Power analysis on skew normal fitting for left tail

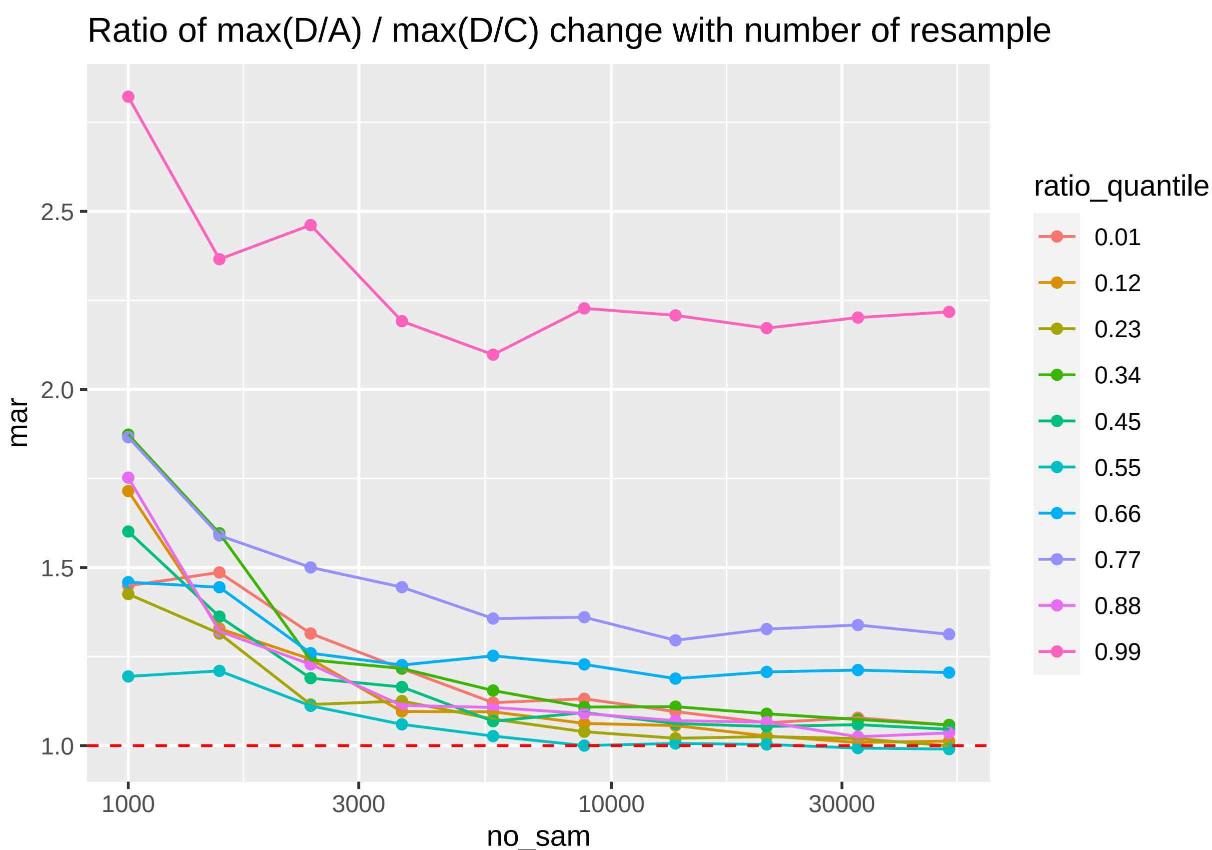
Some notation

In this report, define A = subsample fitted probability; B = subsample empirical tail probability; C = full sample fitted probability and D = full sample empirical tail probability. I consider the analysis for $A/D(D/A)$ and $A/B(B/A)$.

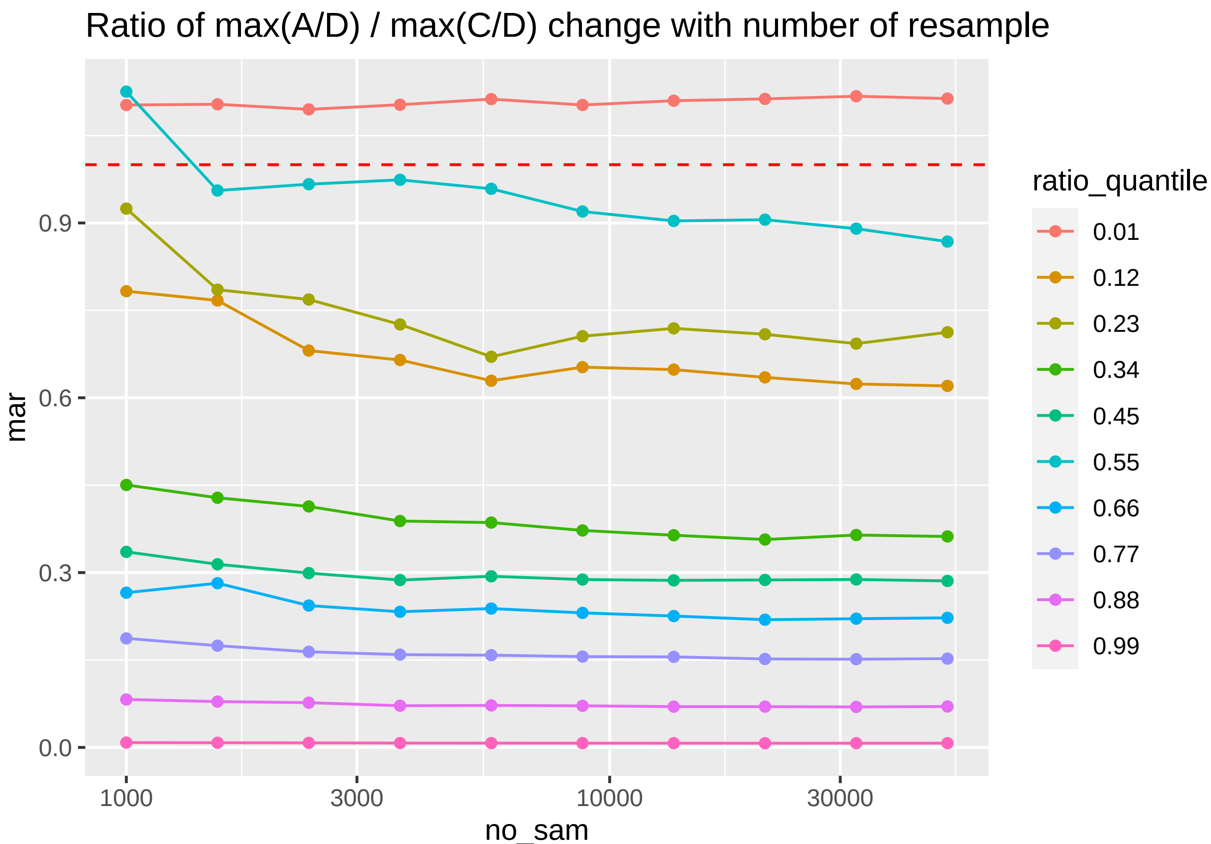
1.Load results for fitting comparison

```
library(dplyr)
library(ggplot2)
library(tidyverse)
undershoot <- read_csv("undershoot_left_fit.csv")[,-1]
overshoot <- read_csv("overshoot_left_fit.csv")[,-1]
quantile_list <- seq(0.01, 0.99, length.out = 10)
no_sam <- round(exp(seq(log(1e3), log(5e4), length.out = 10)))
# rearrange the data frame
B <- 100
undershoot_df <- data.frame(id = rep(1:B, 10*10),
                            ratio_value = 0,
                            no_sam = 0,
                            ratio_quantile = 0)
overshoot_df <- data.frame(id = rep(1:B, 10*10),
                           ratio_value = 0,
                           no_sam = 0,
                           ratio_quantile = 0)

# i: quantile; j: no of sample
for (i in 1:10) {
  for (j in 1:10) {
    start <- (j - 1 + (i-1)*10)*B +1
    end <- (j + (i-1)*10)*B
    undershoot_df[start:end, 2] <- as.vector(undershoot[(((j-1)*B+1) :(j*B)), (i-1)*3+2])[[1]]
    undershoot_df[start:end, 3] <- rep(no_sam[j], B)
    undershoot_df[start:end, 4] <- rep(quantile_list[i], B)
    overshoot_df[start:end, 2] <- as.vector(overshoot[(((j-1)*B+1) :(j*B)), (i-1)*3+2])[[1]]
    overshoot_df[start:end, 3] <- rep(no_sam[j], B)
    overshoot_df[start:end, 4] <- rep(quantile_list[i], B)
  }
}

# load the oracle ratio
param_nc <- read_csv("figures/power_exploration/sknorm_tail_prob_500000_resamples_0.96_percentile/param_
param_twosides <- t(param_nc[,-1])
overshoot_ratio <- as.numeric(param_twosides[, 6])
undershoot_ratio <- as.numeric(param_twosides[, 7])
quantile_list <- seq(0.01, 0.99, length.out = 10)
```

```r
overshoot_set <- data.frame(index = numeric(10), ratio = numeric(10))
undershoot_set <- data.frame(index = numeric(10), ratio = numeric(10))

# find distributions based on right tail
for (r in 1:10){
  dist <- abs(overshoot_ratio[1:330] - quantile(overshoot_ratio[1:330], quantile_list[r]))
  overshoot_set[r, 1] <- which(dist == min(dist))
  overshoot_set[r, 2] <- overshoot_ratio[which(dist == min(dist))]
  dist <- abs(undershoot_ratio[1:330] - quantile(undershoot_ratio[1:330], quantile_list[r]))
  undershoot_set[r, 1] <- which(dist == min(dist))
  undershoot_set[r, 2] <- undershoot_ratio[which(dist == min(dist))]
}

# accuracy matrix for undershoot matrix
undershoot_acc <- matrix(abs(undershoot_df$ratio_value / rep(rep(undershoot_set$ratio, each = 10), each
undershoot_ame <- data.frame(mar = apply(undershoot_acc, 2, mean),
                             no_sam = rep(no_sam, 10),
                             ratio_quantile = as.character(round(rep(quantile_list, each = 10), 2)))
undershoot_ame |>
  ggplot(aes_string(x = "no_sam", y = "mar", colour = "ratio_quantile")) +
  scale_x_log10() +
  geom_point() +
  geom_line() +
  geom_hline(yintercept = 1, linetype = "dashed", colour = "red") +
  labs(title = "Ratio of max(D/A) / max(D/C) change with number of resample")
```



```r
# accuracy matrix for overshoot matrix
overshoot_acc <- matrix(abs(overshoot_df$ratio_value / rep(rep(overshoot_set$ratio, each = 10), each =
overshoot_ame <- data.frame(mar = apply(overshoot_acc, 2, mean),
                            no_sam = rep(no_sam, 10),
                            ratio_quantile = as.character(round(rep(quantile_list, each = 10), 2)))
overshoot_ame |>
  ggplot(aes_string(x = "no_sam", y = "mar", colour = "ratio_quantile")) +
  scale_x_log10() +
  geom_point() +
  geom_line() +
  geom_hline(yintercept = 1, linetype = "dashed", colour = "red") +
  labs(title = "Ratio of max(A/D) / max(C/D) change with number of resample")
```

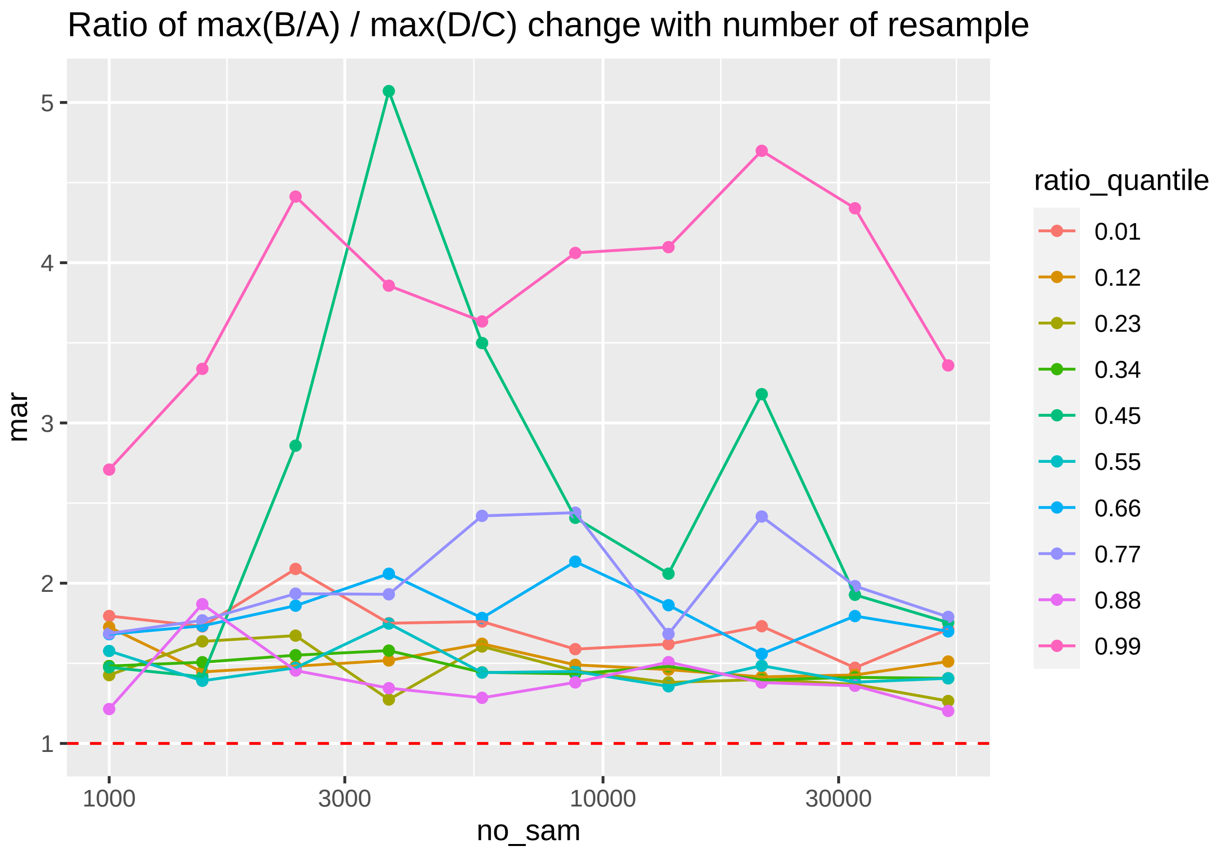


Here we clearly see the ratio $\max[A/D](max[D/A])$ approaches $\max[C/D](\max[D/C])$ very fast. This indicates the error decrease rather fast in estimating the tail probability at least in average sense. In the next section, we mainly consider the changes for the ratio $\max[A/B](max[B/A])$ and how does this approach $\max[C/D](\max[D/C])$.

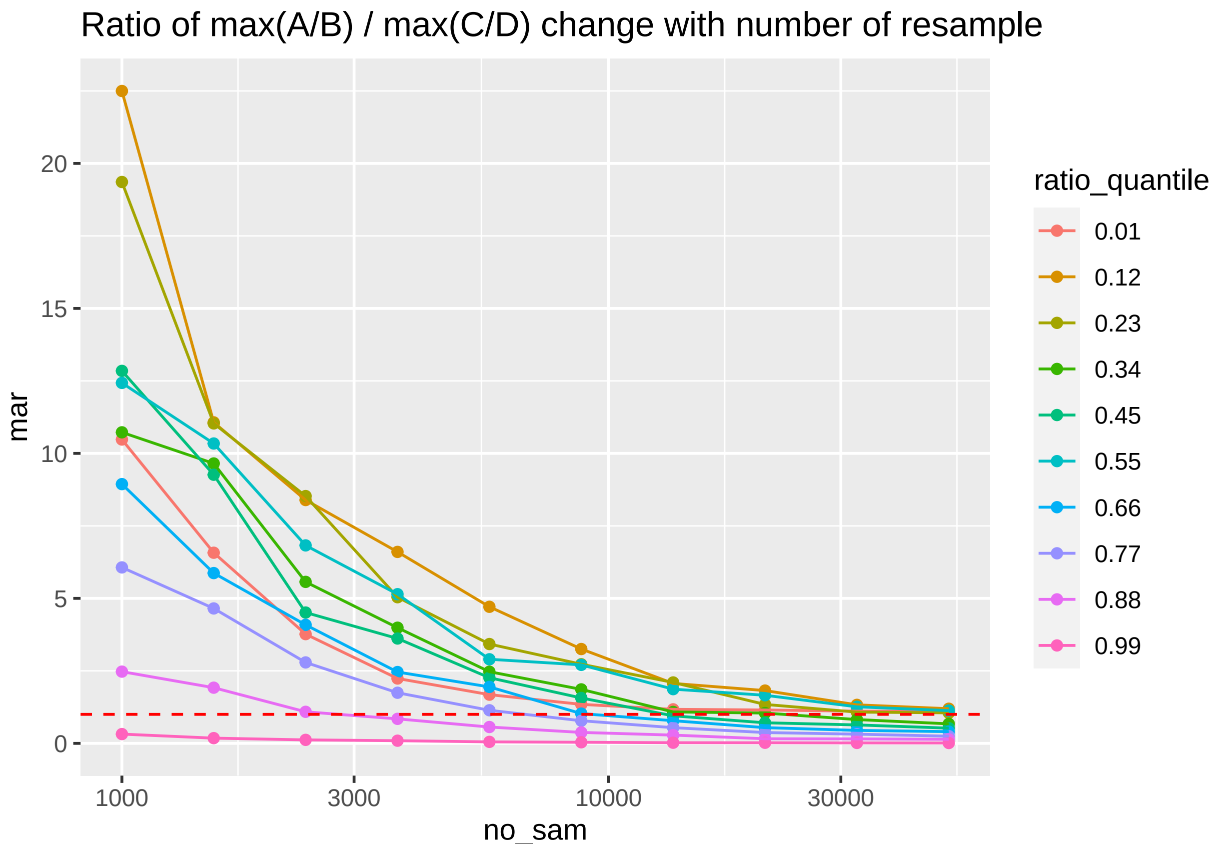
1.Load results for power comparison

```r
undershoot <- read_csv("undershoot_left_power.csv")[,-1]
overshoot <- read_csv("overshoot_left_power.csv")[,-1]
quantile_list <- seq(0.01, 0.99, length.out = 10)
no_sam <- round(exp(seq(log(1e3), log(5e4), length.out = 10)))
# rearrange the data frame
B <- 100
```

```r
undershoot_df <- data.frame(id = rep(1:B, 10*10),
                           ratio_value = 0,
                           no_sam = 0,
                           ratio_quantile = 0)
overshoot_df <- data.frame(id = rep(1:B, 10*10),
                          ratio_value = 0,
                          no_sam = 0,
                          ratio_quantile = 0)

# i: quantile; j: no of sample
for (i in 1:10) {
  for (j in 1:10) {
    start <- (j - 1 + (i-1)*10)*B +1
    end <- (j + (i-1)*10)*B
    undershoot_df[start:end, 2] <- as.vector(undershoot[(((j-1)*B+1) :(j*B)), (i-1)*3+2])[[1]]
    undershoot_df[start:end, 3] <- rep(no_sam[j], B)
    undershoot_df[start:end, 4] <- rep(quantile_list[i], B)
    overshoot_df[start:end, 2] <- as.vector(overshoot[(((j-1)*B+1) :(j*B)), (i-1)*3+2])[[1]]
    overshoot_df[start:end, 3] <- rep(no_sam[j], B)
    overshoot_df[start:end, 4] <- rep(quantile_list[i], B)
  }
}

param_nc <- read_csv("figures/power_exploration/sknorm_tail_prob_500000_resamples_0.96_percentile/param_
param_twosides <- t(param_nc[,-1])
overshoot_ratio <- as.numeric(param_twosides[, 6])
undershoot_ratio <- as.numeric(param_twosides[, 7])
quantile_list <- seq(0.01, 0.99, length.out = 10)
overshoot_set <- data.frame(index = numeric(10), ratio = numeric(10))
undershoot_set <- data.frame(index = numeric(10), ratio = numeric(10))

# find distributions based on right tail
for (r in 1:10){
  dist <- abs(overshoot_ratio[1:330] - quantile(overshoot_ratio[1:330], quantile_list[r]))
  overshoot_set[r, 1] <- which(dist == min(dist))
  overshoot_set[r, 2] <- overshoot_ratio[which(dist == min(dist))]
  dist <- abs(undershoot_ratio[1:330] - quantile(undershoot_ratio[1:330], quantile_list[r]))
  undershoot_set[r, 1] <- which(dist == min(dist))
  undershoot_set[r, 2] <- undershoot_ratio[which(dist == min(dist))]
}

# load the oracle ratio
undershoot_acc <- matrix(abs(undershoot_df$ratio_value / rep(rep(undershoot_set$ratio, each = 10), each
undershoot_ame <- data.frame(mar = apply(undershoot_acc, 2, mean),
                            no_sam = rep(no_sam, 10),
                            ratio_quantile = as.character(round(rep(quantile_list, each = 10), 2)))
undershoot_ame |>
  ggplot(aes_string(x = "no_sam", y = "mar", colour = "ratio_quantile")) +
  scale_x_log10() +
  geom_point() +
  geom_line() +
  geom_hline(yintercept = 1, linetype = "dashed", colour = "red") +
  labs(title = "Ratio of max(B/A) / max(D/C) change with number of resample")
```




```r
# accuracy matrix for overshoot matrix
overshoot_acc <- matrix(abs(overshoot_df$ratio_value / rep(rep(overshoot_set$ratio, each = 10), each = 
overshoot_ame <- data.frame(mar = apply(overshoot_acc, 2, mean),
                            no_sam = rep(no_sam, 10),
                            ratio_quantile = as.character(round(rep(quantile_list, each = 10), 2)))
overshoot_ame |>
  ggplot(aes_string(x = "no_sam", y = "mar", colour = "ratio_quantile")) +
  scale_x_log10() +
  geom_point() +
  geom_line() +
  geom_hline(yintercept = 1, linetype = "dashed", colour = "red") +
  labs(title = "Ratio of max(A/B) / max(C/D) change with number of resample")
```



From the above figure, we notice it is much harder to estimate the ratio of $\max[C/D]$($\max[D/C]$) without using the oracle knowledge of D which we had in the last analysis. Due to the large fluctuation of B in the estimate $\max[A/B]$($\max[B/A]$), the convergence as the increase of sample is not obvious in mean absolute ratio (mar).