5.2: Sampling Distribution of Sample Means

Learning Objectives

- Motivate, state, and apply the Central Limit Theorem (CLT)
- State the expected value (mean) and standard deviation of the sampling distribution of sample means
- Establish guides regarding sufficiently large sample sizes

The Utility of Sampling Distributions

To construct a sampling distribution, we must consider all possible samples of a particular size, n, from a given population. In reality, this is more complicated than studying the entire population since considering every possible sample requires studying every member of the population. If we have all the population data, why mess with all the samples? It is a valid question. The truth is that, in practice, statisticians do not construct sampling distributions by brute force; instead, they deduce key properties of the distribution. Inferential statistics are used to learn about a population by studying a sample, a subset of the population, not the entire population itself.

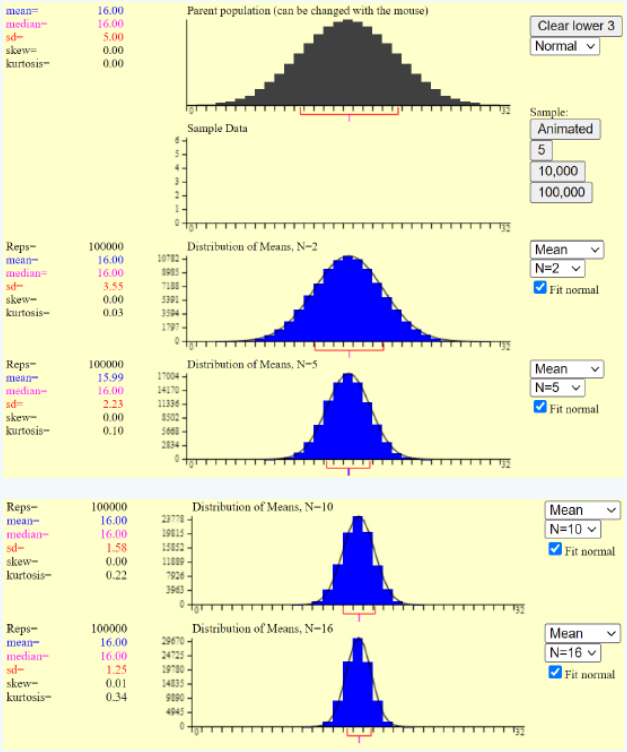
The Sampling Distribution of Sample Means

Using the computer simulation from the last section, we will consider the progression of sampling distributions of sample means from several populations as the sample size increases. Our previous work shows that the sampling distribution of sample means will be centered on the population mean and that the spread will decrease as the sample size increases. What can we say about the general shape of the sampling distributions of sample means regardless of the parent population?

? Text Exercise 5.2.1

The parent population (the distribution in black) is centered above 6 sampling distributions of sample means (the distributions in blue), starting with a sample size of 2 and ending with a sample size of 25. A normal curve has been fit to each of the sampling distributions. Which sampling distributions seem to fit the normal curve better? What trend do you notice across parent populations?

Parent Population: Normal



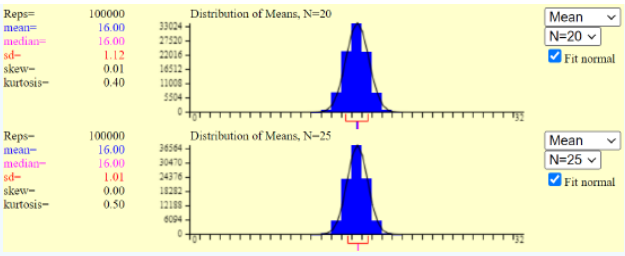


Figure 5.2.1: Sampling distributions of sample means for various sample sizes taken from a normal population

Parent Population: Uniform

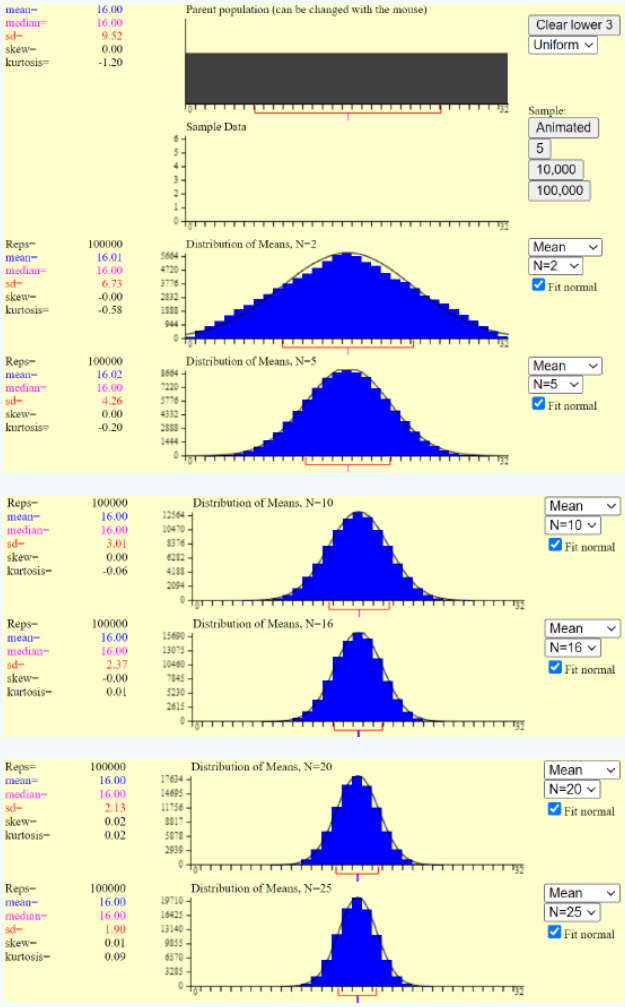


Figure 5.2.2 Sampling distributions of sample means for various sample sizes taken from a uniformly distributed population

Parent Population: Skewed

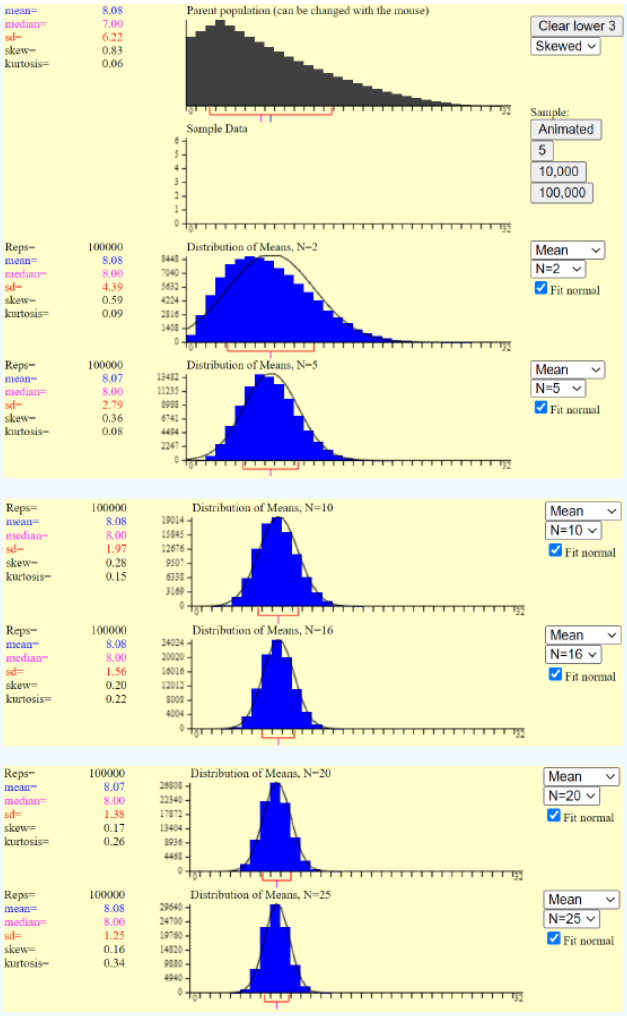


Figure 5.2.3 Sampling distributions of sample means for various sample sizes taken from a skewed population

Answer

Each sampling distribution from the normal parent population fits the normal curve well. All of the sampling distributions except the first sampling distribution with a sample size of 2 from the uniform parent population also fit the normal curve well. For the skewed parent population, it was not until the sample size reached 16 or 20 that the normal curve fit well. We see a trend that the sampling distributions of sample means eventually appear normal regardless of the parent distribution. For some parent distributions, larger sample sizes were necessary for the sampling distribution of sample means to appear to fit the normal curve.

Hopefully, we understand that the sampling distribution of sample means and the normal distribution are connected; furthermore, the sample size used to construct the sampling distribution also plays a role. If not, that is okay. We continue to learn and develop the relationship more formally. It is quite a remarkable result.

Note: Assessing Normality

Up to this point, we have assessed how well a distribution fits a normal curve visually. This level of discussion serves our purposes in an introductory statistics textbook. However, we want to alert the reader that there are analytical methods of assessing how well a normal curve fits a distribution. This process is commonly called assessing normality and is essential to serious statistical study and work.

Central Limit Theorem (CLT)

Given any infinite population with population mean μ and non-zero population variance of σ^2, as the sample size, n, increases, the sampling distribution of sample means approaches a normal distribution with mean $\mu_{\bar{x}} = \mu$ and variance $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$.

Recall that standard deviation is the square root of variance. We can also assert the normal distribution that the sampling distribution of sample means approaches as n increases has a standard deviation of $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. We will utilize this formula quite frequently.

Note: Infinite Populations

Notice that the Central Limit Theorem says, "given any infinite population." We have framed statistical inquiry in terms of understanding the world and people around us. At any given time, there are only finitely many humans, animals, or even atoms existing in our world. So, the statement of the Central Limit Theorem seems to exclude all these populations in which we may have interest, but hope is not lost.

The Central Limit Theorem, as stated above, is a beautiful work of mathematical and statistical theory. There is often a gap between theory and practice that can be bridged satisfactorily. We are in such a case, and our bridge is the notion of a practically infinite population relative to the sample size in consideration. A population may be understood as practically infinite if the sample size of interest is less than 5% of the population size (recall how we saw this earlier with the assumption of independence). The Central Limit Theorem holds in practice for practically infinite populations. Indeed, this rule of thumb of 5% also serves as our threshold at which we treat simple random sampling and sampling with replacement as interchangeable regarding the probability distributions of sample statistics.

Let us think about what the Central Limit Theorem is saying. It claims that if you pick any population (regardless of shape) and look at all samples of size n (for n sufficiently large), their means will be (approximately) normally distributed. We could start with any population, even the craziest of shapes, and an orderly bell curve emerges from the chaos of random selection. This should come as a surprise to someone hearing this for the first time. One would expect that nothing can be said about the sampling distribution; if the population shape is chaotic and we are selecting samples from it at random, then we would expect the sampling distribution to be chaotic as well. After all, the sampling distributions of the range, median, mode, and many other statistics do not follow such nice behavior in general. They behave exactly as the default intuition would expect: chaotically and unpredictably. However, there is something special about the mean. We will see later that the mean is not the only statistic to exhibit nice behavior.

The Central Limit Theorem is the reason the field of inferential statistics exists. The fact that we always get a normal distribution enables us to answer questions in inferential statistics intelligently and precisely, as we shall see now.

Applying the Central Limit Theorem

The Central Limit Theorem clearly states the ideas we have been exploring over the last two sections.

- The sampling distribution of sample means has an expected value (mean), the population mean.
- The spread of the sampling distribution of sample means decreases as n increases because $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. The population standard deviation, σ, is fixed; so, as n increases, we have a fixed number divided by larger and larger numbers making the quotient smaller.
- Finally, the sampling distribution of sample means gets closer and closer to the normal curve as n increases.

As we have seen, the rate at which the sampling distribution's shape becomes normal differs based on the parent population. If the parent population is normal, every sampling distribution appears approximately normal. In the other cases, we needed the sample size to be larger. How can we know if a given sample size is large enough to say that the sampling distribution of sample means is approximately normal? We now provide, without proof, the current knowledge and standards of the statistical community in this regard.

If the parent population is normally distributed, the sampling distribution of sample means will be normally distributed for every sample size, n.

Statisticians have long agreed that for many of the distributions commonly found in medicine, the social sciences, and the natural sciences, a sample size larger than 30 would produce a sampling distribution of sample means that is approximately normal.

Our statistical research may find a population in which a smaller number produces an approximately normal sampling distribution of sample means. We would not know this when we first began studying the population. On the other hand, we may find a population in which a much larger sample would be necessary to produce an approximately normal sampling distribution of sample means. Such distributions are being studied in economics and finance. How can we feel confident in our practice of statistics, especially if we conduct statistical research as part of our profession?

The parent distributions that require larger sample sizes to obtain approximately normal sampling distributions of sample means are most likely extremely skewed or have outliers. If there is no such intuition, we recommend using an initial sample size larger than 30 and testing the data for the presence of outliers or extreme skew (a histogram will probably suffice). If either outliers or extreme skew are detected, proceed by increasing the sample size and repeating the data collection process.

? Text Exercise 5.2.2

The grade distribution for a particular instructor's statistics course (over many years with thousands of students) is negatively skewed with a mean of 71% and a standard deviation of 20%. Compute the probability that the average of a random sample for the indicated sample size is within 2 percentage points of the population mean. What do you notice about the probabilities as n increases?

1. $n = 36$

Answer

Since the random sample is larger than 30, the sampling distribution of sample means is approximately normal with a mean $\mu_{\bar{x}} = 71$ percent and a standard deviation $\sigma_{\bar{x}} = \frac{20}{\sqrt{36}}$ percent. We are determining the probability that the average of the sample is within 2 percentage points of 71 percent: $P(69 < \bar{x} < 73)$. Sketch a picture and compute the probability with technology.

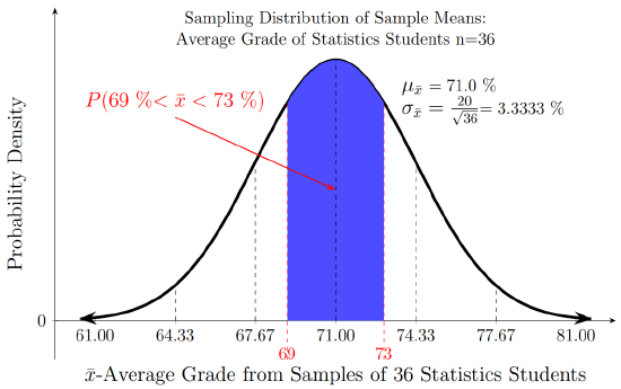


Figure 5.2.4 Sampling distribution of sample means

$$P(69 < \bar{x} < 73) = \text{NORM.DIST}(73, 71, \frac{20}{\sqrt{36}}, 1) - \text{NORM.DIST}(69, 71, \frac{20}{\sqrt{36}}, 1) \approx 45.1494\%$$

2. $n = 72$

Answer

The problem setup remains the same; we update the sample size to 72.

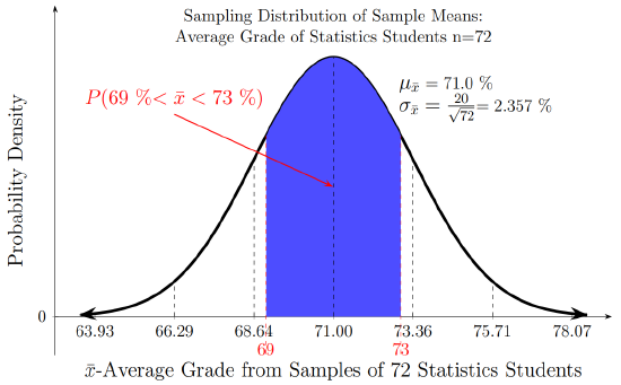


Figure 5.2.5 Sampling distribution of sample means

$$P(69 < \bar{x} < 73) = \text{NORM.DIST}(73, 71, \frac{20}{\sqrt{72}}, 1) - \text{NORM.DIST}(69, 71, \frac{20}{\sqrt{72}}, 1) \approx 60.3856\%$$

3. $n = 144$

Answer

The problem setup remains the same; we update the sample size to 144.

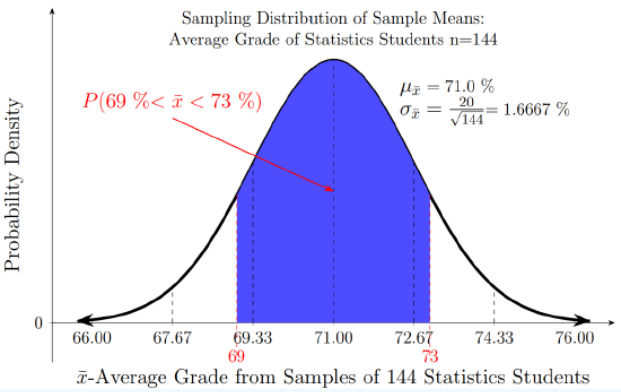


Figure 5.2.6 Sampling distribution of sample means

$$P(69 < \bar{x} < 73) = \text{NORM.DIST}(73, 71, \frac{20}{\sqrt{144}}, 1) - \text{NORM.DIST}(69, 71, \frac{20}{\sqrt{144}}, 1) \approx 76.9861\%$$

4. $n = 288$

Answer

The problem setup remains the same; we update the sample size to 288.

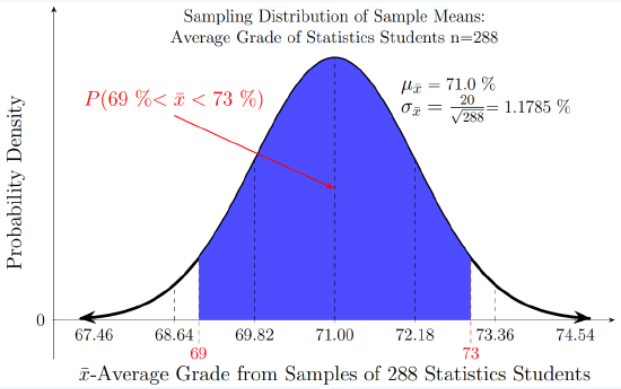


Figure 5.2.7 Sampling distribution of sample means

$$P(69 < \bar{x} < 73) = \text{NORM.DIST}(73, 71, \frac{20}{\sqrt{288}}, 1) - \text{NORM.DIST}(69, 71, \frac{20}{\sqrt{288}}, 1) \approx 91.0314\%$$

As the sample size increases, the standard deviation of the sampling distribution decreases. The interval from 69 to 73 encompasses more standard deviations around the mean, and the probability of the sample mean falling in that interval increases. If n were to increase to the size of the population, we would see a 100% chance of the sample mean being within 2 points of the actual mean. The larger n, the more likely the sample mean is close to the population mean.

Text Exercise 5.2.3

The heights of adult females are normally distributed with a mean of 64 inches and a standard deviation of 2.5 inches.

1. Determine the probability of randomly selecting four adult females whose average height is less than 5 feet 2 inches.

Answer

We randomly selected four adult females from the population and considered their average height. We took a sample of size $n = 4$ and considered the average height, $\bar{x}$, of the sample. We are interested in the following probability: $P(\bar{x} < 62)$. Note that we want to use the same units throughout 5 feet 2 inches is $5 \cdot 12 + 2 = 62$ inches. We must turn to the sampling distribution of sample means to answer the probability question. To compute probabilities, we must know what the probability distribution is. Constructing the probability distribution is out of the question here. We want to utilize the Central Limit Theorem (CLT). When considering the CLT, we ensure our sample size is large enough to assert that the sampling distribution of sample means is approximately normal. Usually, this means we want $n > 30$. This, however, is not the case in this scenario as $n = 4$. To proceed in this scenario, we note that the problem states that the heights of adult females are normally distributed. If the parent population is normally distributed, so are all the sampling distributions of sample means. We know that the sampling distribution of sample means is normally distributed with a mean $\mu_{\bar{x}} = \mu = 64$ inches and a standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2.5}{\sqrt{4}} = 1.25$ inches. We sketch a picture and compute the probability using technology.

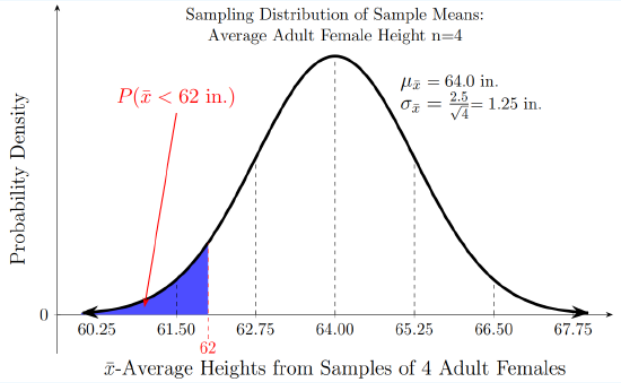


Figure 5.2.8 Sampling distribution of sample means

$$P(\bar{x} < 62) = \text{NORM.DIST}(62, 64, 1.25, 1) \approx 5.4799\%$$

2. Determine the probability of randomly selecting two adult females with an average height within 3 inches of the population mean.

Answer

We are in the context of randomly selecting multiple adult females from the population and considering their average height. We are only sampling 2 adult females; so, $n = 2$. We utilize the fact that the parent population is normal and that the sampling distribution of sample means when $n = 2$ is normal with a mean $\mu_{\bar{x}} = \mu = 64$ inches and a standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2.5}{\sqrt{2}} \approx 1.7678$ inches. An average height is within 3 inches of the population mean if it is larger than $64 - 3 = 61$ inches and smaller than $64 + 3 = 67$ inches. So, we are interested in $P(61 < \bar{x} < 67)$. We sketch a picture and compute the probability using technology.

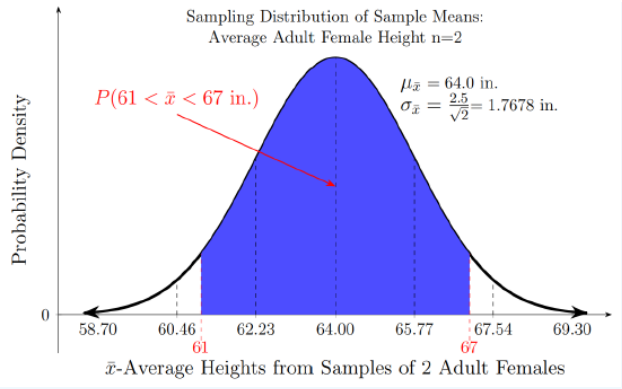


Figure 5.2.9 Sampling distribution of sample means

$$P(61 < \bar{x} < 67) = \text{NORM.DIST}(67, 64, \frac{2.5}{\sqrt{2}}, 1) - \text{NORM.DIST}(61, 64, \frac{2.5}{\sqrt{2}}, 1) \approx 0.955157 - 0.044843 \approx 91.0314\%$$

Text Exercise 5.2.4

Recall from Text Exercise 4.5.1 that the daily growth in the height of wheat plants during a particular stage of development is believed to be uniformly distributed between $\frac{1}{2} = 0.5$ and $\frac{5}{4} = 1.25$ inches and as such has a mean of $\frac{0.5+1.25}{2} = 0.875$ inches and the standard deviation is \(\frac{1.25-0.5}{\sqrt{12}}\) $= \frac{0.75}{\sqrt{12}}$ inches. Determine the probability of randomly selecting 48 wheat plants (during that particular stage of development) with an average daily growth that is greater than 0.9 inches.

Answer

We are randomly selecting 48 wheat plants from the population and considering their average daily growth. We are taking a sample of size $n = 48$ and considering the average height, $\bar{x}$, of the sample. We are interested in the following probability: $P(\bar{x} > 0.9)$. Since the sample size is greater than 30, we can apply the CLT to say that the sampling distribution of sample means is approximately normal with a mean $\mu_{\bar{x}} = \mu$ inches and a standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ inches. Thus, $\mu_{\bar{x}} = 0.875$ inches and $\sigma_{\bar{x}} = \frac{\frac{0.75}{\sqrt{12}}}{\sqrt{48}} = \frac{0.75}{24} = 0.03125$inches. Let us sketch a picture and compute the probability of interest.

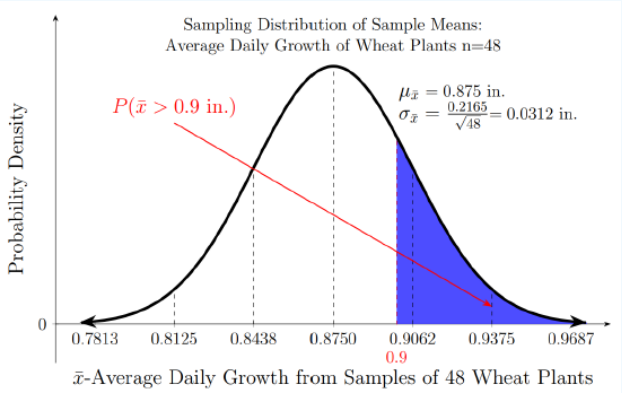


Figure 5.2.10 Sampling distribution of sample means

$$P(\bar{x} > 0.9) = 1 - \text{NORM.DIST}(0.9, 0.875, 0.03125, 1) \approx 1 - 0.788145\% \approx 21.1855\%$$