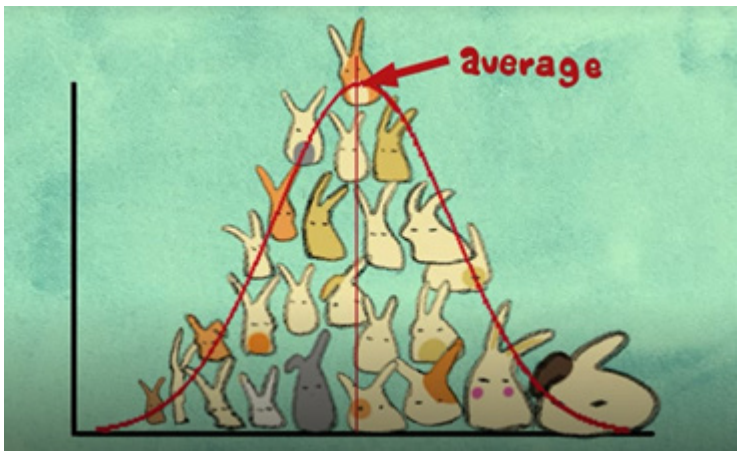
Minitab®

SPC

LESSON: Xbar Chart

Quality Methods
Dr. Diane Evans
ROSE-HULMAN INSTITUTE OF TECHNOLOGY

The Central Limit Theorem



Reference: The video, Bunnies, Dragons and the 'Normal' World, is a New York Times animation that illustrates the Central Limit Theorem (CLT).

Interesting Fact about Averages:

Spring 2019:

- **LeBron James** has a career average of **27-7-7 (points-rebounds-assists)**.
- He's now played in **1,437 NBA games** and has **never finished a game with exactly 27-7-7**.

Sampling Distribution of $\bar{X}$

The video, Bunnies, Dragons and the 'Normal' World, is a New York Times animation that illustrates the Central Limit Theorem

Sampling Distributions

- **The sample mean $\bar{X}$** is a statistic whose value is the **average of sample data from a population**.
- For random samples of size n taken from a given population, **the random variable $\bar{X}$** is **the collection of these sample means, the $\bar{X}$'s**. Like any random variable, $\bar{X}$ has a probability distribution associated with it.
- The **probability distribution of the sample means $\bar{X}$'s** is the **sampling distribution of the mean $\bar{X}$**.
- The **sampling distribution of $\bar{X}$** depends on:
  - The **distribution of the original population** (e.g., normal, skewed, uniform, symmetric),
  - the **sample size n**, and
  - the **method of sample selection**.

The Central Limit Theorem

- **The Central Limit Theorem** is the heart of probability theory.
- The theorem states that the **sampling distribution of $\bar{X}$** can be approximated by a **normal distribution** when the **sample size n is "sufficiently large,"** irrespective of the shape of the original population distribution.
- As the **sample size n increases**, the corresponding **distribution of the sample mean $\bar{X}$ will "converge" around the true population mean μ**. The effect of an "outlier" is diminished when averaged with a large sample.
- A better name for the CLT would be the "**normal convergence theorem**" since the distribution converges about the population mean as the sample size n increases.
- The symbolic explanation of the Central Limit Theorem is:

Let $X_1, X_2, X_3, \ldots, X_n$ be **independent and identically distributed (IID) non-normal random variables** with mean μ and standard deviation σ. If n is "large" enough ($n > 30$ suggested in most texts), then:

The distribution of the sample mean $\bar{X}$ **is *approximately* normally distributed** with **mean μ** and **standard dev** $\frac{\sigma}{\sqrt{n}}$.

The larger the sample size n, the more the distribution appears normal and more tightly converges about the mean μ.

In addition to the CLT, let $X_1, X_2, X_3, \ldots, X_n$ be **IID normally distributed random variables** with mean μ and standard deviation σ.

The distribution of the sample mean $\bar{X}$ is ***exactly* normally distributed** with **mean μ** and **standard deviation** $\frac{\sigma}{\sqrt{n}}$.

The larger the sample size n, the tighter the curve converges about the true population mean μ.

The CLT is difficult to understand without graphics of $\bar{X}$ for increasing sample sizes.

The following example considers averaging **non-normal random variables** for **increasing sample sizes**. Histograms will be used to display the shape of the distribution of $\bar{X}$.

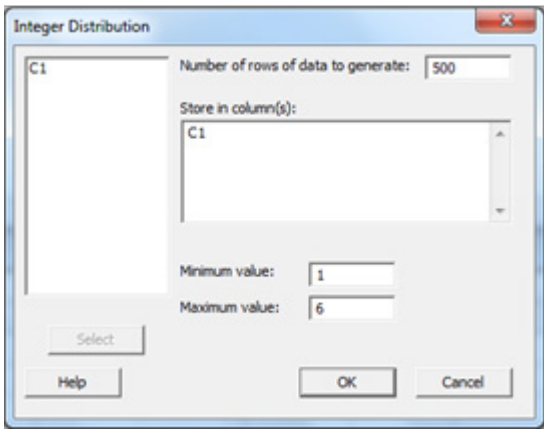
Example 1.

- Let **X1** be random variable representing the **outcome when a fair 6-sided die is rolled**. The random variable can take on the values **x = 1, 2, 3, 4, 5, 6 with probability 1/6** for each.
- In order to **graph the distribution of X1**, we'll use **Minitab to sample 500 data points** from the **distribution of X1** and put them in a Minitab column.

Example 1 continued. To sample data from the integer distribution x = 1, 2, 3, 4, 5, 6 with equally likely probabilities choose:

Minitab desktop (20 or higher):

1. Choose **Calc > Random Data > Integer Distribution**.
2. For **Number of rows of data to generate**, type 500.
3. For **Store in columns**, type C1.
4. For **Minimum value**, type 1.
5. For **Maximum value**, type 6.
6. Click **OK**.



Minitab web app:

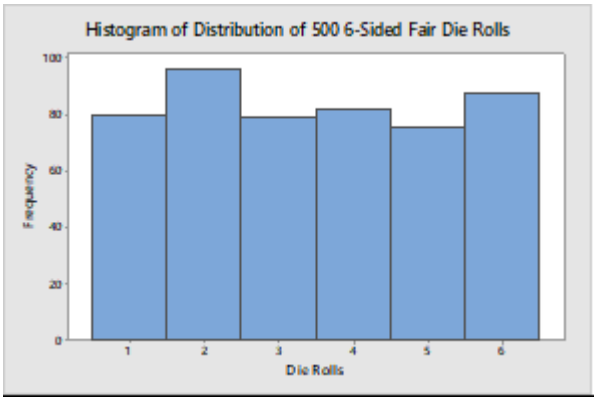
1. Choose **Calc > Random Data**.
2. For **Number of rows of data to generate**, type 500.
3. For **Store in columns**, type C1.
4. From **Distribution**, select **Integer**.
5. For **Minimum value**, type 1.
6. For **Maximum value**, type 6.
7. Click **OK**.

Column **C1** now contains **500 randomly sampled values** with equally likely probabilities chosen from x = 1, 2, 3, 4, 5, 6. Look down the column to make sure the distribution of values makes sense. As you look down column C1, realize that these are the outcomes from someone rolling a fair 6-sided die 500 times.

Below is a histogram of the data in column C1. We expect to see approximately the same heights for values x = 1, 2, 3, 4, 5, 6 around the value $\frac{1}{6} \cdot 500 \cong 83.3$. Everyone's histogram will look slightly different.

Minitab desktop (20 or higher):

1. Choose **Graph** > **Histogram** > **Simple**.
2. For **Graph Variable**, enter C1.
3. Click **OK**.



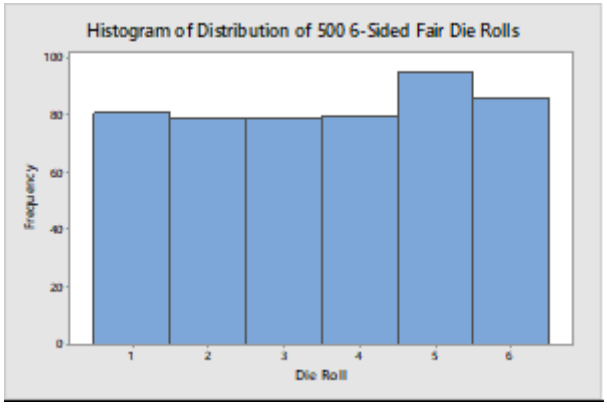
Minitab web app:

1. Choose **Graph** > **Histogram** > **One Y Variable** > **Simple**.
2. For **Y-variable**, enter C1.
3. Click **OK**.

My histogram is fairly flat with the exception of the value x = 2 having a higher frequency than expected. This is not unusual for a random process, such as rolling a die. We can see, though, that the distribution does not have the shape of a normal curve.

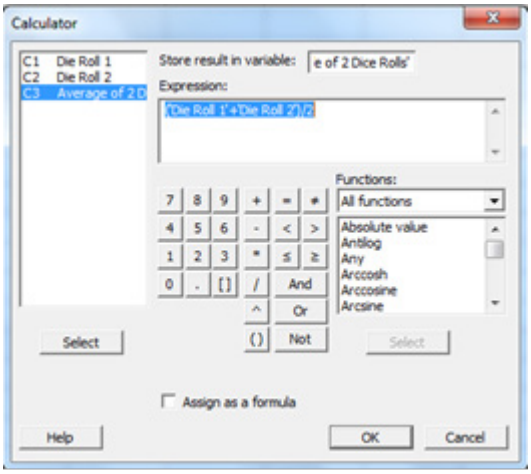
It's also worth noting that the mean of the random variable X1 is **μ = 3.5**. Although not obvious, its standard deviation is $\boldsymbol{\sigma = \sqrt{105}/6 \cong 1.708}$.

In order to graph the average $\bar{X}$ of two die rolls on the next page, we'll create **another 500 die rolls** in the **column C2**. Its **histogram** is:



Sampling Distribution for $\bar{X}$ when n = 2 and initial distribution is discrete uniform for x = 1, 2, 3, 4, 5, 6.

First, we need to use Minitab's calculator to average columns C1 and C2. I renamed these columns as "Die Roll 1" and "Die Roll 2." Here's how to average Die Roll 1 and Die Roll 2



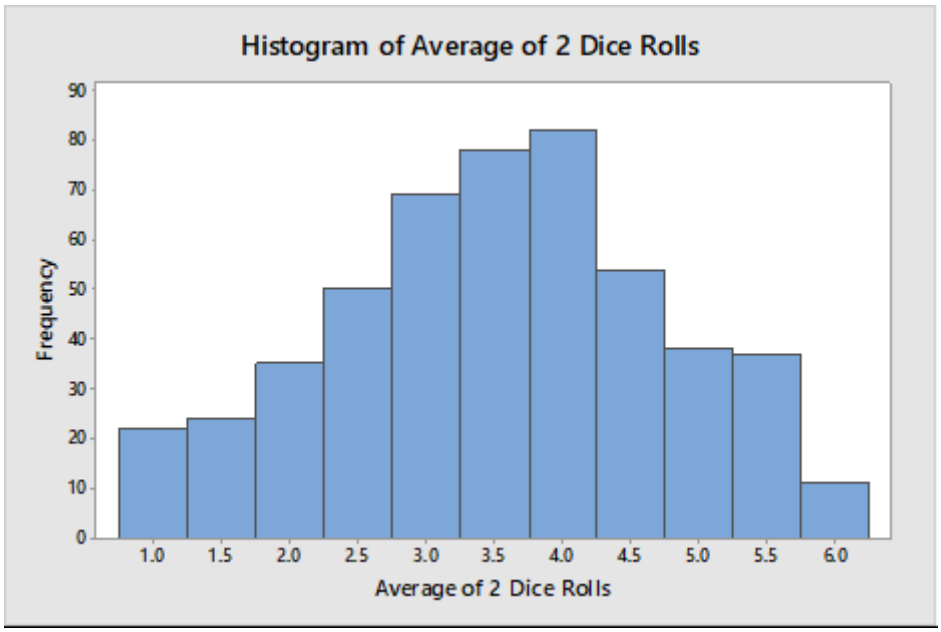
Minitab:

1. Name column C3 "Average of 2 Dice Rolls."
2. Choose **Calc > Calculator**.
3. In **Store result in variable**, enter 'Average of 2 Die Rolls'.
4. In **Expression**, enter ('Die Roll 1' + 'Die Roll 2') / 2.
5. Click **OK**.

The Minitab column "Average of 2 Die Rolls" contains 500 averages of two die rolls. Look down the column to make sure the distribution of values makes sense. These are the outcomes from someone rolling two dice and averaging their values. We'll probably see a lot of values around the average 3.5.

Below is a histogram for the column "Average of 2 Die Rolls." We expect the **highest bin to be around the value 3.5** since 6 die roll averages, namely for rolls (1, 6), (2, 5), (3, 4), (4, 3), (5, 2), and (6, 1), yield an average of 3.5.

The mean of $\overline{X}_2$ is (X¬_1+X_2)/2, which is **μ = 3.5**, and its standard deviation is $\sigma = \sqrt{\mathbf{210}}/\mathbf{12} \cong \mathbf{1.208}$.



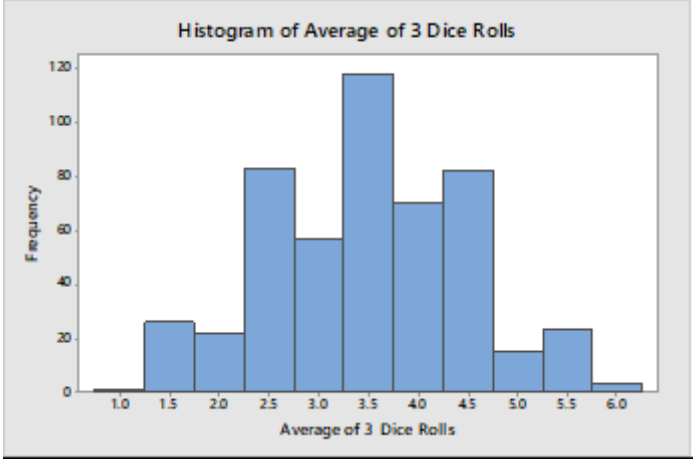
Facts about $\overline{X}$ **when n = 2**:

- The mean μ of the distribution $\overline{X}$ is 3.5, which is the same as the means of X_1 or X_2, which are also 3.5.
- The standard deviation of $\overline{X}$ is 1.208, which is less than the standard deviations of X_1 or X_2, which are 1.708.
- The shape of $\overline{X}$ is converging toward the mean 3.5 and is taking on more of a normal distribution shape.

Sampling Distribution for $\bar{X}$ when n = 3 and initial distribution is discrete uniform for x = 1, 2, 3, 4, 5, 6.

Now we will average three random variable X_1, X_2, and X_3, where each is the outcome from rolling a fair 6-sided die. In Minitab, make a new column of die tosses. Then **average the 3 random variables** in a new column in Minitab. Last, make a histogram of $\bar{X}$ when n = 3.

The mean and standard deviation for $X_3=(X_1+X_2+X_3)/3$ is μ = **3.5** and the standard deviation is $\boldsymbol{\sigma = \sqrt{35}/6 \cong 0.986}$. Notice that the mean μ = 3.5 is the same for averaging 3 die rolls as 2 die rolls. The standard deviation $\bar{X}$ for averaging 3 is smaller though.



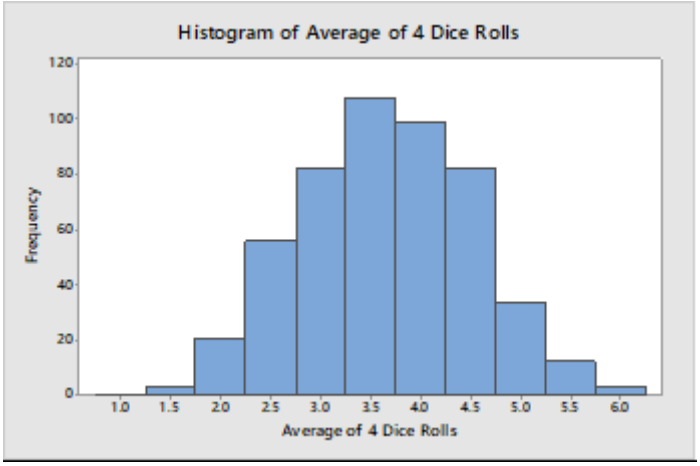
Facts about **$\bar{X}$ when n = 3**:

- The mean of the distribution $\bar{X}$ is 3.5, which is the same as the means of X_1, X_2, X_3, and $\bar{X}_2$.
- The standard deviation of $\bar{X}$ is 0.986, which is less than the standard deviations of X_1, X_2, X_3, and $\bar{X}_2$.
- The shape of $\bar{X}$ is converging toward the mean μ = 3.5 and is taking on more of a normal distribution shape.

Sampling Distribution for $\bar{X}$ when n = 4 and initial distribution is discrete uniform for x = 1, 2, 3, 4, 5, 6.

Now average 4 die roll random variables X_1, X_2, X_3, and X_4. For $\bar{X}_4$, where $\bar{X}_4=(X_1+X_2+X_3+X_4)/4$, you can anticipate what the mean and shape of the distribution will be. We'll confirm with another histogram.

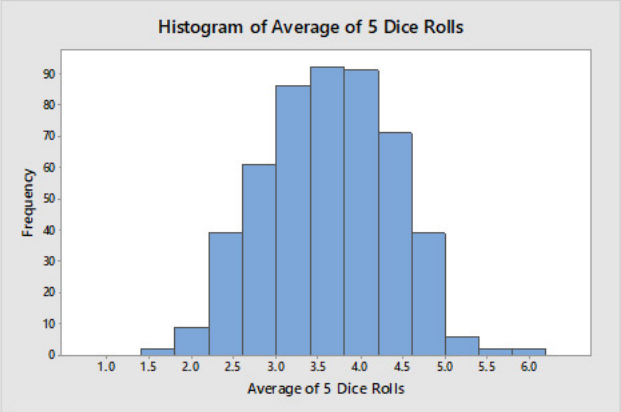
For the case n = 4, the mean is still **μ = 3.5** and the standard deviation is $\boldsymbol{\sigma = \sqrt{105}/12 \cong 0.854}$.



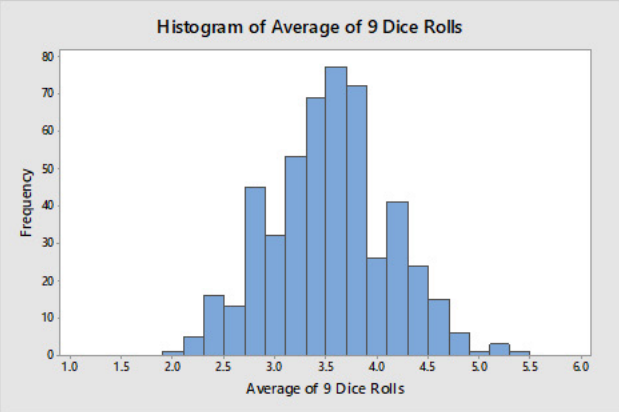
In the case n = 4, we can see a relationship between the standard deviation of X_i and $\bar{X}_4$. For X_i, the standard deviation is √105/6, while the standard deviation of $\bar{X}_4$ is half of that: √105/12.

There **is** a relationship between the standard deviation of the original distribution and the standard deviation of the sampling distribution $\bar{X}$, which is dependent on the sample size n.

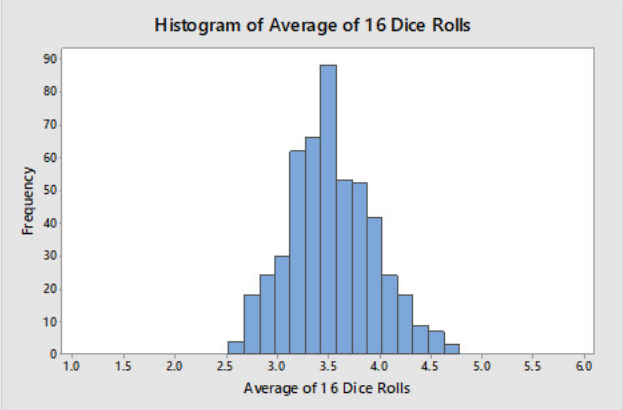
Below are histograms for $\bar{X}$ for n = 5, 9, 16, and 25. For each histogram, the mean and standard deviation are reported. You can see the relationship between the standard deviations with these additional plots.



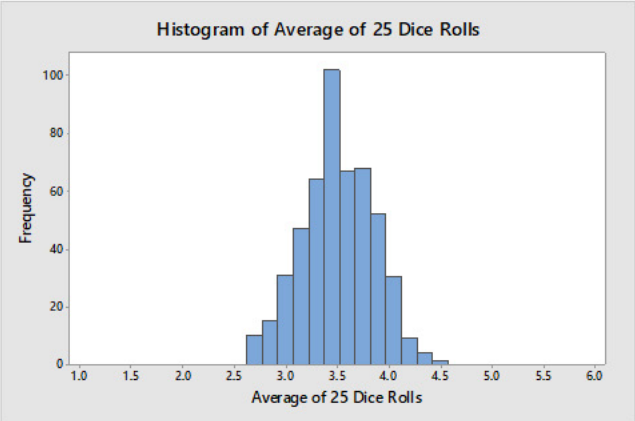
Mean μ = 3.5, Standard Deviation σ = $\sqrt{105}/(6\sqrt{5})$



Mean μ = 3.5, Standard Deviation σ = $\sqrt{105}/(6\sqrt{9})$



Mean μ = 3.5, Standard Deviation $\sqrt{105}/(6 \cdot \sqrt{16})$



Mean μ = 3.5, Standard Deviation σ = $\sqrt{105}/(6\sqrt{25})$

The standard deviations for the distribution of X, $\bar{X}_4$, $\bar{X}_9$, $\bar{X}_{16}$, and $\bar{X}_{25}$ are shown below:

	Original Distribution	Distribution of $\bar{X}_4$	Distribution of $\bar{X}_9$	Distribution of $\bar{X}_{16}$	Distribution of $\bar{X}_{25}$
Standard Deviation	$\sqrt{105}/6$	$\sqrt{105}/(6 \cdot 2)$	$\sqrt{105}/(6 \cdot 3)$	$\sqrt{105}/(6 \cdot 4)$	$\sqrt{105}/(6 \cdot 5)$

Given that **σ is the original population standard deviation**, the **standard deviation of $\bar{X}_n$ is $\sigma/\sqrt{n}$.**

RECAP: Let $X_1, X_2, X_3, \ldots, X_n$ be independent non-normal random variables with identical distributions with mean μ and standard deviation σ. If n is "large" enough ($n > 30$ suggested in most texts), then:

The distribution of the sample mean $\bar{X}$ is *approximately* normally distributed with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

The larger the sample size n, the more the distribution appears normal and more tightly converges about the mean μ.

Note: Although textbooks suggest n > 30 for $\bar{X}$ to approach normality, you can see from the plots on this page that the histograms are displaying a normal shape for n < 30.

- The convergence to normality is dependent on the shape of the original distribution.
- Fairly flat or symmetric distributions tend to have normal $\bar{X}$ plots for smaller sample sizes (e.g. n < 30).

Example 2. Animations of averages of normal, uniform, and right skewed random variables. Examine Rice applet

Advantages of normally distributed $\bar{X}$'s

If we can assume that $\bar{X}$ is normally distributed, and we know the mean μ and standard deviation X of the original distribution, then we can easily determine probabilities for $\bar{X}$ using Minitab (or integrating in Maple).

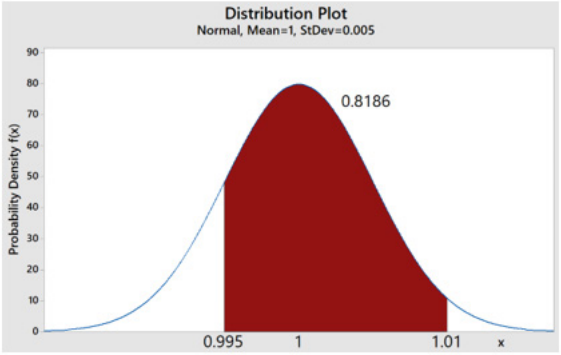
Example 3. Let $X_1, X_2, X_3, \ldots, X_{100}$ denote the weights of 100 independent and identically distributed (IID) bags of candy corn. If the mean weight of each bag is μ = 1 lb and its standard deviation is o = 0.05 lb, determine the probability that the average weight X for 100 bags is between 0.995 lb and 1.01 lb.

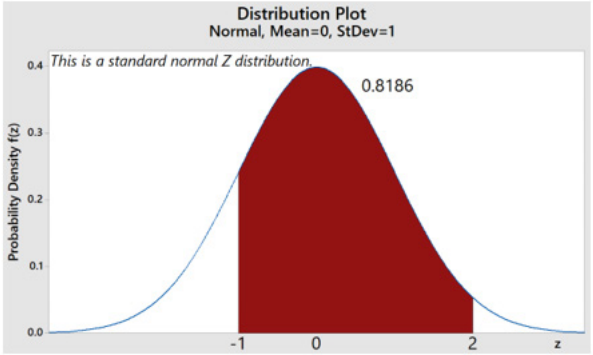
According to the **Central Limit Theorem**, the distribution for $\bar{X}$ for n = 100 is approximately normal, regardless of the distribution of each Xi. We are told that each X_i has the same distribution with mean μ = 1 lb and standard deviation o = 0.05 lb.

Using the formulas that we determined in this lesson, the mean and standard deviation of $\bar{X}$, respectively, are μ = 1 lb and o = 0.05/√100=0.005 lb. Since $\bar{X}$ is approximately normal, then we can use Minitab to compute the required probability. Recall that Z represents a standard normal random variable with mean μ = 0 and std dev o = 1.

$$P(0.995 < \bar{X} < 1.01) = P\left(\frac{0.995 - 1}{0.005} < Z < \frac{1.01 - 1}{0.005}\right) = P(-1 < Z < 2) \cong \mathbf{0.81859}$$

We can determine this probability in Minitab using either $\bar{X}$'s distribution or the standard normal Z distribution.



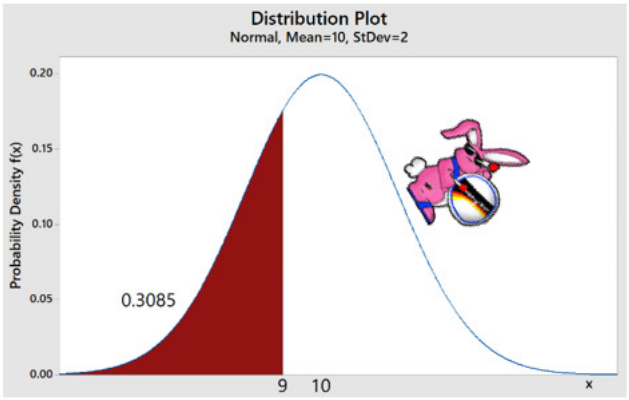


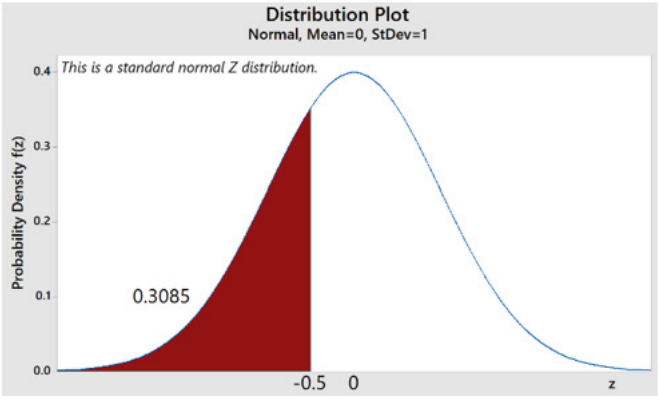
Example 4. Consider a pack of n = 4 Energizer batteries. The lifetime of this brand of battery is normally distributed with a mean of μ = 10 hours and a standard deviation of σ = 2 hours.

(a) If we randomly select **one battery** from the pack, what is the probability that the battery's lifetime is less than 9 hours?

Solution: Let X represent the lifetime of one battery selected from the pack. We can use Minitab to compute this probability using X's distribution or the standard normal Z distribution.

$$P(X < 9) = P\left(\frac{X-10}{2} < \frac{9-10}{2}\right) = P(Z < -0.5) \cong \mathbf{0.30854}.$$



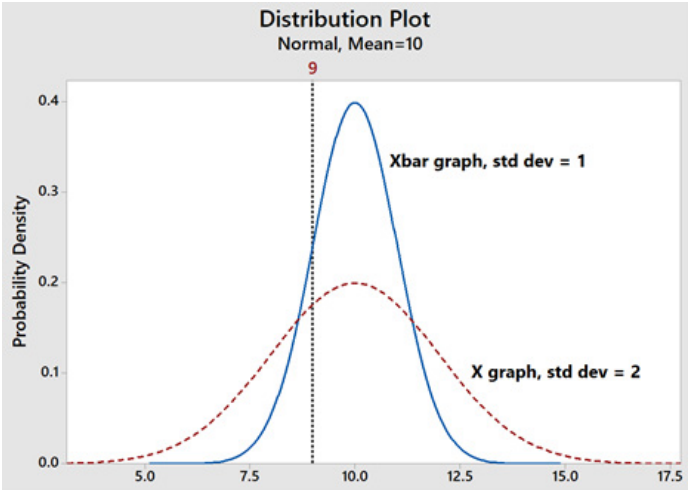


(b) What is the probability that the **average lifetime** $\bar{X}$ of four batteries in a pack is less than 9 hours? For example, one battery could last 9.5 hours, another 8.2 hours, another 11.6 hours, etc.

Solution: Let $\bar{X}$ represent the average lifetime of the pack of 4 batteries. Since X is normally distributed, then $\bar{X}$ is also normally distributed with mean μ = 10 and standard deviation σ = 2/√4 = 1.

$$P(\bar{X} < 9) = P\left(Z < \frac{9-10}{1}\right) = P(Z < -1) \cong \mathbf{0.15866}$$

Below are the graphs of X and $\bar{X}$ for n = 4 on the same plot. You can visually see the difference in the probabilities by looking at the area under the curve to the left of the value 9.



Example 4. (True story) In North Carolina in 2003, an overweight plane crashed in part due to the weight of the passengers' luggage. Below is a scenario about the probability of exceeding luggage requirements.

A small commuter flight leaves from University Park Airport headed to the O'Hare Airport in Chicago with n = 20 passengers. By FAA weight standards for carry-on luggage for this flight, a passenger's carry-on luggage should not exceed 40 pounds. Let's assume that the weight of a passenger's carry-on luggage is normally distributed with a mean weight of μ = 25 pounds and a standard deviation of o = 10 pounds and only **1 carry-on** is allowed per passenger. What's the probability that the **average luggage weight** for the 20 passengers exceeds 40 pounds?

Solution: Since the luggage weights are normally distributed, then the mean of the luggage weights X is also normally distributed with mean $\mu_{\bar{X}}$ =25 lbs. and standard deviation $\sigma_{\bar{X}}$=10/√20 lbs.

Thankfully, the desired probability is approximately 0:

$$P(\bar{X} > 40) = P\left(Z > \frac{40 - 25}{\frac{10}{\sqrt{20}}}\right) \cong P(Z > 6.71) \cong 0$$