

**Sampling Distribution of the Mean:**

- Suppose that the random variable  $X$  has a probability distribution that has mean  $\mu = E[X]$  and standard deviation  $\sigma$  (variance  $\sigma^2$ ). Let's use  $\mu = 75$  and  $\sigma = 10$  ( $\sigma^2 = 100$ ) to illustrate the essential facts.
- Suppose in addition that we draw a sample of **size  $n$**  from this population and find the mean  $\bar{x}$  of the sample. Let's use  $n = 25$  to illustrate the essential facts.
- Suppose further that we look at all possible samples of size  $n$  that we could potentially draw from this population (a gigantic number) and calculate the sample mean  $\bar{x}$  for each of these samples.
- The following facts will be true for the probability distribution of the sample means  $\bar{x}$ .

- The average value of all the sample means will be equal to the mean of the original population. In other words the mean of all the  $\bar{x}$  is equal to the mean  $\mu$  of the original  $X$ s. (Note: this indicates that  $\bar{x}$  is an unbiased estimator of  $\mu$ .) We can state this mathematically as follows:

$$E[\bar{x}] = \mu \quad \text{or} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}, \text{ so for our simple example } \sigma_{\bar{x}} = \frac{10}{\sqrt{25}} = 2$$

- If the samples are drawn with replacement or if they are drawn from an infinite population, the variance and standard deviation of  $\bar{x}$  can be found from the original population values as follows:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}, \text{ so for our example } \sigma_{\bar{x}} = \frac{10}{\sqrt{25}} = 2$$

- If the sample are drawn without replacement from a finite population of size  $N$  then the variance and standard deviation of  $\bar{x}$  can be found from the original population values but the above formulas must be modified as follows:

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

- The expression  $\sqrt{\frac{N-n}{N}}$  is referred to as the finite population correction factor.
- When the sample size  $n$  is much smaller than the population size  $N$ , the finite population correction factor is essentially equal to  $1$ . Since multiplying by  $1$  leaves the remaining expression unchanged the finite population correction factor can be omitted in such situations. Suppose  $n = 50$  and  $N = 20000$ , then:

$$\sqrt{\frac{20000-50}{20000}} \approx \sqrt{\frac{19950}{20000}} \approx \sqrt{0.9975} \approx 0.99875$$

- One only has to be diligent in using the finite population factor when the sample size  $n$  exceeds 5% of the population size  $N$ . Suppose  $N = 1000$  here  $n = 50$  and  $n = 100$  represent 5% and 10 % of the population respectively. For these two values of  $n$  the finite population correction factor becomes:

$$\begin{aligned} &\sqrt{\frac{1000-50}{1000}} \approx \sqrt{\frac{950}{1000}} \approx \sqrt{0.95} \approx 0.9747 \\ &\sqrt{\frac{1000-100}{1000}} \approx \sqrt{\frac{900}{1000}} \approx \sqrt{0.9} \approx 0.9487 \end{aligned}$$

**An Example to Practice the Calculations:**

Suppose that we have a uniform distribution with  $X$  between  $-3$  and  $3$  ( $-3 \leq X \leq 3$ ). For this distribution the mean, standard deviation and variance are:

$$\begin{aligned} & \text{Mean} = \frac{a+b}{2} = \frac{-3+3}{2} = 0 \\ & \text{Standard Deviation} = \sqrt{\frac{(b-a)^2}{12}} = \sqrt{\frac{(3-(-3))^2}{12}} = \sqrt{\frac{36}{12}} = \sqrt{3} \\ & \text{Variance} = \frac{(b-a)^2}{12} = \frac{(3-(-3))^2}{12} = \frac{36}{12} = 3 \end{aligned}$$

If sample sizes of size  $n$  are taken, fill in the following table for the mean, standard deviation and variance of the sampling distribution of the mean.

Sample Size	Mean	Standard Deviation	Variance
$n = 1$	0	$\sqrt{3}$	3
$n = 5$			
$n = 10$			
$n = 15$			
$n = 20$			
$n = 25$			
$n = 30$			

**Simulation:**

Another way to view these results is through simulation. This is similar to the relative frequency method of estimating a probability. In simulation you have a computer perform the experiment multiple times and average your results. For the data summarized below 5000 samples of size  $n$  were generated with the following results:

**Descriptive Statistics: Single X, X-Bar-5, X-Bar-10, X-Bar-15, X-Bar-20, X-Bar-25, X-Bar-30**

Variable	N	Mean	Median	TrMean	StDev	SE Mean
Single X	5000	0.0123	0.0339	0.0142	1.7252	0.0244
X-Bar-5	5000	-0.0014	0.0016	0.0015	0.7751	0.0110
X-Bar-10	5000	-0.01247	-0.02443	-0.01474	0.54824	0.00775
X-Bar-15	5000	0.00340	0.00632	0.00435	0.45061	0.00637
X-Bar-20	5000	0.00311	0.00482	0.00377	0.38478	0.00544
X-Bar-25	5000	-0.00007	0.00178	-0.00071	0.34470	0.00487
X-Bar-30	5000	-0.00812	-0.00648	-0.00810	0.31629	0.00447

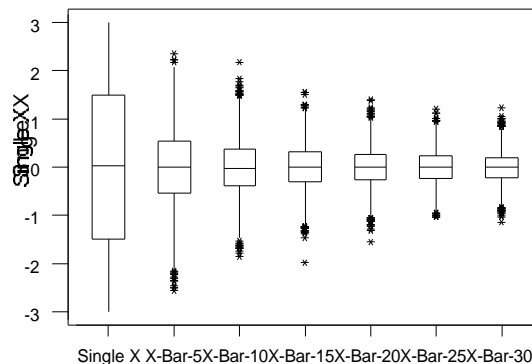
  

Variable	Minimum	Maximum	Q1	Q3
Single X	-2.9966	2.9997	-1.4841	1.4920
X-Bar-5	-2.5646	2.3503	-0.5359	0.5425
X-Bar-10	-1.84958	2.18876	-0.38840	0.35900
X-Bar-15	-1.98505	1.56577	-0.29591	0.31718
X-Bar-20	-1.55609	1.40131	-0.25548	0.26514
X-Bar-25	-1.04075	1.22242	-0.23067	0.23683
X-Bar-30	-1.14929	1.22696	-0.21738	0.19815

The simulated values for the mean and standard deviations and those for the predicted values are shown in the table below.

Sample Size	Predicted Mean	Simulation Mean	Predicted Standard Deviation	Simulation Standard Deviation
$n = 1$	0	0.0123	1.7321	1.7252
$n = 5$	0	-0.0014	0.7746	0.7751
$n = 10$	0	-0.01247	0.5477	0.54824
$n = 15$	0	0.00340	0.4472	0.45061
$n = 20$	0	0.00311	0.3873	0.38478
$n = 25$	0	-0.00007	0.3464	0.34470
$n = 30$	0	-0.00812	0.3162	0.31629

The box plots on the right show how the variability shrinks as the sample sizes decrease. In addition, these box plots show that the mean is the same for each sample size.



**Central Limit Theorem:**

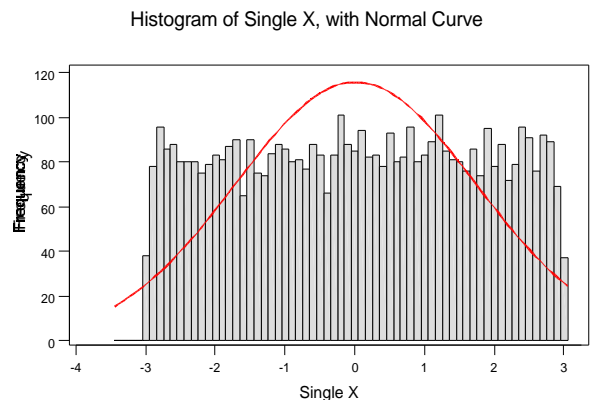
- Suppose that the random variable  $X$  has a probability distribution that may or may not follow a normal probability distribution.
- Even if  $X$  is not normal,  $\bar{x}$  will be approximately normal if the sample size is large enough.
- How large must the sample size be for  $\bar{x}$  to be approximately normal?
  - If the distribution of  $X$  itself is already normal, then  $\bar{x}$  will be normal for any sample size  $n$ .
  - If the distribution of  $X$  is not normal, or if we do not know whether or not  $X$  is normal,  $\bar{x}$  will be approximately normal provided the sample size is at least 30 ( $n \geq 30$ ). Note: for some non-normal distributions, the approximate normality of  $\bar{x}$  can be achieved with even smaller sample sizes, but you must know the correct distribution, which is rare.
- What is the consequence of this theorem?
  - You can find probabilities involving  $\bar{x}$  with z-scores using a formula similar to that for a single  $X$ . The formulas for a single  $X$  and those for  $\bar{x}$  are as follows:

$$z = \frac{x - \mu}{\sigma}$$

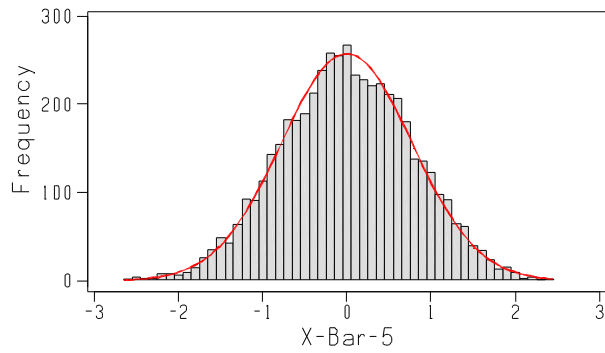
$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

**Histograms Show the Central Limit Theorem at Work:**

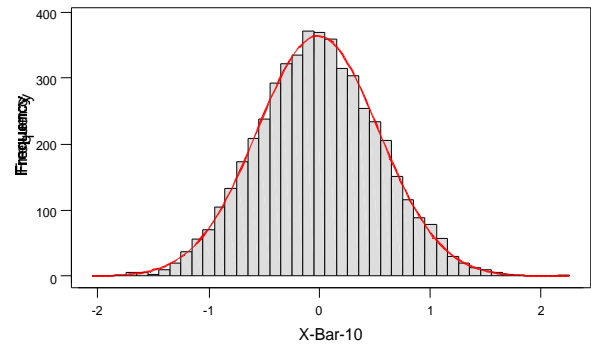
The graph on the right is a histogram of a simulation of 1000  $X$ s with a uniform distribution using  $X$  between  $-3$  and  $3$  ( $-3 \leq X \leq 3$ ). A normal distribution with the same mean and standard deviation is superimposed over the histogram. The rest of this table (see next page) gives similar histograms for samples of size 5, 10, 15, 20, 25, & 30. Notice how the histograms appear to get close to the normal curve as the sample sizes get larger. This shows the Central Limit Theorem at work.



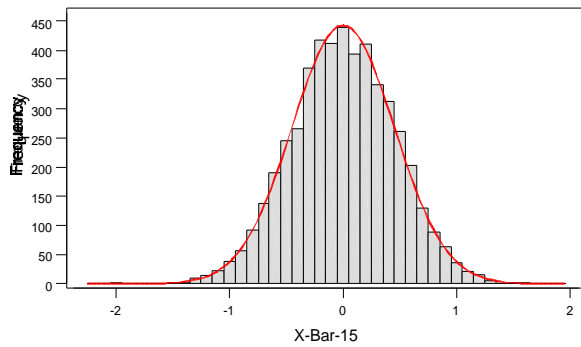
Histogram of X-Bar-5, with Normal Curve



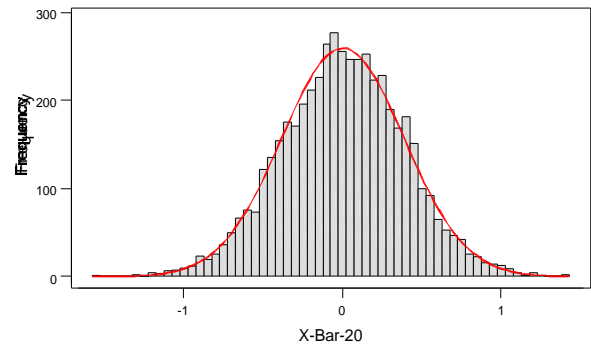
Histogram of X-Bar-10, with Normal Curve



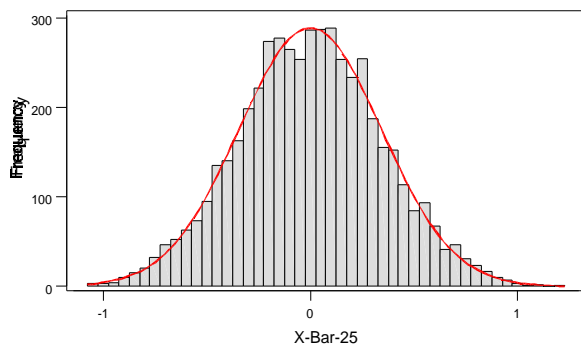
Histogram of X-Bar-15, with Normal Curve



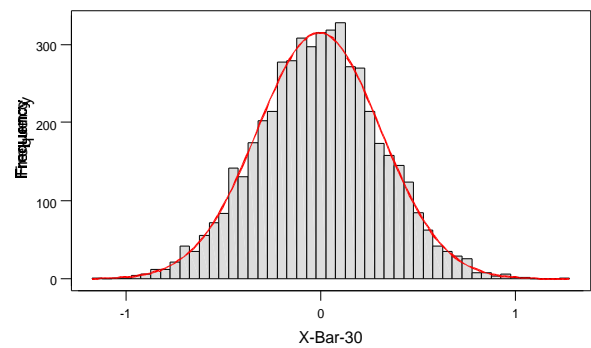
Histogram of X-Bar-20, with Normal Curve



Histogram of X-Bar-25, with Normal Curve

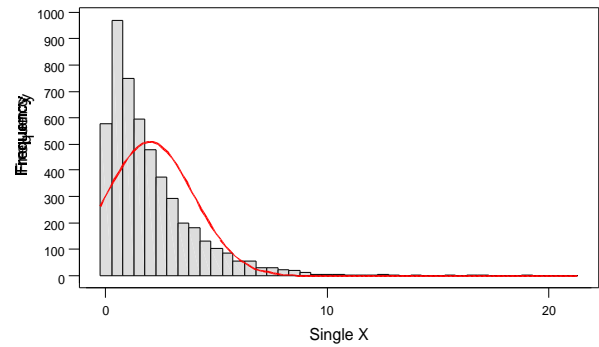


Histogram of X-Bar-30, with Normal Curve

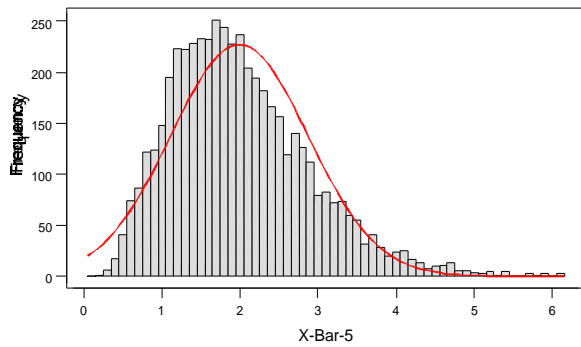


Here is another distribution called the Exponential Distribution. Note that it is not symmetrical. It is very skewed to the right. The mean and standard deviation for the Exponential Distribution always equal each other. For the distribution shown, the mean and standard deviation were chosen to be 2.

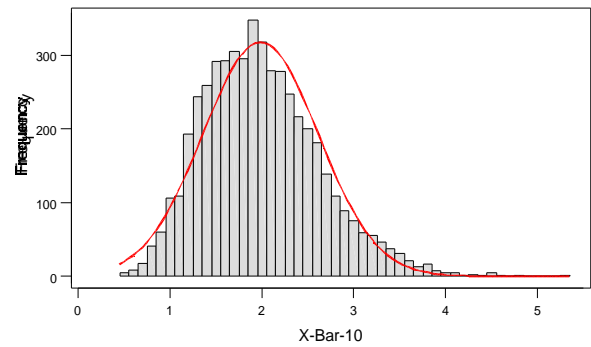
Histogram of Single X, with Normal Curve



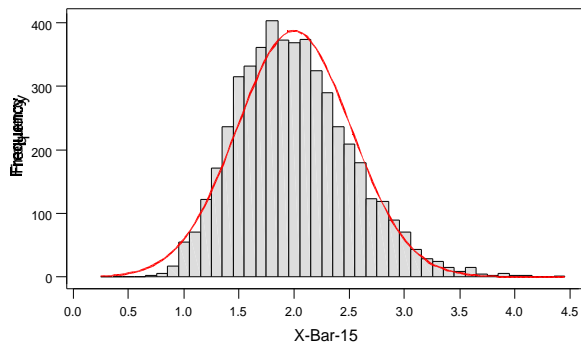
Histogram of X-Bar-5, with Normal Curve



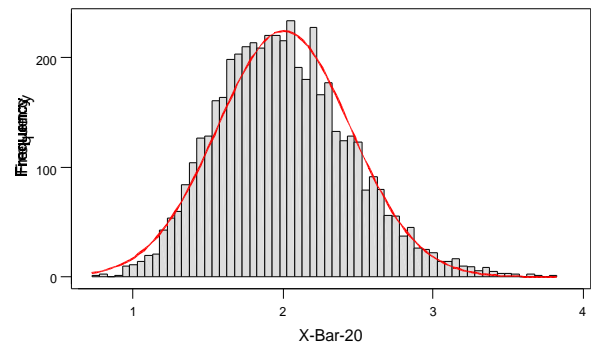
Histogram of X-Bar-10, with Normal Curve

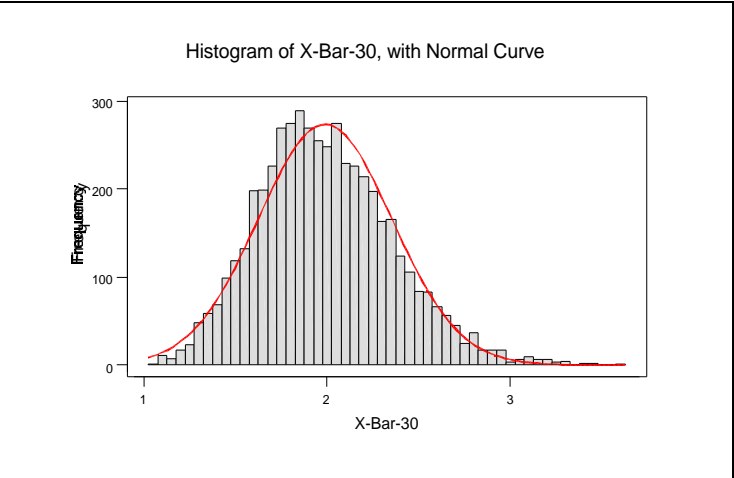
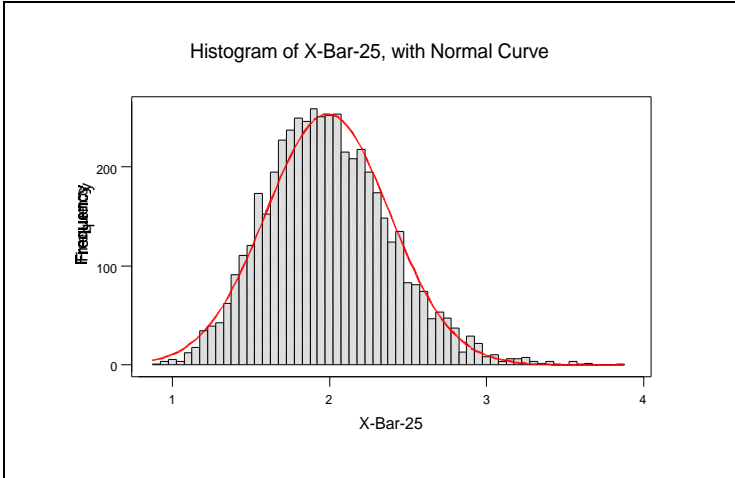


Histogram of X-Bar-15, with Normal Curve



Histogram of X-Bar-20, with Normal Curve





**Descriptive Statistics: Single X, X-Bar-5, X-Bar-10, X-Bar-15, X-Bar-20, X-Bar-2**

Variable	N	Mean	Median	TrMean	StDev	SE Mean
Single X	5000	1.9950	1.4205	1.7796	1.9652	0.0278
X-Bar-5	5000	1.9942	1.8707	1.9486	0.8792	0.0124
X-Bar-10	5000	1.9874	1.9284	1.9641	0.6296	0.0089
X-Bar-15	5000	2.0007	1.9570	1.9849	0.5153	0.0073
X-Bar-20	5000	2.0056	1.9789	1.9944	0.4438	0.0063
X-Bar-25	5000	1.9920	1.9674	1.9822	0.3931	0.0056
X-Bar-30	5000	1.9902	1.9615	1.9804	0.3650	0.0052

Variable	Minimum	Maximum	Q1	Q3
Single X	0.0003	18.7539	0.5846	2.7814
X-Bar-5	0.2098	6.0921	1.3466	2.5033
X-Bar-10	0.4506	5.2675	1.5247	2.3771
X-Bar-15	0.6915	4.4441	1.6280	2.3226
X-Bar-20	0.7268	3.8244	1.6864	2.2865
X-Bar-25	0.9061	3.8271	1.7155	2.2450
X-Bar-30	1.0400	3.6127	1.7323	2.2234

You can see in the box plots below how skewed this distribution is, but as the sample size increases, the variability shrinks and the distribution of  $\bar{x}$  becomes more symmetrical as n increase. The second graph below leaves off the Single X box plot, so that you can see the others in more detail.

