

Bag of little bootstraps

```
library(tidyverse)

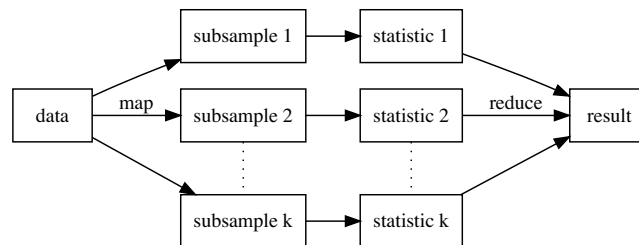
## -- Attaching packages ----- tidyverse 1.3.0 --

## v ggplot2 3.2.1    v purrr  0.3.3
## v tibble  2.1.3    v dplyr  0.8.3
## v tidyr   1.0.0    v stringr 1.4.0
## v readr   1.3.1    v forcats 0.4.0

## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Divide and conquer a.k.a. mapreduce

Divide and conquer allows a single task operation to be executed parallelly.



We have seen that in assignment 3 how we could use map and reduce to compute the mean.

```
# we first random split `flights` into 10 files
library(nycflights13)
set.seed(141)
m <- 10
groups <- sample(seq_len(m), nrow(flights), replace = TRUE)
dir.create("flights/", showWarnings = FALSE)
for (i in seq_len(m)) {
  write_csv(filter(flights, groups == i), str_c("flights/", i, ".csv"))
}
```

```
file_names <- file.path("flights", list.files("flights"))
mean_list <- file_names %>% map(~ mean(read_csv(.)$dep_delay, na.rm = TRUE))
(mean_dep_delay <- mean_list %>% reduce(`+`) / m)
```

```
## [1] 12.63861
```

You may wonder if you could do the same for confidence intervals.

```
ci_list <- file_names %>% map(~ t.test(read_csv(.)$dep_delay)$conf.int)
(mean_ci <- ci_list %>% reduce(`+`) / m)
```

```
## [1] 12.20391 13.07331
## attr("conf.level")
## [1] 0.95
```

Yeah, it gives us a result. But wait, it doesn't look right. Though the mapreduce procedure speeds up the computation, it should give similar result as if we work on the whole dataset.

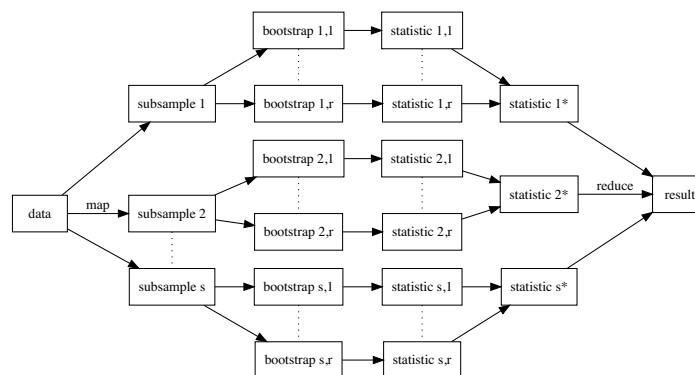
```
t.test(flights$dep_delay)$conf.int
```

```
## [1] 12.50157 12.77657
## attr("conf.level")
## [1] 0.95
```

Lesson learned: we cannot combine any statistics in the reduce step by simply taking the average. We may need to scale the statistics analytically which could be hard or impossible.

The bag of little bootstraps (BLB)

It is a procedure which incorporates features of both the bootstrap and subsampling to yield a robust, computationally efficient means of assessing the quality of estimators



- sample without replacement the sample s times into sizes of b
- for each subsample

- resample each until sample size is n , r times
- compute the bootstrap statistic (e.g., the mean) for each bootstrap sample
- compute the statistic (e.g., confidence interval) from the bootstrap statistics
- take the average of the statistics

Basically, the bag of little bootstraps = subsample + bootstrap. However, for each bootstrap, we sample n from b with replacement instead of sample b from b as in ordinary bootstrap.

A naive (single core) implementation

```
r <- 10 # r should be at least a few thousands, we are using 10 for demo
n <- nrow(flights)

each_boot <- function(i, data) {
  mean(sample(data, n, replace = TRUE), na.rm = TRUE)
}

ci_list <- file_names %>% map(~ {
  sub_dep_delay <- read_csv(.)$dep_delay
  map_dbl(seq_len(r), each_boot, data = sub_dep_delay) %>%
    quantile(c(0.025, 0.975))
})

reduce(ci_list, `+`) / length(ci_list)

##      2.5%      97.5%
## 12.52957 12.75403
```

The sample above is not memory and computationally efficient.

```
# the frequency table of selecting 1000 items from 1:10 with replacement
table(sample(1:10, 100, replace = TRUE))

##
##  1  2  3  4  5  6  7  8  9 10
##  8 10 16 10  4  4 17  7 13 11
```

A more efficient way is to first generate the repetitions by multinomial distribution.

```
rmultinom(1, 100, rep(1, 10))

##      [,1]
## [1,]    13
## [2,]    11
## [3,]    12
## [4,]    10
## [5,]    10
## [6,]     7
## [7,]    12
## [8,]    11
## [9,]     4
## [10,]   10
```

Compute the mean with the frequencies

```
sub_dep_delay <- read_csv(file_names[1])$dep_delay
# it's important to remove the missing values in this step
sub_dep_delay <- sub_dep_delay[!is.na(sub_dep_delay)]
freqs <- rmultinom(1, n, rep(1, length(sub_dep_delay)))
sum(sub_dep_delay * freqs) / n
```

```
## [1] 12.54856
```

Put everything back

```
r <- 10 # r should be at least a few thousands, we are using 10 for demo
n <- nrow(flights)
```

```
each_boot2 <- function(i, data) {
  non_missing_data <- data[!is.na(data)]
  freqs <- rmultinom(1, n, rep(1, length(non_missing_data)))
  sum(non_missing_data * freqs) / n
}
```

```
ci_list <- file_names %>% map(~ {
  sub_dep_delay <- read_csv(.)$dep_delay
  map_dbl(seq_len(r), each_boot2, data = sub_dep_delay) %>%
    quantile(c(0.025, 0.975))
})
```

```
reduce(ci_list, `+`) / length(ci_list)
```

```
##      2.5%      97.5%
## 12.54252 12.73115
```

A parallel version using furrr.

```
library(furrr)
plan(multiprocess, workers = 5)
```

```
ci_list <- file_names %>% future_map(~ {
  sub_dep_delay <- read_csv(.)$dep_delay
  map_dbl(seq_len(r), each_boot2, data = sub_dep_delay) %>%
    quantile(c(0.025, 0.975))
})
reduce(ci_list, `+`) / length(ci_list)
```

```
##      2.5%      97.5%
## 12.53485 12.72271
```

Of course, it is in general a better idea to read the data in the workers than sending the data from master to the workers.

Comparsion

```
r <- 500
naive <- function() {
  file_names %>% map(~ {
    sub_dep_delay <- read_csv(.)$dep_delay
    map_dbl(seq_len(r), each_boot, data = sub_dep_delay) %>%
      quantile(c(0.025, 0.975))
  })
}
improve <- function() {
  file_names %>% map(~ {
    sub_dep_delay <- read_csv(.)$dep_delay
    map_dbl(seq_len(r), each_boot2, data = sub_dep_delay) %>%
      quantile(c(0.025, 0.975))
  })
}
multi_core <- function() {
  file_names %>% future_map(~ {
    sub_dep_delay <- read_csv(.)$dep_delay
    map_dbl(seq_len(r), each_boot2, data = sub_dep_delay) %>%
      quantile(c(0.025, 0.975))
  })
}

# system.time(naive()) # [skipped] take forever
system.time(improve()) # 4x seconds
system.time(multi_core()) # 1x seconds
```

Another example

We want to compute a confidence interval between the correlation of `dep_delay` and `arr_delay`

```
cor.test(flights$dep_delay, flights$arr_delay)

##
## Pearson's product-moment correlation
##
## data: flights$dep_delay and flights$arr_delay
## t = 1295.9, df = 327344, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9142422 0.9153599
## sample estimates:
##      cor
## 0.9148028

r <- 100 # in practice, r shoule be at least 1000 - 10000

boot_cor <- function(i, data) {
```

```

# this function bootstrap data and compute the correlation
b <- nrow(data)
weights <- rmultinom(1, n, rep(1, b))
x <- data$arr_delay
y <- data$dep_delay
mux <- sum(weights * x) / n
muy <- sum(weights * y) / n
sxx <- sum(weights * (x - mux)^2)
syy <- sum(weights * (y - muy)^2)
sxy <- sum(weights * (y - muy) * (x - mux))
sxy / sqrt(sxx * syy)
}

```

```

ci_list<- file_names %>% future_map(~ {
  data <- read_csv(.) %>%
    select(arr_delay, dep_delay) %>%
    drop_na()
  map_dbl(seq_len(r), boot_cor, data = data) %>%
    quantile(c(0.025, 0.975))
})

reduce(ci_list, `+`) / length(ci_list)

```