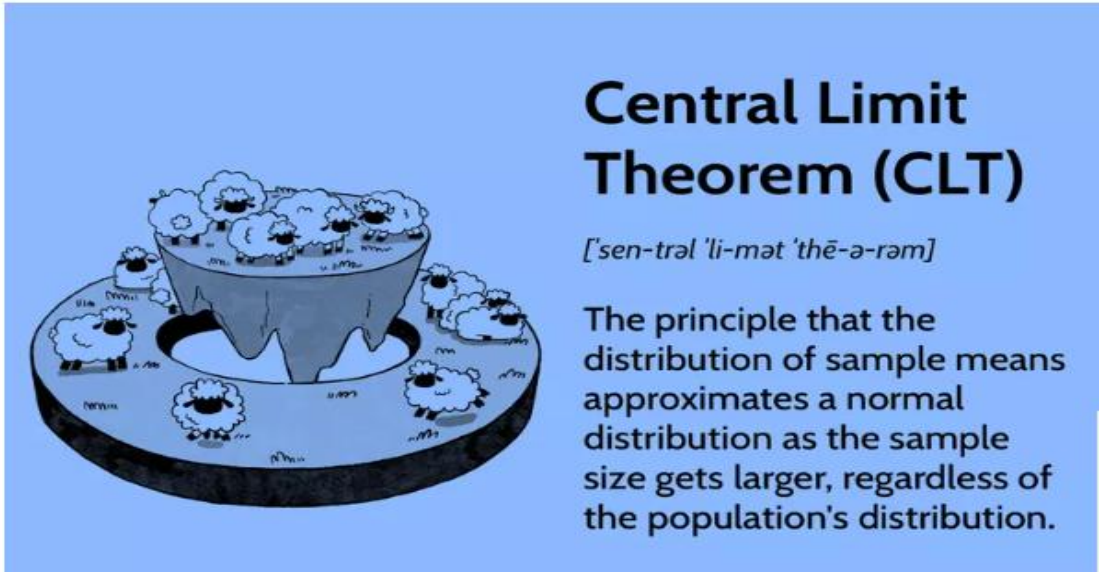


KEY TAKEAWAYS

- The central limit theorem (CLT) states that the distribution of sample means approximates a normal distribution as the sample size gets larger, regardless of the population's distribution.
- A sufficiently large sample size can predict the characteristics of a population more accurately.
- Sample sizes equal to or greater than 30 are often considered sufficient for the CLT to hold.
- A key aspect of CLT is that the average of the sample means and standard deviations will equal the population mean and standard deviation.
- CLT is useful in finance and investing when analyzing a large collection of securities to estimate portfolio distributions and traits for returns, risk, and correlation.

Key Components of the Central Limit Theorem

The central limit theorem has several key components. They largely revolve around sampling technique.

1. **Sampling is successive**: This means some sample units are common with sample units selected on previous occasions.
2. **Sampling is random**: All samples must be selected at random so that they have the same statistical possibility of being selected.
3. **Samples should be independent**: The selections or results from one sample should have no bearing on future samples or other sample results.
4. **Large sample size**: As sample size increases, the sampling distribution should come ever closer to the normal distribution.

Sample size and standard deviations

By convention, we consider a sample size of 30 to be “sufficiently large.” The sample size affects the standard deviation of the sampling distribution. Standard deviation is a measure of the variability or spread of the distribution (i.e., how wide or narrow it is).

When sample size is low (*n* < 30), the central limit theorem doesn’t apply. The sampling distribution will follow a similar distribution to the population. Therefore, the sampling distribution will only be normal if the population is normal. The standard deviation is high. There’s a lot of spread in the samples’ means because they aren’t precise estimates of the population’s mean.

When sample size is high (*n* ≥ 30), the central limit theorem applies. The sampling distribution will approximately follow a normal distribution. The standard deviation is low. There’s not much spread in the samples’ means because they’re precise estimates of the population’s mean.

Central limit theorem formula

Fortunately, you don't need to actually repeatedly sample a population to know the shape of the sampling distribution. The parameters of the sampling distribution of the mean are determined by the parameters of the population:

- The mean of the sampling distribution is the mean of the population.

$$\mu_{\bar{x}} = \mu$$

- The standard deviation of the sampling distribution is the standard deviation of the population divided by the square root of the sample size.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

We can describe the sampling distribution of the mean using this notation:

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

Where:

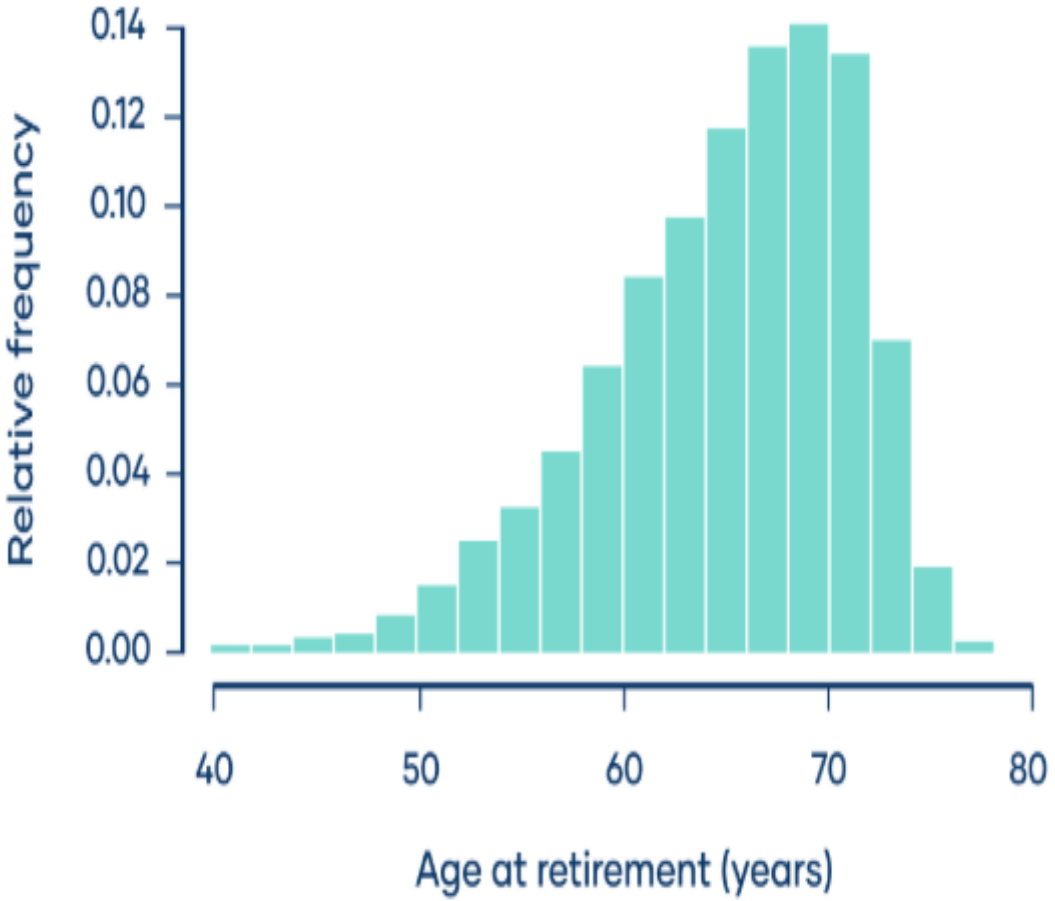
- $\bar{X}$ is the sampling distribution of the sample means
- ~ means "follows the distribution"
- *N* is the normal distribution
- μ is the mean of the population
- σ is the standard deviation of the population
- *n* is the sample size

Central limit theorem examples

Applying the central limit theorem to real distributions may help you to better understand how it works.

Continuous distribution

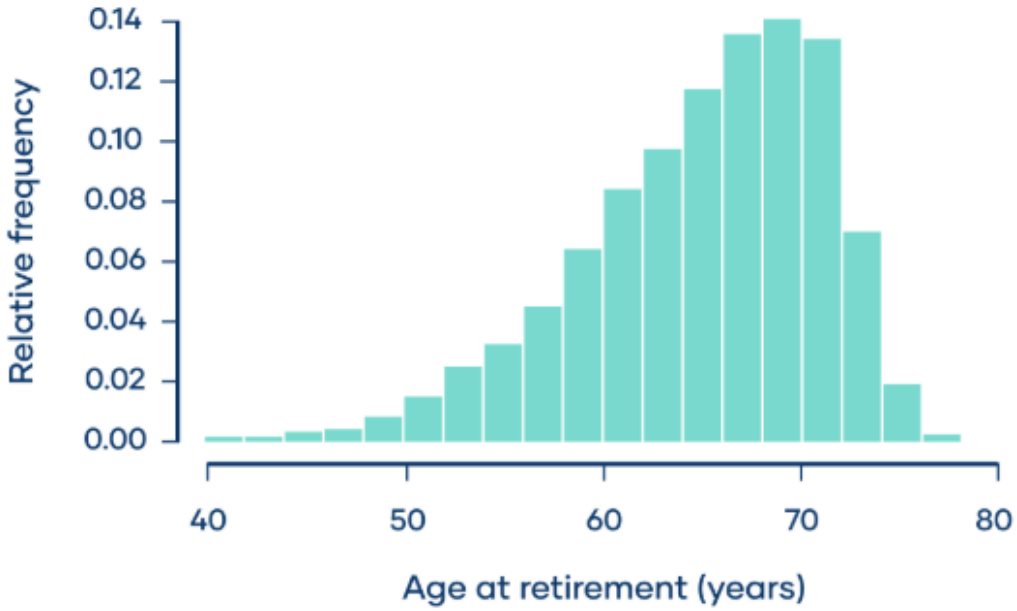
Suppose that you're interested in the age that people retire in the United States. The population is all retired Americans, and the distribution of the population might look something like this:



Continue...

Age at retirement follows a left-skewed distribution. Most people retire within about five years of the mean retirement age of 65 years. However, there's a "long tail" of people who retire much younger, such as at 50 or even 40 years old. The population has a standard deviation of 6 years.

Imagine that you take a small **sample** of the population. You randomly select five retirees and ask them what age they retired.



The mean of the sample is an estimate of the population mean. It might not be a very precise estimate, since the sample size is only 5.

Example: Central limit theorem; mean of a small sample

mean = (68 + 73 + 70 + 62 + 63) / 5

mean = 67.2 years

Example: Central limit theorem; sample of *n* = 5

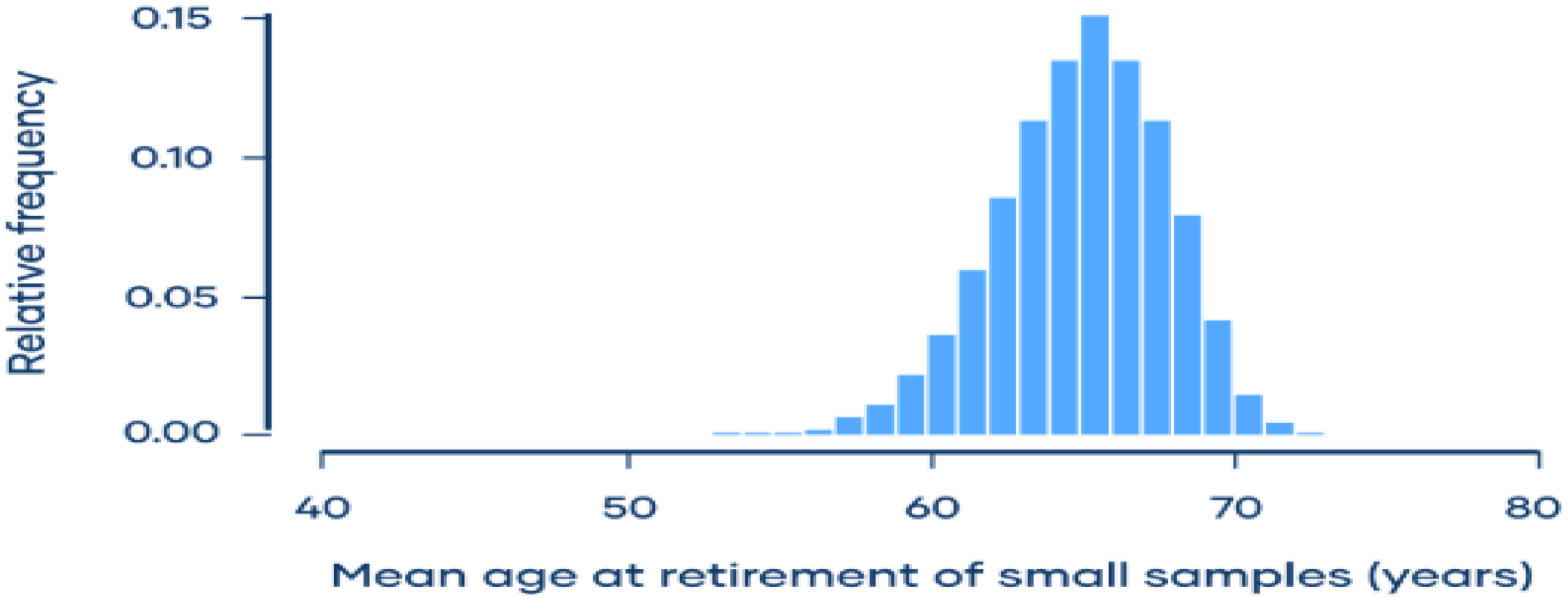
68	73	70	62	63

Suppose that you repeat this procedure 10 times, taking samples of five retirees, and calculating the mean of each sample. This is a **sampling distribution of the mean**.

Example: Central limit theorem; means of 10 small samples

60.8	57.8	62.2	68.6	67.4	67.8	68.3	65.6	66.5	62.1

If you repeat the procedure many more times, a histogram of the sample means will look something like this:



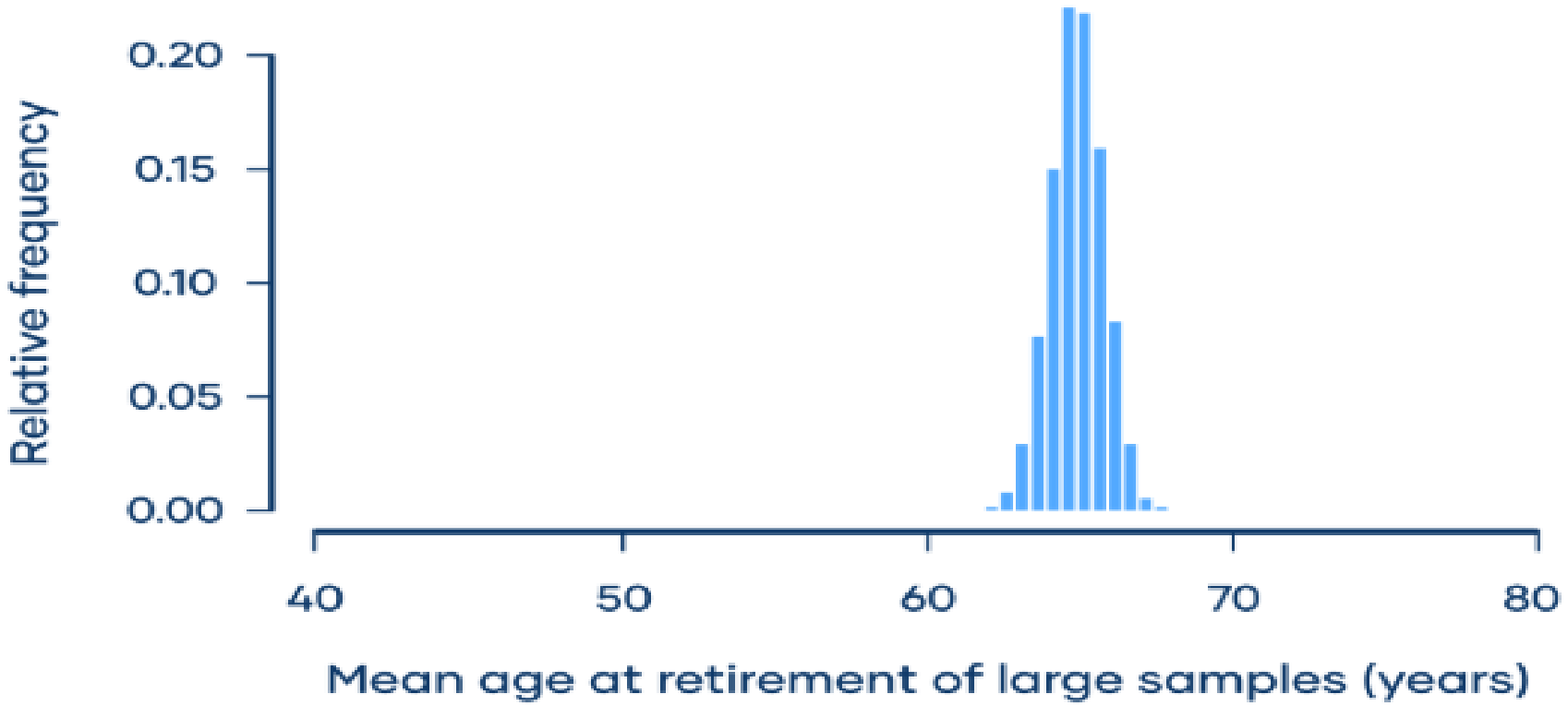
Although this sampling distribution is more normally distributed than the population, it still has a bit of a left skewd. Notice also that the spread of the sampling distribution is less than the spread of the population.

Example: Central limit theorem; sample of $n = 50$

73	49	62	68	72	71	65	60	69	61
62	75	66	63	66	68	76	68	54	74
68	60	72	63	57	64	65	59	72	52
52	72	69	62	68	64	60	65	53	69
59	68	67	71	69	70	52	62	64	68

mean = 64.8 years

Again, you can repeat this procedure many more times, taking samples of fifty retirees, and calculating the mean of each sample:



In the histogram, you can see that this sampling distribution is normally distributed, as predicted by the central limit theorem.

We can use the central limit theorem formula to describe the sampling distribution:

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

μ = 65

σ = 6

n = 50

$$\bar{X} \sim N(65, \frac{6}{\sqrt{50}})$$

$$\bar{X} \sim N(65, 0.85)$$

Deriving Distribution of Sample Mean and Sample Variance

Properties of Random Variables

To derive the sampling distributions for the sample mean and variance we need to define some properties random variables that we will encounter. Let a be a constant number and X be a r.v.:

- $E(a + X) = a + E(X)$
- $Var(a + X) = Var(X)$
- $E(aX) = aE(X)$
- $Var(aX) = a^2 Var(X)$
- $E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i)$ for independent X_i
- $Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i)$ for independent X_i

$\bar{X}$ is an Unbiased Estimator for the Mean (μ)

Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ and variance σ^2, then

$$
\begin{aligned}
E[\bar{X}] &= E\left[\sum_{i=1}^{n} \frac{X_i}{n}\right] \\
&= \frac{1}{n} E\left[\sum_{i=1}^{n} X_i\right] \\
&= \frac{1}{n} \sum_{i=1}^{n} E[X_i] \\
&= \frac{1}{n} \sum_{i=1}^{n} \mu \\
&= \frac{1}{n} n\mu
\end{aligned}
$$

Thus, in any given sample we "expect" the sample average to be near the true population mean.

Variance of $\bar{X}$

Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ and variance σ^2, then

$$
\begin{aligned}
V[\bar{X}] &= V\left[\sum_{i=1}^{n} \frac{X_i}{n}\right] \\
&= \frac{\dot{1}}{n^2} V\left[\sum_{i=1}^{n} X_i\right] \\
&= \frac{1}{n^2} \sum_{i=1}^{n} V[X_i] \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2 \\
&= \frac{1}{n^2} n\sigma^2 \\
&= \frac{\sigma^2}{n}
\end{aligned}
$$

The larger the sample size, the more precise an estimator $\bar{X}$ will be.

We have now shown that $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.

We can see that as n increases, the variability of the sample mean will decrease.

In other words, as $n \to \infty$, $\sigma_{\bar{X}} \to 0$. If we had all n data points from the population in every sample we draw, then $\bar{X}$ would be equal to the true mean μ every time we sampled.

Standard Error of $\bar{X}$

The **standard error of the mean (SEM)** measures the precision with which $\bar{X}$ estimates μ over repeated (i.e. all possible) samples of size n from a population with underlying variance σ^2:

$$\text{SEM} = \sqrt{V[\bar{X}]} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}} = \sigma_{\bar{X}}$$

In practice, we estimate the SEM by $\frac{s}{\sqrt{n}}$, where s is the standard deviation estimated from our sample.

More generally, we use the term *standard error* to refer to the sampling standard deviation of estimators of population parameters. For example, the standard deviation of (the sampling distribution of) $\hat{p}$ is the standard error of $\hat{p}$.

Example: Find the standard error of the given observations:

10,20,30,40,50

Solution: Given,

x= 10, 20,30,40,50

Number of observations, n = 5

Hence, Mean = Total of observations/Numbe

Mean = (10+20+30+40+50)/5

Mean = 150/5 = 30

By the formula of standard error, we know;

SEM = SD/√N

Now, we need to find the standard deviation here.

By the formula of standard deviation, we get;

$$SD = \sqrt{(1/N-1) \times ((x_1 - x_m)^2) + (x_2 - x_m)^2) + \ldots + (x_n - x_m)^2)}$$

$$SD = \sqrt{(1/5-1) \times ((10-30)^2) + (20-30)^2) + (30-30)^2 + (40-30)^2 + (50-30)^2}$$

$$SD = \sqrt{1/4((-20)^2 + (-10)^2 + (0)^2 + (10)^2 + (20)^2}$$

$$SD = \sqrt{1/4(400+100+0+100+400)}$$

$$SD = \sqrt{250}$$

$$SD = 15.811$$

Therefore, putting the values of standard deviation and root of number of observations, we get;

Standard error of mean, SEM = SD/√N

SEM = 15.811/√5

SEM = 15.8114/2.2361

SEM = 7.0711

Ques. Out of the 100 responses, a random sampling procedure was employed to create a sample of 5 responses. 3, 2, 5, 3 and 4 are the chosen responses. Calculate the statistic's standard error based on the responses the person has chosen.

$(x) = (3 + 2 + 5 + 3 + 4)/5 = 3.4$

The formula for calculating standard deviation (s) is given below.

$s = \sqrt{\Sigma ni(xi-x\grave{}_{,,})^2 / n-1}$

The formula for calculating standard deviation (s) is given below.

$s = \sqrt{\Sigma ni(xi-x\grave{}_{,,})^2 / n-1}$

Standard Deviation $= \sqrt{[\{(3 - 3.4)^2 + (2 - 3.4)^2 + (5 - 3.4)^2 + (3 - 3.4)^2 + (4 - 3.4)^2\} / (5 - 1)]}$

Standard Deviation = 1.14

Standard Error is calculated using the formula given below

Standard Error $= s / \sqrt{n}$

Standard Error $= 1.14 / \sqrt{5}$

Standard Error = 0.51