Red and Black : id Masked

coop711

2018-09-16

Data

```
class_roll <- read.table("../data/class_roll_masked.txt",
                         header = TRUE,
                         stringsAsFactors = FALSE,
                         encoding = "UTF-8")
str(class_roll)
```

```
## 'data.frame':    160 obs. of  6 variables:
##  $ dept   : chr  "○○학과" "○○학과" "○○학과" "○○학과" ...
##  $ id     : int  20119999 20119999 20179999 20149999 20169999 20129999 20149999 20169999 20179999 2
0129999 ...
##  $ name   : chr  "강○○" "강○○" "강○○" "강○○" ...
##  $ year   : int  4 4 1 4 2 3 4 2 1 3 ...
##  $ email  : chr  "user_name@naver.com" "user_name@hanmail.net" "user_name@naver.com" "user_name@han
mail.net" ...
##  $ cell_no: chr  "010-9164-xxxx" "010-8574-xxxx" "010-6435-xxxx" "010-2066-xxxx" ...
```

Randomization

```
# set.seed(107)
N <- nrow(class_roll)
class_roll$group <- sample(1:N) %% 2 + 1
class_roll$group <- factor(class_roll$group,
                           labels = c("Red", "Black"))
red_id <- which(class_roll$group == "Red")
black_id <- which(class_roll$group == "Black")
```

학번

```
ID_16 <- factor(ifelse(substr(class_roll$id, 1, 4) >= 2016,
                       "younger_16", "older_16"),
                levels = c("younger_16", "older_16"))
kable(table("그룹" = class_roll$group,
            "16학번 기준" = ID_16))
```

	younger_16	older_16
Red	48	32
Black	39	41

```
ID_15 <- factor(ifelse(substr(class_roll$id, 1, 4) >= 2015,
                       "younger_15", "older_15"),
                levels = c("younger_15", "older_15"))
kable(table("그룹" = class_roll$group,
            "15학번 기준" = ID_15))
```

	younger_15	older_15
Red	52	28
Black	48	32

```
ID_14 <- factor(ifelse(substr(class_roll$id, 1, 4) >= 2014,
                       "younger_14", "older_14"),
                levels = c("younger_14", "older_14"))
kable(table("그룹" = class_roll$group,
            "14학번 기준" = ID_14))
```

	younger_14	older_14
Red	60	20
Black	60	20

```
ID_13 <- factor(ifelse(substr(class_roll$id, 1, 4) >= 2013,
                       "younger_13", "older_13"),
                levels = c("younger_13", "older_13"))
kable(table("그룹" = class_roll$group,
            "13학번 기준" = ID_13))
```

	younger_13	older_13
Red	74	6
Black	72	8

email 서비스업체

```
email_list <- strsplit(class_roll$email, "@", fixed = TRUE)
mail_com <- sapply(email_list, `[`, 2)
kable(table("그룹" = class_roll$group,
            "e-mail" = mail_com))
```

	daum.net	gmail.com	hanmail.net	nate.com	naver.com
Red	0	2	4	4	70
Black	2	4	5	3	65

성씨 분포

```
f_name <- substring(class_roll$name,
                    first = 1, last = 1)
kable(table("Group" = class_roll$group,
            "Family Name" = f_name))
```

	강	고	구	권	김	나	명	문	박	반	방	배	서	성	손	송	신	심	안	양	우	유	윤	이	임	장	전	정	조	차	최	하	한	황
Red	6	0	0	1	16	1	0	1	10	1	0	0	3	0	1	1	1	1	2	0	0	2	3	11	1	3	1	4	4	0	4	0	0	2
Black	0	1	1	3	20	0	1	0	3	0	1	2	1	1	1	2	2	0	3	1	1	3	2	10	2	0	1	4	3	2	5	1	2	1

많이 나오는 성씨

```
f_name_f <- factor(ifelse(f_name %in% c("김", "이", "박"),
                          f_name, "기타"),
                   levels = c("김", "이", "박", "기타"))
kable(table("Group" = class_roll$group,
            "Family Name" = f_name_f))
```

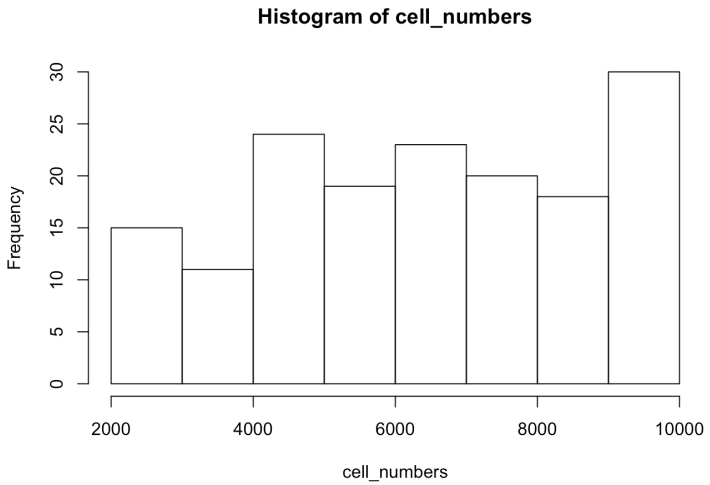
	김	이	박	기타
Red	16	11	10	43
Black	20	10	3	47

전화번호의 분포

```
cell_numbers <- sapply(substr(class_roll$cell_no, 5, 8),
                       as.numeric)
# cut_label <- c("1000~1999", "2000~2999", "3000~3999", "4000~4999", "5000~5999", "6000~6999",
#                "7000~7999", "8000~8999", "9000~9999")
cut_label <- paste(paste0(1:9, "000"), paste0(1:9, "999"), sep = "~")
kable(t(table(cut(cell_numbers,
                  labels = cut_label,
                  breaks = seq(1000, 10000, by = 1000)))))
```

1000~1999	2000~2999	3000~3999	4000~4999	5000~5999	6000~6999	7000~7999	8000~8999	9000~9999
0	15	11	24	19	23	20	18	30

```
hist(cell_numbers)
```



출석부에서 8명 비복원 랜덤 표집

```
# set.seed(1)
kable(class_roll[sample(1:nrow(class_roll), size = 8), ])
```

	dept	id	name	year	email	cell_no	group
50	○○학과	20179999	명○○	1	user_name@naver.com (mailto:user_name@naver.com)	010-6646-xxxx	Black
145	○○학과	20179999	차○○	1	user_name@naver.com (mailto:user_name@naver.com)	010-8616-xxxx	Black
38	○○학과	20169999	김○○	2	user_name@naver.com (mailto:user_name@naver.com)	010-5457-xxxx	Black
70	○○학과	20179999	서○○	1	user_name@naver.com (mailto:user_name@naver.com)	010-2934-xxxx	Red
112	○○학과	20139999	이○○	3	user_name@daum.net (mailto:user_name@daum.net)	010-2910-xxxx	Black
54	○○학과	20169999	박○○	2	user_name@naver.com (mailto:user_name@naver.com)	010-6866-xxxx	Red
95	○○학과	20179999	윤○○	1	user_name@naver.com (mailto:user_name@naver.com)	010-9250-xxxx	Red
4	○○학과	20149999	강○○	4	user_name@hanmail.net (mailto:user_name@hanmail.net)	010-2066-xxxx	Red

`set.seed()` 의 용법

`set.seed()` 를 이용하면 랜덤넘버에 의존하는 실험을 재현할 수 있다. 다음 코드를 반복 수행하거나 다른 사람들의 수행결과와 비교해 보라.

세 결과가 모두 다른 경우

```
sample(1:6, size = 2)
```

```
## [1] 1 5
```

```
sample(1:6, size = 2)
```

```
## [1] 6 1
```

```
sample(1:6, size = 2)
```

```
## [1] 2 5
```

세 번의 수행 결과가 똑같이 반복되는 경우

```
set.seed(1)
sample(1:6, size = 2)
```

```
## [1] 2 6
```

```
sample(1:6, size = 2)
```

```
## [1] 4 5
```

```
sample(1:6, size = 2)
```

```
## [1] 2 5
```

```
set.seed(1)
sample(1:6, size = 2)
```

```
## [1] 2 6
```

```
sample(1:6, size = 2)
```

```
## [1] 4 5
```

```
sample(1:6, size = 2)
```

```
## [1] 2 5
```

동일한 결과를 반복적으로 얻는 경우

```
set.seed(1)
sample(1:6, size = 2)
```

```
## [1] 2 6
```

```
set.seed(1)
sample(1:6, size = 2)
```

```
## [1] 2 6
```

```
set.seed(1)
sample(1:6, size = 2)
```

```
## [1] 2 6
```