Law of Large Numbers (LLN) and Central Limit Theorem (CLT)

Let $X_1, X_2, X_3, ..., X_n$ be a sequence of independent and identically distributed (iid) observations of size n. The sample is drawn at random from a population of size N with mean μ and variance σ^2. The sample mean and variance of the **random sample** are given by

$$\bar{X_n} \quad = \quad \frac{1}{n}\sum_{i=1}^{n} X_i \tag{1}$$

and

$$s_1^2 \quad = \quad \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X_n})^2 \tag{2}$$

respectively.

$\bar{X_n}$ is only an estimate or a **statistic**, not the true mean μ. This is simply due to the fact that the random sample is a sample of the population, not the whole population. The estimate could be made more accurate by increasing n. A larger n would produce a more accurate sample mean. However, acquiring a larger sample size is time consuming and expensive. By contrast, a smaller n is cheaper and less time consuming, but it would produce a less accurate estimate. Either way, we can never claim to know the true mean unless the entire population is measured.

Given that the random sample is drawn from the population at random, then $\bar{X_n}$ is also a random variable. Suppose we draw M random samples of size n. We won't get the same n samples repeatedly. We'll get a different set of observations every time. Consequently, there will be a variation in the estimate of the mean from one sample to another. This variation from sample to sample will be larger if the sample size is smaller, and the variation will be smaller if the sample size is larger. If we plot the M $\bar{X_n}$, $\bar{X_n}$ has a distribution of size M that shows how the mean varies from sample to sample. This distribution is called the **sampling distribution**. The LLN and CLT are statements about the **center** and **shape of the sampling distribution** respectively. As $n \longrightarrow \infty$:

$$\bar{X_n} \quad \longrightarrow \quad \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right). \tag{3}$$

In other words, as the sample size of the M random sample drawn increases, the sampling distribution approaches a Normal distribution with mean given by the LLN and variance given by the CLT:

$$\frac{\sqrt{n}(\bar{X_n} - \mu)}{\sigma} \quad \longrightarrow \quad \mathcal{N}(0, 1). \tag{4}$$

As n increases, the **sampling mean** approaches the true mean μ. The larger the sample size, the greater the accuracy. The **sampling standard deviation** is also known as the **standard error of the mean** since it implicitly measures how precise the sampling mean is: the smaller the standard error, the greater the precision. In essence, if the sample size is sufficiently large, the sampling mean is accurate and precise.

The theorem makes an statement about the convergence if the sample size is sufficiently large. It says nothing about the convergence rate, or how many random samples need to be drawn. If the population distribution is skewed, n must be very large before CLT conditions are reached. The closer the population distribution is to normal, the faster the convergence and the distribution over a smaller n may be approximated by a Normal distribution. A generally accepted rule is if $n \geq 30$, then the sample size is sufficiently large to apply the CLT and LLN with reasonable accuracy.

$Gamma(1,1)$ and $n = 1000$

Figure 1 displays an example of the entire process. The population distribution (Figure 1a) is Gamma distributed with $\alpha = \beta = 1$ and therefore far from Normal. The population mean and standard deviation are given by $\mu = \alpha\beta = 1$ and $\sigma = \sqrt{\alpha}\beta = 1$ respectively. Figure 1b,c show two random samples from this population, each of size $n = 1000$. The sample means and variances vary from sample to sample. Sample 1 has $\bar{x}_1 = 1.01$ and $s_1 = 0.94$; sample 2 has $\bar{x}_2 = 0.98$ and $s_2 = 0.93$. Each random sample is also Gamma distributed. Due to the large n, their shape and estimates are representative of the population.

```r
M <- 1000
N <- 1000000
n <- 1000
set.seed(1)
population_gamma <- rgamma(N,shape = 1,scale = 1)
population_mean <- mean(population_gamma)
population_sd <- sd(population_gamma)

random_sample1 <- sample(population_gamma,size = n,replace = TRUE)
sample1_mean <- mean(random_sample1)
sample1_sd <- sd(random_sample1)

set.seed(2)
random_sample2 <- sample(population_gamma,size = n,replace = TRUE)
sample2_mean <- mean(random_sample2)
sample2_sd <- sd(random_sample2)


par(mfrow=c(2,2))
hist(population_gamma,
 col="lightgreen",
 border="black",
 prob = TRUE,
 xlab = "",
 main = "Population")
lines(density(population_gamma),lwd = 2, col = "red")
mtext("(a)", side=1, line = 4)

hist(random_sample1,
 col="lightblue",
 border="black",
 prob = TRUE,
 xlab = "",
 main = "Random sample 1")
lines(density(random_sample1), lwd = 2, col = "red")
mtext("(b)", side=1, line = 4)

hist(random_sample2,
 col="lightblue",
 border="black",
 prob = TRUE,
 xlab = "",
 main = "Random sample 2")
lines(density(random_sample2), lwd = 2, col = "red")
mtext("(c)", side=1, line = 4)

set.seed(1)
random_sample <- replicate(M, sample(population_gamma, size = n, replace=TRUE))
sampling_dist <-  apply(random_sample, 2, mean)
sampling_mean <- mean(sampling_dist, na.rm = TRUE)
sampling_sd <- sd(sampling_dist, na.rm = TRUE)
theoretical_sampling_sd <- population_sd/sqrt(n)

hist(sampling_dist,
 col="yellow",
 border="black",
 prob = TRUE,
 xlab = "",
 main = paste0("M=",M,"; n=",n,""))
curve( dnorm(x, mean=population_mean,sd=theoretical_sampling_sd), add=T,col="blue", lwd = 3)
lines(density(sampling_dist), lwd = 2, col = "red")
mtext("(d)", side=1, line = 4)
```

$M = 1000$ samples are randomly drawn from the population and their samples means are calculated. The distribution of 1000 sample means is the **sampling distribution** (Figure 1d). The blue curve represents the theoretical sampling distribution given by $\mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right) = \mathcal{N}(1, 0.03)$. The empirical sampling distribution is close to Normal with sampling standard deviation given by $\frac{\sigma}{\sqrt{n}} = s = 0.03$, as expected due to the CLT since n is very large. The sampling mean is very accurate. It is equal to the population mean $\mu = \bar{x} = 1.00$. This is also expected due to the LLN. Even though the population is Gamma distributed, n is sufficiently large to apply the CLT and LLN with reasonable accuracy.

The $\bar{x}$ is an accurate estimate of the true population mean, and the standard error of the mean given by the sampling standard deviation tells us $\bar{x}$ is very precise. The sampling distribution is approximately Normal.

$Gamma(1, 1)$ and $n = 10$

For $n = 10$ (Figure 2), the random samples are less representative of the poputation (Figure 2b,c). The blue curve represents the theoretical sampling distribution $\mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right) = \mathcal{N}(1, 0.31)$. The empirical sampling distribution is further from Normal with sampling standard deviation given by $\frac{\sigma}{\sqrt{n}} \simeq s = 0.32$. The sampling mean is very close to the population mean $\mu \simeq \bar{x} = 0.99$. Notice the sampling standard deviation here is much larger than on Figure 1d. Given the small n, the sample-to-sample variability is much larger and so is the standard error of the mean. The larger the standard error of the mean, the lower the precision.

The LLN converged faster than the CLT. The $\bar{x}$ is an accurate estimate of the true population mean. However, the standard error of the mean shows us $\bar{x}$ is not very precise. Finally, the sampling distribution is approximately Normal.

```
M <- 1000
N <- 1000000
n <- 10
set.seed(1)
population_gamma <- rgamma(N,shape = 1,scale = 1)
```

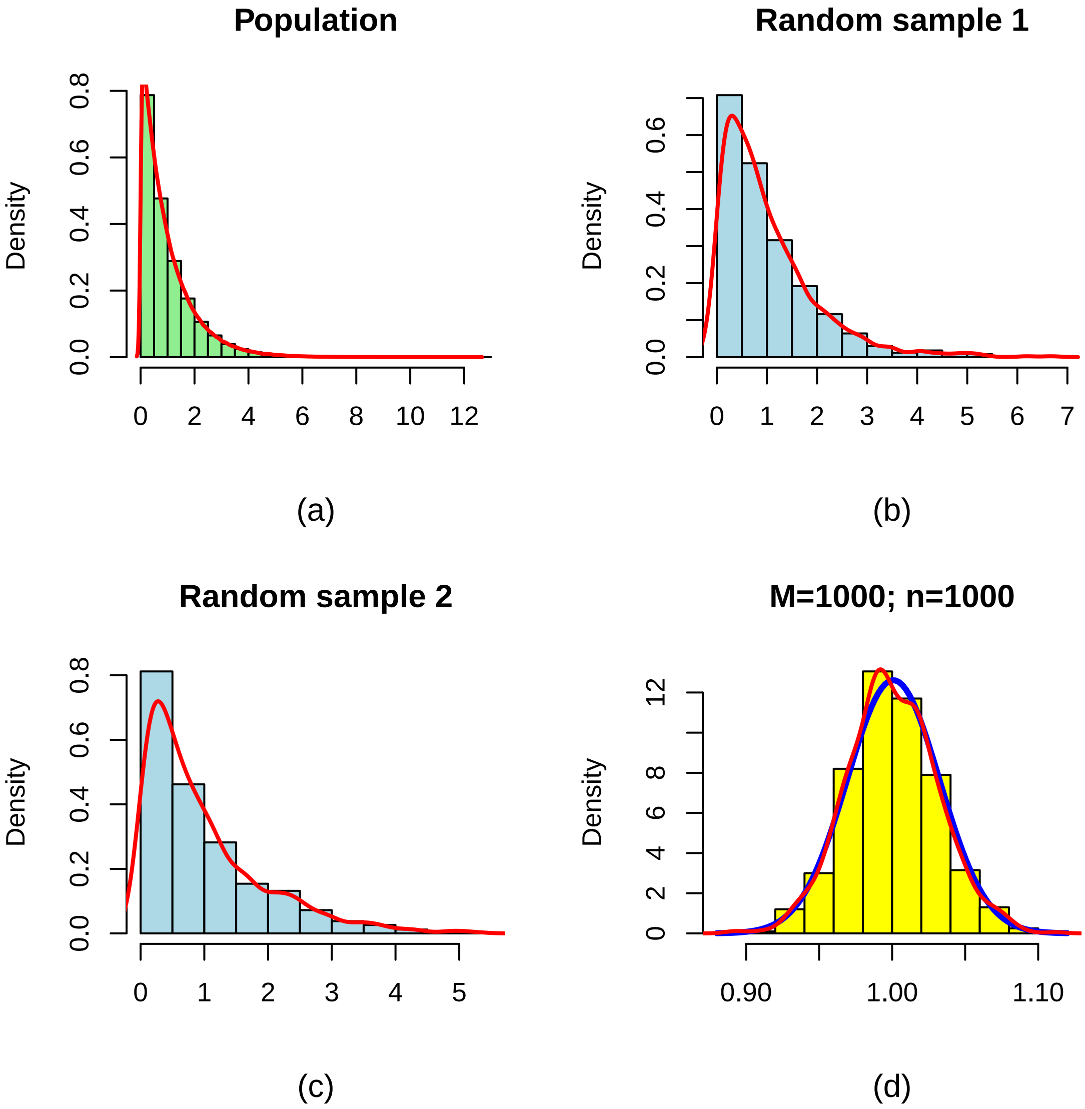


Figure 1: (a) population parameters: $\alpha = 1, \beta = 1, \mu = 1, \sigma = 1$. (b) random sample 1: n $=$ 1000, $\bar{x}_1 = 1.01, s_1 = 0.94$. (c) random sample 2: n $=$ 1000, $\bar{x}_2 = 0.98, s_2 = 0.93$. (d) sampling distribution: $\bar{x} = 1.00, s = 0.03$

```r
population_mean <- mean(population_gamma)
population_sd <- sd(population_gamma)

random_sample1 <- sample(population_gamma,size = n,replace = TRUE)
sample1_mean <- mean(random_sample1)
sample1_sd <- sd(random_sample1)

set.seed(2)
random_sample2 <- sample(population_gamma,size = n,replace = TRUE)
sample2_mean <- mean(random_sample2)
sample2_sd <- sd(random_sample2)


par(mfrow=c(2,2))
hist(population_gamma,
 col="lightgreen",
 border="black",
 prob = TRUE,
 xlab = "",
 main = "Population")
lines(density(population_gamma), lwd = 2, col = "red")
mtext("(a)", side=1, line = 4)

hist(random_sample1,
 col="lightblue",
 border="black",
 prob = TRUE,
 xlab = "",
 main = "Random sample 1")
lines(density(random_sample1),lwd = 2, col = "red")
mtext("(b)", side=1, line = 4)

hist(random_sample2,
 col="lightblue",
 border="black",
 prob = TRUE,
 xlab = "",
 main = "Random sample 2")
lines(density(random_sample2), lwd = 2, col = "red")
mtext("(c)", side=1, line = 4)

set.seed(1)
random_sample <- replicate(M, sample(population_gamma, size = n, replace=TRUE))
sampling_dist <-  apply(random_sample, 2, mean)
sampling_mean <- mean(sampling_dist, na.rm = TRUE)
sampling_sd <- sd(sampling_dist, na.rm = TRUE)
theoretical_sampling_sd <- population_sd/sqrt(n)

hist(sampling_dist,
 col="yellow",
 border="black",
 prob = TRUE,
 xlab = "",
 main = paste0("M=",M,"; n=",n,""))
curve( dnorm(x, mean=population_mean,sd=theoretical_sampling_sd), add=T,col="blue", lwd=3)
lines(density(sampling_dist), lwd = 2, col = "red")
mtext("(d)", side=1, line = 4)
```

$\mathcal{N}(0, 1)$ and $n = 10$

For a normally distributed population with $\mu = 0$ and $\sigma = 1$ (Figure 3a), the two random samples from this population, each of size $n = 10$ are unrepresentative of the population due to the small sample size. Sample 1 has $\bar{x_1} = -0.77$ and $s_1 = 1.19$; sample 2 has $\bar{x_2} = -0.30$ and $s_2 = 0.58$.

The theoretical sampling distribution given by the blue curve is $\mathcal{N}(1, 0.32)$. The empirical sampling distribution is Normal with sampling standard deviation given by $\frac{\sigma}{\sqrt{n}} \simeq s = 0.31$. The sampling mean is close to the population mean $\mu \simeq \bar{x} = -0.01$. The LLN and CLT convergence is very fast here over a smaller n due to the fact that the population is normally distributed. However, the standard error of the mean is still large due to the small sample size.

The $\bar{x}$ is an accurate estimate of the true population mean. However, the standard error of the mean shows us $\bar{x}$ is not very precise. If n is increased, the standard error decreases, and the precision increases. The sampling distribution is Normal.

```
M <- 1000
N <- 1000000
n <- 10
population_gauss <- rnorm(N, mean = 0, sd = 1)
population_mean <- mean(population_gauss)
population_sd <- sd(population_gauss)

set.seed(1)
random_sample1 <- sample(population_gauss,size = n,replace = TRUE)
sample1_mean <- mean(random_sample1)
sample1_sd <- sd(random_sample1)

set.seed(2)
random_sample2 <- sample(population_gauss,size = n,replace = TRUE)
sample2_mean <- mean(random_sample2)
sample2_sd <- sd(random_sample2)

par(mfrow=c(2,2))
hist(population_gauss,
 col="lightgreen",
 border="black",
 prob = TRUE,
 xlab = "",
 main = "Population")
lines(density(population_gauss), lwd = 2, col = "red")
mtext("(a)", side=1, line = 4)

hist(random_sample1,
 col="lightblue",
 border="black",
 prob = TRUE,
 xlab = "",
 main = "Random sample 1")
```

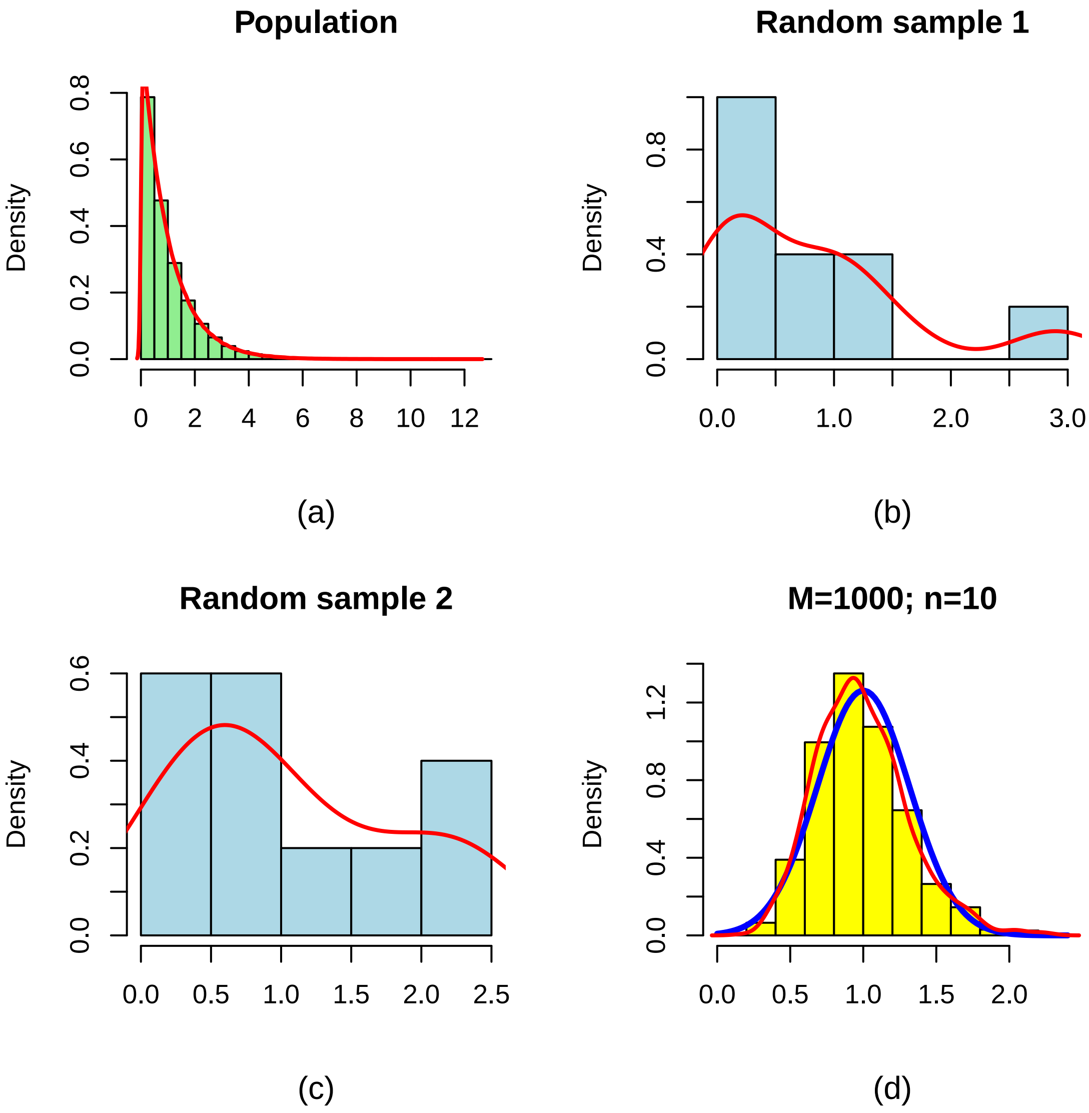


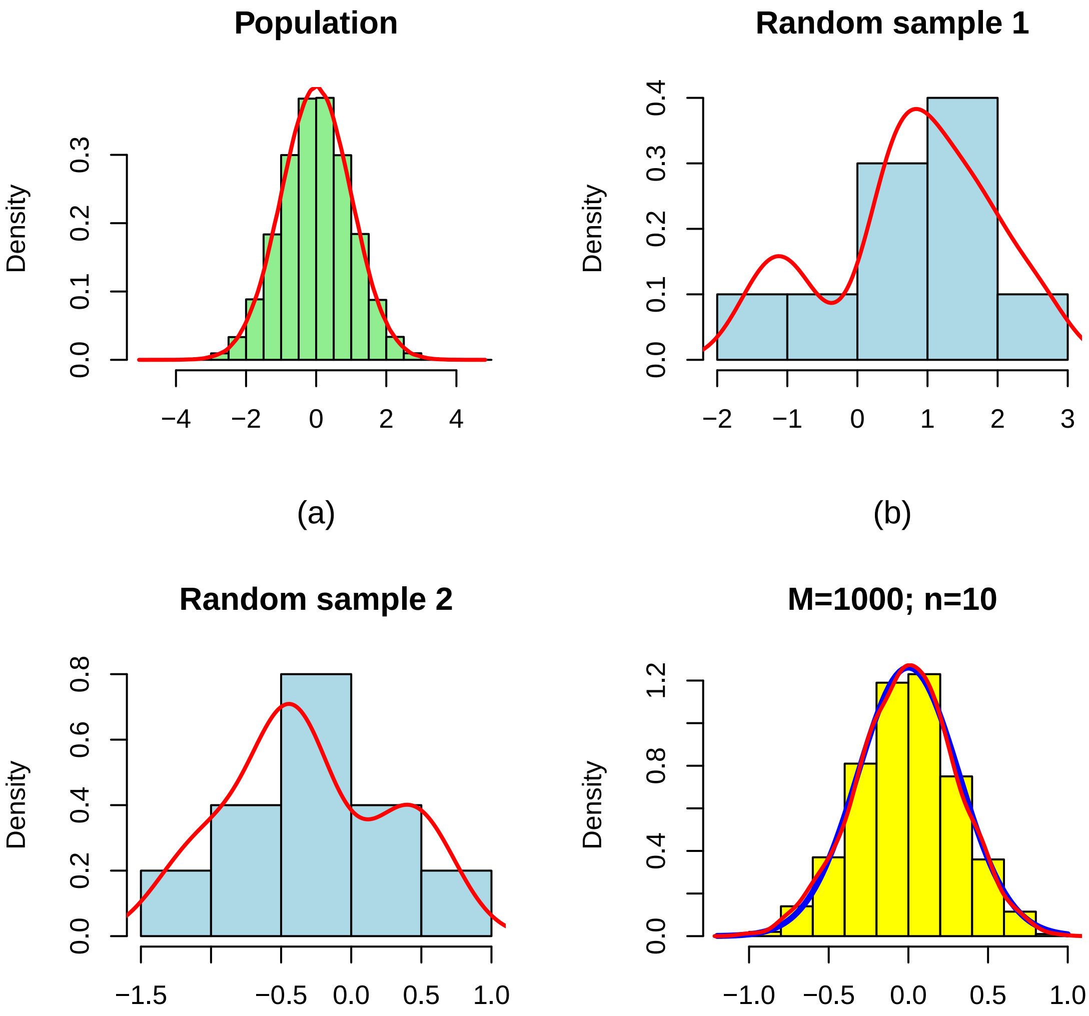
Figure 2: (a) population parameters: $\alpha = 1, \beta = 1, \mu = 1, \sigma = 1$. (b) random sample 1: n = 10, $\bar{x}_1 = 0.81, s_1 = 0.89$. (c) random sample 2: n = 10, $\bar{x}_2 = 1.04, s_2 = 0.81$. (d) sampling distribution: $\bar{x} = 0.98, s = 0.31$

```r
lines(density(random_sample1), lwd = 2, col = "red")
mtext("(b)", side=1, line = 4)

hist(random_sample2,
 col="lightblue",
 border="black",
 prob = TRUE,
 xlab = "",
 main = "Random sample 2")
lines(density(random_sample2), lwd = 2, col = "red")
mtext("(c)", side=1, line = 4)

set.seed(1)
random_sample <- replicate(M, sample(population_gauss, size = n, replace=TRUE))
sampling_dist <-  apply(random_sample, 2, mean)
sampling_mean <- mean(sampling_dist, na.rm = TRUE)
sampling_sd <- sd(sampling_dist, na.rm = TRUE)
theoretical_sampling_sd <- population_sd/sqrt(n)

hist(sampling_dist,
 col="yellow",
 border="black",
 prob = TRUE,
 xlab = "",
 main = paste0("M=",M,"; n=",n,""))
curve( dnorm(x, mean=population_mean,sd=theoretical_sampling_sd), add=T,col="blue", lwd=3)
lines(density(sampling_dist), lwd = 2, col = "red")
mtext("(d)", side=1, line = 4)
```



(a)

(b)

(c)

(d)

Figure 3: (a) population parameters: $\mu = 0, \sigma = 1$. (b) random sample 1: n $=$ 10, $\bar{x}_1 = 0.77, s_1 = 1.19$. (c) random sample 2: n $=$ 10, $\bar{x}_2 = -0.30, s_2 = 0.58$. (d) sampling distribution: $\bar{x} = -0.01, s = 0.31$