Simulation Study to Understand Sampling Distribution

Problem 2

Sulagna Barat | MDS202244

Problem assignee: Sulagna Barat

Reviewer(s): Rohan Karthikeyan and Adarsha Mondal

Problem 2 : Simulation Study to Understand Sampling Distribution

Part A Suppose $X_1, X_2, \cdots, X_n \overset{iid}{\sim} Gamma(\alpha, \sigma)$, with pdf as

$$f(x|\alpha, \sigma) = \frac{1}{\sigma^\alpha \Gamma(\alpha)} e^{-x/\sigma} x^{\alpha-1}, \quad 0 < x < \infty,$$

The mean and variance are $E(X) = \alpha\sigma$ and $Var(X) = \alpha\sigma^2$. Note that `shape` $= \alpha$ and `scale` $= \sigma$.

1. Write a `function` in `R` which will compute the MLE of $\theta = \log(\alpha)$ using `optim` function in `R`. You can name it `MyMLE`
2. Choose `n=20`, and `alpha=1.5` and `sigma=2.2`
   (i) Simulate $\{X_1, X_2, \cdots, X_n\}$ from `rgamma(n=20,shape=1.5,scale=2.2)`
   (ii) Apply the `MyMLE` to estimate θ and append the value in a vector
   (iii) Repeat the step (i) and (ii) 1000 times
   (iv) Draw histogram of the estimated MLEs of θ.
   (v) Draw a vertical line using `abline` function at the true value of θ.
   (vi) Use `quantile` function on estimated θ's to find the 2.5 and 97.5-percentile points.
3. Choose `n=40`, and `alpha=1.5` and repeat the (2).
4. Choose `n=100`, and `alpha=1.5` and repeat the (2).
5. Check if the gap between 2.5 and 97.5-percentile points are shrinking as sample size `n` is increasing?

Hint: Perhaps you should think of writing a single `function` where you will provide the values of `n`, `sim_size`, `alpha` and `sigma`; and it will return the desired output.

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Let $x_1, x_2, \cdots, x_n$ be a random sample of a popuation with pdf $f(x;\theta)$, where θ is a parameter. Consider a function,

$$f(x_1, x_2, ..., x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

When $x_1, x_2, \cdots, x_n$ are given, then $f(x_1, x_2, ..., x_n; \theta)$ is a function of θ only, that we call likelihood function of θ, denoted by $L(\theta)$. An estimate of θ for which $L(\theta)$ is maximum (consequently $logL(\theta)$ is also maximum) is suggested by the 'Maximum Likelihood'. Maximum Likelihood Estimator of a parameter θ is a consistent estimator of θ.
Here we are using *optim* function to compute MLE of $\theta = \log(\alpha)$ and naming it as MyMLE.
1.

```r
mle <- function(log_alpha, data, sigma) {
    l = sum(log(dgamma(data, shape = exp(log_alpha), scale = sigma)))
    return(-l)
}
MyMLE <- function(data, sigma) {
    log_alpha_init <- log(mean(data)^2/var(data))
    estimator <- optim(log_alpha_init,
                 mle,
                 data = data,
                 sigma = sigma)
    log_alpha_cap <- estimator$par
    return(log_alpha_cap)
}
```

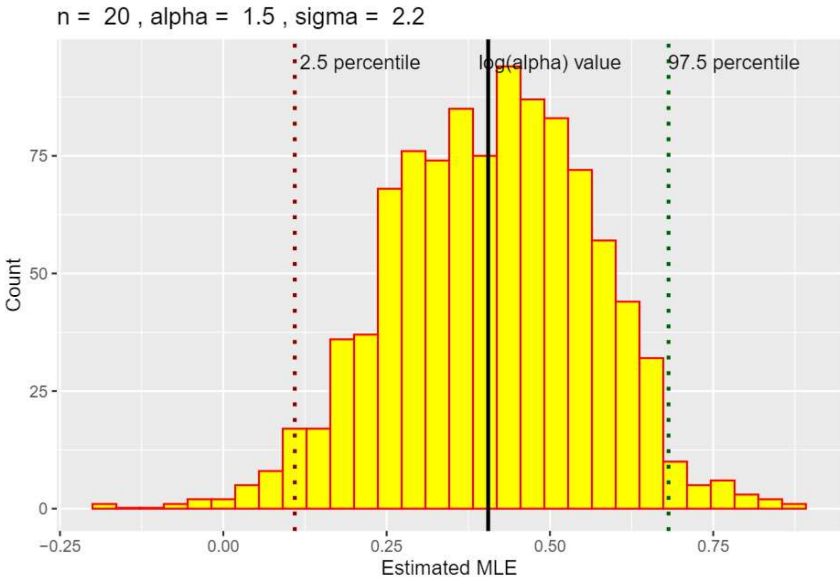
```r
get_estim <- function(n, alpha, sigma) {
    estim <- c()
    for (i in 1:1000) {
        samples <- rgamma(n, shape = alpha, scale = sigma)
        estim <- append(estim, MyMLE(data = samples, sigma = sigma))
    }
    return(estim)
}
```

2

.

```
n = 20
alpha = 1.5
sigma = 2.2
estim_mle <- tibble(get_estim(n = n, alpha = alpha, sigma = sigma))
colnames(estim_mle) <- c("estim")
percentile_2.5 <- quantile(estim_mle$estim, probs = 0.025, names = FALSE)
percentile_97.5 <- quantile(estim_mle$estim, probs = 0.975, names = FALSE)
estim_mle %>%
    ggplot(aes(estim)) +
    geom_histogram(color = "red", fill = "yellow") +
    geom_vline(xintercept = log(alpha),
               size = 1,
               linetype = "solid") +
    annotate("text", label = "log(alpha) value", x = 0.5, y = 95, color = "black") +
    geom_vline(xintercept = percentile_2.5,
               color = "dark red", size = 1, linetype = "dotted") +
    annotate("text", label = "2.5 percentile", x = percentile_2.5 + 0.1, y = 95, color = "black") +
    geom_vline(xintercept = percentile_97.5,
               color = "dark green", size = 1, linetype = "dotted") +
    annotate("text", label = "97.5 percentile", x = percentile_97.5 + 0.1, y = 95, color = "black") +
    labs(title = paste("n = ", n, ", alpha = ", alpha, ", sigma = ", sigma),
         x = "Estimated MLE",
         y = "Count")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



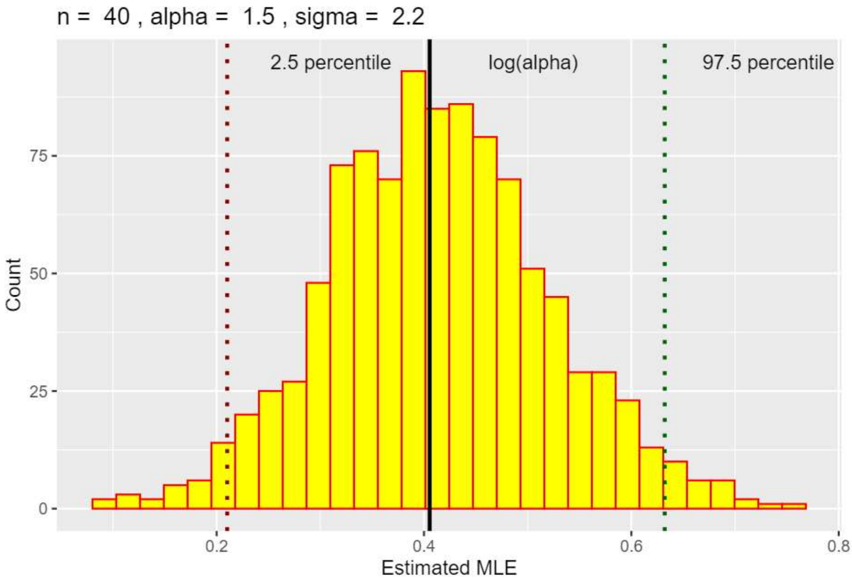

```
diff_20 <- percentile_97.5 - percentile_2.5
```

3

.

```
n = 40
alpha = 1.5
sigma = 2.2
estim_mle <- tibble(get_estim(n = n, alpha = alpha, sigma = sigma))
colnames(estim_mle) <- c("estim")
percentile_2.5 <- quantile(estim_mle$estim, probs = 0.025, names = FALSE)
percentile_97.5 <- quantile(estim_mle$estim, probs = 0.975, names = FALSE)
estim_mle %>%
    ggplot(aes(estim)) +
    geom_histogram(color = "red", fill = "yellow") +
    geom_vline(xintercept = log(alpha),
               size = 1,
               linetype = "solid") +
    annotate("text", label = "log(alpha)", x = log(alpha) + 0.1, y = 95, color = "black") +
    geom_vline(xintercept = percentile_2.5,
               color = "dark red", size = 1, linetype = "dotted") +
    annotate("text", label = "2.5 percentile", x = percentile_2.5 + 0.1, y = 95, color = "black") +
    geom_vline(xintercept = percentile_97.5,
               color = "dark green", size = 1, linetype = "dotted") +
    annotate("text", label = "97.5 percentile", x = percentile_97.5 + 0.1, y = 95, color = "black") +
    labs(title = paste("n = ", n, ", alpha = ", alpha, ", sigma = ", sigma),
         x = "Estimated MLE",
         y = "Count")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



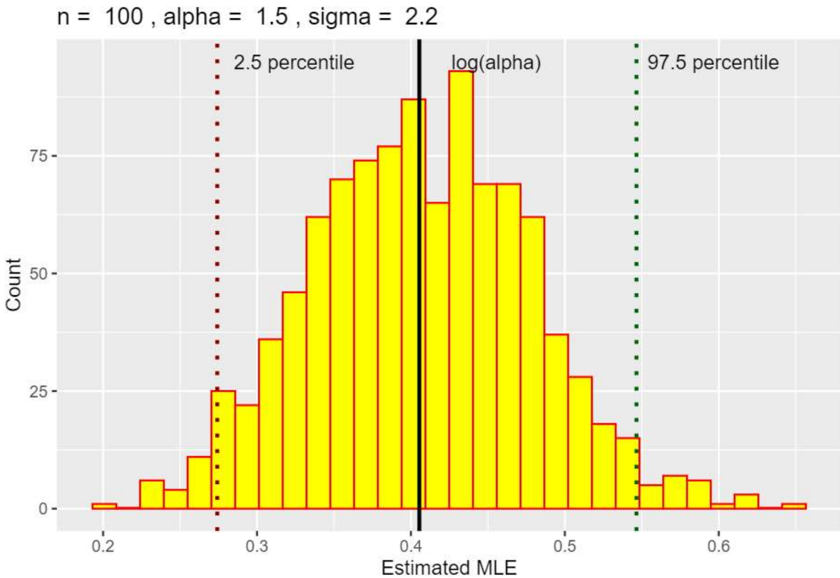

```
diff_40 <- percentile_97.5 - percentile_2.5
```

4

.

```
n = 100
alpha = 1.5
sigma = 2.2
estim_mle <- tibble(get_estim(n = n, alpha = alpha, sigma = sigma))
colnames(estim_mle) <- c("estim")
percentile_2.5 <- quantile(estim_mle$estim, probs = 0.025, names = FALSE)
percentile_97.5 <- quantile(estim_mle$estim, probs = 0.975, names = FALSE)
estim_mle %>%
    ggplot(aes(estim)) +
    geom_histogram(color = "red", fill = "yellow") +
    geom_vline(xintercept = log(alpha),
               size = 1,
               linetype = "solid") +
    annotate("text", label = "log(alpha)", x = log(alpha) + 0.05, y = 95, color = "black") +
    geom_vline(xintercept = percentile_2.5,
               color = "dark red", size = 1, linetype = "dotted") +
    annotate("text", label = "2.5 percentile", x = percentile_2.5 + 0.05, y = 95, color = "black") +
    geom_vline(xintercept = percentile_97.5,
               color = "dark green", size = 1, linetype = "dotted") +
    annotate("text", label = "97.5 percentile", x = percentile_97.5 + 0.05, y = 95, color = "black") +
    labs(title = paste("n = ", n, ", alpha = ", alpha, ", sigma = ", sigma),
         x = "Estimated MLE",
         y = "Count")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
diff_100 <- percentile_97.5 - percentile_2.5
```

5

.

```
diff_20
```

```
## [1] 0.5721496
```

```
diff_40
```

```
## [1] 0.4219604
```

```
diff_100
```

```
## [1] 0.2723683
```

Conclusion: Clearly, the gap between the percentile points is decreasing as the sample size increases.