The R Book

Chapter 2: Essentials of the R Language
Session 8

Writing functions in R

objects that carry out operations on arguments

Syntax:
function (argument « variable ») body

Keyword: **function** indicates to R that you want to create a function.

Argument - comma-separated list of formal arguments.
- variable symbol (x or y),
- symbol = expression (pch=16)
- ...

Body - any valid R expression / set of R expressions.
- often contained in curly brackets { }, each expression on a separate line.

Arithmetic mean

```
> arithmetic.mean<-function(x) sum(x)/length(x)
> y<-c(3,3,4,5,5)
> arithmetic.mean(y)
[1] 4
```

Arithmetic mean

```
> arithmetic.mean<-function(x) sum(x)/length(x)
> y<-c(3,3,4,5,5)
> arithmetic.mean(y)
[1] 4
```

Median

Arithmetic mean

```
> arithmetic.mean<-function(x) sum(x)/length(x)
> y<-c(3,3,4,5,5)
> arithmetic.mean(y)
[1] 4
```

Median

```
> sort(y)[ceiling(length(y)/2)]
```

[1] 4 but what if y has even numbers of values ?

Arithmetic mean

```
> arithmetic.mean<-function(x) sum(x)/length(x)
> y<-c(3,3,4,5,5)
> arithmetic.mean(y)
[1] 4
```

Median

```
> sort(y)[ceiling(length(y)/2)]
```

[1] 4 but what if y has even numbers of values ?

```
> med<-function(x) {
+ odd.even<-length(x)%%2
+ if (odd.even == 0) (sort(x)[length(x)/2]+sort(x)[1+ length(x)/2])/2
+ else sort(x)[ceiling(length(x)/2)]
+ }
> a<-c(1,2,3,4)
> med(a)
[1] 2.5
```

Modulo, remainder after division

geometric mean

- For processes that change multiplicatively rather than additively

- the nth root of the product of the data

$$\hat{y} = \sqrt[n]{\prod y}$$

geometric mean

- For processes that change multiplicatively rather than additively

- the nth root of the product of the data

$$\hat{y} = \sqrt[n]{\prod y}$$

Example:

Numbers of insects on 5 plants: 10, 1, 1000, 1, 10

10 x 1 x 1000 x 1 x 10 = 100000
=> 100000^(1/5) = 10 (geometric mean)

Compare to the arithmetic mean:

```
> insects<-c(10,1,1000,1,10)
> mean(insects)
[1] 204.4
```

geometric mean

- the nth root of the product of the data

$$\hat{y} = \sqrt[n]{\prod y}$$

Example:

Numbers of insects on 5 plants: 10, 1, 1000, 1, 10

10 x 1 x 1000 x 1 x 10 = 100000
=> 100000^(1/5) = 10 (geometric mean)

Alternative calculation:

arithmetic mean of the logarithm-transformed values

$$\log \bar{x}_{\text{geom}} = \frac{1}{n} \sum_{i=1}^{n} \log x_i$$

geometric mean

- the nth root of the product of the data

$$\hat{y} = \sqrt[n]{\prod y}$$

Example:

Numbers of insects on 5 plants: 10, 1, 1000, 1, 10

10 x 1 x 1000 x 1 x 10 = 100000
=> 100000^(1/5) = 10 (geometric mean)

Alternative calculation using logarithms

```
> insects<-c(10,1,1000,1,10)
> exp(mean(log(insects)))
[1] 10
```

geometric mean

- the nth root of the product of the data

$$\hat{y} = \sqrt[n]{\prod y}$$

write the function

geometric mean

- the nth root of the product of the data

$$\hat{y} = \sqrt[n]{\prod y}$$

write the function

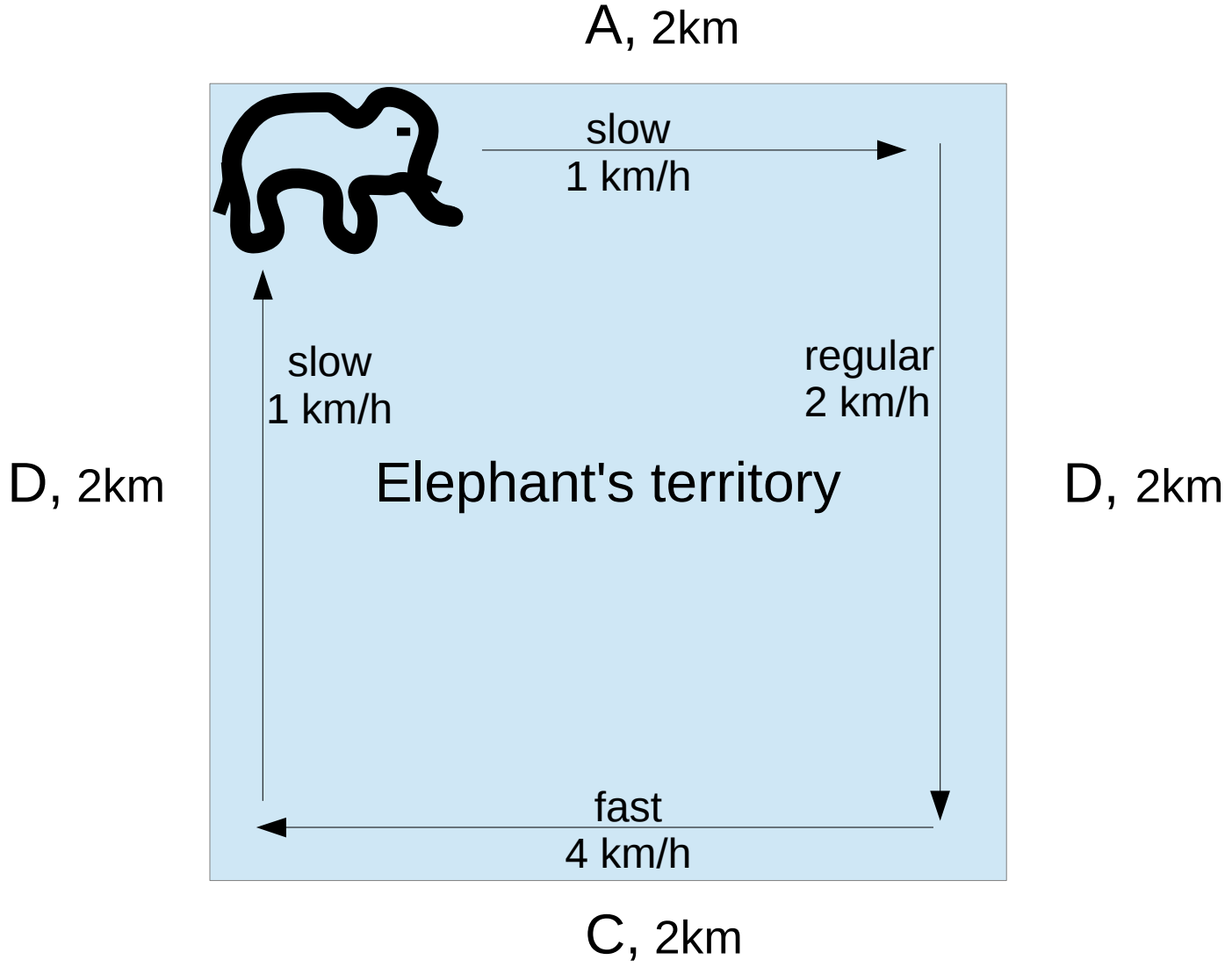
```
> geometric<-function(x) exp(mean(log(x)))
> geometric(insects)
[1] 10
```

Harmonic mean (average of rates)

- What is the average speed of the elephant?

Mean:(1km/h + 2km/h + 4km/h + 1km/h) /4 = 8/4 = 2 km/h

Median : 1km/h + (2km/h – 1km/h)/2 = 1.5 km/h



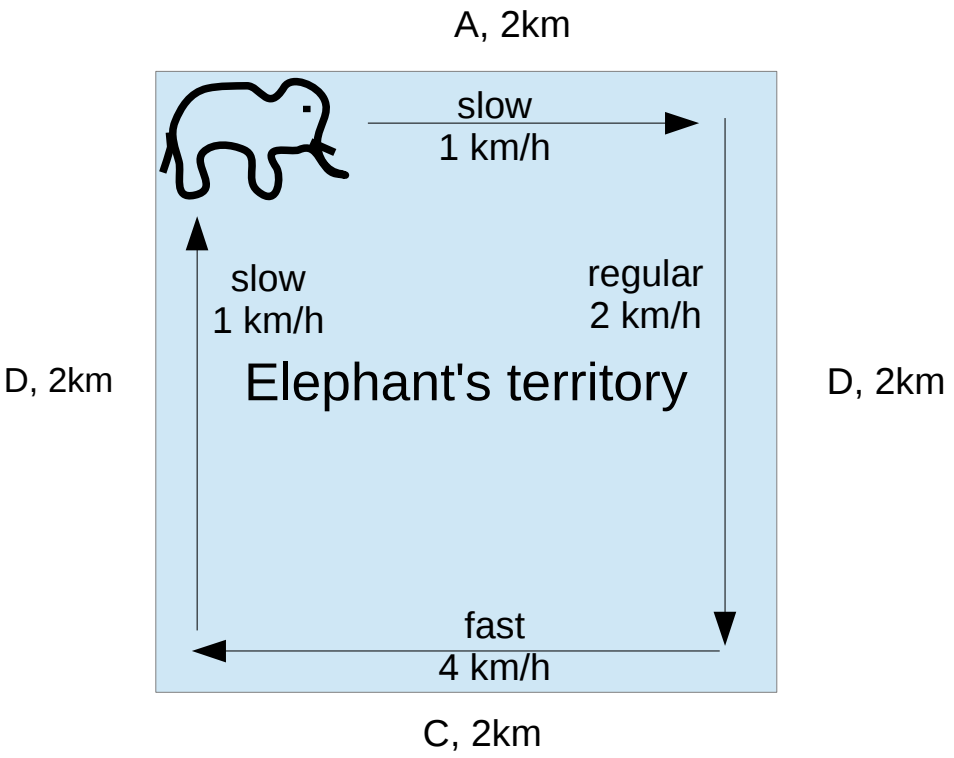
Harmonic mean (average of rates)

- What is the average speed of the elephant?
  => total distance travelled 4 × 2 = 8 km
  => total time spent 2h + 1h + 0.5h + 2h = 5.5h

  => 8km/5.5h = 1.4545 km/h

Mean: 2 km/h
Median: 1.5 km/h

harmonic mean: reciprocal of the average of reciprocals



Harmonic mean (average of rates)

- The harmonic mean is the reciprocal of the average of the reciprocals

$$\tilde{y} = \frac{1}{(\Sigma(1/y))/n} = \frac{n}{\Sigma(1/y)}$$

write the function:

Harmonic mean (average of rates)

- The harmonic mean is the reciprocal of the average of the reciprocals

$$\tilde{y} = \frac{1}{(\Sigma(1/y))/n} = \frac{n}{\underbrace{\Sigma(1/y)}}$$

write the function:

1/mean(1/x)

Harmonic mean (average of rates)

- The harmonic mean is the reciprocal of the average of the reciprocals

$$\tilde{y} = \frac{1}{(\Sigma(1/y))/n} = \frac{n}{\Sigma(1/y)}$$

write the function:

```
> harmonic<-function (x) 1/mean(1/x)
> harmonic(c(1,2,4,1))
[1] 1.454545
```

Compare to:

Mean: 2km/h
Median: 1.5km/h
Total average: 8km/5.5h = 1.4545 km/h

Variance

- sum of the squares of the difference between the data and the arithmetic mean

$$SS = \Sigma(y - \bar{y})^2$$

- we need to calculate the sample mean $\bar{y}$ from the data

- degrees of freedom $d.f. = n - k$

  the number of values in the final calculation of a statistic that are free to vary.

  (sample size, n, minus the number of parameters, k, estimated from the data.)

- For the variance, one parameter from the data, $\bar{y}$, and => n − 1 degrees of freedom.

Variance

- **sum of the squares** $SS = \Sigma(y - \bar{y})^2$

- **degrees of freedom** $d.f. = n - k$

- **variance**

$$\text{variance} = s^2 = \frac{\text{sum of squares}}{\text{degrees of freedom}}$$

$\sqrt{\text{s2}}$ = **standard deviation S.D.**

Variance

$$\text{variance} = s^2 = \frac{\text{sum of squares}}{\text{degrees of freedom}}$$

write the function:

```
> y<-c(13,7,5,12,9,15,6,11,9,7,12)
```

Variance

$$\text{variance} = s^2 = \frac{\text{sum of squares}}{\text{degrees of freedom}}$$

write the function:

```
> y<-c(13,7,5,12,9,15,6,11,9,7,12)
> variance <-function(x) sum((x – mean(x))∧ 2)/(length(x)-1)
> variance(y)
 [1] 10.25455

Built in function var():
> var(y)
 [1] 10.25455
```

Variance ratio test

Fisher's F test

dividing the larger variance by the smaller variance

s^2_x / s^2_y, for x>y

Test if the value falls into the 95 % confidence interval of the F-distribution

F-distribution: (chapter 7)

(density function)

d.f in the numerator

d.f in the denominator

$$f(x) = \frac{r\Gamma(1/2(r+s))}{s\Gamma(1/2r)\Gamma(1/2s)} \frac{(rx/s)^{(r-1)/2}}{[1+(rx/s)]^{(r+s)/2}}$$

Will be explained later

Variance ratio test

Fisher's F test

dividing the larger variance by the smaller variance

s^2_x / s^2_y, **for x>y**

write the function:

Variance ratio test

Fisher's F test

dividing the larger variance by the smaller variance

s^2_x / s^2_y, **for x>y**

write the function:

```
variance.ratio<-function(x,y) {
v1<-var(x)
v2<-var(y)
```

Variance ratio test

Fisher's F test

dividing the larger variance by the smaller variance

s^2_x / s^2_y, **for x>y**

write the function:

```
variance.ratio<-function(x,y) {
v1<-var(x)
v2<-var(y)
if (var(x) > var(y)) {
vr<-var(x)/var(y)
df1<-length(x)-1
df2<-length(y)-1}
else { vr<-var(y)/var(x)
df1<-length(y)-1
df2<-length(x)-1}
```

nested expressions
also within brackets

Variance ratio test

Fisher's F test

dividing the larger variance by the smaller variance

s^2_x / s^2_y, for x>y

write the function:

```
variance.ratio<-function(x,y) {
v1<-var(x)
v2<-var(y)
if (var(x) > var(y)) {
vr<-var(x)/var(y)
df1<-length(x)-1
df2<-length(y)-1}
else { vr<-var(y)/var(x)
df1<-length(y)-1
df2<-length(x)-1}
2*(1-pf(vr,df1,df2)) }
```

nested expression
also within brackets

probability of an F-ratio ≥ vr
pf = in-built function : F-distribution

Variance ratio test

Fisher's F test

dividing the larger variance by the smaller variance

s^2_x / s^2_y, **for x>y**

Use some data:

(n, mean, sd)

```
> a<-rnorm(10,15,2)
> b<-rnorm(10,15,4)

> variance.ratio(a,b)
 [1] 0.01593334
```

Built-in function:

```
var.test(a,b)

F test to compare two variances
data: a and b
F = 0.1748, num df = 9, denom df = 9, p-value = 0.01593
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
0.04340939 0.70360673
```

Using Variance

Variance is used in two main ways:
- **measuring unreliability (e.g. Confidence intervals)**
- **testing hypotheses (e.g. Student's t test, Chapter 8)**

<u>Measure of unreliability should:</u>
- **go up as variance increases**
- **go down as sample size increases**
- **be of the same unit of measurement as the sample mean**

Standard error:

$$se_{\bar{y}} = \sqrt{\frac{s^2}{n}}$$

variance

Sample size

square-root:
same unit dimension as the mean

Using Variance

Variance is used in two main ways:
- **measuring unreliability (e.g. Confidence intervals)**
- **testing hypotheses (e.g. Student's t test, Chapter 8)**

<u>Measure of unreliability should:</u>
- **go up as variance increases**
- **go down as sample size increases**
- **be of the same unit of measurement as the sample mean**

Standard error:

$$se_{\bar{y}} = \sqrt{\frac{s^2}{n}}$$

variance

Sample size

square-root:
same unit dimension as the mean

<u>function:</u>

```
se<-function(x) sqrt(var(x)/length(x))
```