runif → simulations of a $\mathcal{U}([0;1])$

punif → evaluation of the distribution f^{cr}

dunif → evaluation of the density f^{ct}

```
help.start()

runif(10)

[1]

> set.seed(3)
> runif(10)
> set.seed(3)
> runif(10)
```

9. Simulate 10 samples of size 100 from $B(0.2)$

a) with the R function ⇒ create a matrix

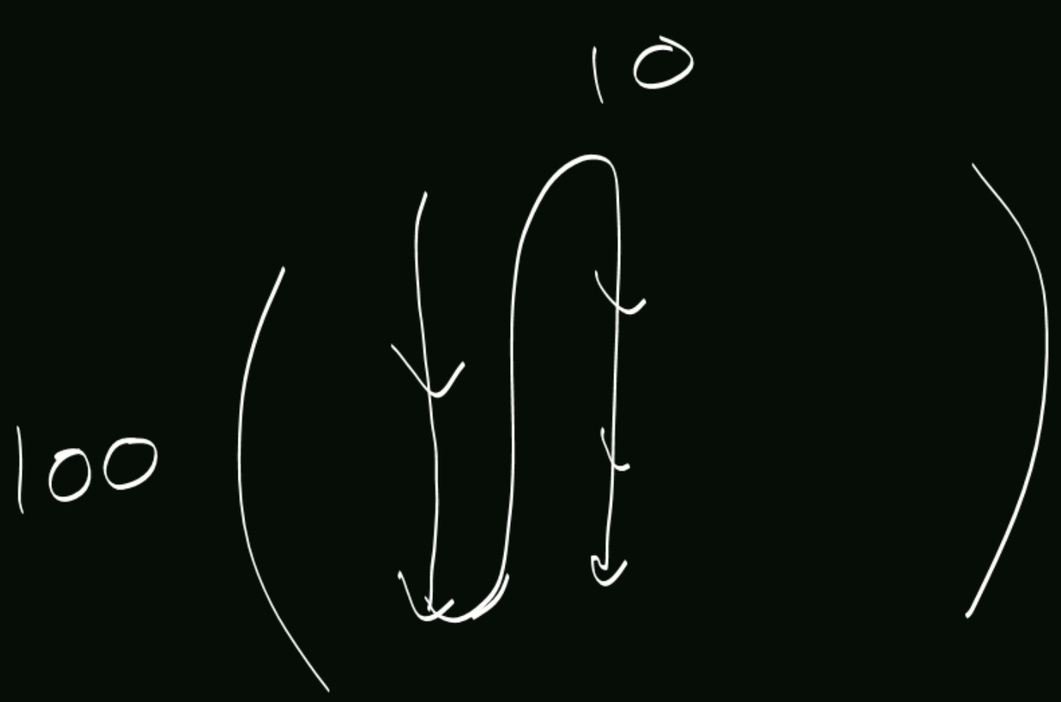
b) with the runif fct

↳ create a fct with several arguments (n: the size of the sample, k: the number of samples)

a)

rbinom(m, n, p) → m réalisations de $B(n; p)$

```
A = matrix(data = rbinom(1000, 1, 0.2),
           ncol = 10)
```



$B(p) = B(1; p)$

A[1:10,] → extraction of the rows number 1, 2, .., 10

for rows for column

A[c(1,3,4),] → extraction of the rows number 1, 3 and 4

c(1,3,4) → create a vector of length 3 whose elements are 1, 3, 4

c(seq(1,2,0.1), seq(7,5,-0.5))

↑ vector ↑ vector

A[, 3:4] → extraction of the columns number 3 and 4

A[2:4 , 1:3] → extraction of the observations of A which are on line 2, 3, 4 and column 1, 2, 3.

A[-2,] → we delete the row number 2

E = A[-c(2,4),]

J = A[-2,]

K = J[-4,]

⚠ K ≠ E

a = 2

a <- 2

2 -> a

data.frame → quite similar to a matrix

list → quite similar to a vector

```
L=list(a=seq(1,10,0.5),b=A)
L[[1]] ⇔ L$a
```

```
t:sample(c("blue","red","yellow"),
         500, replace=TRUE,
         prob=c(0.1,0.7,0.2))

p2=factor(t)
```

f2 → is a factor
associated to qualitative
values

levels: red blue yellow

c(f2, 1:10) bad

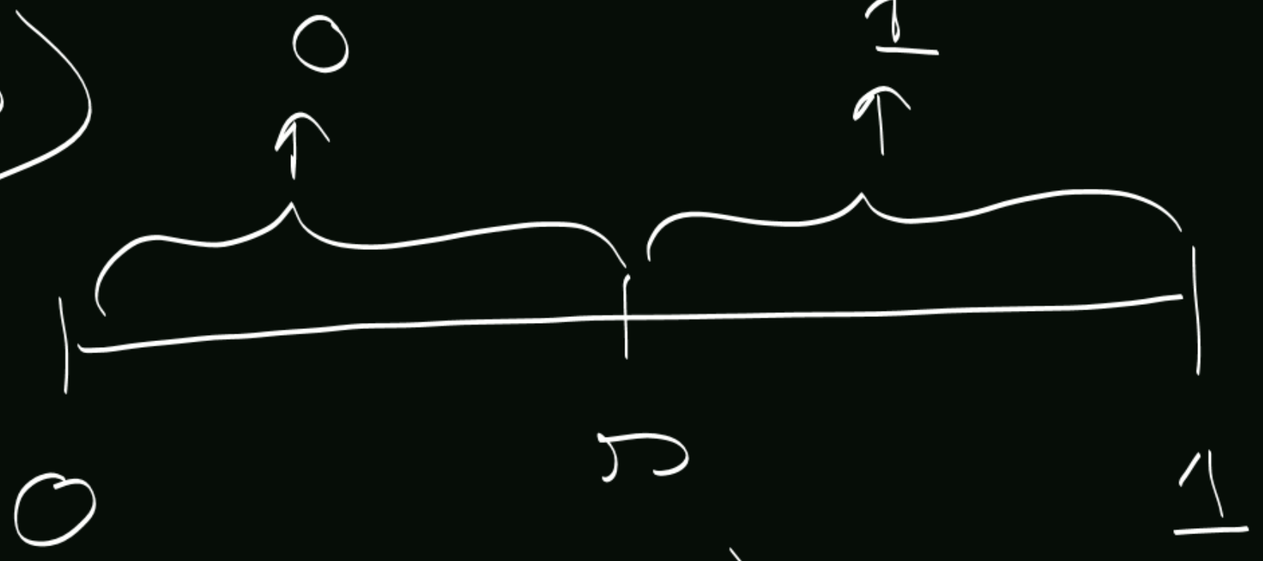
list(f2, 1:10)

```
K=matrix(data=0, ncol=2, nrow=3)
K[,1]=l2[1:3]
K[,2]=1:3
is.matrix(K)
K
```

R&:

```
B = matrix(data = 0, ncol = 10,
           nrow = 10)
```

b)



$$P(X=1) = p$$

$$P(X=1) = P(U \geq s) = \int_{s=1-s}^{1} dt$$

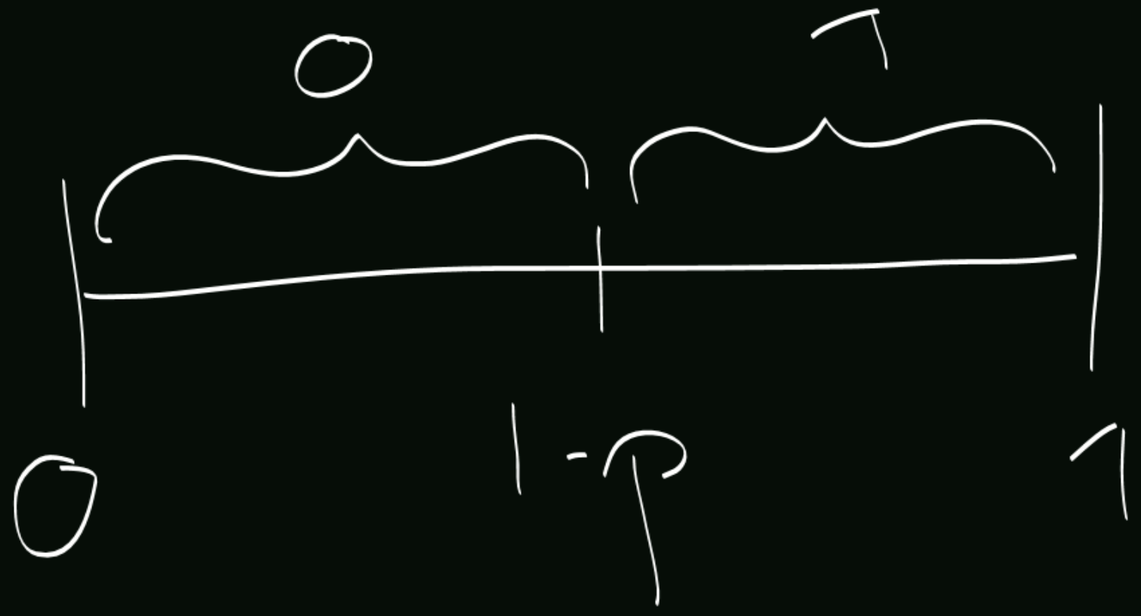
let $U \sim \mathcal{U}([0,1])$

We create X by:

if $U < s \Rightarrow X = 0$

if $U \geq s \Rightarrow X = 1$

$\Rightarrow \quad x = 1 - p$



I simulate u with respect to a $\mathcal{U}([0;1])$

if $u < 1-p$, I create $x = 0$

otherwise I create $x = 1$

```
simnber <- function(n,p)
{
  x=c()    # empty vector
for (i in 1:n)
  {
  u=runif(1)
  if (u<(1-p))    {x=c(x,0)}
    else {x=c(x,1)}
  }
```

simuber = x

}

Chapter : Descriptive statistic for variable in dimension 1

I Introduction

We speak about a variable for which we have some observation.

population: all the "statistical individuals" that can be concerned by the study.

sample: a subset of the population which is really concerned by the study

observations $\Rightarrow$ determine the nature of the variable

qualitative		quantitative	
categorial	ordinal	discrete	continuous
no notion of rang	notion of rank on the values	finite set of values or in bijection with $\mathbb{N}$	the set of values can just be expressed as interval
ex: color of eyes, professions, ...	ex: judgment on movies, grade at bachelor (MB, B, TB)	ex: number of children, ...	ex: height, weight,

II Graphics

a) Categorial variable

table of representation

modalities	frequency	percent-	angle
a_1	n_1	f_1	α_1
a_2	n_2	f_2	α_2
⋮	⋮	⋮	⋮
a_K	n_K	f_K	α_K
	n	1	360

let $x_1, \dots, x_n$ be the observation

n_i = number of j in $\{1, \dots, n\}$ such that $x_j = a_i$

$$f_i = \frac{n_i}{n} \in [0;1]$$

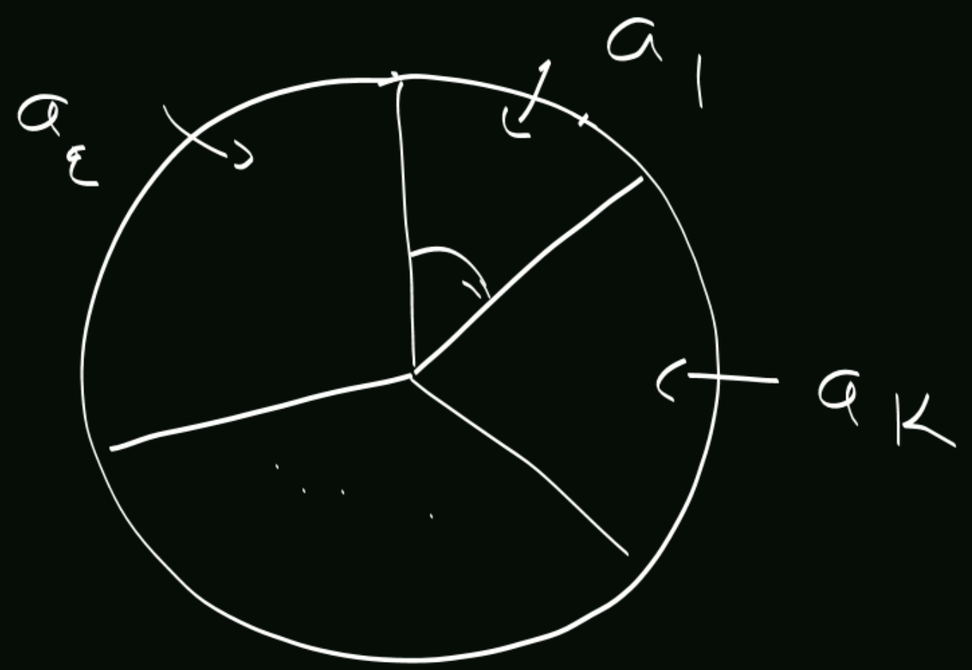
$$\alpha_i = 360 \times f_i$$

ex:

black black brow blue green
green blue grey black green

	frequency
a_1 = black	3
brown	1
blue	2
green	3
grey	1

The good representation is:



Title ← not forget

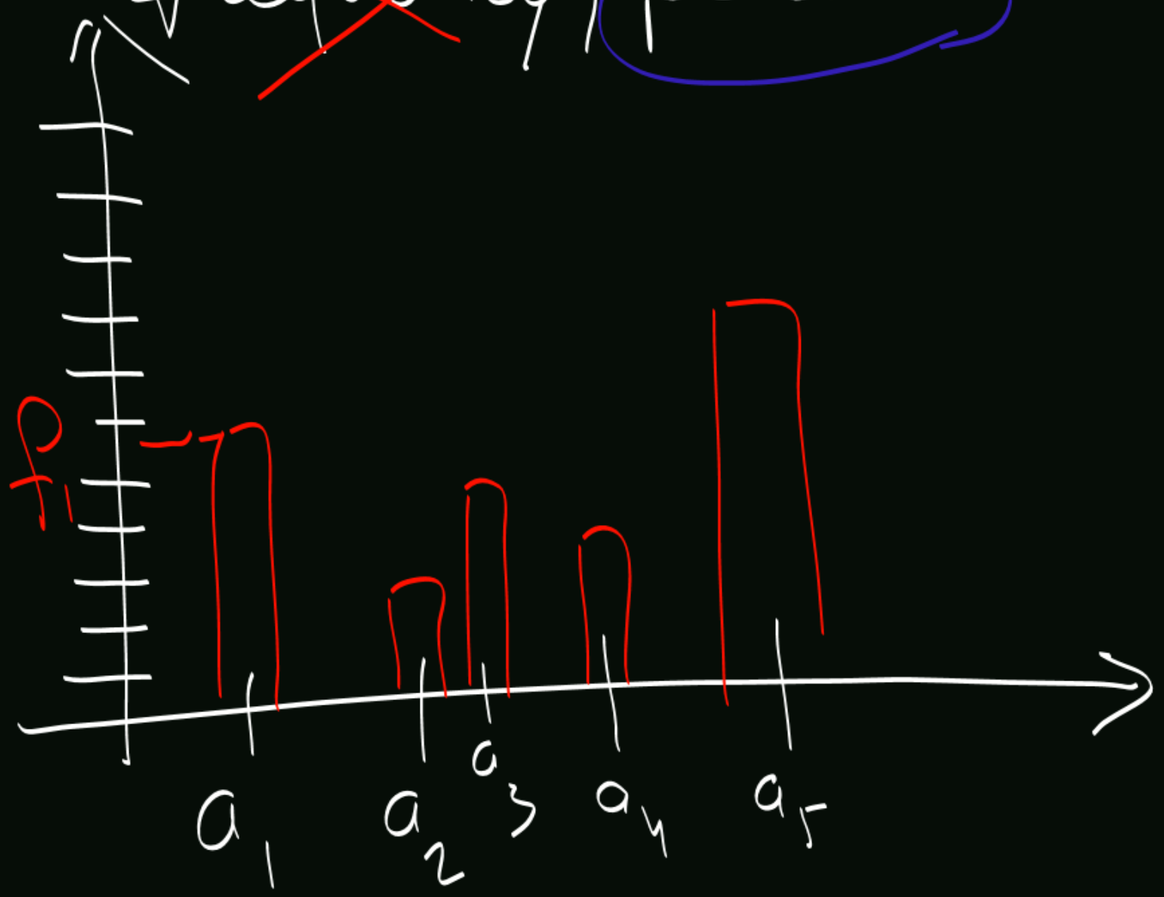
b) ordinal variable

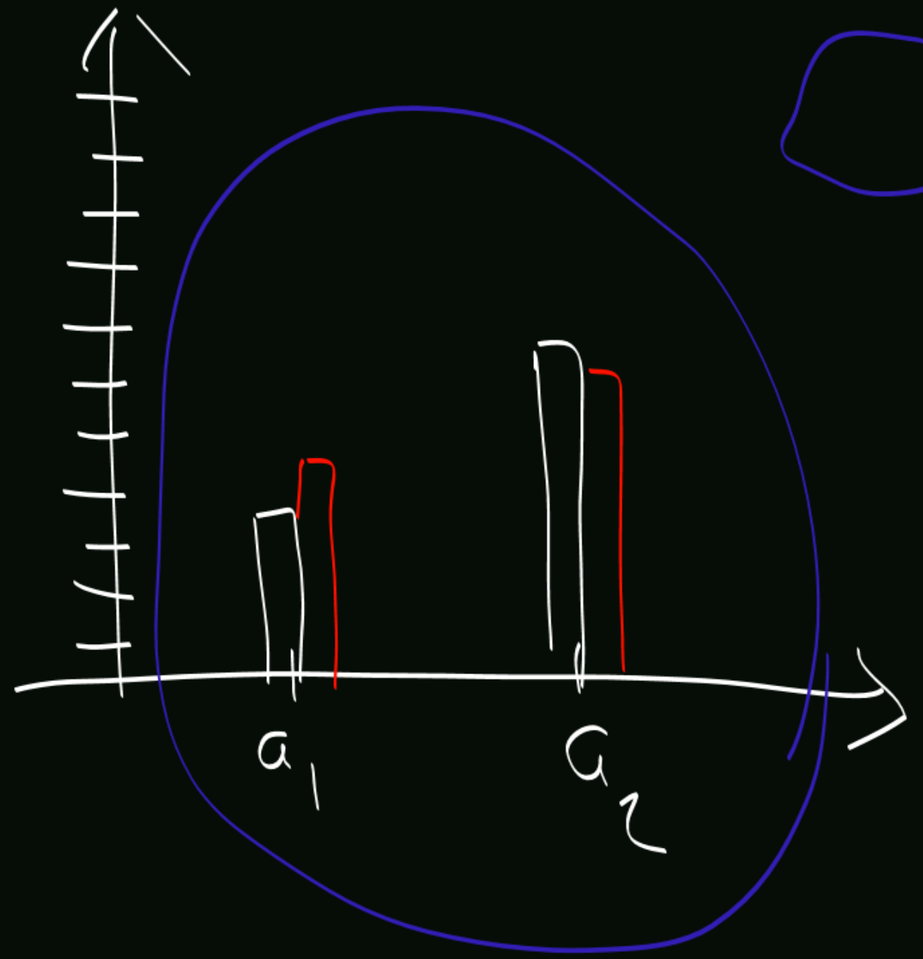
Table of representation

modalities	frequency	percent
a_1	n_1	f_1
a_2	n_2	f_2
⋮	⋮	⋮
a_K	n_K	f_K
	n	1

The good representation

~~frequency~~ / percent





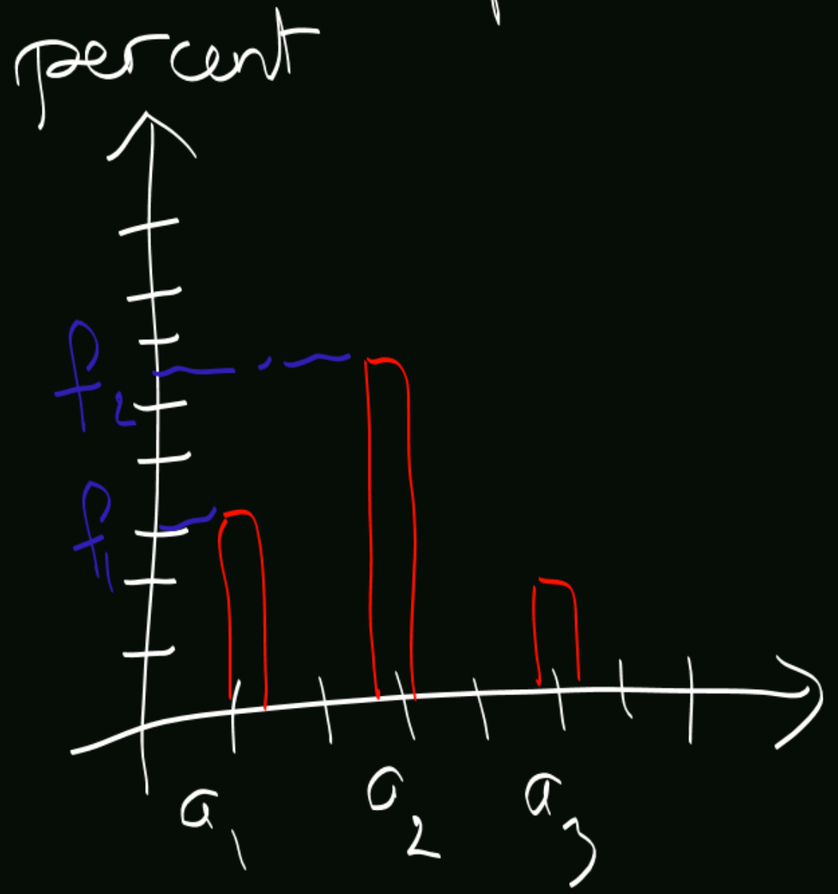
c) Discrete variable

$x_1, \ldots, x_n$: n observations

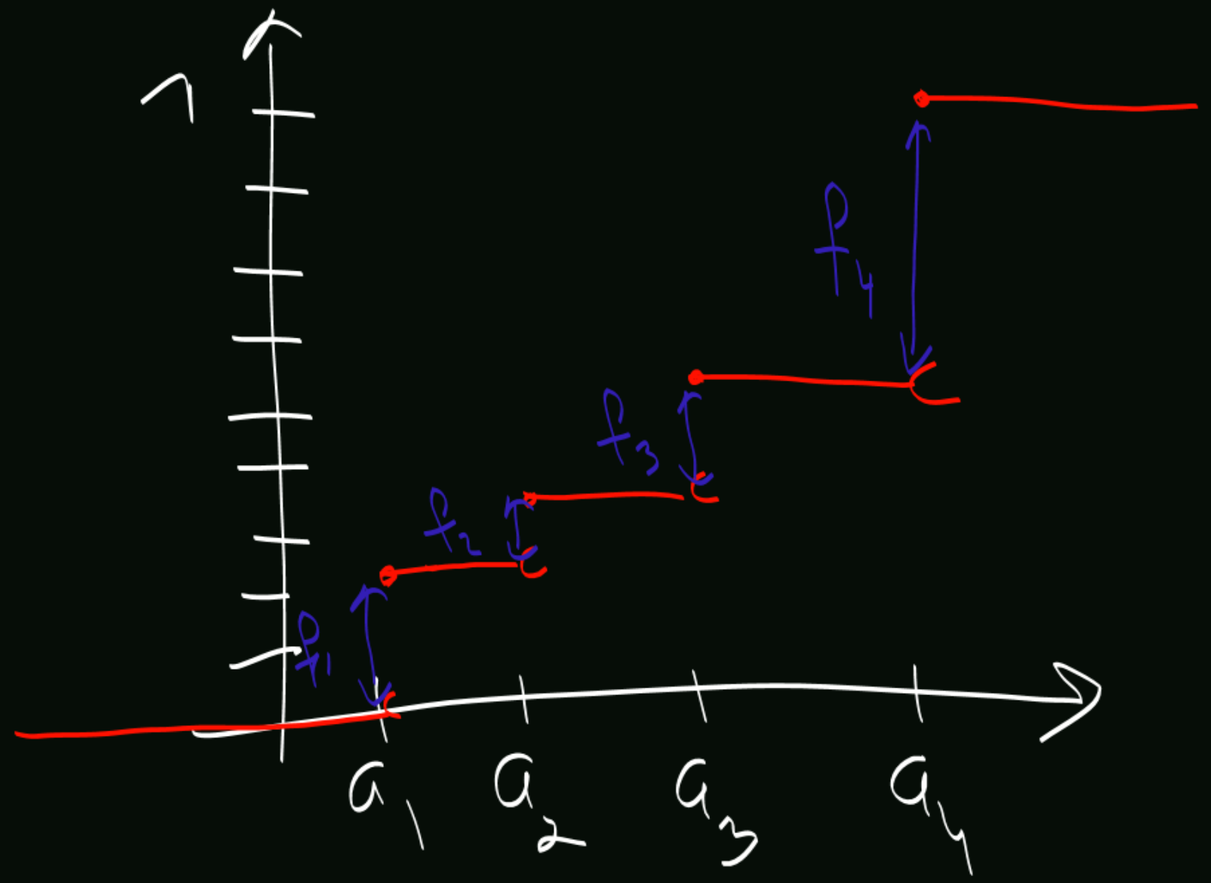
values	frequency	percent	cumulative percent
a_1	n_1	f_1	$F_1 = f_1$
	n_2	f_2	$F_2 = f_1 + f_2$
$\vdots$	$\vdots$	$\vdots$	$F_i = F_{i-1} + f_i$
a_K	n_K	f_K	$F_K = 1$
	n	1	

Graphics:

barplot



plot of cumulative percent



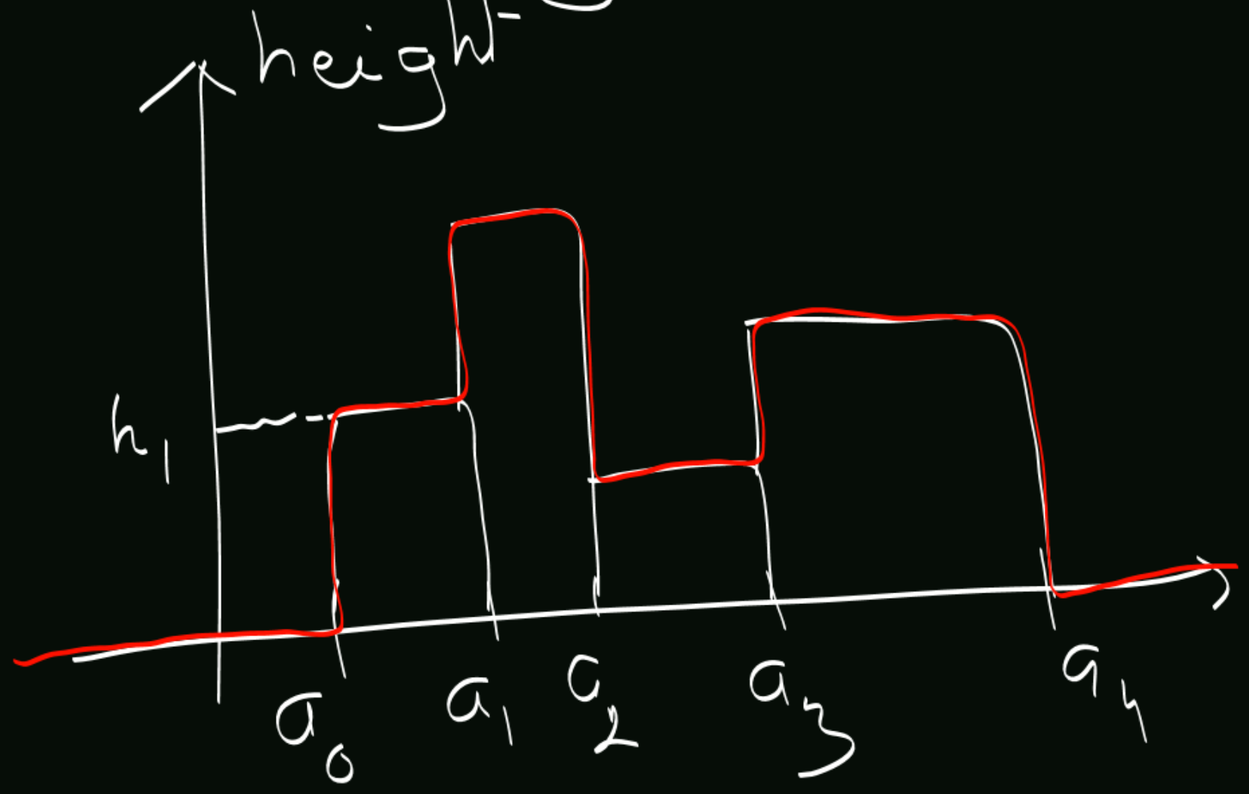
d) Continuous

classes	frequency	percent	cumulative percent	height
		f_1	$F_1 = f_1$	h_1
$[a_0 ; a_1[$	n_1	f_2	$F_2 = F_1 + f_2$	h_2
$[a_1 ; a_2[$	n_2			
$\vdots$	$\vdots$	$\vdots$		
$[a_{k-1} ; a_k[$	n_k	f_k	$F_k = 1$	h_k
	n	1		

$$h_i = \frac{f_i}{a_i - a_{i-1}}$$

Graphics

histogram



area of the histogram

$$= \sum_{i=1}^{k} (a_i - a_{i-1}) \times h_i$$

$$= \sum_{i=1}^{k} (a_i - a_{i-1}) \times \frac{f_i}{(a_i - a_{i-1})}$$

$$= 1$$

2 questions:

- How many classes to create?
- How to determine the limits of the class?

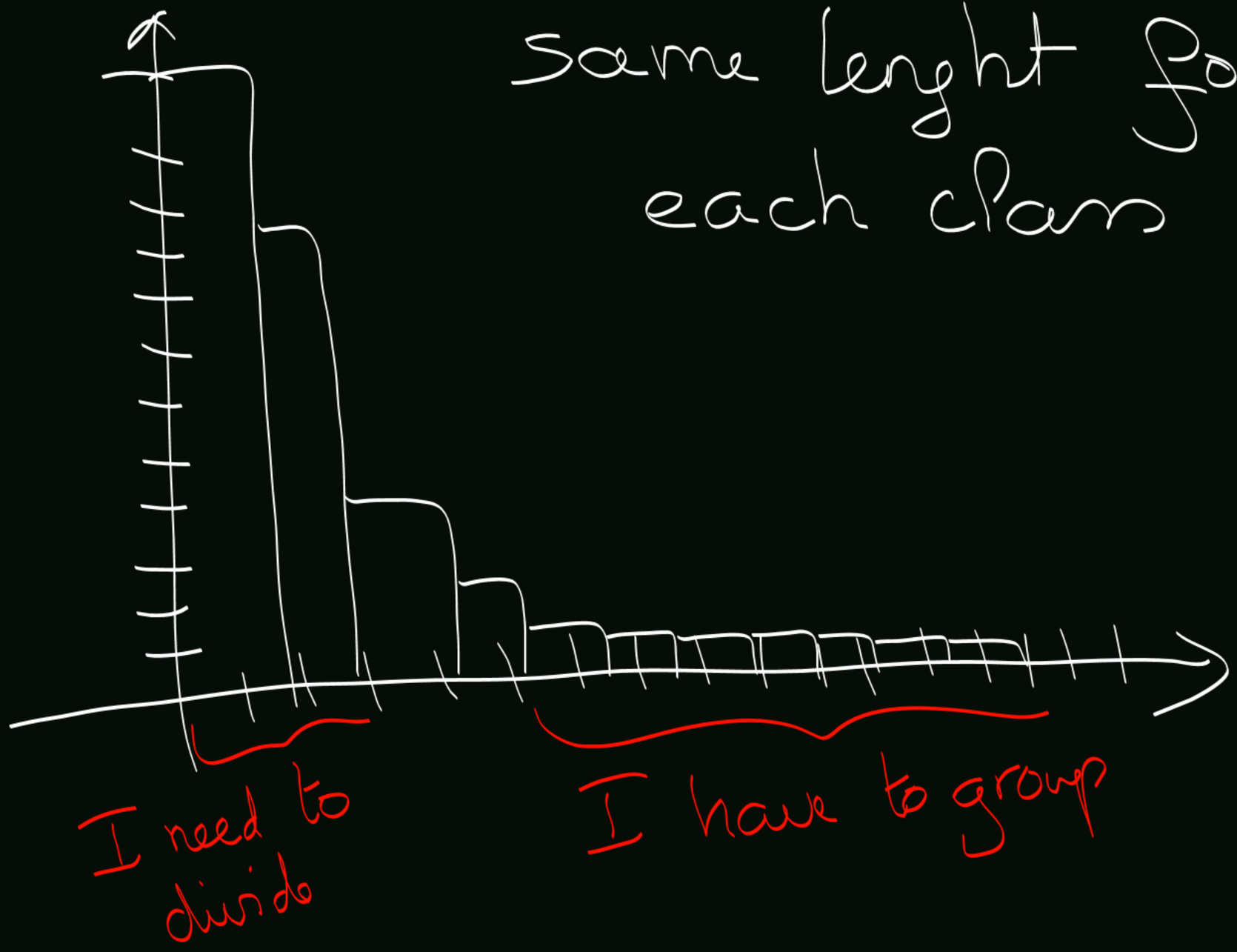
Sturge's method:

1) if n denotes the number of observations, the number K of classes is defined by:

$$K \approx 1 + 3.22 \times \log_{10}(n)$$

$$\text{range} = \frac{a_K - a_0}{K}$$

$$\forall i \in \{1, \ldots, K-1\},\ a_i = a_0 + i \times \text{range}$$


same lenght for each class
I need to divide
I have to group

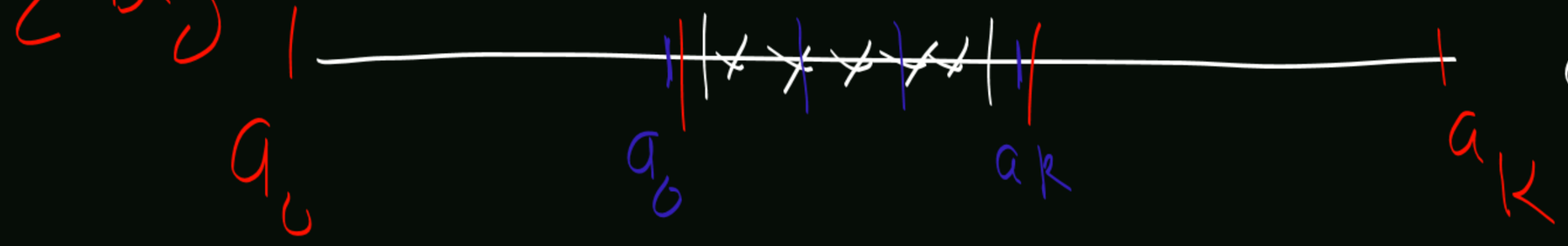
2) We choose ε, with ε small with respect to the data.

$$a_0 = \min(\text{observations}) - \varepsilon$$

$$a_K = \max(\text{observations}) + \varepsilon$$

In practice:

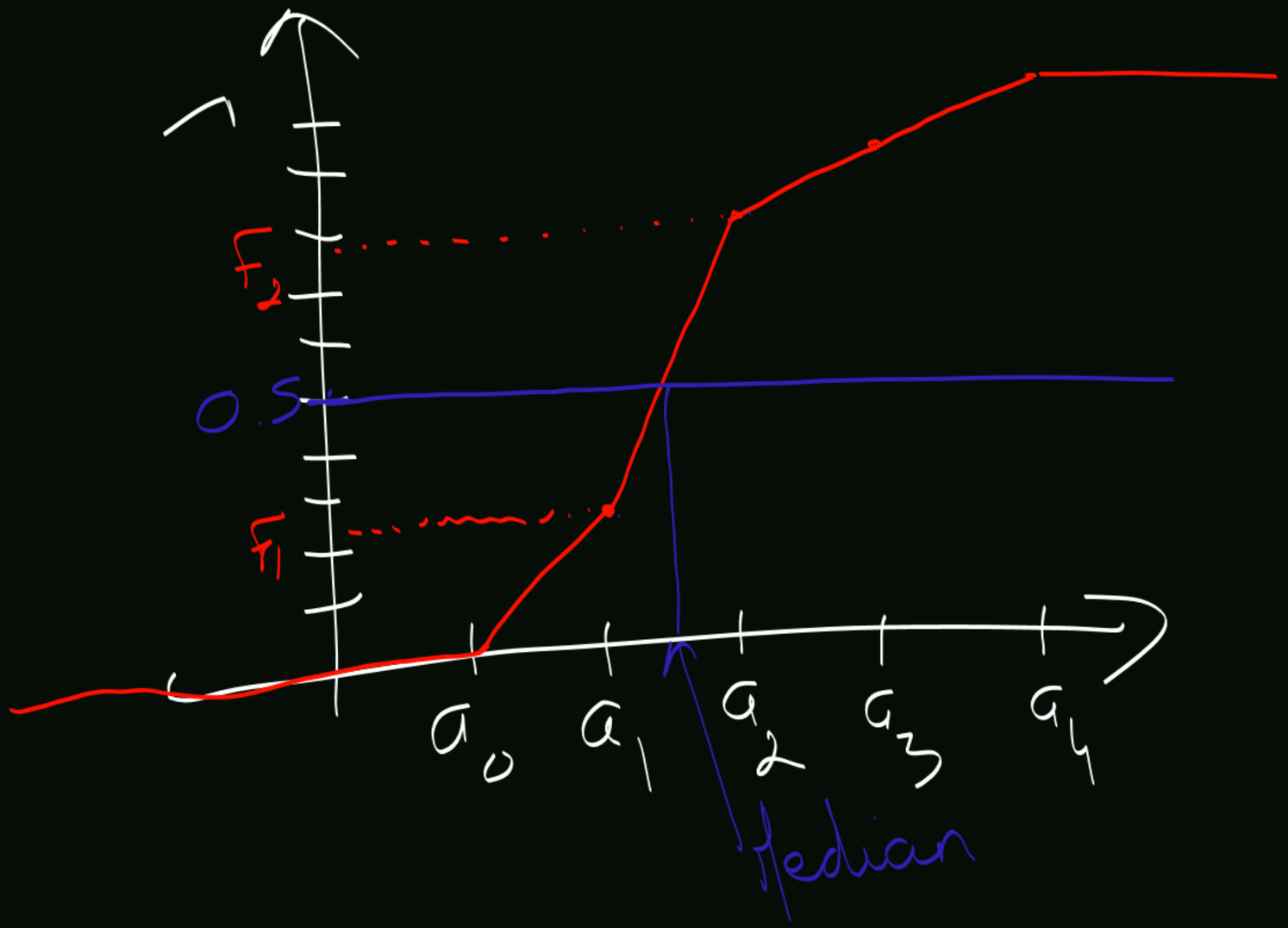
$$\varepsilon = \frac{\max(\text{obs}) - \min(\text{obs})}{10^3}$$



RR:

The density f^{cl} is an extrapolation of the histogram.

plot of cumulative percent



⇒ assumption:
the observations
are uniform
on each class!

? Take the datasets data1, data2, data3, data4 are do the good graphics

9 Take the file named extra and try to illustrate the fact that the density is the limit of an histogram

To load the data in the R console, use the read.table function

> setwd() → to change the work directory
> getwd()

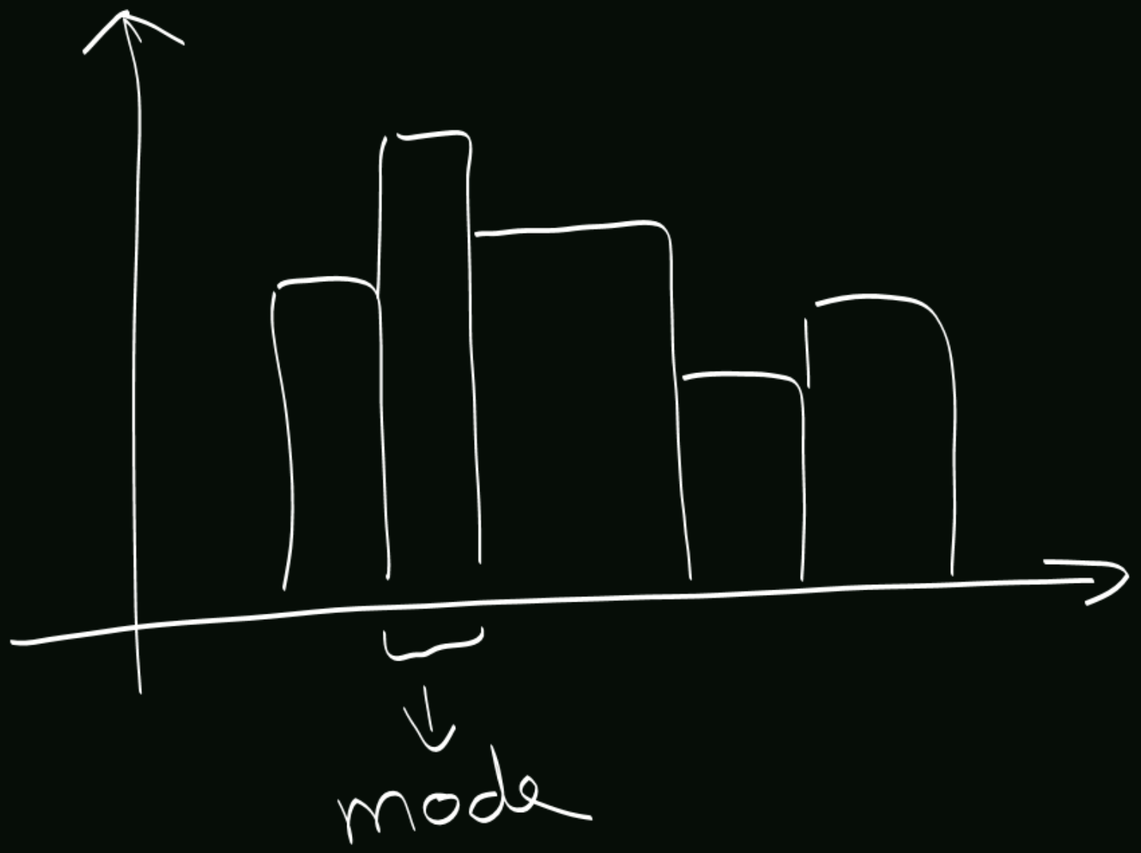
I Statistical parameters

- parameters of position (quantile)
- _____ dispersion (variance/sd, range, coeffi-cient of variation)
- _____ centrality (mean, median, mode)

a) parameters of locality

⇒ all is done for a quantitative variable

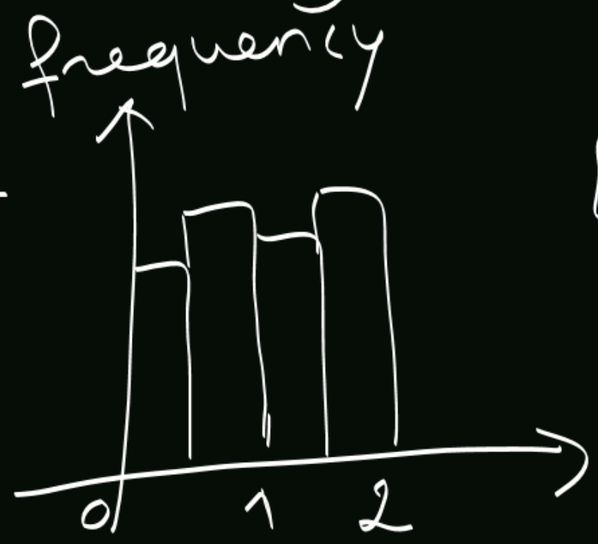
mode = | value associated to the biggest frequency — discrete
| class associated to the Biggest height — continuous


mode

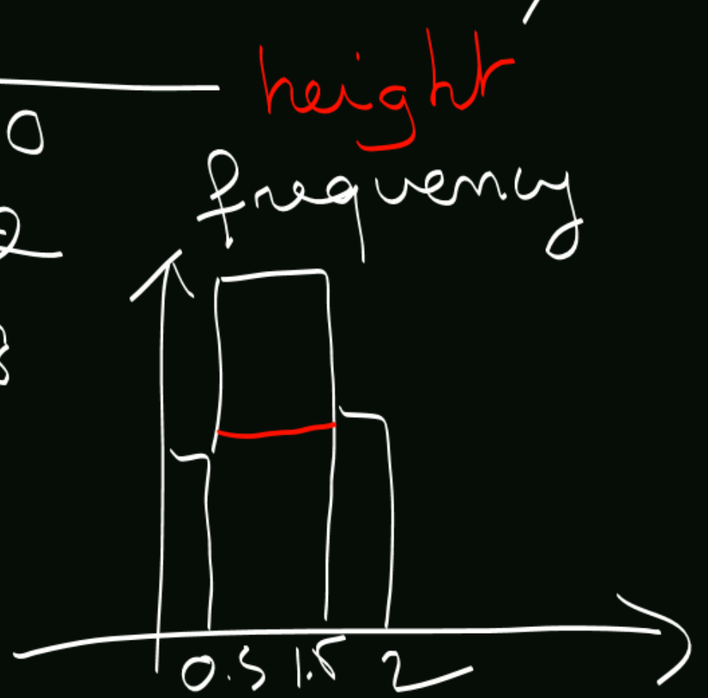
Rk :

Suppose you have observations that come from a $\mathcal{U}([0;2])$

$[0;0.5[$	20
$[0.5;1[$	28
$[1;1.5[$	24
$[1.5;2[$	28



$[0;0.5[$	20
$[0.5;1.5[$	52
$[1.5;2[$	28



discrete

	frequency
a_1	n_1
⋮	
a_k	n_k

- **mean**

2 formula:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{k} n_i a_i$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i c_i$$

where c_i is the midle of the class $[a_{i-1}; a_i[$

- median

property: the median should be a possible value.

ex 0 0 1 1 1 2 2 3 4 4

⇒ ~~med = 1.5~~

- discrete

Median is the value associated to cumulative percent which is immediately bigger or equal to 0.5.

values	cumulative values
a_1	0.1
a_2	0.4
a_3	0.6
a_4	0.7
a_5	1

Median $= a_3$

- continuous

× to find the find the class where the median is

This class is the one associated to the cumulative percent which is immediately bigger or equal to 0.5

Let F be the function associated to the cumulative percent.

Let $[a, b[$ the class where the median is:

$$F(a) < 0.5 \leq F(b)$$

$$\| \quad \|$$

$$F(\text{median})$$

- to evaluate the median we perform some interpolation by a linear function on $[a, b[$

$$\begin{cases} F(a) = p \times a + q \\ F(b) = p \times b + q \end{cases}$$

p and q are unknown!

We solve:

$$0.5 = p \times \text{Median} + q$$

We find the Median

$$\Rightarrow \text{Median} = a + (b-a) \times \frac{0.5 - F(a)}{F(b) - F(a)}$$