The Sorting Hat

Alex Alekseev

```
library(tidyverse)
```

Intro

Imagine the following simulation based on the Harry Potter series. In particular, on a scene in the Philosopher's Stone when the Sorting Hat assigns students to houses. For simplicity, let's just take two houses: Gryffindore and Slytherine. We are interested in the causal effect of being assigned to Gryffindore on a student's academic performance. We assume the perspective of the Sorting Hat that knows each student's potential performance in each house.

Data

```
df <- tribble(
  ~name, ~perf_slytherin, ~perf_gryffindor,
  "Harry", 40, 50,
  "Draco", 55, 35,
  "Goyle", 15, 10,
  "Ron", 5, 18,
  "Pancy", 40, 10,
  "Hermione", 40, 70,
  "Blaise", 25, 7,
  "Fred", 10, 25,
  "George", 10, 30,
  "Crabbe", 20, 5
)

df
```

name	perf_slytherin	perf_gryffindor
Harry	40	50
Draco	55	35
Goyle	15	10
Ron	5	18
Pancy	40	10
Hermione	40	70

name	perf_slytherin	perf_gryffindor
Blaise	25	7
Fred	10	25
George	10	30
Crabbe	20	5

True effects

Using these data we can compute the individual effect of being assigned to Gryffindore for every student, as well as the average effect.

```
df <-
  df %>%
  mutate(delta = perf_gryffindor - perf_slytherin)
df
```

name	perf_slytherin	perf_gryffindor	delta
Harry	40	50	10
Draco	55	35	-20
Goyle	15	10	-5
Ron	5	18	13
Pancy	40	10	-30
Hermione	40	70	30
Blaise	25	7	-18
Fred	10	25	15
George	10	30	20
Crabbe	20	5	-15

For some students, being assigned to Gryffindor has a positive effect, for some it has a negative effect.

```
mean(df$delta)
```

```
## [1] 0
```

But on average, the effect of Gryffindor on academic performance is zero.

Assignment

Knowing all these effects, the Sorting Hat makes the assignment based on whether the individual effect is positive (assigned to Gryffindor) or negative (assigned to Slytherine).

```
df <-
  df %>%
  mutate(gryffindor_sh = 1*(perf_gryffindor >= perf_slytherin))
df
```

name	perf_slytherin	perf_gryffindor	delta	gryffindor_sh
Harry	40	50	10	1
Draco	55	35	-20	0
Goyle	15	10	-5	0
Ron	5	18	13	1
Pancy	40	10	-30	0
Hermione	40	70	30	1
Blaise	25	7	-18	0
Fred	10	25	15	1
George	10	30	20	1
Crabbe	20	5	-15	0

Let's have a look at the average Gryffindor effect for those whom the Hat assigned to Gryffindore and for those whome it assigned to Slytherine.

```
mean(df$delta[df$gryffindor_sh == 1])
```

```
## [1] 17.6
```

```
mean(df$delta[df$gryffindor_sh == 0])
```

```
## [1] -17.6
```

These two effects are different and opposite in signs. The Gryffindor effect for those who went to Gryffindore is positive, while the Gryffindor effect for those who went to Slytherine is negative.

Let's also have a look at the average performance in both houses conditional on whether a student would be assigned to Gryffindor or Slytherin.

```
df %>%
  group_by(gryffindor_sh) %>%
  summarize(
    mean_perf_slyth = mean(perf_slytherin)
    , mean_perf_gryf = mean(perf_gryffindor)
  )
```

gryffindor_sh	mean_perf_slyth	mean_perf_gryf
0	31	13.4
1	21	38.6

Naive estimates

Now let's assume an outsider's perspective who does not have the same insights as the Sorting Hat. We can only observe the assignment and the actual performance of each student in their respective house.

```r
df <-
  df %>%
  mutate(
    perf_sh =
      perf_gryffindor*gryffindor_sh + perf_slytherin*(1 - gryffindor_sh)
  )
df
```

name	perf_slytherin	perf_gryffindor	delta	gryffindor_sh	perf_sh
Harry	40	50	10	1	50
Draco	55	35	-20	0	55
Goyle	15	10	-5	0	15
Ron	5	18	13	1	18
Pancy	40	10	-30	0	40
Hermione	40	70	30	1	70
Blaise	25	7	-18	0	25
Fred	10	25	15	1	25
George	10	30	20	1	30
Crabbe	20	5	-15	0	20

What is the Gryffindore effect now?

```r
df %>%
  group_by(gryffindor_sh) %>%
  summarize(mean_perf_sh = mean(perf_sh)) %>%
  mutate(diff = diff(mean_perf_sh))
```

gryffindor_sh	mean_perf_sh	diff
0	31.0	7.6
1	38.6	7.6

The observed Gryffindore effect is positive now! But we know that on average, the Gryffindore effect is actually zero. We got a biased estimate. Why did that happen?

Random assignment

Suppose that instead of using the Hat, each student was randomly assigned to a house.

```r
set.seed(42)
df <-
  df %>%
  mutate(gryffindor_rnd = 1*(runif(10) <= 0.5)) %>%
  mutate(
    perf_rnd =
      perf_gryffindor*gryffindor_rnd + perf_slytherin*(1 - gryffindor_rnd)
  )
```

```r
df %>%
  group_by(gryffindor_rnd) %>%
  summarize(mean_perf_rnd = mean(perf_rnd)) %>%
  mutate(diff = diff(mean_perf_rnd))
```

gryffindor_rnd	mean_perf_rnd	diff
0	29.375	-11.875
1	17.500	-11.875

Now the effect is negative. However, this is the result of a single possible random assignment. Let's see what happens if we repeat the random assignment many times and look at the average.

```r
perf_rnd_fn <- function() {
  df <-
    df %>%
    mutate(gryffindor_rnd = 1*(runif(10) <= 0.5)) %>%
    mutate(
      perf_rnd =
        perf_gryffindor*gryffindor_rnd + perf_slytherin*(1 - gryffindor_rnd)
    )

  res <-
    mean(df$perf_rnd[df$gryffindor_rnd == 1]) -
    mean(df$perf_rnd[df$gryffindor_rnd == 0])

  return(res)
}

sim <- replicate(5000, perf_rnd_fn(), simplify = T)

t.test(sim)
```

```
##
##  One Sample t-test
##
## data:  sim
## t = -0.83925, df = 4991, p-value = 0.4014
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.4420456  0.1770256
## sample estimates:
## mean of x
##  -0.13251
```

The estimated effect is not statistically different from zero, which is what we should get because the true effect is zero.

What do you think would happen if we repeat this analysis for the average Gryffindor effect for those who got assigned to Gryffindore and for those who got assigned to Slytherin?