Statistical Inference

Kalpana S

OVERVIEW

A simulation experiment was conducted, in an attempt to experimentally verify three fundamental propositions for the sampling mean of a random sample with n independent identically distributed (i.i.d) random variables from any distri-bution with finite expected value μ and variance σ^2:

(P1): The expected value of the sampling mean is μ.

(P2): The variance of the sampling mean is $\frac{\sigma^2}{n}$.

(P3): As n → ∞ the distribution of the sampling mean asyptotically converges to the Normal distribution.

1. SOFTWARE

R programming language was used with the library:

```
library(tidyverse)
```

The code for the figures can be found at the APPENDIX.

2. SIMULATION

The number "7095332" was set as the random seed.

```
set.seed(7095332)
```

In this experiment, $m = 1000$ random samples had been produced, each of which with $n = 40$ observations.

```
number_of_simulations <- 1000
sample_size <- 40
```

All the observations were produced from an Exponential distribution with rate $\lambda = 0.2$.

```
lambda <- 0.2
```

With the function 'rexp()' the 1000 random samples were obtained, each with 40 observations from the Exponential distribution with rate, $\lambda = 0.2$.

SUPPLEMENTARY INFORMATION
Further details can be found at the
GitHub repository:https://github.com/Kalpana912k/Statistical-Inference

```
simulated_samples <- tibble(
  "sim_id" = seq_len(number_of_simulations),
  "random_sample" = lapply(
    X = sim_id,
    FUN = function(i) rexp(sample_size, lambda)
  )
) %>%
  unnest(random_sample) %>%
  mutate(
    "obs_id" = rep(
      seq_len(sample_size),
      times = number_of_simulations
    )
  ) %>%
  select(sim_id, obs_id, "value" = random_sample)
```

3. SIMULATED SAMPLES

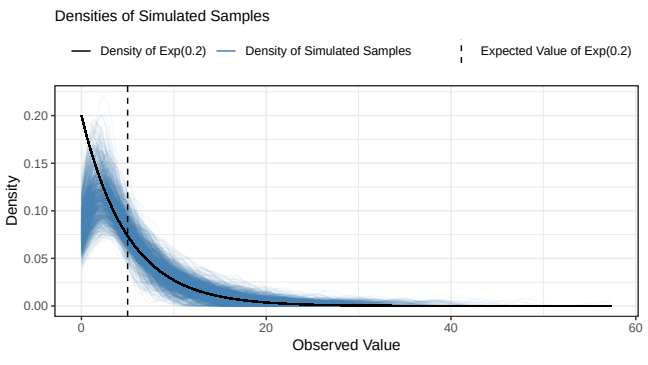
From the simulation, a 1000 samples y_i were obtained, which contain observations that are the realizations of the 40000 i.i.d. random variables $Y_{i,j} \sim Exp(0.2)$ where $i = 1, 2, \ldots, 1000$ and $j = 1, 2, \ldots 40$.

A r.v. $Y_{i,j} \sim Exp(\lambda)$, where $\lambda \in (0, \infty)$, has an expected value $\mu_Y := E(Y_{i,j}) = \frac{1}{\lambda}$ and variance $\sigma_Y^2 := Var(Y_{i,j}) = \frac{1}{\lambda^2}$.

So for $\lambda = 0.2$ it is $\mu_Y = 5$ and $\sigma_Y^2 = 25$.

```
origin_distr_____expected_value <- 1/lambda
origin_distr_____variance <- 1/(lambda^2)
```

Figure 1



A comparison of the densities from the 1000 simulated samples versus the Exponential distribution from which their observations were sampled.

4. SAMPLE OF SAMPLE MEANS

For each random sample $\underline{Y}_i = (Y_{i,1}, Y_{i,2}, \ldots, Y_{i,40})$, the r.v. $\bar{Y}_i = \frac{1}{40} \cdot \sum_{j=1}^{40} Y_{ij}$, is the sampling mean.

The sampling means $\bar{Y}_i$, of the the 1000 random samples, are i.i.d. random variables because they are linear transformations of the i.i.d. random variables $Y_{i,j}$.

Define $X_i := \bar{Y}_i$, so the r.v X_i has expected value $\mu_X = E(X_i) = E(\bar{Y}_i)$ and variance $\sigma_X^2 = Var(X_i) = Var(\bar{Y}_i)$.

The sample $\underline{x} = (x_1, x_2, \ldots, x_{1000}) \equiv (\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_{1000})$ where $x_i = \bar{y}_i = \frac{1}{40} \cdot \sum_{j=1}^{40} y_{i,j}$, consists of the observed sample means of 1000 original samples $\underline{y}_1, \underline{y}_2, \ldots, \underline{y}_{1000}$.

```
sample_means <- simulated_samples %>%
  group_by(sim_id) %>%
  summarise("value" = mean(value))
```

5. MEAN OF SAMPLE MEANS

According to the proposition (P1), the r.v. X_i defined as equivalent to the sampling mean $\bar{Y}_i$ of the i-th random sample with n observations, has an expected value equal to the expected value of the distribution from which the sample was taken.

So $\mu_X = E(X_i) = E(\bar{Y}_i) = E(Y_{ij}) = \mu_Y = 5$.

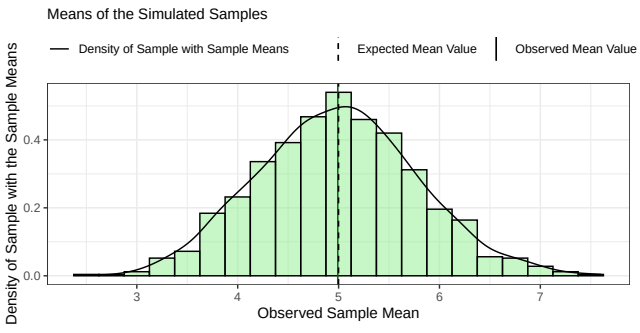
```
expected_mean_of_sampling_mean <-
  origin_distr_____expected_value
```

Indeed the mean $\bar{x} = 4.992119$, of the sample $\underline{x}$, is 'very close' to the theoretical expected value $\mu_X = 5$.

```
(mean_of_sample_means <-
   mean(sample_means$value))
```

```
## [1] 4.992119
```

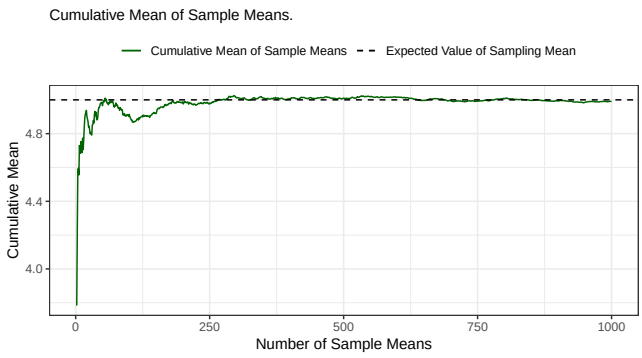
Figure 2



The Histogram and the Density of Sample with the Means of the Simulated Samples.
The Mean of all Sample Means is 'close' to the Expected Value
of the Exponential Distribution from which the samples were produced
but the density curve is not similar.

Furthermore as the number of means from sample means that are taking into account increases the cumulative mean converges to the expected value $\mu_x = 5$ (Figure 3).

Figure 3



The cumulative mean of i.i.d. sample means, converges to the theoretical expected value
of the sampling mean, as the number of the observed means increases.

6. VARIANCE OF SAMPLE MEANS

According to the proposition (P2), the r.v. X_i defined as equivalent to the sampling mean $\bar{Y}_i$ of the i-th random sample with n observations, has a variance equal to the variance of the distribution from which the sample was taken divided by the sample size n.

So $\sigma_X^2 = Var(X_i) = Var(\bar{Y}_i) = \frac{Var(Y_{ij})}{n} = \frac{\sigma_Y^2}{n} = 0.625$.

```
expected_variance_of_sampling_mean <-
  origin_distr_____variance / sample_size
```

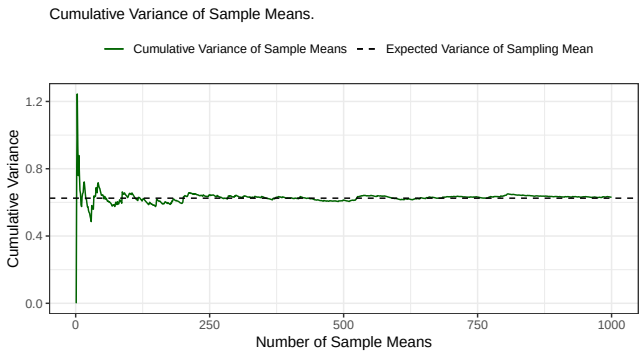
Indeed the sample variance $s^2 = 0.625151$, of the sample $\underline{x}$, is 'very close' to the theoretical variance $\sigma_X^2 = 0.625$.

```
(variance_of_sample_means <-
   var(sample_means$value))
```

```
## [1] 0.625151
```

Furthermore as the number of means from sample means that are taking into account increases the cumulative variance converges to the expected value $\sigma_X^2 = 5$ (Figure 4).

Figure 4



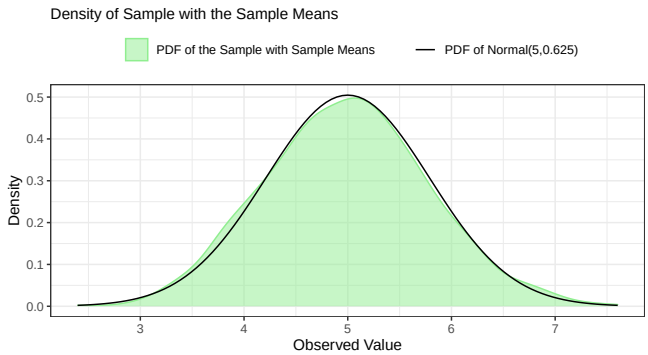
The cumulative variance of i.i.d. sample means, converges to the theoretical variance
of the sampling mean, as the number of the observed means increases.

7. DISTRIBUTION OF SAMPLE MEANS

According to the proposition (P3), the distribution of the r.v. X_i defined as equivalent to the sampling mean $\bar{Y}_i$ of the i-th random sample with n observations, is approximately Normal with expected value μ_X and variance σ^2_X.

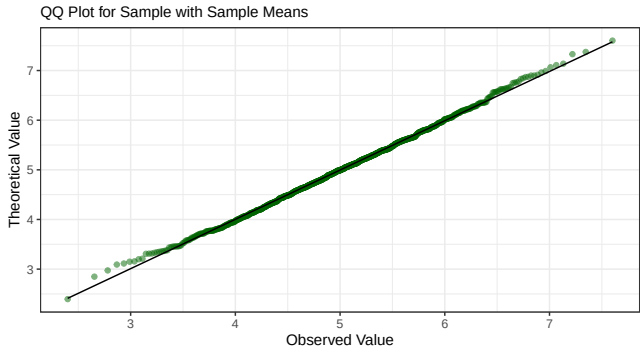
From a visual examination of the density (Figure 5), and the QQ plot (Figure 6) of the sample $\underline{x}$ with the means of the simulated samples it is clear that the observed values fit 'very well' those from a Normal distribution with expected value $\mu_X = 5$ and variance $\sigma^2_X = 0.625$.

Figure 5



The Density of the Sample with Sample Meansseems to fit very well
the PDF of the Normal Distribution with expected value 5 and variance 0.625.

Figure 6



The Observed Values of the Sample with Sample Means seems to correspond very well
to the Theoretical values from the Normal Distribution with expected value 5 and variance 0.625.

The normality assumption was also verified with the Shapiro-Wilk test, from which a p-value 0.6803 was obtained, indicating that there are not enough evidence to discard the null hypothesis, that the sample comes from a Normal Distribution.

```
shapiro.test(sample_means$value)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sample_means$value
## W = 0.99869, p-value = 0.6803
```

8. REMARKS

It is quite possible, especially for the proposition (P3), to obtain samples that may not pass the 'Shapiro-Wilk' test for normality. By increasing the number of simulated samples and/or the sample size of each of them, the observations will eventually conform with the claims of Central Limit Theorem (CTL) and the Law of Large Numbers (LLN).

9. REFERENCES

Caffo, B. (2016). Statistical inference for data science. Retrived from `https://leanpub.com/LittleInferenceBook`

S. S. SHAPIRO, M. B. WILK, An analysis of variance test for normality (complete samples), Biometrika, Volume 52, Issue 3-4, December 1965, Pages 591–611, `https://doi.org/10.1093/biomet/52.3-4.591`

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL `http://www.rstudio.com/`.

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, `https://doi.org/10.21105/joss.01686`

JJ Allaire and Yihui Xie and Jonathan McPherson and Javier Luraschi and Kevin Ushey and Aron Atkins and Hadley Wickham and Joe Cheng and Winston Chang and Richard Iannone (2020). rmarkdown: Dynamic Documents for R. R package version 2.3. URL `https://rmarkdown.rstudio.com`.

Yihui Xie and J.J. Allaire and Garrett Grolemund (2018). R Markdown: The Definitive Guide. Chapman and Hall/CRC. ISBN 9781138359338. URL `https://bookdown.org/yihui/rmarkdown`.

Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.28.

Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963

Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595

JJ Allaire, Yihui Xie, R Foundation, Hadley Wickham, Journal of Statistical Software, Ramnath Vaidyanathan, Association for Computing Machinery, Carl Boettiger, Elsevier, Karl Broman, Kirill Mueller, Bastiaan Quast, Randall Pruim, Ben Marwick, Charlotte Wickham, Oliver Keyes, Miao Yu, Daniel Emaasit, Thierry Onkelinx, Alessandro Gasparini, Marc-Andre Desautels, Dominik Leutnant, MDPI, Taylor and Francis, Oğuzhan Öğreden, Dalton Hance, Daniel Nüst, Petter Uvesten, Elio Campitelli, John Muschelli, Zhian N. Kamvar, Noam Ross, Robrecht Cannoodt, Duncan Luguern and David M. Kaplan (2020). rticles: Article Formats for R Markdown. R package version 0.14. `https://CRAN.R-project.org/package=rticles`

10. APPENDIX

10.1 Code For Figure 1

The code below was used to create the Figure 1.

```
figure_1 <- ggplot(
  simulated_samples[1:400,],
  aes(x = value, group = sim_id)
) +
  geom_line(
    aes(color = "Density of Simulated Samples"),
    stat = "density",
    alpha = 0.05
  ) +
  stat_function(
    aes(color = "Density of Exp(0.2)"),
    fun = dexp,
    args = list("rate" = lambda),
    geom = "line"
  ) +
  scale_color_manual(
    values = c("black", "steelblue")
  ) +
  geom_vline(
    aes(
      xintercept = origin_distr_____expected_value,
      linetype = "Expected Value of Exp(0.2)"
    )
  ) +
  scale_linetype_manual(values = "dashed") +
  labs(
    title = "Figure 1\n",
    subtitle = "Densities of Simulated Samples",
    x = "Observed Value",
    y = "Density",
    color = "",
    linetype = "",
    caption = paste0(
      "\n",
      "A comparison of the densities ",
      "from the 1000 simulated samples ", "\n",
      "versus the Exponential distribution ",
      "from which their observations were sampled."
    )
  ) +
  theme_bw() +
  theme(
    plot.title = element_text(hjust = 0.5),
    plot.caption = element_text(hjust = 0.5),
    legend.position = "top"
  )
```

10.2 Code For Figure 2

The code below was used to create the Figure 2.

```
figure_2 <- ggplot(
  sample_means,
  aes(x = value, y = ..density..)
) +
  geom_histogram(
    binwidth = 0.25,
    color = "black",
    fill = "lightgreen",
    alpha = 0.5
  ) +
  geom_line(
    aes(color =
        "Density of Sample with Sample Means"
    ),
    stat = "density"
  ) +
  scale_color_manual(values = "black") +
  geom_vline(
    aes(
      xintercept = mean_of_sample_means,
      linetype = "Observed Mean Value"
    ),
    color = "darkgreen"
  ) +
  geom_vline(
    aes(
      xintercept = origin_distr_____expected_value,
      linetype = "Expected Mean Value"
    ),
    color = "black"
  ) +
  scale_linetype_manual(
    values = c("dashed", "solid")
  ) +
  labs(
    title = "Figure 2\n",
    subtitle = "Means of the Simulated Samples",
    x = "Observed Sample Mean",
    y = "Density of Sample with the Sample Means",
    color = "",
    linetype = "",
    caption = paste0(
      "\n",
      "The Histogram and the Density of Sample ",
      "with the Means of the Simulated Samples.",
      "\n",
      "The Mean of all Sample Means is 'close' " ,
      "to the Expected Value ", "\n",
      "of the Exponential Distribution from ",
      "which the samples were produced ", "\n",
      "but the density curve is not similar."
    )
  ) +
  theme_bw() +
  theme(
    plot.title = element_text(hjust = 0.5),
    plot.caption = element_text(hjust = 0.5),
    legend.position = "top",
    legend.direction = "horizontal"
  )
```

10.3 Code For Figure 3

The code below was used to create the Figure 3.

```
figure_3 <- ggplot(
  sample_means,
  aes(x = sim_id, y = cummean(value))
) +
  geom_line(
    aes(
      color = "Cumulative Mean of Sample Means"
    )
  ) +
  geom_hline(
    aes(
      yintercept = expected_mean_of_sampling_mean,
      color = "Expected Value of Sampling Mean"
    ),
    linetype = "dashed"
  ) +
  scale_color_manual(
    values = c("darkgreen", "black")
  ) +
  guides(
    color = guide_legend(
      override.aes = list(
        linetype = c("solid", "dashed")
      )
    )
  ) +
  labs(
    title = "Figure 3\n",
    subtitle =
      "Cumulative Mean of Sample Means.",
    x = "Number of Sample Means",
    y = "Cumulative Mean",
    color = "",
    caption = paste(
      "\n",
      "The cumulative mean of i.i.d. ",
      "sample means, converges to the ",
      "theoretical expected value ", "\n",
      "of the sampling mean, as the number ",
      "of the observed means increases."
    )
  ) +
  theme_bw() +
  theme(
    plot.title = element_text(hjust = 0.5),
    plot.caption = element_text(hjust = 0.5),
    legend.position = "top"
  )
```

10.4 Code For Figure 4

The code below was used to create the Figure 4.

```
figure_4 <- ggplot(
  sample_means,
  aes(
    x = sim_id,
    y = cumsum((value - cummean(value))^2)/sim_id)
) +
  geom_line(
    aes(
      color = "Cumulative Variance of Sample Means"
    )
  ) +
  geom_hline(
    aes(
      yintercept =
        expected_variance_of_sampling_mean,
      color = "Expected Variance of Sampling Mean"
    ),
    linetype = "dashed"
  ) +
  scale_color_manual(
    values = c("darkgreen", "black")
  ) +
  guides(
    color = guide_legend(
      override.aes = list(
        linetype = c("solid", "dashed")
      )
    )
  ) +
  labs(
    title = "Figure 4\n",
    subtitle =
      "Cumulative Variance of Sample Means.",
    x = "Number of Sample Means",
    y = "Cumulative Variance",
    color = "",
    caption = paste(
      "\n",
      "The cumulative variance of i.i.d. ",
      "sample means, converges to the ",
      "theoretical variance ", "\n",
      "of the sampling mean, as the number ",
      "of the observed means increases."
    )
  ) +
  theme_bw() +
  theme(
    plot.title = element_text(hjust = 0.5),
    plot.caption = element_text(hjust = 0.5),
    legend.position = "top"
  )
```

10.5 Code For Figure 5

The code below was used to create the Figure 5.

```
figure_5 <- ggplot(
  sample_means,
  aes(x = value)
) +
  geom_density(
    aes(
      fill = "PDF of the Sample with Sample Means"
    ),
    color = "lightgreen",
    alpha = 0.5
  ) +
  scale_fill_manual(
    values = "lightgreen"
  ) +
  stat_function(
    aes(color = "PDF of Normal(5,0.625)"),
    fun = dnorm,
    args = list(
      "mean" = expected_mean_of_sampling_mean,
      "sd" =
        sqrt(expected_variance_of_sampling_mean)
    ),
    geom = "line"
  ) +
  scale_color_manual(values = "black") +
  labs(
    title = "Figure 5\n",
    subtitle =
      "Density of Sample with the Sample Means",
    x = "Observed Value",
    y = "Density",
    color = "",
    fill = "",
    caption = paste0(
      "\n",
      "The Density of the Sample with Sample Means",
      "seems to fit very well ", "\n",
      "the PDF of the Normal Distribution with ",
      "expected value 5 and variance 0.625."
    )
  ) +
  theme_bw() +
  theme(
    plot.title = element_text(hjust = 0.5),
    plot.caption = element_text(hjust = 0.5),
    legend.position = "top"
  )
```

10.6 Code For Figure 6

The code below was used to create the Figure 6.

```
figure_6 <- ggplot(
  sample_means,
  aes(sample = value)
) +
  geom_qq(
    distribution = qnorm,
    dparams = list(
      "mean" = expected_mean_of_sampling_mean,
      "sd" =
        sqrt(expected_variance_of_sampling_mean)
    ),
    color = "darkgreen",
    alpha = 0.5
  ) +
  geom_qq_line(
    distribution = qnorm,
    dparams = list(
      "mean" = expected_mean_of_sampling_mean,
      "sd" =
        sqrt(expected_variance_of_sampling_mean)
    )
  ) +
  labs(
    title = "Figure 6\n",
    subtitle =
      "QQ Plot for Sample with Sample Means",
    x = "Observed Value",
    y = "Theoretical Value",
    caption = paste0(
      "\n",
      "The Observed Values of the Sample with ",
      "Sample Means seems to correspond very well ",
      "\n",
      "to the Theoretical values from the Normal " ,
      "Distribution with expected value 5 ",
      "and variance 0.625."
    )
  ) +
  theme_bw() +
  theme(
    plot.title = element_text(hjust = 0.5),
    plot.caption = element_text(hjust = 0.5)
  )
```