The Central Limit Theorem (CLT)

Basically, the Central Limit Theorem states that the distribution of the sum (or average) of a large number of i.i.d. random variables converges in law to a normal distribution regardless of the original distribution.

Let $X_1, X_2, ..., X_n$ be a sequence of independent and identically distributed ($i.i.d$) random variables with $E[X] = \mu, < \infty$ (finite first moment), and $Var(X) = \sigma^2$ exists.

(CLT) As n becomes large, then the distribution of $\frac{1}{n} \sum_{i=1}^{n} x_i$ (the sample mean) is approximately a Normal distribution $N\left(\mu, \frac{\sigma^2}{n}\right)$.

Useful direct consequence

As a consequence, we also have that

$$\frac{\sum_{i=1}^{n}(x_i - \mu)}{\sqrt{n}\ \sigma} \xrightarrow{L} N(0, 1)$$

Indeed, we have that

$$\sum_{i=1}^{n}(x_i - \mu) \xrightarrow{L} N(0, n\sigma^2)$$

$$\sum_{i=1}^{n}x_i \xrightarrow{L} N(n\mu, n\sigma^2)$$

$$\frac{1}{n}\sum_{i=1}^{n}x_i \xrightarrow{L} N\left(\mu, \frac{\sigma^2}{n}\right)$$

Example 1: Exponential variables

Example 1   Let $y_1, y_2, ..., y_n$ be $i.i.d$ exponential realizations of $Y_i \sim \mathcal{E}(\lambda = 5)$ and $n = 1,000$. The theoretical mean is $E[Y] = 1/\lambda = 1/5 = 0.2$ and the theoretical variance is $Var(Y) = 1/5^2$. Then we expect the sample average $\frac{1}{n}\sum_{i=1}^{n} y_i$ to converge to a Normal distribution $N\left(E[Y] = \mu = \frac{1}{5}, \frac{\sigma^2}{n} = \frac{1}{25,000}\right)$
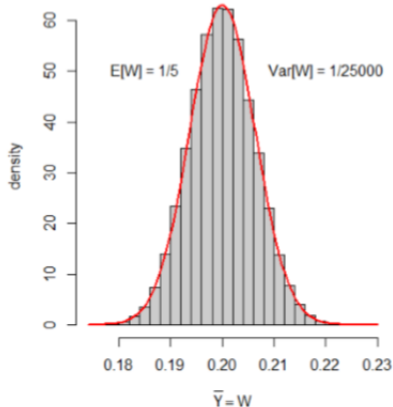
In other words: The distribution of the samples means of Exponentially distributed samples will obey a Normal distribution.

Example 1: Visualization in R


Distribution of Exponential random variates

$E[Y] = 1/5$

$Var[Y] = 1/25$

Distribution of the empirical mean

$E[W] = 1/5$ $Var[W] = 1/25000$
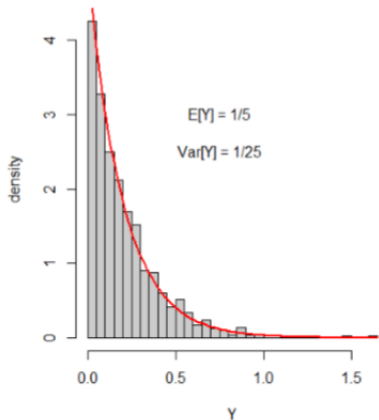
R code

```
1  set.seed(2023)
2  par(mfrow=c(1,2))
3  lambda = 5
4  n = 1000
5  x = rexp(n, lambda)  # random exponential variates
6  hist(x, breaks=30, main="Distribution of Exponential random variates", col="
      gray80",
7      xlab="Y", ylab="density", prob=TRUE )
8  # obviously the distribution is very skewed (to the right)
9  curve(dexp(x, rate=lambda),
10      col="red", lwd=2, add=TRUE, yaxt="n")
11 text(x=0.75, y=3, expression("E[Y] = 1/5"), cex=1)
12 text(x=0.75, y=2.5, expression("Var[Y] = 1/25"), cex=1)
13
14 mean(x) #[1] 0.2023026 ; var(x) # [1] 0.03969075
15
16 # Illustration of the CLT
17 nsim = 100000; x1 = rep(0,nsim)
18 for(i in 1:nsim){
19   x1[i] <- mean(rexp(n, 5))
20 }
21
22 hist(x1, breaks=20, main="Distribution of the empirical mean", col="gray80",
23      xlab=expression(bar(Y)==W), ylab="density", prob=TRUE )
24 curve(dnorm(x, mean=1/lambda, sd=1/lambda/sqrt(n)),
25      col="red", lwd=2, add=TRUE, yaxt="n")
26
27 # the distribution of the empirical mean tends to a nice N(0,1) distr.
28 mean(x1) # [1] 0.1999819
29 var(x1) # [1] 3.994372e-05
30 text(x=0.185, y=50, expression("E[W] = 1/5") , cex=1)
31 text(x=0.22, y=50, expression("Var[W] = 1/25000"), cex=1)
```

Example 1: Visualization in Python
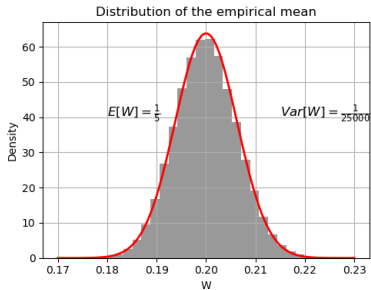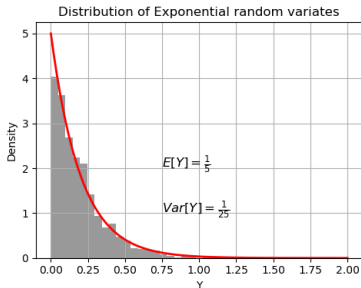
Python code

```python
1  import numpy as np
2  from scipy.stats import norm
3  import matplotlib.pyplot as plt
4
5  np.random.seed(2023)
6  lambda_val = 5
7  n = 1000
8  x = np.random.exponential(scale=1/lambda_val, size=n)
9
10 plt.figure(figsize=(10, 4))
11 plt.subplot(1, 2, 1)
12 plt.hist(x, bins=30, density=True, color='gray', alpha=0.8)
13 plt.plot(np.linspace(0, 2, 100), lambda_val *
14          np.exp(-lambda_val * np.linspace(0, 2, 100)), 'r', linewidth=2)
15 plt.xlabel('Y'); plt.ylabel('Density')
16 plt.title('Distribution of Exponential random variates')
17 plt.text(0.75, 2, r'$E[Y] = \frac{1}{5}$', fontsize=12)
18 plt.text(0.75, 1, r'$Var[Y] = \frac{1}{25}$', fontsize=12); plt.grid()
19
20 sample_means = np.zeros(100000)
21 for i in range(100000):
22     sample_means[i] = np.mean(np.random.exponential(scale=1/lambda_val, size=n))
23
24 plt.subplot(1, 2, 2)
25 plt.hist(sample_means, bins=30, density=True, color='gray', alpha=0.8)
26 xs = np.linspace(0.17, 0.23, 100)
27 pdf = norm.pdf(xs, 0.2, 25/4000)
28 plt.plot(xs, pdf, 'k', linewidth=2, color = 'red')
29 plt.xlabel('W'); plt.ylabel('Density')
30 plt.title('Distribution of the empirical mean')
31 plt.text(0.18, 40, r'$E[W] = \frac{1}{5}$', fontsize=12)
32 plt.text(0.215, 40, r'$Var[W] = \frac{1}{25000}$', fontsize=12); plt.grid()
33 plt.tight_layout(); plt.show()
```

Example 2: Application 1/3

Example 2 According to a certain maritime organization, the distribution of the length of a specific fish species at maturity in the Mekong delta has mean $\mu = 83$ cm. and standard deviation $\sigma = 7$ cm. The random variable Y denotes the length of these fishes.

(i) Suppose we sample 25 individuals. What is the probability that the sample average is above 86 cm. ?

(i) From the CLT, we know that the distribution of the sample average $\frac{1}{25} \sum_{i=1}^{25} y_i$ has a Normal distribution $N\left(\mu = 83, \frac{\sigma}{\sqrt{n}} = \frac{7}{5}\right)$.

So, $P\left(\overline{y} > 86\right) = P\left(Z > \frac{\overline{y} - \mu}{\sigma/\sqrt{n}}\right) = P\left(Z > \frac{86 - 83}{1.4}\right) = P\left(Z > 2.142857\right) = \mathbf{0.01606 \approx 1.6\%}$

```
1 pnorm(q=86, mean=83, sd=7/5, lower.tail = FALSE)
2 # 0.01606229
```

```
1 1 - norm.cdf(86, loc=83, scale=7/5)
2 # 0.016062285603828275
```

Example 2: Application 2/3

Example 2 (continued) According to a certain maritime organization, the distribution of the length of a specific fish species at maturity in the Mekong delta has mean $\mu = 83$ cm. and standard deviation $\sigma = 7$ cm. The random variable Y denotes the length of these fishes.

(ii) Give a 95% Confidence Interval for the sample average length.

(ii) From the CLT, we know that the distribution of the sample average $\frac{1}{25}\sum_{i=1}^{25} y_i$ has a Normal distribution $N\left(\mu = 83, \frac{\sigma}{\sqrt{n}} = \frac{7}{5}\right)$.

So, $P\left(-1.96 > Z > 1.96\right) = 95\%$ $\Leftrightarrow$ $P\left(\mu - 1.96 * \sigma/\sqrt{n} > \overline{y} > \mu + 1.96 *\right.$

$\left.\sigma/\sqrt{n}\right) = 95\%$ $\Leftrightarrow$ $\left[\mathbf{80.26}, \mathbf{85.74}\right]$

```
1 c((83 - 1.96*(7/5)), (83 + 1.96*(7/5)))
2 # [1] 80.256 85.744
```

```
1 [(83 - 1.96*(7/5)), (83 + 1.96*(7/5))]
2 # [80.256, 85.744]
```

Example 2: Application 3/3

Example 2 (continued) According to a certain maritime organization, the distribution of the length of a specific fish species at maturity in the Mekong delta has mean $\mu = 83$ cm. and standard deviation $\sigma = 7$ cm. The random variable Y denotes the length of these fishes.

(iii) How many mature tuna must we sample if we want a 95 % confidence that the sample average is within 1 cm. from the population mean ?

(iii) From the CLT, we know that the distribution of the sample average $\frac{1}{25} \sum_{i=1}^{n} y_i$ has a Normal distribution

$$N\left(\mu = 83, \frac{\sigma}{\sqrt{n}} = \frac{7}{\sqrt{n}}\right).$$

So, $Z = \frac{\bar{y} - \mu}{\sigma / \sqrt{n}} \quad \Leftrightarrow \quad 1.96 = \frac{1}{7/\sqrt{n}} \quad \Leftrightarrow \quad n = (1.96 * 7)^2 \approx \mathbf{189}$ fishes

So we replace all we know and solve for n, the sample size.

References

R.V.Hogg and E.A.Tanis: Probability and Statistical Inference, Sixth Edition, Prentice Hall, Upper Saddle River, N.J., 2001.

Ross, S., A First Course in Probability, Eighth Edition, Pearson, 2010

The R Project for Statistical Computing:
https://www.r-project.org/

Python:
https://www.python.org/

course notes