Simulation Exercise

ThinkersPark

2023-02-12

Mean of 40 exponentials vs. CLT (Central Limit Theorem)

This report presents Part 1 of the Statistical Inference project: The simulation exercise.

In this section, we will investigate the exponential distribution in R and compare it with the Central Limit Theorem. A sample of exponential distribution is simulated in R using function $rexp(n, \lambda)$, where λ is the rate parameter. The theoretical mean of exponential distribution is $\mu = 1/\lambda$ and the theoretical standard deviation is also $\sigma = 1/\lambda$.

In this simulation ("Simulation 1", please see Annex for simulation code), the investigated distribution is that of the **sample mean $\bar{X}_{nexp}$ of 40 exponentials**, $nexp = 40$. Namely:

- In every simulation, a sample of 40 exponentials is simulated using $rexp(nexp, \lambda)$,
- For every sample of 40 exponentials, the sample mean is calculated (vector *xnexp*),
- The mean and the standard deviation of the investigated distribution (i.e. that of sample mean), are estimated (vectors *mxnexp* and *sxnexp*, respectively).
- The parameter λ is set at 0.2, for all of the simulations,
- The number of simulations is $n = 1000$.

By the Central Limit Theorem, the distribution of the normalised sample mean $\bar{X}_n$, for the sample size n, is asymptotically normal (i.e. when the sample size n increases to infinity).

$$T_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\text{Estimate} - \text{Mean of estimate}}{\text{Std. Dev. of estimate}} \rightarrow^{D}_{n\rightarrow\infty} N(0,1)$$

where:

- T_n is the statistic representing the normalised sample mean,
- μ is the theoretical distribution mean, in case this case $\mu = 1/\lambda = 5$, and μ is also the asymptotic value of the sample mean,
- σ is the theoretical distribution standard deviation, in case this case $\sigma = 1/\lambda = 5$, and $\sigma/\sqrt{n}$ is the theoretical standard deviation of the sample mean, for the sample size n,
- D means convergence of the probability distribution function.

For an estimate sample mean of a sample with size *nexp*, a useful way of thinking of its distribution is that it is close to normal, with mean μ and variance $\sigma^2/nexp$:

$$\bar{X}_{nexp} \sim N(\mu, \sigma^2/nexp)$$

The below histogram presents the distribution of the sample mean of 40 exponentials (each time), for $n = 1000$ simulations (please refer to Annex for plot generation code).

With more simulations, the distribution of the sample mean is closer to normal, centering around the theoretical distribution mean μ.

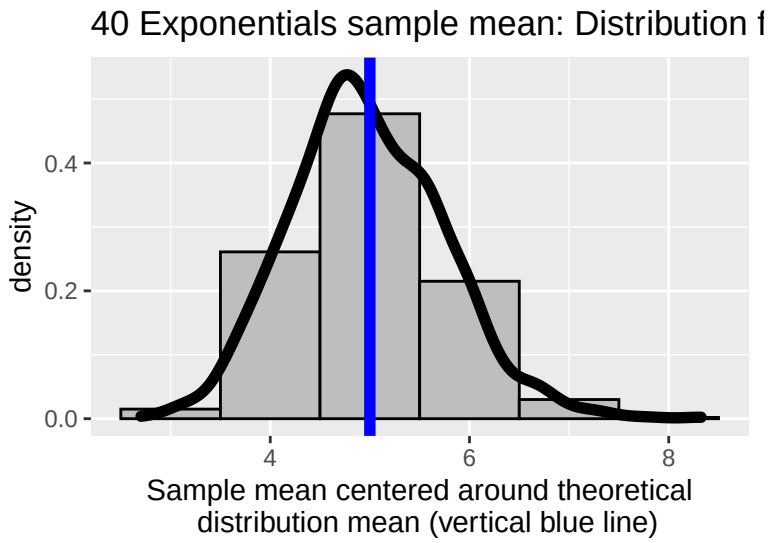

Figure 1: Distribution of the estimate sample mean of 40 exponentials (each time), for n simulations.

Estimate vs. theoretical parameters

In this section, we will show convergence of the parameters (mean and standard deviation) of the investigated distribution (i.e. that of sample mean). For the sample mean and the sample standard deviation of 40 exponentials (each time), the convergence is illustrated on the plots below (again, please refer to Annex for plot generation code).

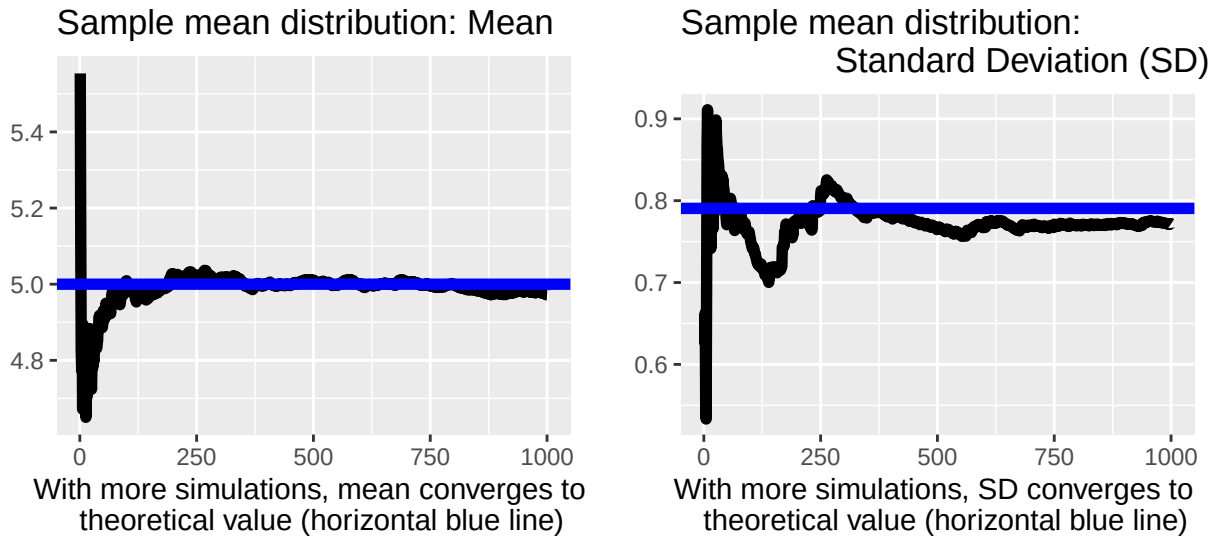

Figure 2: Convergence of the sample mean and the sample standard deviation of 40 exponentials (each time) to the theoretical parameter values.

With more simulations, the mean parameter converges to the theoretical mean μ. However with a constant sample size $nexp = 40$, the standard deviation parameter converges to the theoretical standard deviation **for this sample size**, i.e. to $\sigma/\sqrt{nexp} = 0.79$.

Large collection of averages of 40 exponentials vs. large collection of random exponentials

Now let us look at the distribution of the statistic T_n. From the Central Limit Theorem presented above, we expect this distribution to be asymptotically normal $N(0,1)$. In this simulation ("Simulation 2" - please refer to Annex for simulation code), we no longer look at a large collection of averages (of 40 exponentials each time), but at a large collection of random exponentials. Namely:

- When n increases (in theory, $n \to \infty$), the sample is simulated using $rexp(n, \lambda)$, and not $rexp(nexp, \lambda)$ like before.
- For every sample of (increasing) size n, the sample mean is calculated (vector xn),
- The statistic T_n is calculated as well, using (i) the theoretical standard deviation σ (vector $Tstat\_theosd$), and (ii) the estimated standard deviation s (vector $Tstat\_estsd$)
- The parameter λ is set at 0.2, and the number of simulations is $n = 1000$ as before.

The below two plots show how the distribution of the sample mean behaves with the increasing sample size n, first for $n = 40$ (left), then for $n = 1000$ (right). As the sample size increases, the distribution remains centered around the theoretical mean, but its standard deviation visibly decreases ($\sigma/\sqrt{n}$ decreases as n increases).

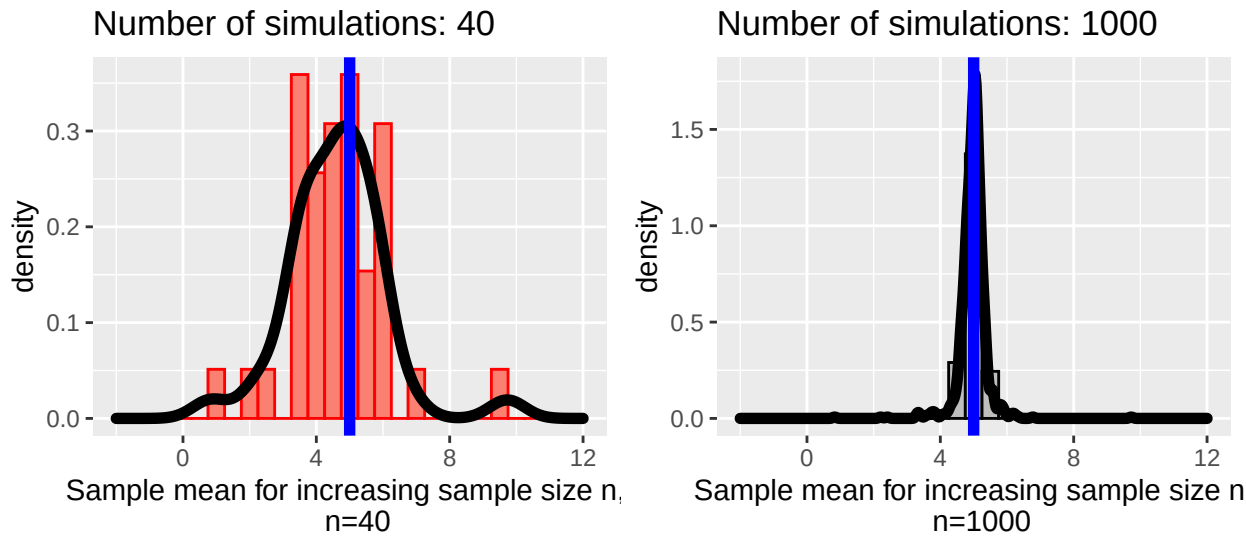

Figure 3: Distribution of the n-sample mean, for n=40 and n=1000 simulations.

Finally, we will have a look at the behaviour of the statistic T_n, calculated using the theoretical standard deviation σ, and the estimated standard deviation s, for $n = 1000$ simulations each time (part of "Simulation 2"). The below plot illustrates the convergence of T_n to standard normal distribution $N(0,1)$, for either standard deviation parameter (standard normal density is shown as reference).

Conclusions

In this exercise, the behaviour of exponential sample mean was investigated over a large number of simulations vs. the mechanics of Central Limit Theorem (CLT). In the first scenario - for the mean of 40-exponentials (each time/ constant for each simulation), in the second scenario - the mean of of an increasing number of exponentials (increasing with every simulation). In both cases, the mean was converging to the theoretical mean of the exponential distribution. However, its standard deviation behaved differently - it was convergent

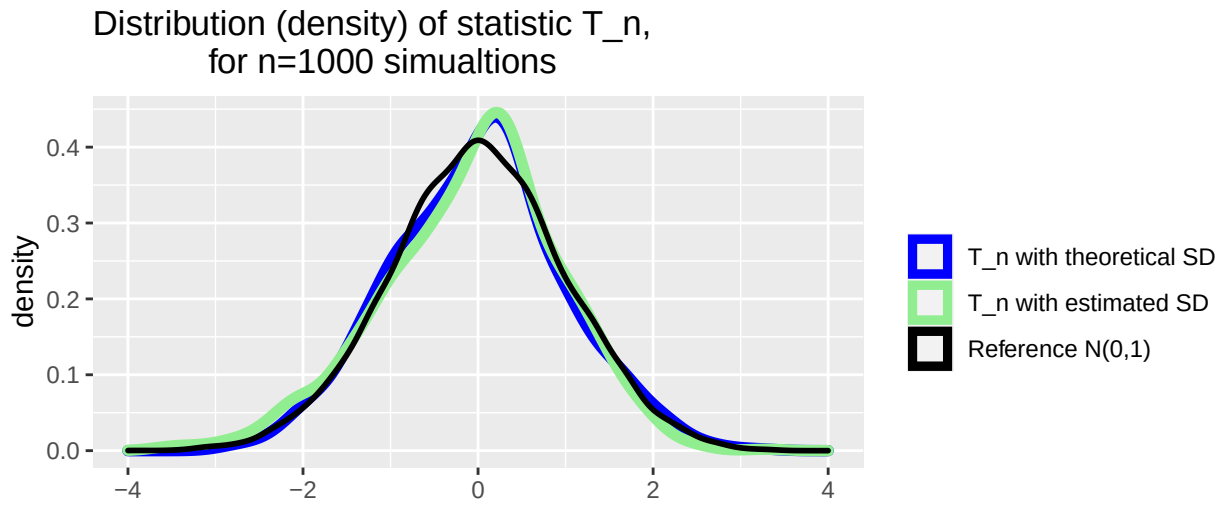

Figure 4: Distribution of the statistic T_n, for n=1000 simulations, vs. standard normal distribution.

to the theoretical standard deviation of 40-exponential sample in the first scenario, and asymptotically decreased in the second scneario. The behaviour of normalised T-statistics was demonstrated to be consistent with CLT.

Annex

Simulation code

Simulation 1: Distribution of the sample mean $\bar{X}_{nexp}$ of 40 exponentials.

```
set.seed(12345)
nsim <- 1000; nexp <- 40; lambda <- 0.2
xnexp = NULL; mxnexp = NULL; sxnexp = NULL
for (n in 1:nsim) {
  sample <- rexp(nexp,lambda)
  xnexp <- c(xnexp,mean(sample))
  mxnexp <- c(mxnexp,mean(xnexp))
  sxnexp <- c(sxnexp,sd(xnexp))
}
xnexp <- as.data.frame(xnexp)
mxnexp <- as.data.frame(mxnexp)
sxnexp <- as.data.frame(sxnexp)
```

Simulation 2: Distribution of the statistic T_n.

```
set.seed(12345)
nsim <- 1000; lambda <- 0.2
xn = NULL; Tstat_theosd = NULL; Tstat_estsd = NULL
for (n in 1:nsim) {
  sample <- rexp(n,lambda)
  xn <- c(xn,mean(sample))
```

```
  Tstat_theosd <- c(Tstat_theosd,(mean(sample)-(1/lambda))/((1/lambda)/sqrt(n)))
  Tstat_estsd <- c(Tstat_estsd,(mean(sample)-(1/lambda))/(sd(sample)/sqrt(n)))
}
xn <- as.data.frame(xn)
Tstat_theosd <- as.data.frame(Tstat_theosd)
Tstat_estsd <- as.data.frame(Tstat_estsd)
```

Plot generation code

Figure 1: "Distribution of the estimate sample mean of 40 exponentials (each time), for n simulations."

```
library(ggplot2)
g1 <- ggplot(as.data.frame(xnexp[,1]),aes(x=xnexp[,1]))
g1 <- g1 + geom_histogram(fill="grey",
      binwidth=1, aes(y=..density..),colour="black")
g1 <- g1 + geom_density(size=2)
g1 <- g1 + geom_vline(xintercept = 1/lambda, size=2, colour="blue")
g1 <- g1 + labs(title="40 Exponentials sample mean: Distribution for 1000 simulations",
x="Sample mean centered around theoretical
  distribution mean (vertical blue line)")
g1
```

Figure 2: "Convergence of the sample mean and the sample standard deviation of 40 exponentials (each time) to the theoretical parameter values."

```
library(gridExtra)
g21 <- ggplot(as.data.frame(mxnexp[,1]),aes(x=1:length(mxnexp[,1]),y=mxnexp[,1]))
g21 <- g21 + geom_line(size=2,colour="black")
g21 <- g21 + geom_hline(yintercept=(1/lambda),size=2,colour="blue")
g21 <- g21 + labs(title="Sample mean distribution: Mean", y="",
  x="With more simulations, mean converges to
  theoretical value (horizontal blue line)")
g22 <- ggplot(as.data.frame(sxnexp[,1]),aes(x=1:length(sxnexp[,1]),y=sxnexp[,1]))
g22 <- g22 + geom_line(size=2,colour="black")
g22 <- g22 + geom_hline(yintercept=(1/lambda)/sqrt(nexp),size=2,colour="blue")
g22 <- g22 + labs(title="Sample mean distribution:
                Standard Deviation (SD)", y="",
  x="With more simulations, SD converges to
  theoretical value (horizontal blue line)")
grid.arrange(g21,g22,ncol=2)
```

Figure 3: "Distribution of the n-sample mean, for n=40 and n=1000 simulations."

```
  g31 <- ggplot(as.data.frame(xn[1:4,1]),aes(x=xn[1:40,1]))
  g31 <- g31+ geom_histogram(fill="salmon",
      binwidth=0.5, aes(y=..density..),colour="red")
  g31 <- g31 + geom_density(size=2) + xlim(-2,12)
  g31 <- g31 + geom_vline(xintercept = 5, size=2, colour="blue")
  g31 <- g31 + labs(title="Number of simulations: 40",
  x="Sample mean for increasing sample size,
  n=40")
```

```
g32 <- ggplot(as.data.frame(xn[,1]),aes(x=xn[,1]))
g32 <- g32+ geom_histogram(fill="grey",
    binwidth=0.5, aes(y=..density..),colour="black")
g32 <- g32 + geom_density(size=2) + xlim(-2,12)
g32 <- g32 + geom_vline(xintercept = 5, size=2, colour="blue")
g32 <- g32 + labs(title="Number of simulations: 1000",
x="Sample mean for increasing sample size n,
n=1000")
grid.arrange(g31, g32, ncol=2)
```

Figure 4: "Distribution of the statistic T_n, for n=1000 simulations, vs. standard normal distribution."

```
Tstat <- cbind(Tstat_theosd[2:length(Tstat_theosd[,1]),1],
               Tstat_estsd[2:length(Tstat_estsd[,1]),1])
colnames(Tstat) <- c("theosd","estsd")
set.seed(12345)
normal <- rnorm(10000)
g4 <- ggplot()
g4 <- g4 + geom_density(data=as.data.frame(Tstat),
      aes(x = theosd, colour="T_n with theoretical SD"), size=2)
g4 <- g4 + geom_density(data=as.data.frame(Tstat),
          aes(x = estsd, colour="T_n with estimated SD"), size=2)
g4 <- g4 + geom_density(aes(x = normal, colour="Reference N(0,1)"), size=1)
g4 <- g4 + labs(title="Distribution (density) of statistic T_n,
          for n=1000 simualtions", x="")
g4 <- g4 + xlim(-4,4) + scale_colour_manual("",
      breaks = c("T_n with theoretical SD", "T_n with estimated SD", "Reference N(0,1)"),
      values = c("T_n with theoretical SD"="blue", "T_n with estimated SD"="lightgreen",
      "Reference N(0,1)"="black"))
g4
```