The Central Limit Theorem (CLT)

The Central Limit Theorem states that the distribution of the sum (or average) of a large number of independent, identically distributed random variables approaches 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

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

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

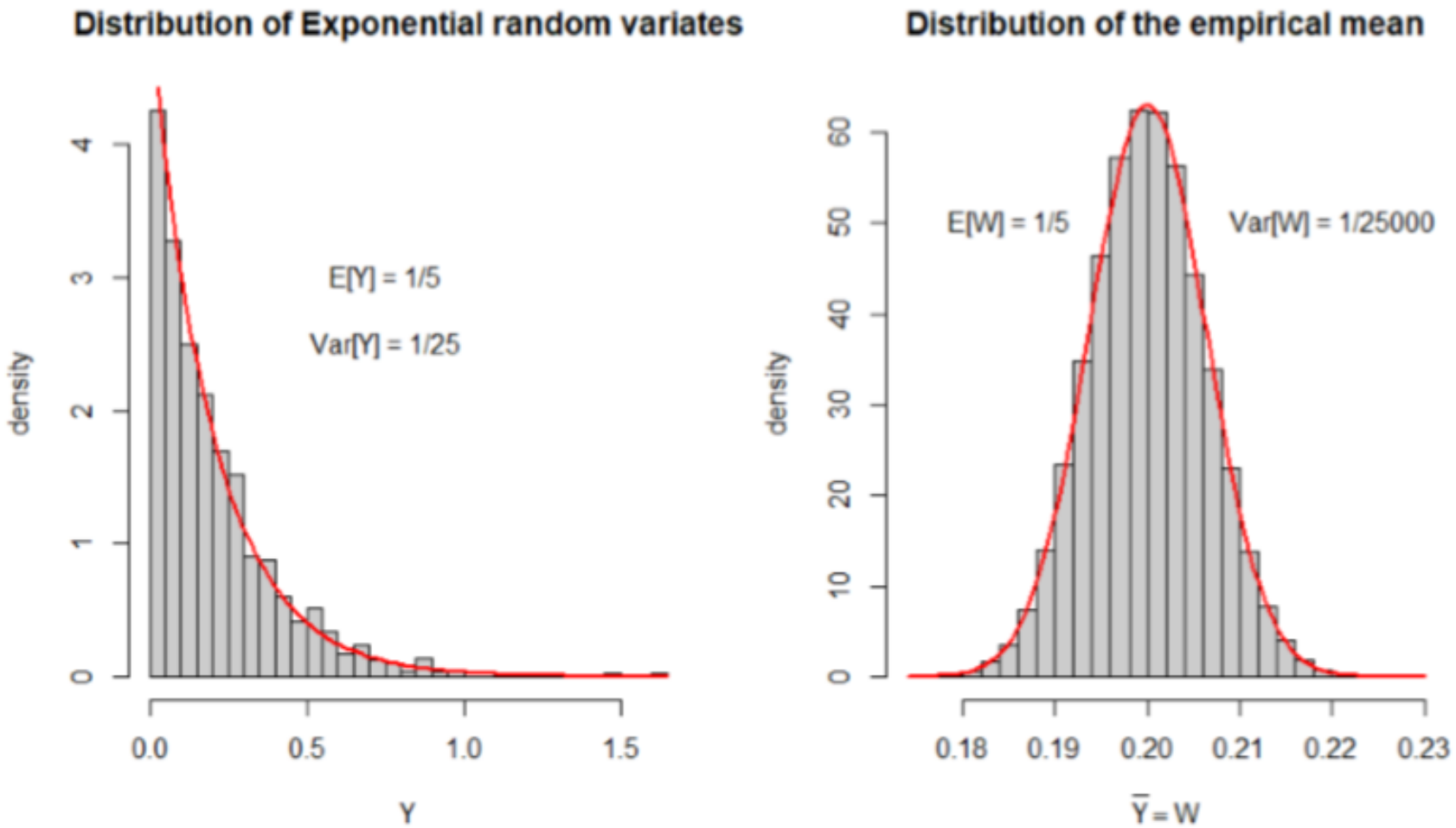
$$\frac{1}{n}\sum_{i=1}^{n} x_i \xrightarrow{L} N\left(\mu, \frac{\sigma^2}{n}\right)$$

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 measn of Exponentially distributed samples will obey a Normal distribution.

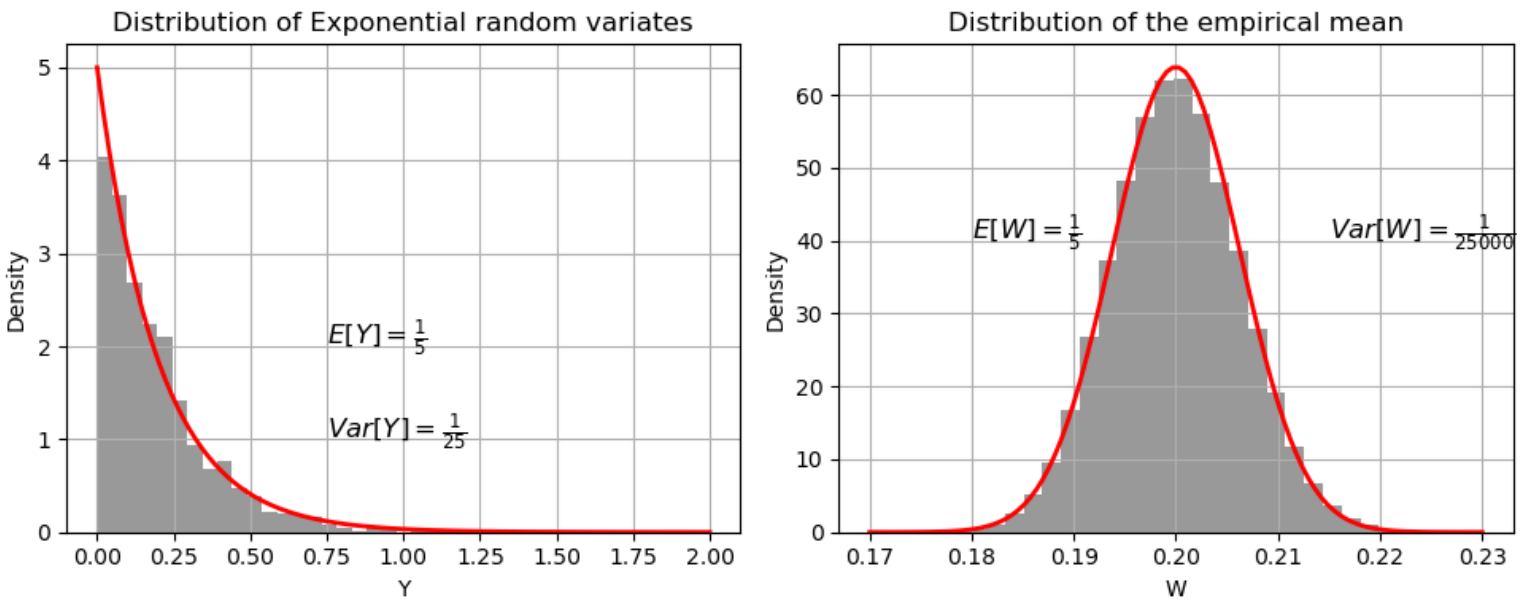
Example 1: Visualization in R



R code

```r
set.seed(2023)
par(mfrow=c(1,2))
lambda = 5
n = 1000
x = rexp(n, lambda)  # random exponential variates
hist(x, breaks=30, main="Distribution of Exponential random variates", col="
     gray80",
     xlab="Y", ylab="density", prob=TRUE )
# obviously the distribution is very skewed (to the right)
curve(dexp(x, rate=lambda),
      col="red", lwd=2, add=TRUE, yaxt="n")
text(x=0.75, y=3, expression("E[Y] = 1/5"), cex=1)
text(x=0.75, y=2.5, expression("Var[Y] = 1/25"), cex=1)

mean(x) #[1] 0.2023026 ; var(x) # [1] 0.03969075

# Illustration of the CLT
nsim = 100000; x1 = rep(0,nsim)
for(i in 1:nsim){
  x1[i] <- mean(rexp(n, 5))
}

hist(x1, breaks=20, main="Distribution of the empirical mean", col="gray80",
     xlab=expression(bar(Y)==W), ylab="density", prob=TRUE )
curve(dnorm(x, mean=1/lambda, sd=1/lambda/sqrt(n)),
      col="red", lwd=2, add=TRUE, yaxt="n")

# the distribution of the empirical mean tends to a nice N(0,1) distr.
mean(x1) # [1] 0.1999819
var(x1) # [1] 3.994372e-05
text(x=0.185, y=50, expression("E[W] = 1/5") , cex=1)
text(x=0.22, y=50, expression("Var[W] = 1/25000"), cex=1)
```

Example 1: Visualization in Python



Python code

```python
import numpy as np
from scipy.stats import norm
import matplotlib.pyplot as plt

np.random.seed(2023)
lambda_val = 5
n = 1000
x = np.random.exponential(scale=1/lambda_val, size=n)

plt.figure(figsize=(10, 4))
plt.subplot(1, 2, 1)
plt.hist(x, bins=30, density=True, color='gray', alpha=0.8)
plt.plot(np.linspace(0, 2, 100), lambda_val *
            np.exp(-lambda_val * np.linspace(0, 2, 100)), 'r', linewidth=2)
plt.xlabel('Y'); plt.ylabel('Density')
plt.title('Distribution of Exponential random variates')
plt.text(0.75, 2, r'$E[Y] = \frac{1}{5}$', fontsize=12)
plt.text(0.75, 1, r'$Var[Y] = \frac{1}{25}$', fontsize=12); plt.grid()

sample_means = np.zeros(100000)
for i in range(100000):
    sample_means[i] = np.mean(np.random.exponential(scale=1/lambda_val, size=n))

plt.subplot(1, 2, 2)
plt.hist(sample_means, bins=30, density=True, color='gray', alpha=0.8)
xs = np.linspace(0.17, 0.23, 100)
pdf = norm.pdf(xs, 0.2, 25/4000)
plt.plot(xs, pdf, 'k', linewidth=2, color = 'red')
plt.xlabel('W'); plt.ylabel('Density')
plt.title('Distribution of the empirical mean')
plt.text(0.18, 40, r'$E[W] = \frac{1}{5}$', fontsize=12)
plt.text(0.215, 40, r'$Var[W] = \frac{1}{25000}$', fontsize=12); plt.grid()
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, $P\left(\overline{y} > 86\right) = P\left(Z > \frac{\overline{y}-\mu}{\sigma/\sqrt{n}}\right) = P\left(Z > \frac{86-83}{1.4}\right) = P\left(Z > 2.142857\right) = \mathbf{0.01606} \approx \mathbf{1.6\%}$

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

```python
1 - norm.cdf(86, loc=83, scale=7/5)
# 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, $P\Big(-1.96 > Z > 1.96\Big) = 95\% \quad \Leftrightarrow \quad P\Big(\mu - 1.96 * \sigma/\sqrt{n} > \overline{y} > \mu + 1.96 * \sigma/\sqrt{n}\Big) = 95\% \quad \Leftrightarrow \quad \Big[\mathbf{80.26, 85.74}\Big]$

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

```
[(83 - 1.96*(7/5)), (83 + 1.96*(7/5))]
# [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, $Z = \frac{\overline{y}-\mu}{\sigma/\sqrt{n}} \quad \Leftrightarrow \quad 1.96 = \frac{1}{7/\sqrt{n}} \quad \Leftrightarrow \quad n = (1.96 * 7)^2 \approx \mathbf{189}$ 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