Introduction to Data Analysis – Week 2 SIT718

Delaram Pahlevani

Median and Mode

The median of a set of data is the 'middlest' number after the data is ordered. For example if I have a set of values 5,9,3,1,4 , the first step would be to organise them in ascending order:

1,3,4,5,9

and the median of these numbers would be 4 as it is the number in the middle.

The mode of a set of values is the number which occurs the most often in a set of data. e.g. the mode of 1,1,2,2,3,3,3,4,5,6,6,7,8 is 3 .

If we have multiple numbers with the same number of occurrences, we can equally consider them all modes of the given set of data

The Geometric and Harmonic Means

The need for alternative measures of centre when there are outliers or skewed data is well known in introductory statistics courses, however there are other reasons why it sometimes just isn’t appropriate to use the arithmetic mean to find the ‘average’.

Consider the following scenario:

Your pay goes up by 20% one year and 10% the next. What is the average pay increase?

If our pay were $10,000 in the beginning find the next two income.

Definition of Geometric Mean

- For two or more arguments, the geometric mean is the value obtained when we multiply all of the inputs together and then take the *n*-th root (where *n* is how many values we have).

- For an input vector $x = |< x_1, x_2, \ldots, x_n >$, the geometric mean is given by

$$GM(x) = \left(\prod_{i=1}^{n} x_i\right)^{1/n} = (x_1 x_2 \ldots x_n)^{\frac{1}{n}}$$

Geometric Mean Continued

- The geometric mean should be used for averaging whenever our values are related by multiplication. In the case of two inputs, we can interpret the product as the area of a rectangle. The geometric mean then can be interpreted as the dimensions of a square that would be required to give the same area.

Question:

Your pay goes up by 20% one year and 10% the next. What is the average pay increase?

Geometric Mean-continued

- The geometric mean will be strictly monotonic except if one of the inputs is zero. If any single input is zero, we will immediately obtain a zero result, even if we have thousands of inputs with high values.

$$GM(x) = \left(\prod_{i=1}^{n} x_i\right)^{1/n} = (x_1 x_2 \ldots x_n)^{\frac{1}{n}}$$

This leads us to some other aggregation properties worth noting.

The Harmonic Mean

Consider the following scenario:

Leia can paint a house in 3 h, Luke can paint one in 5, and we want to know how long it would take them to paint two houses if they work together. This is equivalent to asking how long it would take two people who both worked at Leia and Luke's 'average' pace. What is the arithmetic mean?

Leia paints a house at the rate of 1/3 of a house per hour.
Luke paints at the rate of 1/5 of a house per hour.

The Harmonic Mean-continued

So we need to find out how many houses they would paint together per hour by adding the values (1/3 and 1/5):

$$\frac{1}{3}+\frac{1}{5}=\frac{5}{15}+\frac{3}{15}=\frac{8}{15}$$

Now that we know this rate, we can ask how long it takes to paint 2 houses, which is:

$$2\div\frac{8}{15}=2\times\frac{15}{8}=\frac{30}{8}=3.75$$

Definition of Harmonic Mean

To find the harmonic mean of a set of numbers, we find the average of their reciprocal values (i.e. in fraction form this means we flip the number upside down) and then take the reciprocal of this result.

Example:

$$HM(2,3,4) = \frac{3}{\frac{1}{2}+\frac{1}{3}+\frac{1}{4}} = 2.769$$

Definition of Harmonic Mean-continued

- For an input vector $x = | < x_1, x_2, \ldots, x_n >$, the harmonic mean is given by

$$HM(x) = n(\sum_{i=1}^{n} \frac{1}{x_i})^{-1} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}$$

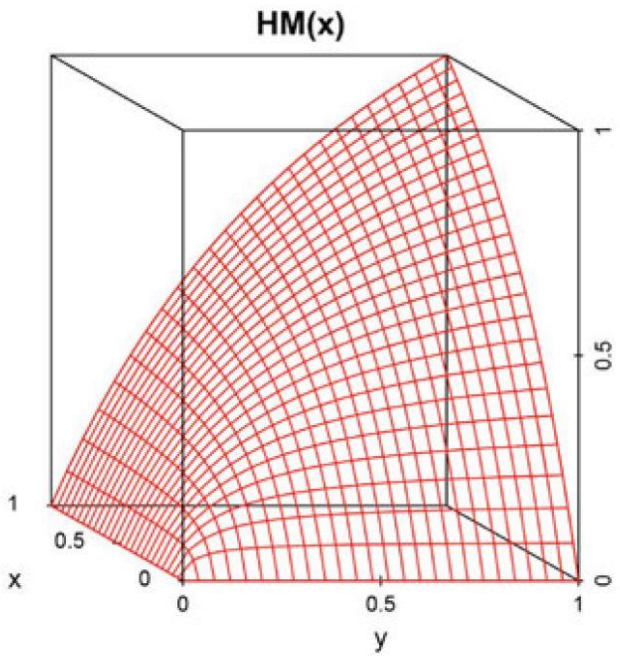
If any of the inputs is 0, we define the HM to give an output of 0.

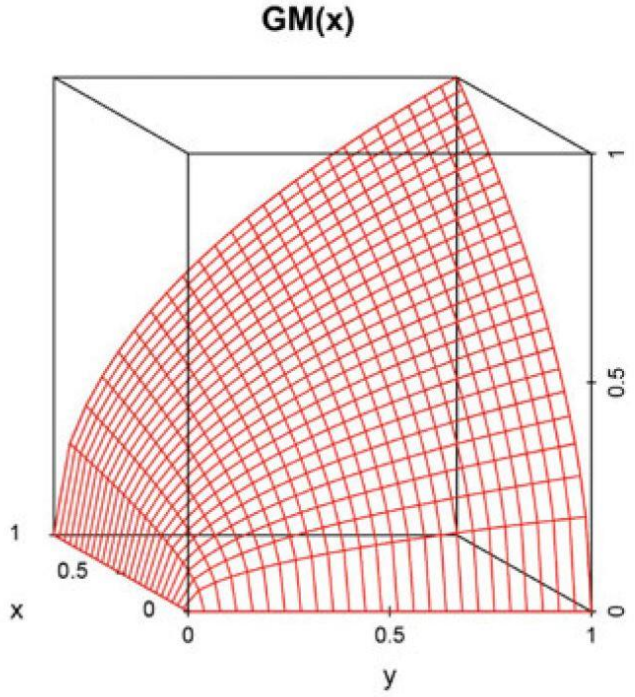
Question

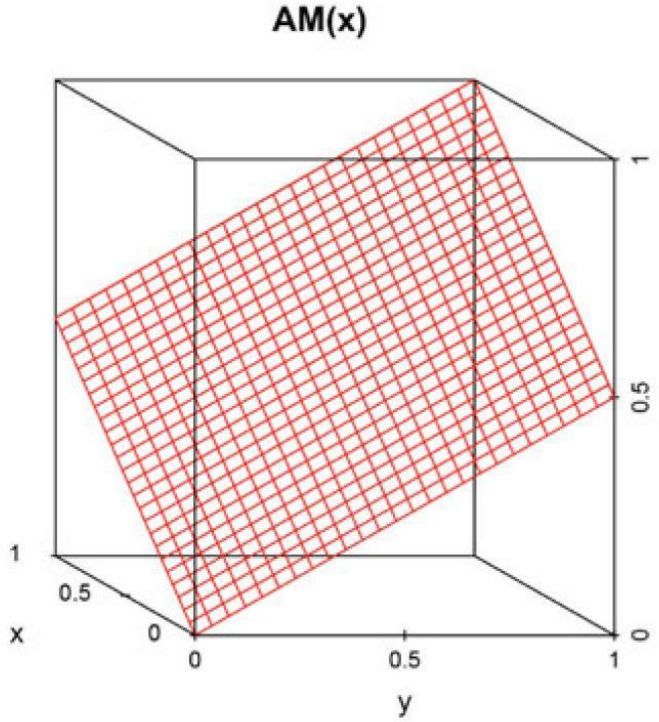
Calculate the arithmetic mean, geometric mean and harmonic mean for the input vector $x = < 0.6, 0.9, 0.26, 0.7 >$. Use pen and paper ☺

Arithmetic mean: 0.615
Geometric mean: 0.560
Harmonic mean: 0.497

$$\boldsymbol{HM(x) \leq GM(x) \leq AM(x)}$$







Arithmetic, Geometric and Harmonic Mean

Control Structures: if

```
if(<condition>) {
## do something
} else {
## do something else
}
if(<condition1>) {
## do something
} else if(<condition2>) {
## do something different
} else {
## do something different
}
```

```
if(x > 3) {
y <- 10
} else {
y <- 0
}
```

```
y <- if(x > 3) {
10
} else {
0
}
```

for

for loops take an iterator variable and assign it successive values from a sequence or vector. For loops are most commonly used for iterating over the elements of an object (list, vector, etc.)

```
for(i in 1:10) {
print(i)
}
```

This loop takes the i variable and in each iteration of the loop gives it values 1, 2, 3, ..., 10, and then
exits.

```
x <- c("a", "b", "c", "d")
for(i in 1:4) {
print(x[i])
}
```

For and If Example

Often times one will be dealing with messy data. This can happen for any number of reasons: faulty equipment, users entering survey information incorrectly etc. Implementing checks at every step in the data analysis process helps to identify these errors that would otherwise lead to inaccurate analyses. This is an example of 'defensive programming' where implementing these checks or tests can save us from future problems

```
sumArray = function(arrayOfNumbers){
  result = 0
  if(typeof(arrayOfNumbers)=="double"){
    for(i in c(1:length(arrayOfNumbers))) {
      result = result + arrayOfNumbers[i]
    }
    result
  } else
    print("Warning: There is a non-numeric
element in your array.")
}
```

Test your updated sumArray with inputs c(1,2,3,4), c(9,4,"t",5) and c("not","a","number") and try to guess what might happen.

Power mean

The power mean is best seen as a generalised mean which encompasses the arithmetic, geometric and harmonic means. In other words, we can obtain the arithmetic/geometric/harmonic mean from the power mean by specifying a specific p value.

$$PM_p(x) = \left(\frac{1}{n}\sum_{i=1}^{n} {x_1}^p\right)^{\frac{1}{p}} = \left(\frac{{x_1}^p + {x_2}^p + \cdots + {x_n}^p}{n}\right)^{\frac{1}{p}}$$

```
PM = function(x, p) {
if (p ==0) {
prod(x)^{1/length(x)}
} else {
mean(x^p)^{1/p}
}
}
```

Special Case	P
Arithmetic Mean	1
Geometric Mean	0
Harmonic Mean	-1

Practice Questions Using R

Suppose you have x= <0.1, 0.2, 0.6, 0.9>, calculate:

(i) The arithmetic mean

(ii) The geometric mean

(iii) The harmonic mean

(iv) The median and compare the results.

```
HM = function(x) {

 if(prod(x)==0) {
  return (0)}
  else {
  length(x)/sum(1/x)
 }}
```

```
GM <- function(x) {
 if(prod(x)==0) {
  return (0)
 }
 else {
  prod(x)^(1/length(x))
}}
```

Aggregation Function Definition

For an input vector $x = | < x_1, x_2, \ldots, x_n >$, a multivariate function *A(x)* can be referred to as an aggregation function if it:

A) is monotone increasing (in either a strict or non-strict sense), i.e. an increase to any of the inputs cannot result in a decrease to the output. We can express this property as:

$$if\ x \leq y \quad then \quad A(x) \leq A(y);$$

B) satisfies the boundary conditions

$$A(a, a, \ldots, a) = a, \qquad A(b, b, \ldots, b) = b,$$

where a and b are the minimum and maximum values possible.

Aggregation Function Example

A function like the product $A(x,y)=xy$ is an aggregation function over the interval [0,1] because it is monotone and $f(0,0)=0$ and $f(1,1)=1$, but not if our interval is [1,10] since $10 \times 10 = 100$

We then consider averaging functions to be a special sub-class of aggregation functions.

Definition: An aggregation function $A(x)$ is averaging if its output is bounded between the minimum and maximum of its inputs:

Averaging Function Definition

An aggregation function A(x) defined over $[a,b]^n$ $is\ avaraging\ if$:

$$\boldsymbol{min}(\boldsymbol{x}) \leq \boldsymbol{A}(\boldsymbol{x}) \leq \boldsymbol{max}(\boldsymbol{x}), \qquad \boldsymbol{for\ all}\ \ \boldsymbol{x} \in [a,b]^n$$

Question

Show that the 3-variate function

$$f(x_1, x_2, x_3) = \frac{x_1}{2}(x_2 + x_3)$$

is an aggregation function for $x_1, x_2, x_3 \in [0,1]$

Use pencil and paper again ☺

Existing Functions Using RStudio

Mean()

```
mean(c(2,3,6,7))
mean(2:7)

a = 5
long.array = c(20,7,4)
the.value = 12
mean(c(4,a,1:3,the.value,long.array))
```

Existing Functions Using RStudio

Median() and Mean()

```
median(1:6)
median(c(1,7,82,2))
mean(c(2,3,6,7))
mean(2:7)
mean(my.seq)
mean(c(4,a,1:3,the.value,long.array))
```

Practice with RStudio: arrays and matrices

cbind and rbind: filling column and rows

Assign the vectors and perform the cbind() and rbind() operations.

```
a<-c(1,2,3,7,9)
a<-cbind(a, c(21,2,1,5,6))
a<-cbind(a, c(2,-1,5,0,-1))
a<- cbind(a, c(1,9,7,2,1),array(6,5))

b <-c(3,6,1,9,2)
b <-rbind(b, c(3,2,1,8,9))
b <- rbind(b,c(4,1,12,1,2))
```

Creating a matrix

1. Using arrays

a = array (0, c(3,4))

matrix (nrow= , ncol=)

Create a 3×4 array and then assign values to different entries using the following:

```
A[3,1]<-4
A[1,]<-c(1,2,3,4)
A[,2]<-c(6,5,4)
A[2:3,3:4]<-array(-1,c(2,2))
```

Practice Questions Using R

Define the function $f(x,y) = \frac{x^2+y^2}{x+y}, (x+y \neq 0 \; and \; f(0,0) = 0)$ and evaluate *f(0.3,0.9)*, and *f(0.4, 0.9)*. Based on your results, can it be stated that f is not aggregation function?

```
a = function (x,y){
  if (x+y == 0){
    return (0)
  }
  else{
    (x^2 + y^2)/(x+y)
  }
}
x = a(0.3, 0.4)
y = a(0.9, 0.9)
y-x
```

Runif Function

The function runif(n, min = a, max = b) generates an array of n random numbers between a and b . Type runif in the console to view the documentation for this function.

Practice:

Try to generate 10 numbers between 5, 100 by using runif function.