output with random data: 0
```

```
## var function output with random data: 0.9972751
```

c) For each subset Xi = {x1,. . . ,xi}, i = 1,. . . ,10000 compute the difference Yi = myvar(Xi) - var(Xi), where var(Xi) is the standard variance estimation function in R. Plot the dependence Yi on i. Draw conclusions from this plot. How well does your function work? Can you explain the behaviour?

The difference between myvar and var function can be attributed to the way floating point arithmetic works in computers. This can lead to precision errors, especially when dealing with a combination of very large and very small numbers.

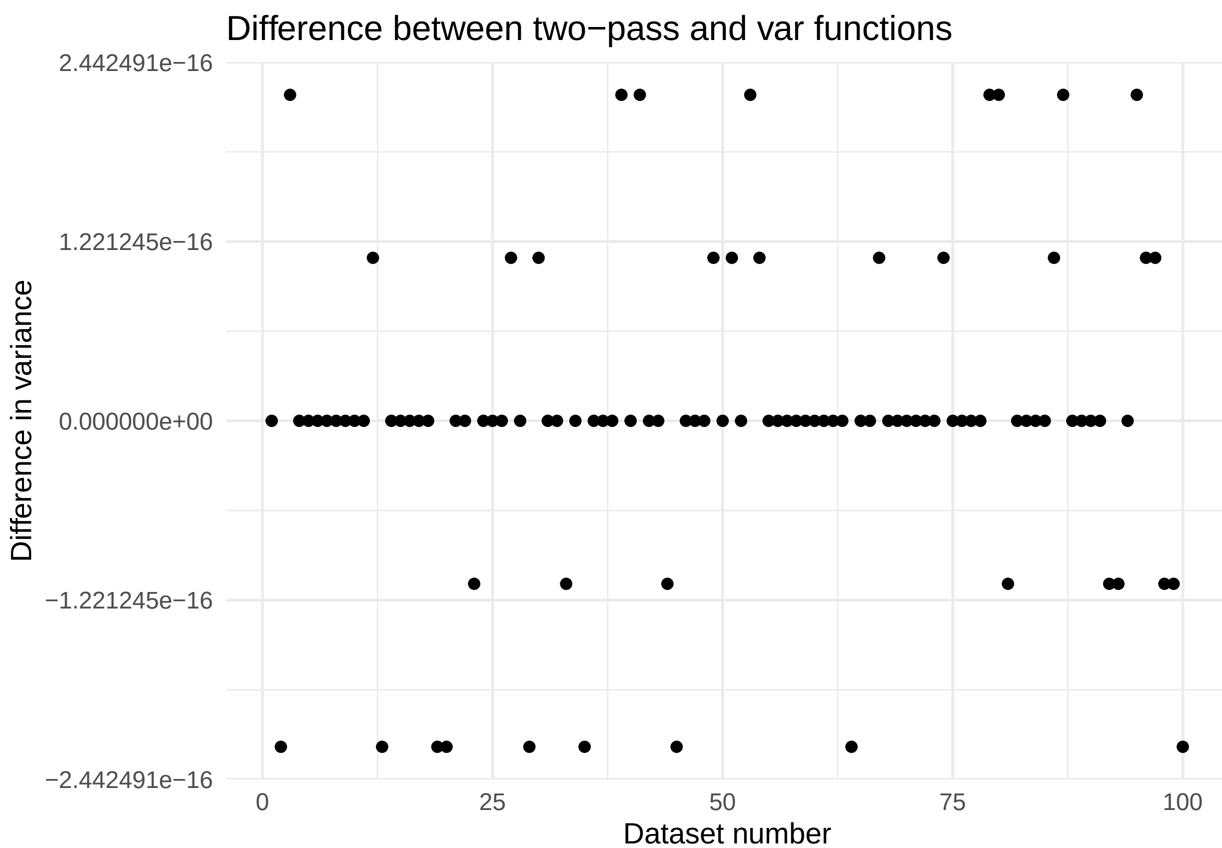
source:https://en.wikipedia.org/wiki/Floating-point_arithmetic#IEEE_754:_floating_point_in_modern_computers



d) How can you better implement a variance estimator? Find and implement a formula that will give the same results as var().

We used two-pass method to calculate variance in another way. In two-pass method, we first compute the sample mean then the sum of the squares of the differences from the mean.

source: https://en.wikipedia.org/wiki/Algorithms_for_calculating_variance



#Appendix

##Appendix A: Code for the questions

Chunk Label: Question2 a

```
myvar <- function(x) {
  n <- length(x)
  return((sum(x^2) - (sum(x)^2) / n) / (n - 1))
}
ex_data<-c(1,2,3,4)
cat("myvar function with the example data(c(1,2,3,4)):",myvar(ex_data), "\n")
cat("var function with the example data(c(1,2,3,4)):",var(ex_data))
```

Chunk Label: Question2 b

```
myvar <- function(x) {
  n <- length(x)
  return((sum(x^2) - (sum(x)^2) / n) / (n - 1))
}
set.seed(123)
n <- 10000
mean_value <- 10^8
```

```r
variance<-1
std_dev <- sqrt(variance)
x <- rnorm(n, mean = mean_value, sd = std_dev)

cat("myvar function output with random data:",myvar(x), "\n")
cat("var function output with random data:",var(x))
```

Chunk Label: Question2 c

```r
myvar <- function(x) {
  n <- length(x)
  return((sum(x^2) - (sum(x)^2) / n) / (n - 1))
}

set.seed(123)
n <- 10000
mean_value <- 10^8
variance<-1
std_dev <- sqrt(variance)
x <- rnorm(n, mean = mean_value, sd = std_dev)

computed_variance <- myvar(x)
difference_fnctns <- numeric(n)
for (i in 1:n) {
  x_sub <- x[1:i]
  difference_fnctns[i] <- myvar(x_sub) - var(x_sub)
}
plot(difference_fnctns)
```

Chunk Label: Question2 d

```r
library(ggplot2)

var_two_pass <- function(x) {
  n <- length(x)
  mean_val <- sum(x) / n #mean calculation
  sum_sq_diffs <- sum((x - mean_val)^2) #sum squared differences from the mean
  return(sum_sq_diffs / (n - 1)) #variance calculate
}

set.seed(123)
n <- 10000
mean_value <- 10^8
variance <- 1
std_dev <- sqrt(variance)

num_datasets <- 100
differences <- numeric(num_datasets)
```

```r
for (i in 1:num_datasets) {
  x <- rnorm(n, mean = mean_value, sd = std_dev)
  differences[i] <- var_two_pass(x) - var(x)
}

df <- data.frame(dataset = 1:num_datasets, difference = differences)
ggplot(df, aes(x = dataset, y = difference)) +
  geom_point() +
  theme_minimal() +
  labs(title = "Difference between two-pass and var functions",
       x = "Dataset number",
       y = "Difference in variance")
```