# Statistical analysis of the Stigmer data

## Contents



Packages needed:

```
require(tidyverse)
require(rpart)
require(rpart.plot)
```

Vocabulary used:

- there are three **rules** : R0, R1 and R2,
- a session is constituted of 15 **rounds**,
- Each round is constituted of 5 **steps**,
- The **gain per step** is equal to 0, 15 (or 50), 99,
- The **gain per round** is the sum of the 5 gains obtained during a round,
- The **final gain** is the sum of the gains obtained by a player during a session,
- A **cell** is one of the 9 cells of the grid,
- Under R1, a player let a **coin** at each cell visited,
- Under R2, a player has the choice to let or not a coin at the visited cell.

# 1 Importation of the data

The function *import_files()* (codes presented in .Rmd file) leads us to import the data depending on the intermediate value in the game (15 or 50).

To import the data, we call the previous function:

```
file_15 <- import_files(15)
file_50 <- import_files(50)
```

## 1.1 Detection of anomalies

These anomalies have been detected in the version of the data given on March the 5th 2018

### 1.1.1 Stigmer 15

- In: Rule 1, session 03, in round "7B", a player is called 4B instead of B4. We correct this anomaly:

```r
file_15[file_15$player == "4B", "player"] <- "B4"
```

- In: Rule 2, session 01, we found two rounds with name "14A" : we kept only one round (the one with the date the earlier).
- In: Rule 2, session 03, there is one missing row in round "11A". A player did not play at step 1. We impute this missing value by considering that he had a higher probability to get a 0. We correct this:

```r
file_15[is.na(file_15$gain_1), c("gain_1", "action_1", "gain")] <-
    c(0, 0, 114)
file_15[is.na(file_15$gain_1), "cell_1"] <- "1_1"
```

- In: Rule 2, session 01bis, in round "14", a player is called "A1" instead of "B4". We correct this :

```r
file_15[file_15$player == "A1" & file_15$session == "session_01Bis" &
        file_15$rule == "R2" & file_15$round == "T14" & file_15$gain_2 == 0 , "player"] <- "B4"
```

- In: Rule 2, session 01bis, in round "15", a player is called "B4" instead of "A1". We correct this :

```r
file_15[file_15$player == "B4" & file_15$session == "session_01Bis" &
        file_15$rule == "R2" & file_15$round == "T15" & file_15$gain_2 == 0 , "player"] <- "A1"
```

- In Rule 1, players should have let a coin at each step. We notice that this is not necessarly the case :
  - player A2 in session 01, round "1A", "3A", "5A", "12A", "13A",
  - player B1 in session 02, round "6B", "7B", "8B", "12B",
  - player B2 in session 01, round "6B",
  - player A2 in session 02, round "11A",
  - player A1 in session 03, round "1A",
  - player B4 in session 03, round "6B".

**Remark:** no corrections have been applied because in Rule 1, we implicitely suppose that all players have indicated their choice.

- We now identify the players who did not play value 99 although they knew where it was located. For doing this, we create the function *detect_bad_player()* (codes presented in .Rmd file) :

We print the bad players :

```r
detec_bad_player(file_15)
```

```
## ** Bad players in R0
## In session_03 , player A5 has played:
##     round gain_1 gain_2 gain_3 gain_4 gain_5
## 155   T01      0     99     15      0      0
## In session_03 , player B3 has played:
##     round gain_1 gain_2 gain_3 gain_4 gain_5
## 158   T01      0      0     15     99     15
## In session_03 , player B4 has played:
##     round gain_1 gain_2 gain_3 gain_4 gain_5
## 179   T03      0     15     15     99      0
## 209   T06     15     99     15      0      0
## 229   T08     15      0     99      0     15
## 249   T10      0     15     99      0      0
## 259   T11      0     15     15     99      0
## 279   T13     99     15      0      0      0
## In session_03 , player B5 has played:
##     round gain_1 gain_2 gain_3 gain_4 gain_5
## 230   T08     99     99     99     99      0
```

```
##
## ** Bad players in R1
## In session_01 , player A3 has played:
##     round gain_1 gain_2 gain_3 gain_4 gain_5
## 3     T01      0      0     99     15      0
## 13    T02      0      0     99      0     15
## 23    T03      0     15     15     99     15
## 33    T04     99     15      0      0      0
## 53    T06     15      0     99      0      0
## 63    T07     15     99      0      0      0
## 73    T08     15     15      0     99      0
## 83    T09      0      0     99     15      0
## 103   T11      0      0     99     15     15
## 113   T12     15     99     15      0     15
## 133   T14     15      0     99      0      0
## 143   T15      0      0     15     99     15
## In session_03 , player B4 has played:
##     round gain_1 gain_2 gain_3 gain_4 gain_5
## 319   T02      0     99      0      0     15
## 366   T07      0     15     15     99      0
## 389   T09      0      0     99     15      0
## 429   T13      0     15     99      0     15
## 449   T15     99     15      0     15      0
##
## ** Bad players in R2
## In session_01 , player A3 has played:
##   round gain_1 gain_2 gain_3 gain_4 gain_5
## 3   T01     15      0     99      0      0
## In session_03 , player B4 has played:
##     round gain_1 gain_2 gain_3 gain_4 gain_5
## 479   T03      0     99      0      0     15
## 489   T04      0     15      0     99      0
## 499   T05      0     99      0     15     15
## 559   T11     99      0      0     15      0
## 569   T12      0      0     99     15      0
## 599   T15     15     99      0      0     15
```

**Remark:** we do not apply any corrections for these individuals and keep them in the analysis.

#### 1.1.2 Stigmer 50

- In : Rule 0 / session 01 / round “3B”, player B4 did not play at all during this game. Correction done: none.
- In Rule 1, players should have indicated all their choices (value 1). We notice that this not necessary the case :
  - player B2 in session 01, round “1B”,
  - player A3 in session 01, round “3A”, “5A”,
  - player A4 in session 01, round “1A”,
  - player A3 in session 01, round “2A”, “3A”, “4A”.

**Remark:** no corrections have been applied because in Rule 1, we implicitely suppose that all players have indicated their choice.

- We now identify the players who did not play value 99 although they know where it is located :

```
detec_bad_player(file_50)
```

```
## ** Bad players in R0
## In session_01 , player A1 has played:
##    round gain_1 gain_2 gain_3 gain_4 gain_5
## 30   T04     99      0     50     99     99
## 40   T05     50      0     99      0      0
## 60   T07      0      0     99      0      0
## In session_02 , player A2 has played:
##     round gain_1 gain_2 gain_3 gain_4 gain_5
## 171   T03      0      0     99     50     50
## 181   T04     50     50     99      0      0
## 201   T06     50      0     99     50     50
## 261   T12     99     50      0      0      0
## 271   T13      0      0      0     99     50
## 291   T15      0     50     99      0     50
## In session_02 , player B3 has played:
##     round gain_1 gain_2 gain_3 gain_4 gain_5
## 187   T04     50      0     99     50     50
## 207   T06      0     50     99      0     50
## 227   T08     99      0     50     50     50
## 237   T09     99     99     50     50     50
##
## ** Bad players in R1
## In session_02 , player B3 has played:
##     round gain_1 gain_2 gain_3 gain_4 gain_5
## 165   T02      0     99      0     99     99
## 218   T07      0     99     99     99     50
##
## ** Bad players in R2
## In session_01 , player B2 has played:
##     round gain_1 gain_2 gain_3 gain_4 gain_5
## 127   T13     99     50     99     99     99
```

**Remark:** we do not apply any corrections for these individuals and keep them in the analysis.

# 2 Statistical analysis of Stigmer 15

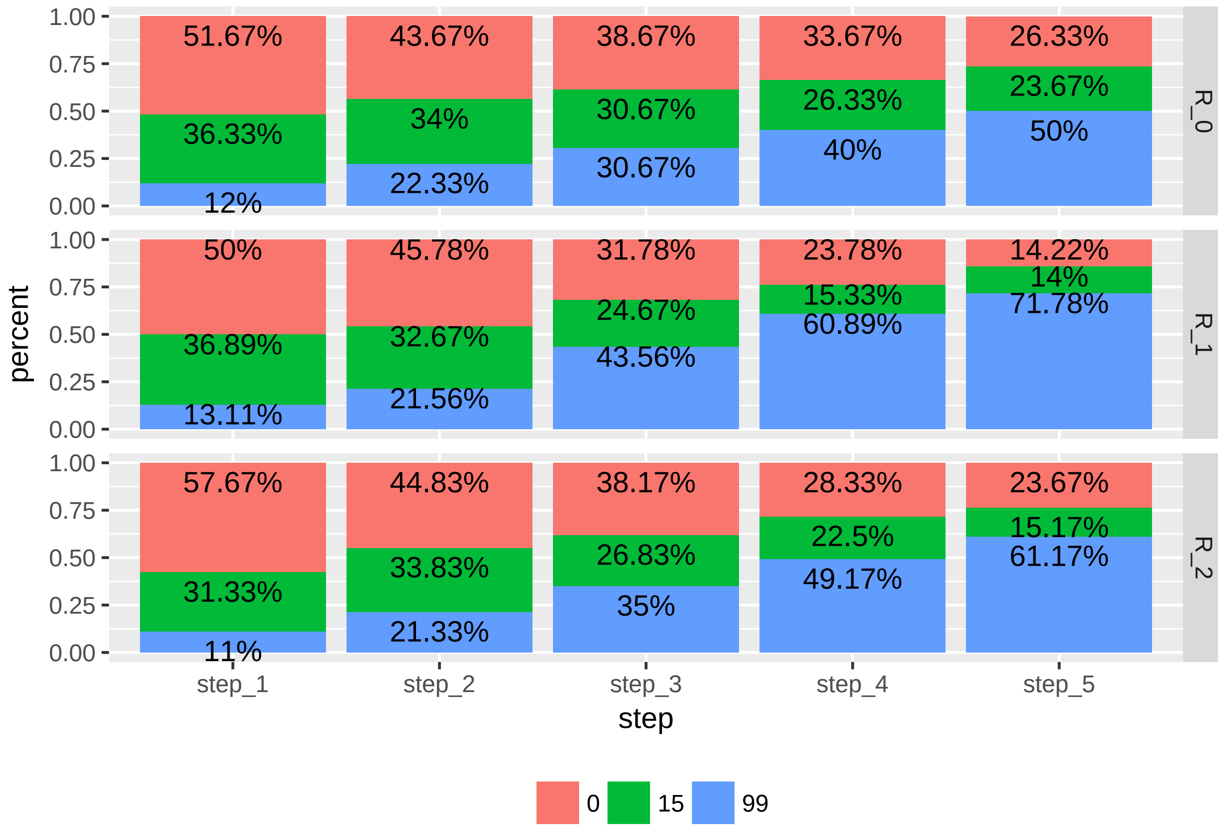
## 2.1 Statistical distribution of the gain per step

The idea here is to represent the empirical probability to get one of the values 0, 15 (or 50), 99 at each of the 5 steps, depending on the rule (0, 1 or 2).

### 2.1.1 Representation

We create the function *distib_per_step()* (codes presented in .Rmd file) which plots the statistical distribution depending on the Rule (R0, R1, R2) and the steps (1, 2, 3, 4 or 5).

```
distib_per_step(fic = file_15, game = 15)
```




**Remark:** at step 1 and 2, the distributions seem to be the same under R0, R1 and R2. From step 3, we remark some differences between the rules. For example, at step 3 under R0, the probablity to find 99 or 15 is the same and slightly smaller than the probability to find 0. On contrary, under R1, the probability to find 99 is higher than the probabilities to get 0 neither 15. The distribution under R2 seem to be a mixture of the distributions under R0 and R2.

We now present the analysis rule by rule.

## 2.2 Analysis under R0

We select the corresponding rows:

```
file_15_R0 <- file_15[file_15$rule == "R0", ]
```

Under R0, there are :

- 2 sessions,
- There are 15 rounds in a session,
- For each session, there are two parallel games: 5 players called $A1, ..., A5$ and 5 players called $B1, ..., B5$, which means that there are 20 different players.

In this section, we try to identify what is the best strategy for players to optimize their profit. Then, we look if the players did adopt such a strategy and if we can see differences in terms of gain between the different strategies.

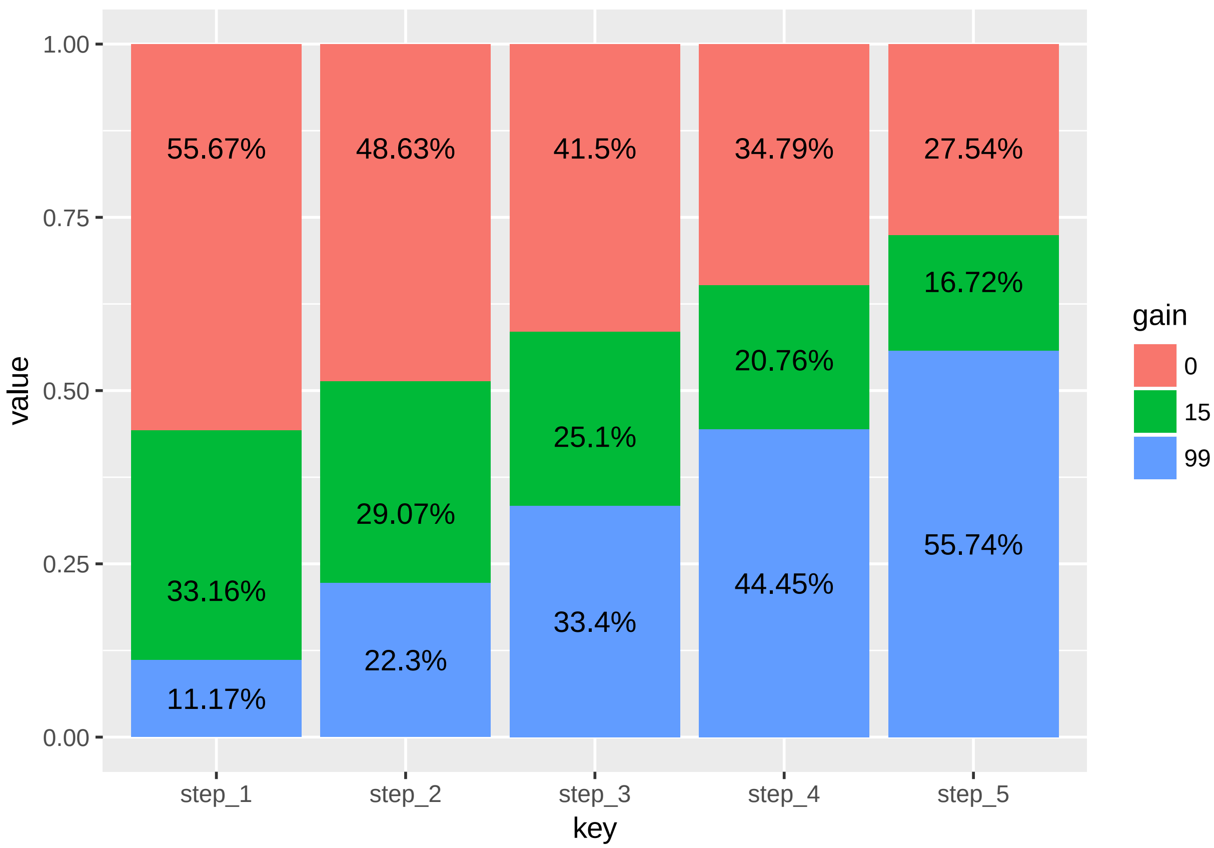
### 2.2.1 Probability to get 0, 15, 99 under R0: comparaison between empirical and expected distribution

Under R0, players have no information about the rest of the group. Thus, to maximize his profit, a player should explore a new cell at each new step until he finds the value 99 (this could be prooved by using theoretical probability). Thus, we can simulate a high number of times this kind of behaviour, such that we have the expected distribution. We program the function *simu_distrib_R0()* (codes presented in .Rmd file) for doing this task.

We simulate the behaviours of 100000 players (codes presented in .Rmd file).

We tidy the data (codes presented in .Rmd file).

We finally plot the theoretical distribution (codes presented in .Rmd file):




We compare below the empirical distribution with the theoretical one obtained previously. We use a $\chi^2$ test which consists in comparing at each step the empirical distribution of 0, 15, 99 to the theoretical one computed previously. It seems that the more we progress in the experience, the less the players seem to behave as players who optimize their profit. This is probably due to the fact that some players prefer to conserve their results by playing the value 15 (when they know where it is located) rather than continuing to explore, especially at the end of a round.

**Interpretation of the $\chi^2$ test:** the null hypothesis is "Distribution of empirical and theoretical are identical". When the p-value is lower than 0.05, we usually reject the null hypothesis. If the p-value is upper than 0.05, we cannot reject it. In our case, the distributions (emprirical VS theoretical) are the same at step 1, 2, 3 and slightly different at step 4, 5.

```
chisq.test(table(as.factor(file_15_R0[, "gain_1"])),
           p = tab_15[, 1])
```

```
##
##  Chi-squared test for given probabilities
##
## data:  table(as.factor(file_15_R0[, "gain_1"]))
## X-squared = 1.9626, df = 2, p-value = 0.3748
```

```
chisq.test(table(as.factor(file_15_R0[, "gain_2"])),
           p = tab_15[, 2])
```

```
##
##  Chi-squared test for given probabilities
##
## data:  table(as.factor(file_15_R0[, "gain_2"]))
## X-squared = 4.0281, df = 2, p-value = 0.1334
```

```
chisq.test(table(as.factor(file_15_R0[, "gain_3"])),
           p = tab_15[, 3])
```

```
##
##  Chi-squared test for given probabilities
##
## data:  table(as.factor(file_15_R0[, "gain_3"]))
## X-squared = 4.9627, df = 2, p-value = 0.08363
```

```
chisq.test(table(as.factor(file_15_R0[, "gain_4"])),
           p = tab_15[, 4])
```

```
##
##  Chi-squared test for given probabilities
##
## data:  table(as.factor(file_15_R0[, "gain_4"]))
## X-squared = 5.9353, df = 2, p-value = 0.05142
```

```
chisq.test(table(as.factor(file_15_R0[, "gain_5"])),
           p = tab_15[, 5])
```

```
##
##  Chi-squared test for given probabilities
##
## data:  table(as.factor(file_15_R0[, "gain_5"]))
## X-squared = 10.587, df = 2, p-value = 0.005024
```

### 2.2.2 Analysis of the gain per round under R0

We are now interested in the gain per round (the sum of the 5 gains obtained during a round).

#### 2.2.2.1 Comparaison between empirical and expected distribution

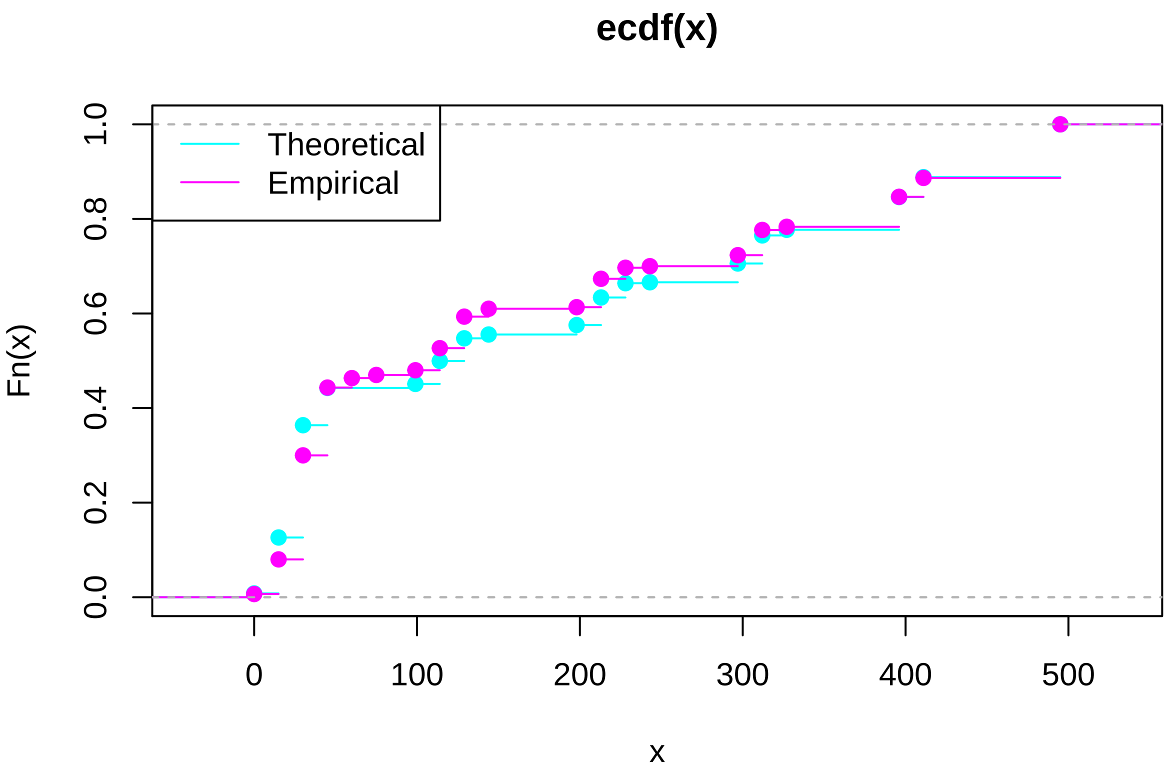
Thanks to the previous simulation, we can obtain the statistical distibution of the gain per round under R0 when players adopt an optimal strategy

Example of interpretation of the empirical cumulative distribution function :

- the probability to obtain a gain per round lower or equal to 30 is 36.41%.
- 49.925% of the population obtained a value lower or equal to 114. On contrary, $100 - 49.925 = 50.075\%$ obtained a gain per round strictly larger than 114.

```r
cumsum(prop.table(table(final_gain_15)))
```

```
##       0      15      30      45      99     114     129     144     198
## 0.00799 0.12628 0.36374 0.44260 0.45107 0.49963 0.54748 0.55552 0.57549
##     213     228     243     297     312     327     396     411     495
## 0.63381 0.66402 0.66599 0.70571 0.76507 0.77700 0.84631 0.88831 1.00000
```




The plot of the empirical cumulative distribution function ("theoretical" versus "empirical") does not show a big difference between the two distributions since the two curves (cyan vs magenta) seem close.

We use the Kolmogorov-Smirnov test to verify that the two distributions (empirical and theoritical) are extracted from the same distribution or not. The null hypothesis which is "the two distributions are identical" canno't be rejected here. In other term, the gain per round obtained by players could be the ones obtained by players who optimize their profit.

```r
ks.test(x = final_gain_15,
        y = file_15_R0[, "gain"],
        exact = F)
```

```
## Warning in ks.test(x = final_gain_15, y = file_15_R0[, "gain"], exact = F):
## p-value will be approximate in the presence of ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  final_gain_15 and file_15_R0[, "gain"]
## D = 0.06374, p-value = 0.1759
## alternative hypothesis: two-sided
```

**Remark:** note that the Kolmogorov-Smirnov test is supposed to be used on continous variable which is not the case here (the gain per round is indeed discrete). A solution could be to use instead a $\chi^2$ test. However, we can not use it for the following reason : in the optimal situation (theoretical distribution), players might not be able to obtain a gain per round equal to 60 or 75 (a player is exploring new cases unless he founds 99, so he can only find the value 15 one, two or three times). Besides, the probability to obtain these values 60 or 75 are equal to 0, although these situations can occur in the experimental context (players who play several times the same cell containing 15). As the $\chi^2$ test can be seen as the sum of the distances between theoretical and empirical values divided by the theoretical probability, we canno't use theoretical probability equal to 0.

#### 2.2.2.2 Exploring or taking no risk ?

We show previously that players seem to behave as players who maximise their profit (exploring until the value 99 is found). However, we are interested in checking if all players behave like that and if not, could we notice differences in the final gain. For each player, we compute the final gain after the 15 rounds and compute the percentage of time they have chosen to continue to explore new cells until they find 99.

For doing that, we program the function *exploring()* (codes presented in .Rmd file) that will be also used in next sections. This functions permits to identify players who have the opportunity to explore at a given step and do it (or not).

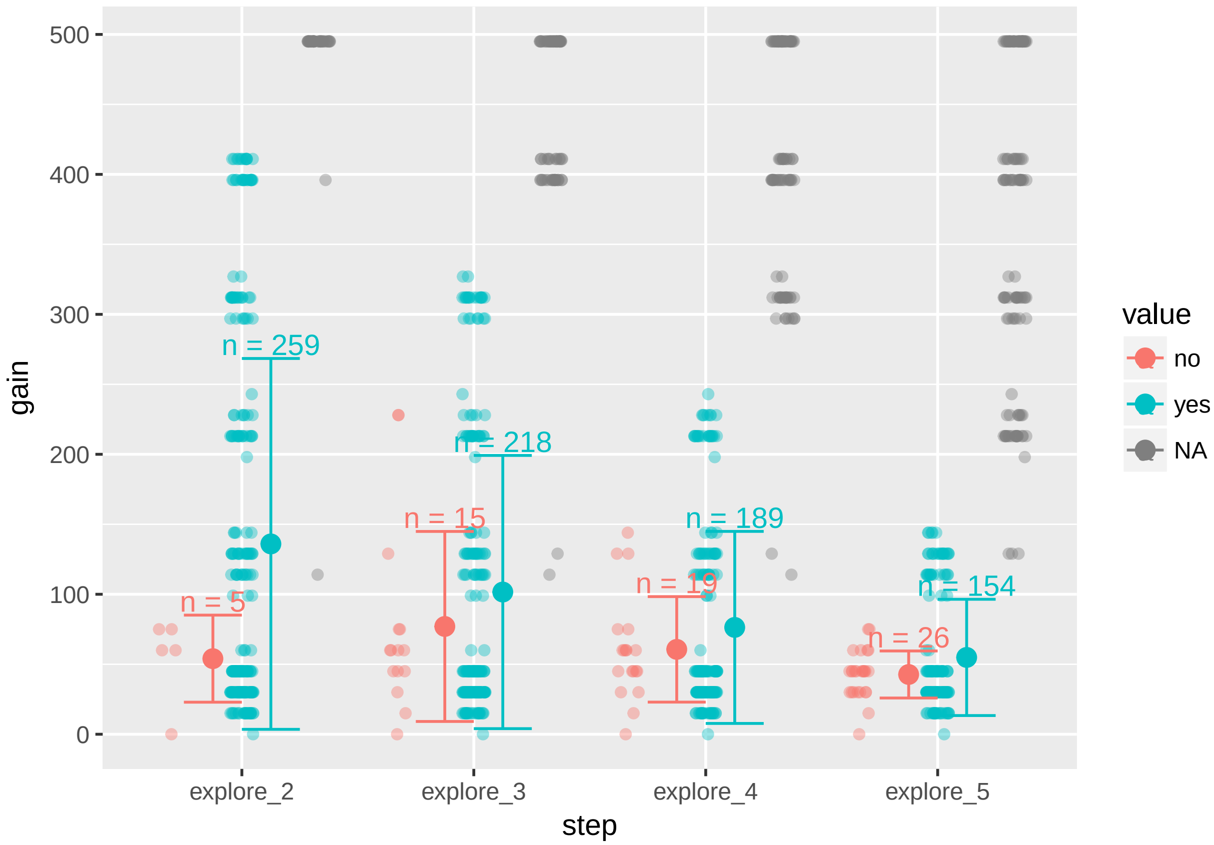
**Remark:** a player who does not explore is necessarly a player who plays a cell which has already been visited and does not contain the value 99. Typically, one could think about a player who prefers to play again the cell 15 rather exploring.

Now, we propose to compare the gain per round obtained depending on the fact that a player had the opportunity to explore new cells.

First, we tidy the data (codes presented in .Rmd file).

Then, we compute the mean and standard deviation obtained at each step and depending on the fact that a player explored new cells or not (codes presented in .Rmd file).

Finally, we plot in y-axis the gain per round and in x-axis the steps. We represent in blue (resp. in red) the gain per round obtained when a player explored new cells (resp. when he did not explore) knowing that the player had the opportunity to do it. We plot the gain per round in grey obtained by people who did not have the opportunity to explore (that means that they know where the cell 99 is located.




**Interpretation of the plot:**

- at each step there are much more people who prefer exploring ($n = 259, 218, 189, 154$ at step 2,3,4,5) rather taking no risk ($n = 5, 15, 19, 26$ at step 2,3,4,5).
- players who explore seem to obtain "in average" a higher gain per round than people who does not explore.
- the range of the standard error of the red part is always included in the range of the standard error of the blue part which seems to indicate that the differences between the means are not significant.

**Statistical test:**

To test the hypothesis of equality of means at each step, we use a non parametric test called "Kruskal-Wallis Rank Sum Test". It shows us that the differences of the means between players who explore and players who do not explore are non significant at any step.

```
(kruskal.test(gather_file_15$gain[gather_file_15$step == "explore_2"] ~
                            factor(gather_file_15$value[gather_file_15$step == "explore_2"])))
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  gather_file_15$gain[gather_file_15$step == "explore_2"] by factor(gather_file_15$value[gather_
## Kruskal-Wallis chi-squared = 0.40938, df = 1, p-value = 0.5223
```

```
(kruskal.test(gather_file_15$gain[gather_file_15$step == "explore_3"] ~
                            factor(gather_file_15$value[gather_file_15$step == "explore_3"])))
```

```
##
##  Kruskal-Wallis rank sum test
```

```
##
## data:  gather_file_15$gain[gather_file_15$step == "explore_3"] by factor(gather_file_15$value[gather_
## Kruskal-Wallis chi-squared = 0.049873, df = 1, p-value = 0.8233
```

```
(kruskal.test(gather_file_15$gain[gather_file_15$step == "explore_4"] ~
                            factor(gather_file_15$value[gather_file_15$step == "explore_4"])))
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  gather_file_15$gain[gather_file_15$step == "explore_4"] by factor(gather_file_15$value[gather_
## Kruskal-Wallis chi-squared = 0.54774, df = 1, p-value = 0.4592
```

```
(kruskal.test(gather_file_15$gain[gather_file_15$step == "explore_5"] ~
                            factor(gather_file_15$value[gather_file_15$step == "explore_5"])))
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  gather_file_15$gain[gather_file_15$step == "explore_5"] by factor(gather_file_15$value[gather_
## Kruskal-Wallis chi-squared = 0.27042, df = 1, p-value = 0.603
```

### 2.2.3 The effect of exploring (or not) on the final gain

In this section, we aggregate the results per player and consider thus the final gain. Indeed, we suspect that when a player adopts the same strategy during the 15 rounds, this probably has a effect on the final gain.

#### 2.2.3.1 The expected "optimal" final gain

Thanks to the following theoretical distribution of the gain per round :

```
## final_gain_15
##       0      15      30      45      99     114     129     144     198
## 0.00799 0.11829 0.23746 0.07886 0.00847 0.04856 0.04785 0.00804 0.01997
##     213     228     243     297     312     327     396     411     495
## 0.05832 0.03021 0.00197 0.03972 0.05936 0.01193 0.06931 0.04200 0.11169
```

we can deducce what is the expected gain per round under R0 for an optimal behaviour. It corresponds simply to $\sum_x xPr[X = x]$ where $x$ corresponds to the possible values of the gain per round. It is here equal to:

```
sum(as.numeric(names(prop_tab)) * prop_tab)
```

```
## [1] 184.108
```

When mutiplying this number by 15, we obtain the theoretical optimal final gain

```
(opti_15_R0 <- 15 * sum(as.numeric(names(prop_tab)) * prop_tab))
```

```
## [1] 2761.62
```

We rank here the final gain obtained by the 20 players :

```
sort(players_15$gain)
```

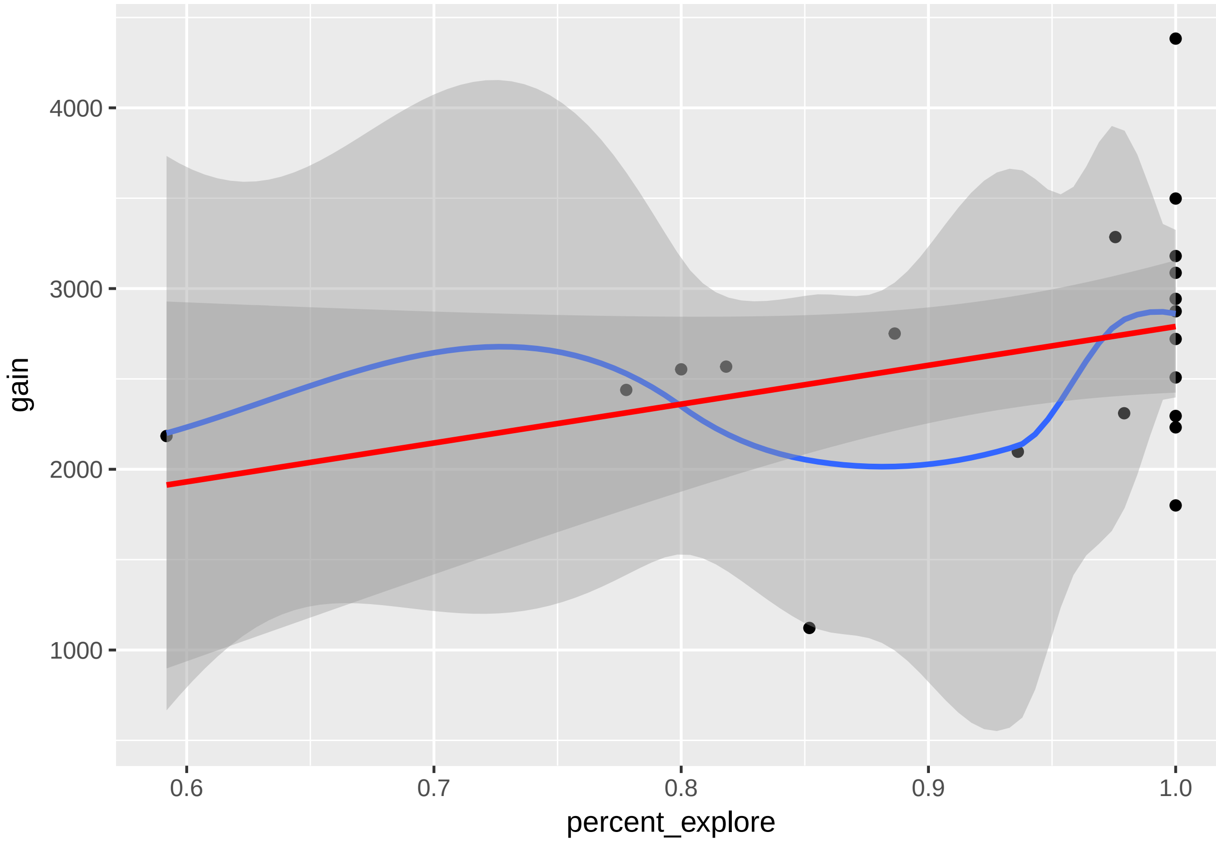
```
##  [1] 1122 1800 2097 2184 2232 2295 2310 2439 2508 2553 2568 2721 2751 2874
## [15] 2943 3087 3180 3285 3498 4383
```

We can remark that there are 7 players who obtained a final gain upper than the theoretical optimal final gain.

#### 2.2.3.2 Linear regression

We plot the final gain depending on the percentage of times a player has explored new cells (knowing that he had the opportunity to do it).

We then plot the data :




We remark that :

- 11 players (on a total of 20) systematically explored when they had the opportunity to do it,
- the player with the highest score is a player who had explored new cells every time he had the opportunity to do it,
- the player with the lowest score is a player who had not explored new cells every time he has the possibility to do it
- the linear regression line seems to indicate that the more the player explore new cells the more his final gain will be higher. However, it seems difficult to give interestings results because globally, players tend to explore new cells when they can do it. Only few of them did not adopt such a strategy.

### 2.2.4 Conclusion under R0

Under R0, players had two main choices :

- exploring until they find 99 (the best strategy),
- takes no risk and plays again the value 15 when they already found it.

It seems that player globally tend to optimize their profit by exploring new cells until they find the value 99. Some of them have a tendency to keep the value 15 rather exploring. In theory, the first strategy is the best. However, because there are only a few observations in our experience, we could not detect significant differences between these two main behaviours at the different levels of observations (per round or per session).

## 2.3 Analysis under R1

We select the corresponding rows:

```r
file_15_R1 <- file_15[file_15$rule == "R1", ]
```

Under R1, there are :

- 3 sessions,
- There are 15 rounds in a session,
- For each session, there are two parallel games: 5 players called $A1, ..., A5$ and 5 players called $B1, ..., B5$,which means that there are 30 different players.

In this section, we try to identify what is the best strategy for players. Under R1, players can use an additionnal information related to the frequency of visited cells. We try here to understand what is the best way to use this additional information and check if players do adopt this strategy.

### 2.3.1 Probability to get 0, 15, 99 under R1: comparaison between empirical and expected distribution

#### 2.3.1.1 Intuitive best strategy under R1

In this section we try to understand what could be the best strategy for players under R1.

##### 2.3.1.1.1 At step 1

Players explore like under R0.

##### 2.3.1.1.2 At step 2

We think about at least 3 different strategies :

- Strategy 1: a player (who did not find 99 yet) explores new cells like under R0 (he does not take into account what the other players have done).
- Strategy 2: a player (who did not find 99 yet) explores new cells among the cells which have not been visited at step 1 by the other players.

We do simulations to understand better what are the differences between these two strategies and awnser to several questions (see codes in the .Rmd file).

**Q1:** what is the probability that at least one player found the value 99 ?

- in strategy 1, this probability is equal to :

```r
mean(res_sim_R1_s1[3, ])
```

```
## [1] 0.71453
```

- in strategy 2, this probability is equal to :

```r
mean(res_sim_R1_s2[3, ])
```

```
## [1] 0.81452
```

**Q2:** what is the probability that the cell which has been the most visited is the cell with value 99?

- in strategy 1, this probability is equal to :

```
sum(table(res_sim_R1_s1[1, ], res_sim_R1_s1[2, ])[,2])/100000
```

```
## [1] 0.28693
```

- in strategy 2, this probability is equal to :

```
sum(table(res_sim_R1_s2[1, ], res_sim_R1_s2[2, ])[,2])/100000
```

```
## [1] 0.28912
```

**Q3:** what is the probability to find 99 when playing the $k$th cell the most visited ?

- interpretation with strategy 1 : the proba to find 99 by playing the most visited cell is equal to 28.7%. After playing the two most visited cell, this prbo is equal to 51.7%

```
##       1       2       3       4       5       6       7       8       9
## 0.28693 0.51725 0.61129 0.70112 0.72275 0.78182 0.90695 0.98790 1.00000
```

- in strategy 2 : the proba to find 99 after playing the two most visited cell is higher than in strategy 1.

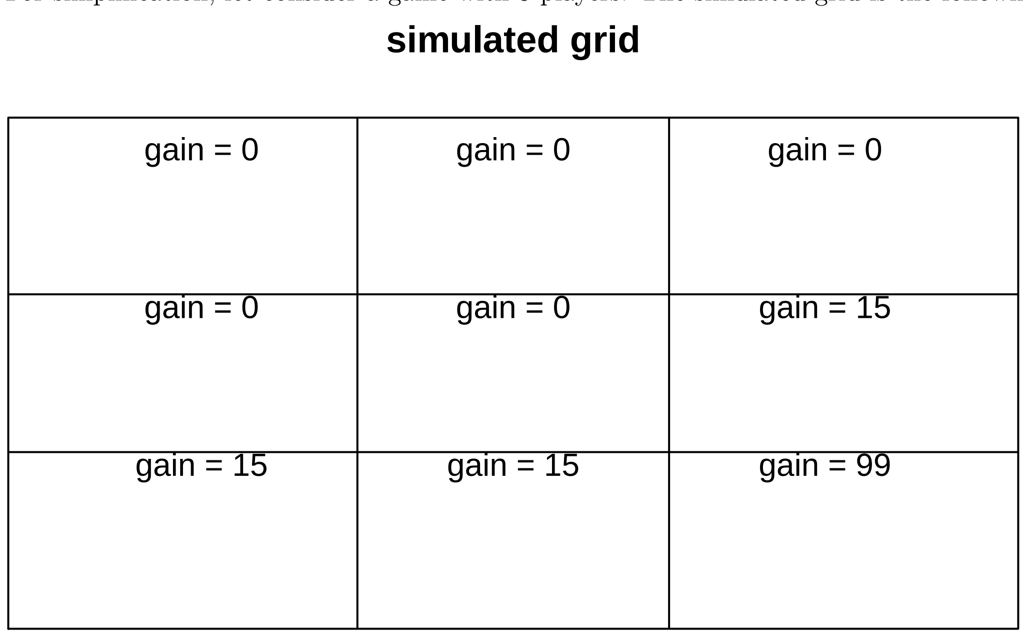
```
##       1       2       3       4       5       6       7       8       9
## 0.28912 0.63899 0.76593 0.81098 0.81485 0.82355 0.87076 0.95918 1.00000
```

**Conclusion:** strategy 2 seems more interesting than strategy 1. However, it could be than a mixture between both strategy (some players who adopt strategy 1 and some others who adopt strategy 2).

#### 2.3.1.1.3 At step 3

A players chooses the cell which has been the most visited after the second step. Here we show why it seems the best solution to do that.

For simplification, let consider a game with 5 players. The simulated grid is the following one :




At the first tour, players are supposed to explore independantly the game. The expectation of the number of coins per cell is the following one (as there are 5 players, the sum of the expectations is equal to 5) :

## Information given after step 1

| | | |
|---|---|---|
| gain = 0<br>coin = 0.556 | gain = 0<br>coin = 0.556 | gain = 0<br>coin = 0.556 |
| gain = 0<br>coin = 0.556 | gain = 0<br>coin = 0.556 | gain = 15<br>coin = 0.556 |
| gain = 15<br>coin = 0.556 | gain = 15<br>coin = 0.556 | gain = 99<br>coin = 0.556 |

At the second step, players who found 99 are supposed to play 99 again. The others players are supposed to explore new cells with a probability equal to 1/8 and the expected number of coins let in each cell is thus equal to $5 \times (1 - 1/9) \times 1/8$. After the second step, we sum the expected coins at the 1st and 2nd tour, such that we obtain the expected information which is given to all players before the 3rd step :

## Proba at step 2

| | | |
|---|---|---|
| gain = 0<br>coin = 0.5555 | gain = 0<br>coin = 0.5555 | gain = 0<br>coin = 0.5555 |
| gain = 0<br>coin = 0.5555 | gain = 0<br>coin = 0.5555 | gain = 15<br>coin = 0.5555 |
| gain = 15<br>coin = 0.5555 | gain = 15<br>coin = 0.5555 | gain = 99<br>coin = 1.1115 |

## Information given after step 2

| | | |
|---|---|---|
| gain = 0<br>coin = 1.1115 | gain = 0<br>coin = 1.1115 | gain = 0<br>coin = 1.1115 |
| gain = 0<br>coin = 1.1115 | gain = 0<br>coin = 1.1115 | gain = 15<br>coin = 1.1115 |
| gain = 15<br>coin = 1.1115 | gain = 15<br>coin = 1.1115 | gain = 99<br>coin = 1.6675 |

We notice that the cell 99 might be the cell with the maximum expected number of coins let at the end of step 2. This suggests that to optimize their profit in R1, players should play the cell which has been the most visited after the second step.

Moreover, the simulation results obtained at step 2 tend to indicate that if a player has already visited the cell the most visited and knows that it does not contain 99, he should try to play the second most visited cell.

Finally, if the player knows than the second most visited is not the good one (he has visited the two most visietd cells ank know that there do not contain 99), than he should visits a new cell.

##### 2.3.1.1.4 At step 4

Like in step 3. If players has already played the three first most visited cell and he knows that they do not contain 99, he explores a new cell (like in strategy 2).

##### 2.3.1.1.5 At step 5

Like in step 4. If players has already played the four first most visited cell and he knows that they do not contain 99, he explores a new cell (like in strategy 2).

#### 2.3.1.2 Simulation of the optimal behaviour of players under R1

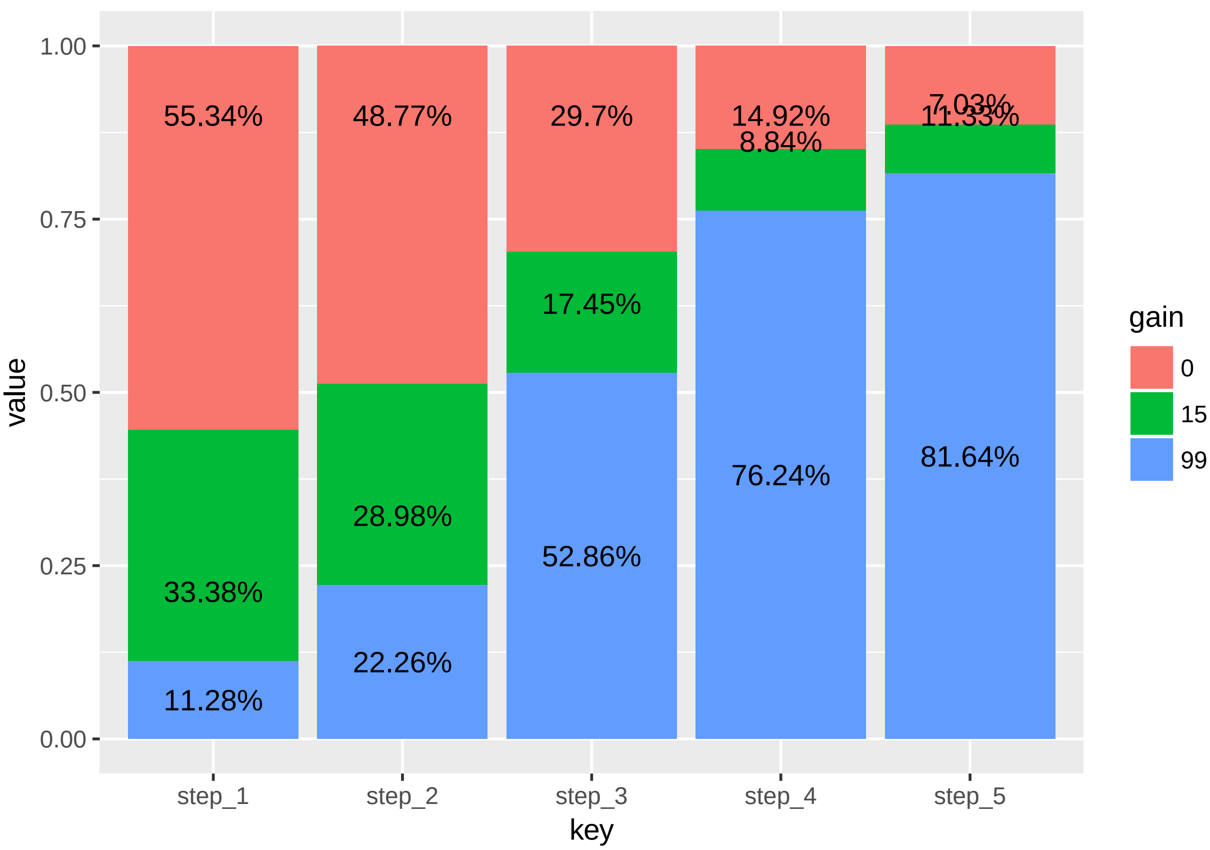
We would like to have the theoretical statistical distribution for players who optimize their profit under R1.

For doing that, we do simulations by reproduccing round per round the behaviours of such players. A party is in 5 steps with 5 players. We suppose that the players adopt these two strategies :

- At step 1 and 2, they explore new cells until 99 has been found,
- After step 2, they choose to play the cell which has been the most visited :
  - If several cells have been visited the same number of times, the cell is choosen randomly.
  - If the most visited cell has already been visited (hence, player knows that it does not contain 99), the player choose to play the second most visited, then 3rd, etc…

This function is called *simu_distrib_R1()* (codes presented in .Rmd file).

We obtain this distribution. Note that at step 1 and 2, distributions are obviously identical to the ones obtained in Rule 0.




We compare below the empirical distribution with the theoretical one obtained previously. We use a $\chi^2$ test which consists in comparing at each step the empirical distribution of 0, 15, 99 to the theoretical one computed previously. At step 2, players behave as we can expect. At step 3, 4 and 5, the p-values of the test are very close to 5% which indicates that players behave not completely as players who optimize their profit (note however that the differences are not so important).

```
chisq.test(table(as.factor(file_15_R1[, "gain_2"])),
           p = tab_15_R1[, 2])
```

```
##
##  Chi-squared test for given probabilities
##
## data:  table(as.factor(file_15_R1[, "gain_2"]))
## X-squared = 3.0361, df = 2, p-value = 0.2191
```

```
chisq.test(table(as.factor(file_15_R1[, "gain_3"])),
           p = tab_15_R1[, 3])
```

```
##
##  Chi-squared test for given probabilities
##
## data:  table(as.factor(file_15_R1[, "gain_3"]))
## X-squared = 21.472, df = 2, p-value = 2.175e-05
```

```
chisq.test(table(as.factor(file_15_R1[, "gain_4"])),
           p = tab_15_R1[, 4])
```

```
##
##  Chi-squared test for given probabilities
##
## data:  table(as.factor(file_15_R1[, "gain_4"]))
## X-squared = 59.069, df = 2, p-value = 1.49e-13
```

```
chisq.test(table(as.factor(file_15_R1[,"gain_5"])),
           p = tab_15_R1[, 5])
```

```
##
##  Chi-squared test for given probabilities
##
## data:  table(as.factor(file_15_R1[, "gain_5"]))
## X-squared = 39.784, df = 2, p-value = 2.296e-09
```

### 2.3.2 Analysis of the gain per round under R1

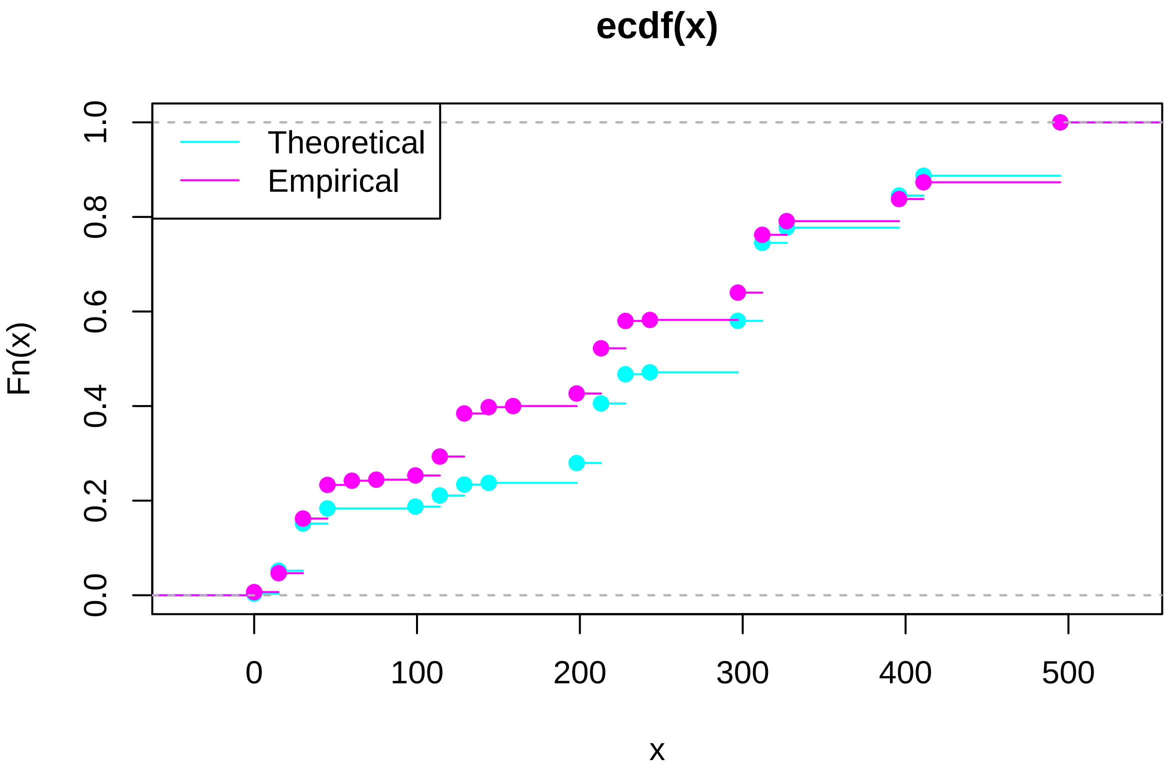
We are now interested in the gain per round (the sum of the 5 gains obtained during a round).

#### 2.3.2.1 Comparaison between empirical and expected distribution

Thanks to the previous simulation, we can obtain the statistical distibution of the final gain under R1 when players adopt an optimal behaviour.

The plot of the empirical cumulative distribution function ("theoretical" versus "empirical") does not show a big difference between the two distributions.

```
##       0      15      30      45      99     114     129     144     198
## 0.00326 0.05205 0.15166 0.18359 0.18756 0.21101 0.23411 0.23756 0.27955
##     213     228     243     297     312     327     396     411     495
## 0.40566 0.46738 0.47144 0.58032 0.74524 0.77744 0.84534 0.88724 1.00000
```




The following test confirms the previous plot. Hereafter, the test rejects the hyptohesis of equality of the two dsitributions, but the this difference is not so big.

```
ks.test(x = final_gain_15_R1,
        y = file_15_R1[, "gain"])
```

```
## Warning in ks.test(x = final_gain_15_R1, y = file_15_R1[, "gain"]): p-value
## will be approximate in the presence of ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  final_gain_15_R1 and file_15_R1[, "gain"]
## D = 0.16244, p-value = 1.08e-10
## alternative hypothesis: two-sided
```

#### 2.3.2.2 Copying, exploring or taking no risk ?

The idea of this section is to determine whose players belong to one of these strategy (copying, exploring, taking no risk) and if we can notice differences in terms of gain between them.

**Algorithm (step 1):** to detect players who copy other players at each step, we need to know which is the cell the most visited at the end of a step. We create a matrix which contains for each of the 9 cells, the number of times it has been visited at the end of a step for a given round. Then, we count the number of visits per cell at each step, for given players (A or B) per round and per session (codes presented in .Rmd file).

After, we create the function *cell_visited()* (codes presented in .Rmd file) which could be useful later. It consists in giving for each step the ranking of the most visited cells in a sense of the players have let a coin.

We apply the previous function to our data (codes presented in the .Rmd file).

**Algorithm (step 2):** once we kwow how many times the cells have been visited at each step, one could say if a player decides "yes" or "not" to play the most visited cell. We do the following assumption: if player has already played the most visited cell and knows that it does not contain the value 99, he is supposed to play the second most visited cell, then the third, etc. For doing this, we create the function *copy_or_not()* (codes presented in .Rmd file) which permits to know at step 2, 3, 4 and 5 if a player belongs to one of these categories:

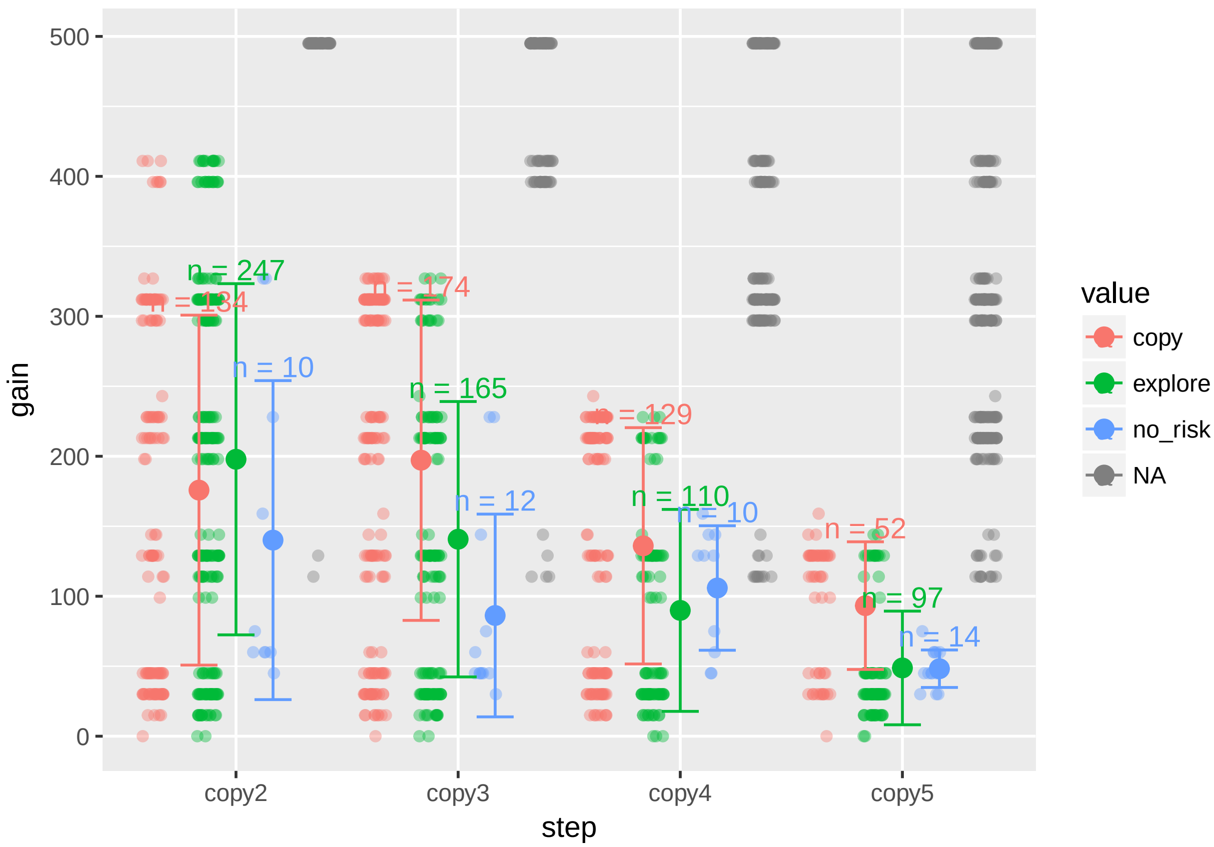
- **copy**: a player visits the cell the most visited (the cell which contains the maximum number of coins),
- **explore**: a player re-played a cell which does contain 15 and that he has already visited,
- **no_risk**: a player does not follow one of the two previous behaviours,

We apply our data to the previous function (codes presented in .Rmd file).

For representing the data, we first tidy the data (codes presented in .Rmd file).

We compute the mean and standard deviation obtained at each step and depending on the fact that a player explored new cells or not (codes presented in .Rmd file).

We plot in y-axis the gain per round and in x-axis the steps. We represent in red (resp. in green resp. in blue) the gain per round obtained when the player played the most visited cell visited (resp. when he explored resp. when he took no risk) knowing that the player had the opportunity to do it. We plot the gain per round in grey obtained by people who did not have the opportunity to visit (that means that they knew where the cell 99 was located).




**Interesting remarks concerning the figure:**

- at step 2, there are more people who prefer exploring ($n = 247$) rather copying ($n = 134$). We can see that at this step, it seems more interesting to explore rather copying. Indeed, at this step, there is absolutely no reason that the most visited cell contains the value 99. People who do copy at this step decreases the probability to find the value 99 at step 2, simply because this causes a decrease of the number of discovered cell.

- at step 3, 4 and 5 it is always more interesting to copy rather exploring or playing the value 15 knowing where it is.
- at step 4 and 5, the differences between "exploring" and "re-playing 15" does not seem as important as it is the case under R0.
- at step 5, there are more players who prefer exploring ($n = 97$) rather copying ($n = 52$).
- The worst case which can happen is a situation where people by copying, do not have the opportunity to explore enough and hence do not sucess to find the value 99. We present here a example where this situation occured :

```
file_15_R1_copy[file_15_R1_copy$round_p == "T06_A_session_02",
                c(paste0("gain_", 1:5), paste0("cell_", 1:5), paste0("copy", 2:5))]
```

```
##     gain_1 gain_2 gain_3 gain_4 gain_5 cell_1 cell_2 cell_3 cell_4 cell_5
## 501      0     15     15      0      0    1_0    0_2    2_2    1_0    0_0
## 502     15      0      0      0     15    0_2    1_2    1_0    2_1    2_2
## 503      0     15      0     15     15    1_1    0_2    1_0    2_2    2_0
## 504     15      0      0      0     15    0_2    0_0    1_0    1_1    2_2
## 505     15      0     15     15      0    2_0    1_0    0_2    2_2    1_1
##       copy2   copy3   copy4   copy5
## 501    copy explore explore explore
## 502 explore    copy explore    copy
## 503    copy    copy    copy explore
## 504 explore    copy explore    copy
## 505 explore    copy    copy    copy
```

To test the hypothesis of equality of the means (here we test simultaneously the equality "copy" = "explore" = "no_risk") at each step, we use a non parametric test called "Kruskal-Wallis Rank Sum Test". It shows us that the differences of the means are significant at step 3, 4 and 5. It is not at step 2.

```
(kruskal.test(gather_file_15_R1$gain[gather_file_15_R1$step == "copy2"] ~
                            factor(gather_file_15_R1$value[gather_file_15_R1$step == "copy2"])))
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  gather_file_15_R1$gain[gather_file_15_R1$step == "copy2"] by factor(gather_file_15_R1$value[ga
## Kruskal-Wallis chi-squared = 1.8803, df = 2, p-value = 0.3906
```

```
(kruskal.test(gather_file_15_R1$gain[gather_file_15_R1$step == "copy3"] ~
                            factor(gather_file_15_R1$value[gather_file_15_R1$step == "copy3"])))
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  gather_file_15_R1$gain[gather_file_15_R1$step == "copy3"] by factor(gather_file_15_R1$value[ga
## Kruskal-Wallis chi-squared = 27.463, df = 2, p-value = 1.088e-06
```

```
(kruskal.test(gather_file_15_R1$gain[gather_file_15_R1$step == "copy4"] ~
                            factor(gather_file_15_R1$value[gather_file_15_R1$step == "copy4"])))
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  gather_file_15_R1$gain[gather_file_15_R1$step == "copy4"] by factor(gather_file_15_R1$value[ga
```

```
## Kruskal-Wallis chi-squared = 21.859, df = 2, p-value = 1.792e-05
```

```r
(kruskal.test(gather_file_15_R1$gain[gather_file_15_R1$step == "copy5"] ~
                            factor(gather_file_15_R1$value[gather_file_15_R1$step == "copy5"])))
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  gather_file_15_R1$gain[gather_file_15_R1$step == "copy5"] by factor(gather_file_15_R1$value[ga
## Kruskal-Wallis chi-squared = 29.096, df = 2, p-value = 4.808e-07
```

### 2.3.3 The effect of copying VS exploring VS re-playing 15 on the final gain

In this section, we aggregate the results per player and consider thus the final gain. There are 30 players and for doing a statistical analysis on these players, we need to define some variables for each player.

#### 2.3.3.1 The expected "optimal" final gain

Thanks to the following theoretical distribution of the gain per round :

```
## final_gain_15_R1
##       0      15      30      45      99     114     129     144     198
## 0.00326 0.04879 0.09961 0.03193 0.00397 0.02345 0.02310 0.00345 0.04199
##     213     228     243     297     312     327     396     411     495
## 0.12611 0.06172 0.00406 0.10888 0.16492 0.03220 0.06790 0.04190 0.11276
```

we can deducce what is the expected gain per round under R1 for an optimal behaviour. It corresponds simply to $\sum_x xPr[X = x]$ where $x$ corresponds to the possible values of the gain per round. It is here equal to:

```r
sum(as.numeric(names(prop_tab)) * prop_tab)
```

```
## [1] 256.1815
```

When mutiplying this number by 15, we obtain the theoretical optimal final gain

```r
(opti_15_R1 <- 15 * sum(as.numeric(names(prop_tab)) * prop_tab))
```

```
## [1] 3842.723
```

We rank here the final gain obtained by the 20 players :

```r
sort(players_15_R1$final_gain)
```

```
##  [1]  954 1761 2295 2745 2844 2928 2934 3033 3108 3171 3201 3225 3354 3384
## [15] 3528 3537 3567 3576 3720 3735 3744 3963 4071 4086 4092 4140 4194 4323
## [29] 4443 4635
```

We can remark that there are 7 players who obtained a final gain upper than the theoretical optimal final gain.

#### 2.3.3.2 Create new variables per player

We create the following variables per player :

- The dependent variable is the final gain.
- The characterestics we are interested in are the behaviours of the players at step 2, 3, 4 and 5. For example, if we look at the player "A1" in session 01, we observe the following behaviour at step 2: he

copied only 2 times whereas he explored 11 times. Hence, this player will be considered as an "explorer" at step 2.

```
##
##    copy explore
##       2      11
```

At step 3, because of the equality (4 times "copy" and 4 times "explore"), we will choose randomly his behaviour :

```
##
##    copy explore
##       4       4
```

**Important remark:** we can suppose that if some players change their behaviour, this is probably due to the fact that they adapt their strategy to the present round. For example, depending on the maximum number of coins let on a cell, the player might play differently from a round to another. Such a behaviour is not taken into account in this part. Indeed, we create a variable based on a "general" behaviour at every step.

At step 4, he will be considered as a "copyer" :

```
##
##    copy explore
##       4       2
```

At step 5, he will be considered as an "explorer" :

```
table(file_15_R1_copy[file_15_R1_copy$session == "session_01" &
                  file_15_R1_copy$player == "A1", "copy5"])
```

```
##
## explore
##       5
```

We do this for all players (codes presented in .Rmd file).

We know that there are 2 players who behave badly (see first section). We delete them for the statistical analysis because they have a too strong influence. Note they probably had an influence during the game.

```
players_15_R1 <- filter(players_15_R1, ! (player == "A3" &  session == "session_01") )
players_15_R1 <- filter(players_15_R1, ! (player == "B4" &  session == "session_03") )
```

#### 2.3.3.3 Exploratory analysis

**At step 2 :** we analyse now the behaviours of the players at step 2. Most of them prefer exploring (18 players) rather copying (9 players). Only one player prefer to keep the value 15 (note that this behaviour does not help the group because other players could then think that the cell 99 is at the cell the player is playing again).

```
table(players_15_R1$step_2)
```

```
##
##    copy explore no_risk
##       8      19       1
```

Besides, the average mean of the final gain obtained by the players who have a tendency to copy (3773) is larger than the average mean of the group who explores (3472) or the group who takes no risk (3108). This result is the opposite of what we have seen previously.

The final gain per group :

```r
tapply(players_15_R1$final_gain, players_15_R1$step_2, mean)
```

```
##     copy  explore  no_risk
## 3804.375 3475.421 3108.000
```

**At step 3 and 4:** there are no players who decide to play 15 again. Besides, we cross the behaviours at step 3 and 4 and create a new variable which consists in "copying" if players copy both at step 3 and 4, "exploring" if they explore at step 3 and 4 and "both" when they copy/explore or explore/copy (codes presented in the .Rmd file).

There is a majority of players who copy (13) then explore (9) and there are 6 players who eplore and copy or copy and explore.

```r
table(players_15_R1$step_3_and_4)
```

```
##
##    both    copy explore
##      10      12       6
```

The average mean of the final gain is higher for players who do copy (3699) rather exploring (3273) or doing both (3537):

```r
tapply(players_15_R1$final_gain, players_15_R1$step_3_and_4, mean)
```

```
##    both    copy explore
##  3558.3  3696.0  3273.5
```

**At step 5:** most of players prefer to explore. The frequency :

```r
table(players_15_R1$step_5)
```

```
##
##    copy explore no_risk
##       6      19       3
```

The final gain per group is higher for people who explore (3613) rather copying (3394) or plays 15 again (3246). Thus, this result is also opposite to the one seen previously.

```r
tapply(players_15_R1$final_gain, players_15_R1$step_5, mean)
```

```
##     copy  explore  no_risk
## 3452.000 3593.684 3528.000
```

**Behaviours adopted across the time:** here we look if there is a better strategy to adopt during the round. For example, is it better to "explore" at step 2, then "copy" at step 3/4 and finally "explore" at step 5.

We present all the observed behaviours at step 2/step 3,4/step 5 and rank them with the average final gain:

```
##                   behaviour final_gain freq
## 1     explore/explore/copy    2794.50    2
## 2     no_risk/copy/no_risk    3108.00    1
## 3     explore/both/explore    3261.00    4
## 4        explore/copy/copy    3354.00    1
## 5        copy/both/no_risk    3384.00    1
## 6  explore/explore/explore    3513.00    4
## 7     explore/copy/explore    3619.20    5
## 8        copy/both/explore    3628.50    2
## 9        explore/both/copy    3903.00    2
## 10       copy/copy/explore    3957.75    4
## 11          copy/copy/copy    3963.00    1
## 12    explore/both/no_risk    4092.00    1
```

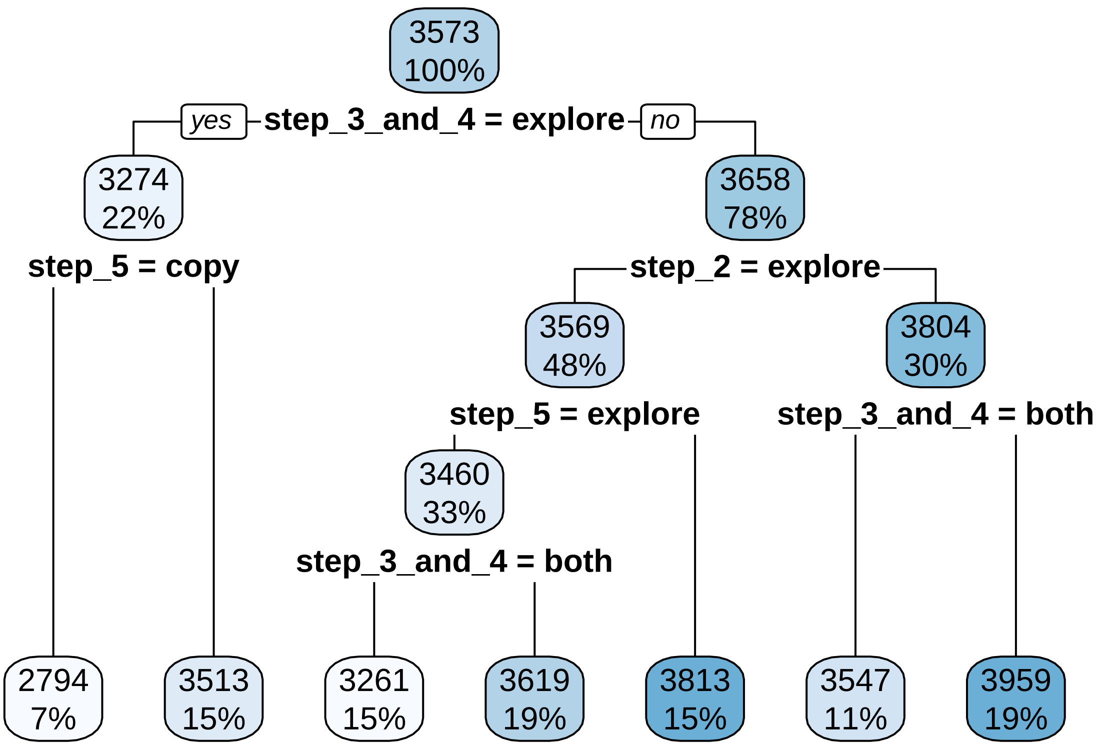
**Interpretation:** by considering our few number of observations, the remarks we are giving here might be used with precautions.

- the two worst strategy consists in explore/explore/copy, no_risk/copy/no_risk.
- the two best strategy consists in explore/both/no_risk and copy/copy/copy.

#### 2.3.3.4 Machine Learning

Finally, we do a regression tree trying to explain the final gain obtained by players depending on their behaviours. Before doing this, we drop the player who took no risk at step 2 and 5 which seems to adopt a behaviour very different from the other players.

```
players_15_R1 <- filter(players_15_R1, step_2 != "no_risk")
```




**Interpretation:** at the first step, the tree discriminates the population into two groups, one group who does explore at step 3 and 4 (on the left) and another group who does not (it means that there are the players who copy or do both). The group on the right has a lower average mean (3658) than the group on the left (3274). The discrimination continues sub-group by sub-group until the final node which gives the percentage of observations and the average mean of the final gain. Usually, we look the observations which fall in the finale nodes with the highest (resp. lowest) predicted values. In our case, the two final nodes with highest values correspond to :

- (last node in dark blue) : players who do copy at step 3 and 4, copy at step 2,
- (5th node in dark blue) : players who copy at step 3 and 4, explore at step 2 and copy at step 5.

### 2.3.4 Conclusion under R1

## 2.4 Analysis under R2

We select the corresponding rows:

```r
file_15_R2 <- file_15[file_15$rule == "R2", ]
```

Under R2, there are :

- 4 sessions,
- For each session, there are two parallel games: 5 players called $A1, ..., A5$ and 5 players called $B1, ..., B5$,
- There are 15 rounds in a session.

At each step, the players have the opportunity to let (or not to let) a coin on the cell they have just played.

### 2.4.1 Optimal strategy for the group

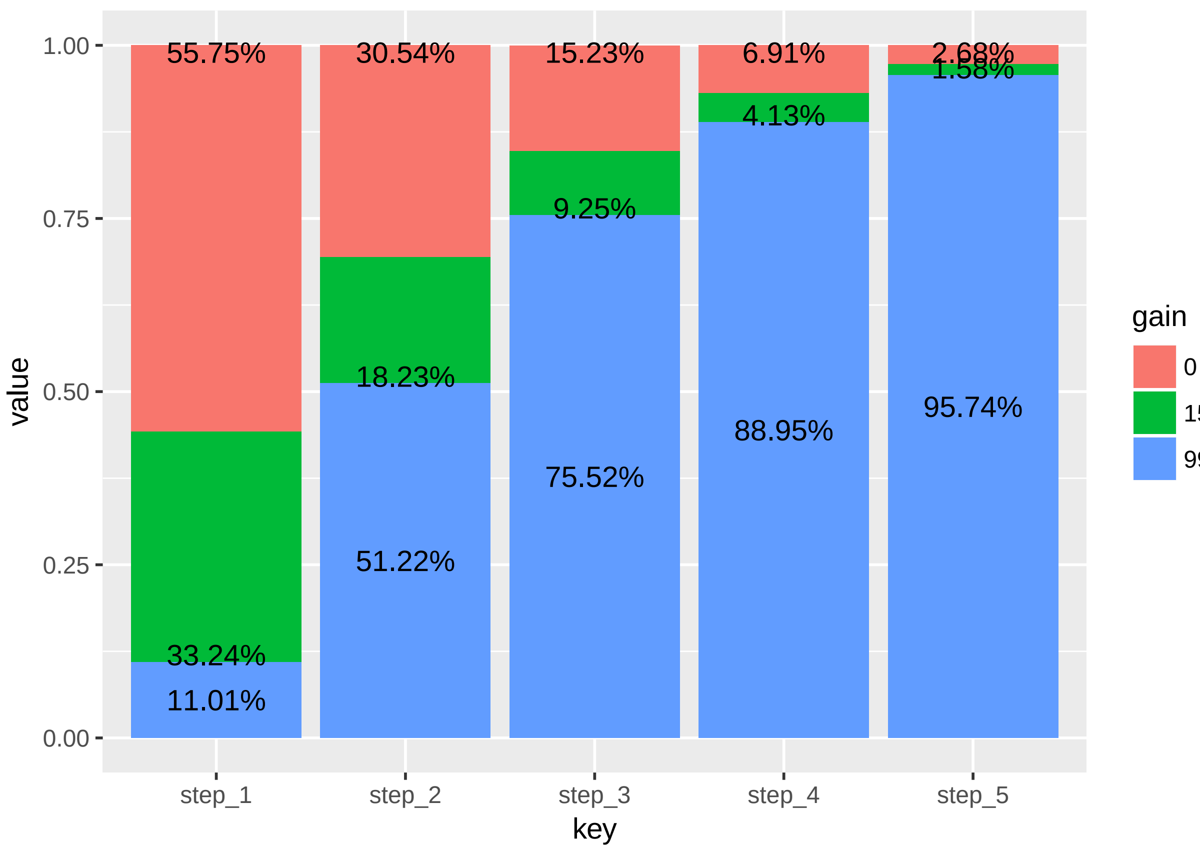
It seems that the best strategy for the group under R2 consists in :

- a player lets a coin only and only if he found the cell 99.
- a player explores new cell until the cell 99 has been discovered by himself or someone else.

Let simulate such a behaviour by using the function *simu_distrib_R2()* (codes are in the .Rmd file).

Simulation for several games:

```r
distrib_th_15_R2 <- NULL
for (k in 1:20000)
  distrib_th_15_R2 <- rbind(distrib_th_15_R2, simu_distrib_R2(15))
```




**Remark:** the distribution is obviously advantageous for the group. However, comparing to the empirical distribution, this is clearly not the behaviour which has been adopted by the players. This can be explained by the fact that if all players behave similarly, they optimize the total gain, but the probability to have the best score at the end of a session among the players is the same for all. Hence, if players first think about how they could obtain a better score than other, they might adopt a strategy which is different from the one adopted by others.

### 2.4.2 Example of an optimal strategy for one player

To illustrate our idea, we simulate a game where 4 players adopt the previous strategy and one player decide to never let a coin although he will play the most visited cell. In this situation, it is interesting to notice that other players will not necessarly remark than player 5 is rigging the game because he does not give a bad information. Here, we are interested to know the final gain this players will obtain at the end of the 15 round and what is his probability to win the session. We program the function *simu_gain_R2_b()* (codes availabes in the .Rmd file).

Simulation for several games:

```
distrib_gain_15_R2 <- NULL
for (k in 1:1000)
  distrib_gain_15_R2 <- rbind(distrib_gain_15_R2, simu_gain_R2_b(15))
```

The probability for player 5 to win the session is much more higher than :

```
table(apply(distrib_gain_15_R2, 1, which.max))/1000
```

```
##
##     1     2     3     4     5
## 0.115 0.083 0.118 0.105 0.579
```

His average mean of the final gain is also quite higher for the player 5:

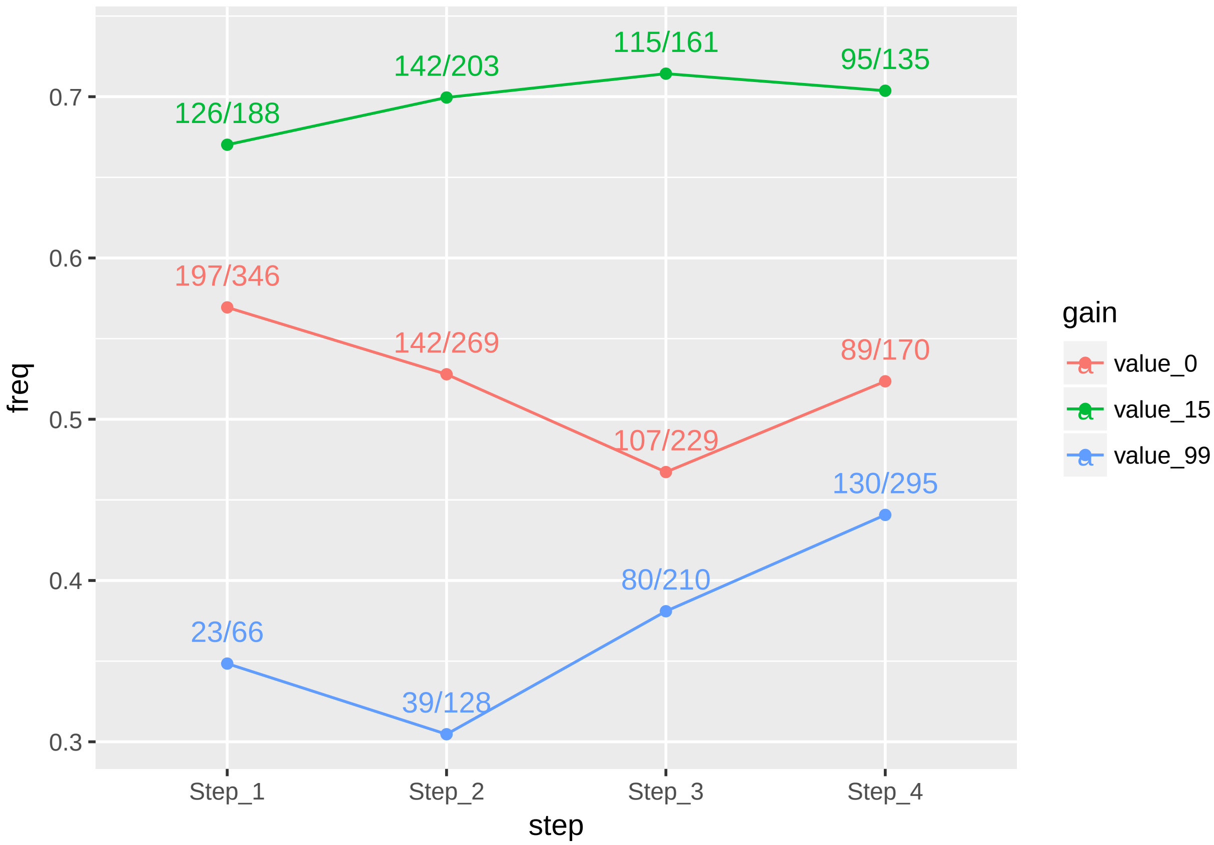
```
apply(distrib_gain_15_R2, 2, mean)
```

```
## [1] 4633.896 4622.319 4619.430 4627.323 4940.607
```

### 2.4.3 Analysis of the behaviours of the players under R2

We compute here the probability to let a coin after playing the cell 99 depending on the step.

We plot the figure:




**Interpretation:** at step 1, 66 players found the value 99. 34.85% of them let a coin on the cell. Hence, it seems that players let most frequently a coin when they found 15 or 0 rather than 99. Thus, it could be that in these conditions, it is not interesting anymore to play the most visited cell. This is what we are going to check now.

#### 2.4.4 Analysis of the number of coins let depending on the values 0, 15 or 99.

From the previous figure, we can deduct the expected number of coins let by players after each step. How did we do that : there are 197 players who let a coin after obtaining the value 0. To get the expected number of coins for any cell which contains 0, we simply divide 197 by 120 ($120 = 15$ rounds $\times$ 2 parallel session $\times$ 4 different sessions). As there are 5 cells which contain the value 0, we divide 197/120 by 5.

**Remark:** as the players have the choice to let a coin or not, the sum of the expected values is now different from 5. For example at step 1, it is equal to 2.88. It is interesting to notice that under R0, a player who would like to use the behaviours of the others, should visit the cell which has been the less visited.

**Expected number of coins let at step 1**

| | | |
|---|---|---|
| gain = 0<br>coin = 0.35 | gain = 0<br>coin = 0.35 | gain = 0<br>coin = 0.35 |
| gain = 0<br>coin = 0.35 | gain = 0<br>coin = 0.35 | gain = 15<br>coin = 0.33 |
| gain = 15<br>coin = 0.33 | gain = 15<br>coin = 0.33 | gain = 99<br>coin = 0.19 |

After the second step, we remark that the cell which contains 99 is still the one where we can find the smallest number of coins:

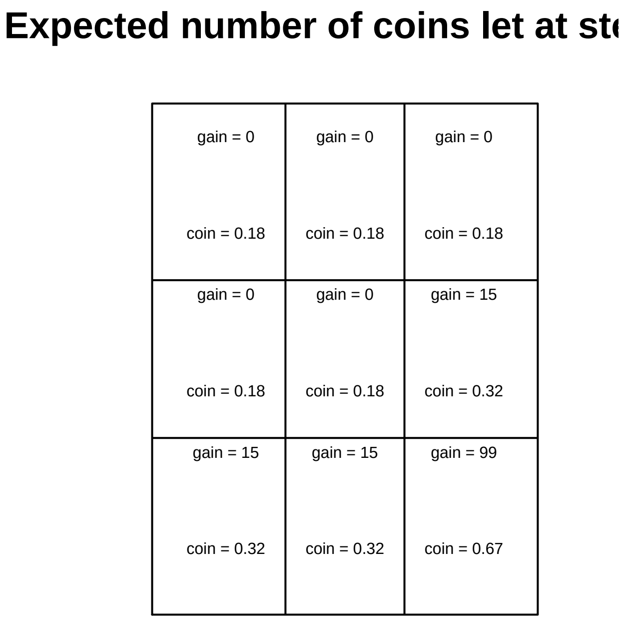
**Expected number of coins let at ste**

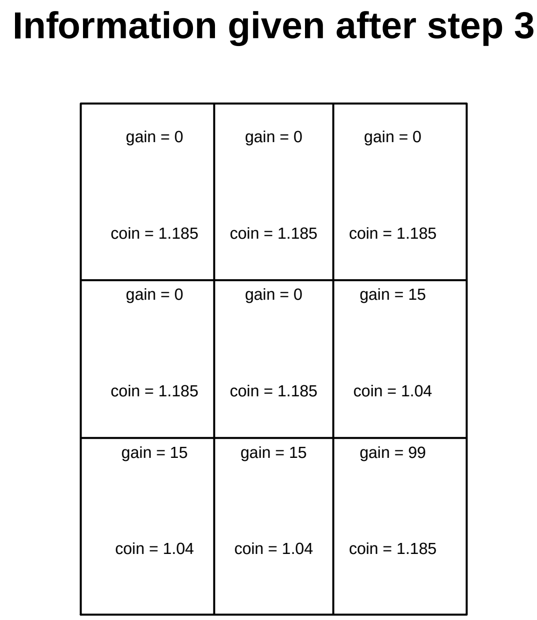
| | | |
|---|---|---|
| gain = 0<br>coin = 0.24 | gain = 0<br>coin = 0.24 | gain = 0<br>coin = 0.24 |
| gain = 0<br>coin = 0.24 | gain = 0<br>coin = 0.24 | gain = 15<br>coin = 0.39 |
| gain = 15<br>coin = 0.39 | gain = 15<br>coin = 0.39 | gain = 99<br>coin = 0.325 |

**Information given after step 2**

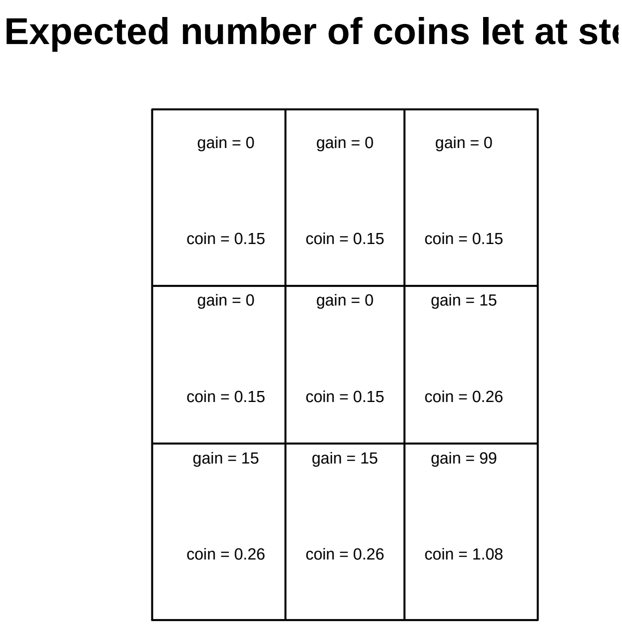
| | | |
|---|---|---|
| gain = 0<br>coin = 0.59 | gain = 0<br>coin = 0.59 | gain = 0<br>coin = 0.59 |
| gain = 0<br>coin = 0.59 | gain = 0<br>coin = 0.59 | gain = 15<br>coin = 0.72 |
| gain = 15<br>coin = 0.72 | gain = 15<br>coin = 0.72 | gain = 99<br>coin = 0.515 |

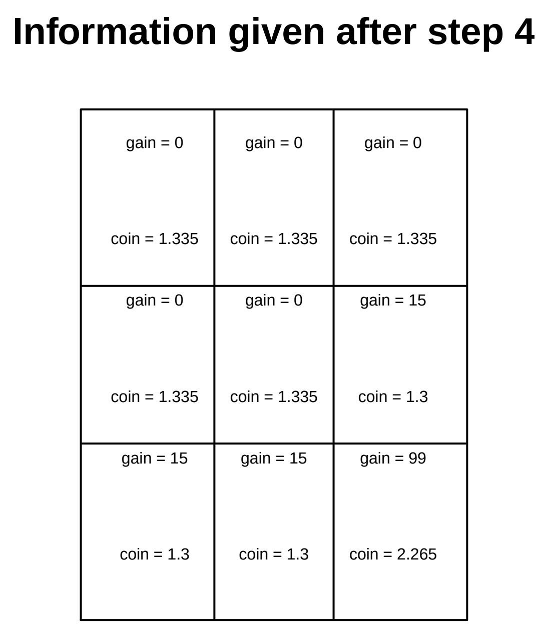
After step 3, the cell which contains 99 is now the one with the highest number of coins:




After step 4, the cell which contains 99 is still the one with the highest number of coins:




**Conclusion:** it seems that a player who played the less visited cell after step 1 and step 2 and the most visited cell after step 3 and step 4, would optimize his chance to find 99. We are now looking if this is really the case on our data.

### 2.4.5 Behaviours among the players under R2

First, we apply the previous function *cell_visited()* which allows us to see how often a cell has been visited per round. For doing this, we need to count the number of visits per cell at each step, for given players (A or B) per round and per session:

```
file_15_R2$round_p <- paste0(file_15_R2$round, "_", substr(file_15_R2$player, 1, 1),
                             "_", file_15_R2$session)
```

```r
cell_visited_R2 <- cell_visited(file_15_R2)
```

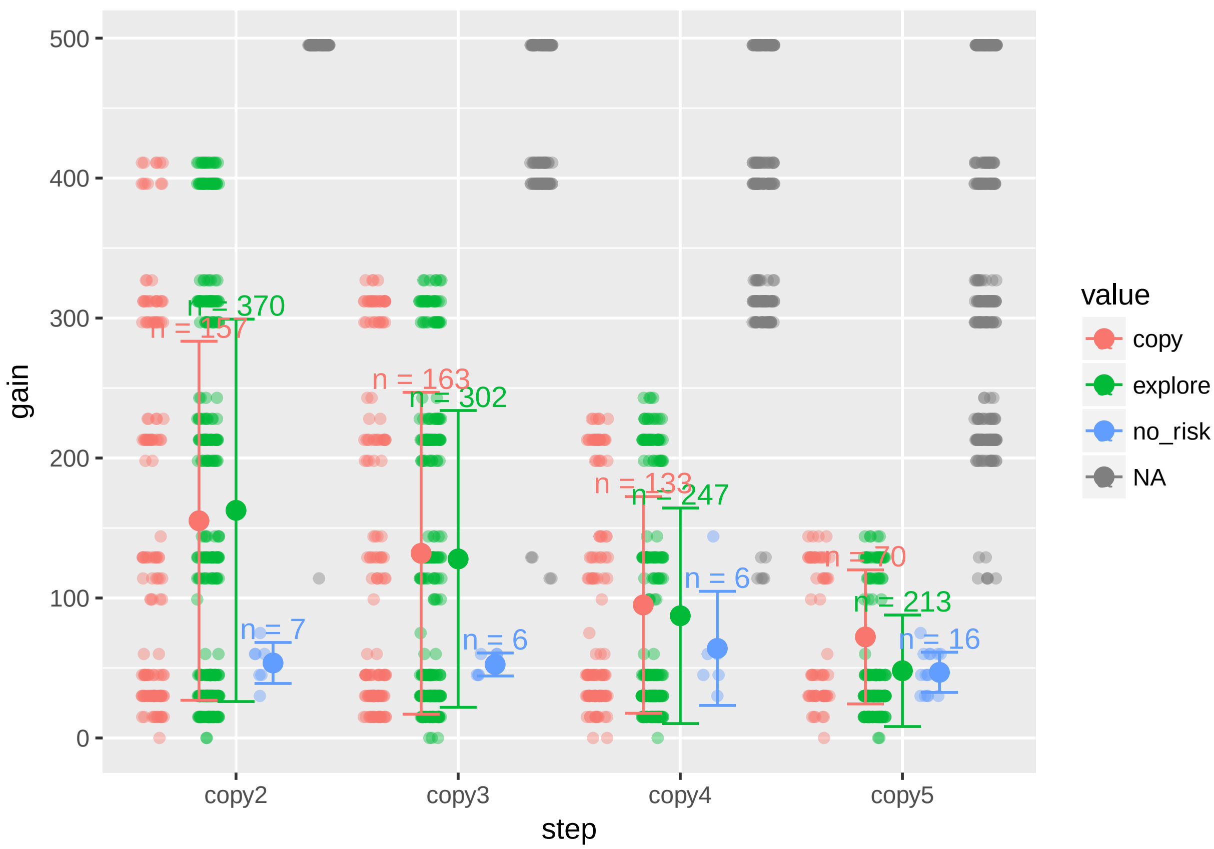
We apply our data to the previous function :

```r
file_15_R2_copy <- copy_or_not(file_15_R2, cell_visited_R2)
```

For representing the data, we first tidy the data (codes presented in the .Rmd file).

We compute the mean and standard deviation obtained at each step and depending on the fact that a player explored new cases or not (codes presented in the .Rmd file).

We plot in y-axis the gain per round and in x-axis the steps. We represent in blue (resp. in red) the gain per round obtained when a player played a visited cell cell (resp. when he did not visit) knowing that the player had the opportunity to do it. We plot the gain per round in grey obtained by people who did not have the opportunity to visit (that means that they know where the cell 99 is located):




**Interpretation:** it seems that the advantageous under R1 which consists in playing the most visited cell is not correct anymore under R2 excepted at step 5. Indeed, as seen previously, the number of coins let on the cell 99 seems to be big enough to detect a potential tendency. This can be confirmed by the following $\chi^2$ test.

```r
(kruskal.test(gather_file_15_R2$gain[gather_file_15_R2$step == "copy2"] ~
factor(gather_file_15_R2$value[gather_file_15_R2$step == "copy2"])))
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  gather_file_15_R2$gain[gather_file_15_R2$step == "copy2"] by factor(gather_file_15_R2$value[ga
## Kruskal-Wallis chi-squared = 1.5139, df = 2, p-value = 0.4691
```

```r
(kruskal.test(gather_file_15_R2$gain[gather_file_15_R2$step == "copy3"] ~
factor(gather_file_15_R2$value[gather_file_15_R2$step == "copy3"])))
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  gather_file_15_R2$gain[gather_file_15_R2$step == "copy3"] by factor(gather_file_15_R2$value[ga
## Kruskal-Wallis chi-squared = 0.31793, df = 2, p-value = 0.853
```

```r
(kruskal.test(gather_file_15_R2$gain[gather_file_15_R2$step == "copy4"] ~
factor(gather_file_15_R2$value[gather_file_15_R2$step == "copy4"])))
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  gather_file_15_R2$gain[gather_file_15_R2$step == "copy4"] by factor(gather_file_15_R2$value[ga
## Kruskal-Wallis chi-squared = 2.3459, df = 2, p-value = 0.3095
```

```r
(kruskal.test(gather_file_15_R2$gain[gather_file_15_R2$step == "copy5"] ~
factor(gather_file_15_R2$value[gather_file_15_R2$step == "copy5"])))
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  gather_file_15_R2$gain[gather_file_15_R2$step == "copy5"] by factor(gather_file_15_R2$value[ga
## Kruskal-Wallis chi-squared = 18.673, df = 2, p-value = 8.816e-05
```

### 2.4.6 The effect of copying/exploring/re-playing 15 on the final gain

We are now looking for each player how he behaves during one session and trying to explain if its behaviours could explain the final gain.

We look at two diffrent behaviours :

- Do the player copy, explore or takes no risk (codes presented in the .Rmd file)
- Do the player give good or bad information to the others (codes presented in the .Rmd file)

We define a player as :

- a **collaborator** if he let majoritarly a coin when he found 99 and no coin when he found 0
- a **liar** if he majoritarly let a coin when he found 0 and no coin when he found 99.

We know that there is 1 player who behave badly (see first section). We delete him for the statistical analysis because they have a too strong influence.

```r
players_15_R2 <- filter(players_15_R2, ! (player == "B4" &  session == "session_03") )
```

We first look at the variable which indicates if a player gave good or bad information. There are more "liar" than "collaborator":

```r
table(players_15_R2$info)
```

```
##
##         both collaborator         liar
##           14           10           15
```

The group of the liar seems to have a better final gain:

```r
tapply(players_15_R2$final_gain, players_15_R2$info, mean)
```

```
##         both collaborator         liar
##     2755.929     2906.700     3229.800
```

Concerning the other variable (copy, explore, or keep 15), we cross it with the previous variable. We can see that most whatever the step, most of the **liar** do not try to copy, probably because they fear that other players do behave similarly. However, the two other groups (**collaborators** or **both**) try more oftenly to copy.

```r
table(players_15_R2$step_2, players_15_R2$info)
```

```
##
##           both collaborator liar
##   copy       4            3    1
##   explore   10            7   14
```

```r
table(players_15_R2$step_3, players_15_R2$info)
```

```
##
##           both collaborator liar
##   copy       5            4    2
##   explore    9            6   13
```

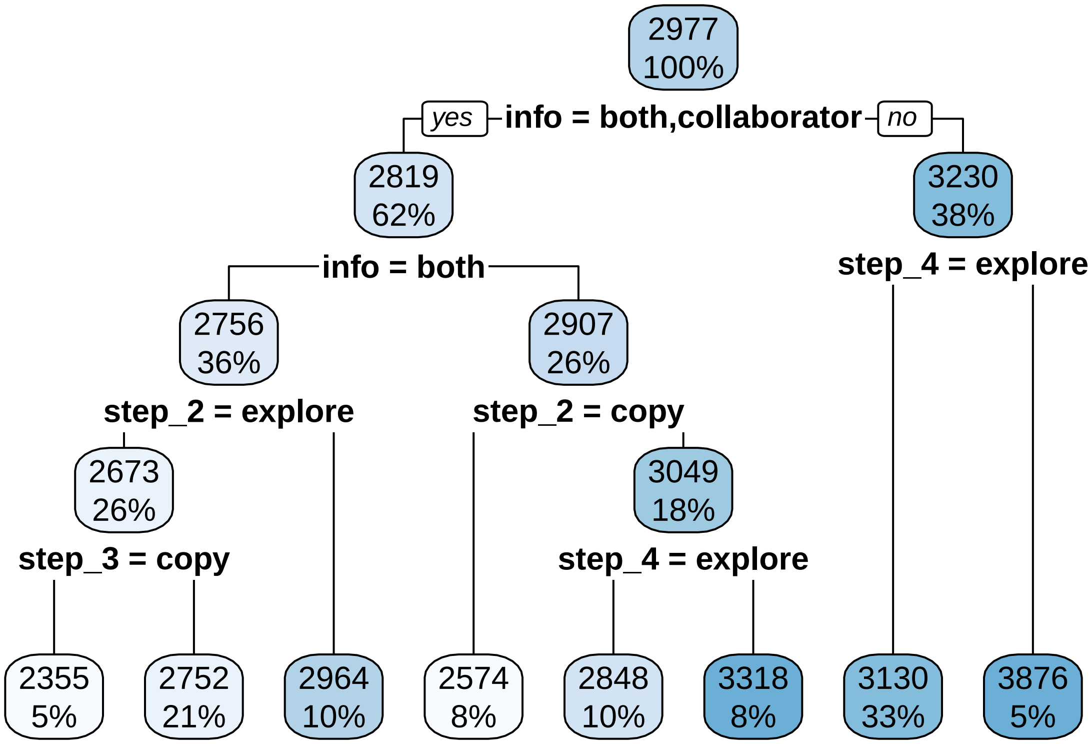
```r
table(players_15_R2$step_4, players_15_R2$info)
```

```
##
##           both collaborator liar
##   copy       6            4    2
##   explore    8            6   13
```

```r
table(players_15_R2$step_5, players_15_R2$info)
```

```
##
##           both collaborator liar
##   copy       3            2    0
##   explore   11            6   14
##   no_risk    0            2    1
```

Finally, we do a regression tree trying to explain the final gain obtained by players depending on their behaviours.




**Interpretation:** the variable which discriminates the most the players is the fact that player give good or bad information. The node with the higher final gain corresponds to players who are **liar** and do copy at the step 4.

# 3 Statistical analysis of Stigmer 50

TBD