The distribution of the sample means.

In this lesson, we shall consider the cases where the population is at least 20 times larger than the size of the sample.
For example:

If we take a simple random sample of size 30 from the employees of a large corporation which has 1100 employees,
$\frac{1100}{30} = 36.66666667$**, the population is 36.67 times more than the sample size**
This example meets the at least 20 times larger criterion.

On the other hand if we take a simple random sample of 50 students from a small college with 700 students
$\frac{700}{50} = 14$ **, the at least 20 times criterion is not met. In such cases, we use what is called "Finite Population Correction Factor"**

When the population is at least 20 times larger than the sample, we have the following Theorem:
If x (the population) has mean μ and standard deviation σ
then $\overline{x}$ mean of simple random samples of size n, has mean μ and st dev $\frac{\sigma}{\sqrt{n}}$

If x is normal, $\overline{x}$ is also normal.

Let us consider the following example for an illustration of the above result:

Example:
Given that the life of a certain make of light bulb shows a normal distribution with

mean 1000 hours and st dev σ of 250 hours.

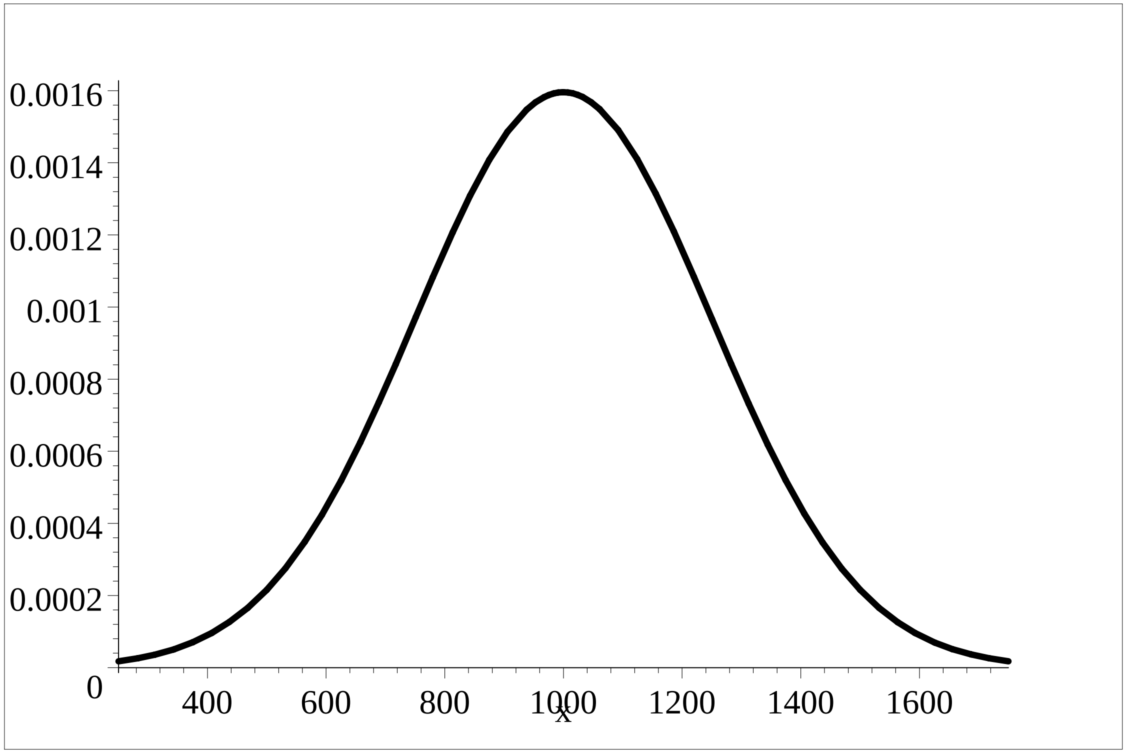
$\mu = $ **1000**

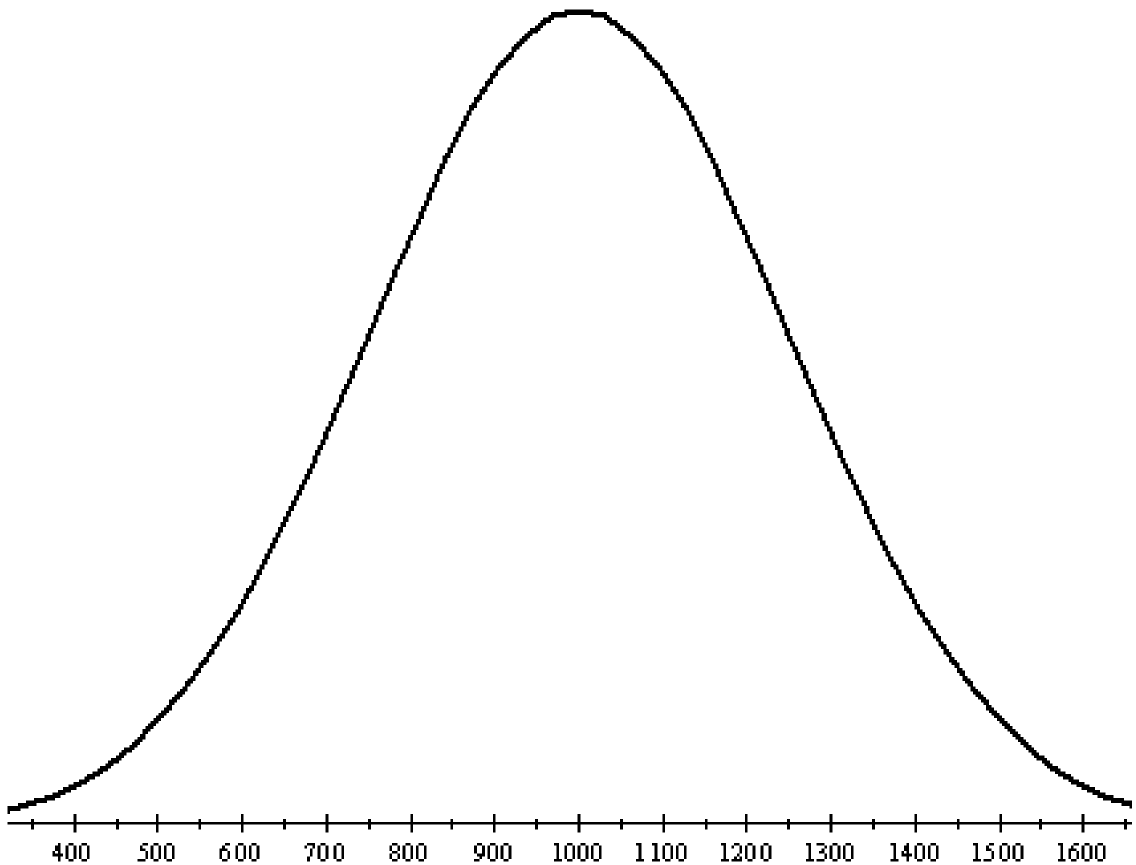
$\sigma = $ **250**

$b(x) = \frac{1}{250\sqrt{2\pi}} e^{-\left((x-1000)^2/(2\times 250^2)\right)}$ **(you may ignore this expression, used only to sketch the graph)**

The distribution of x is given by

$b(x)$



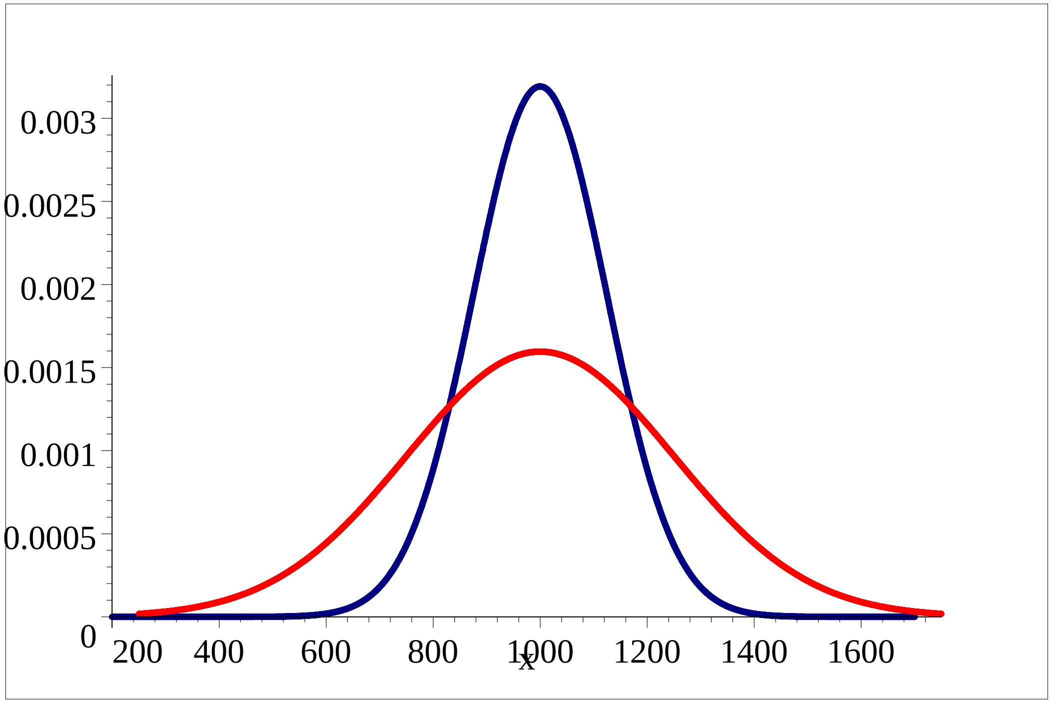


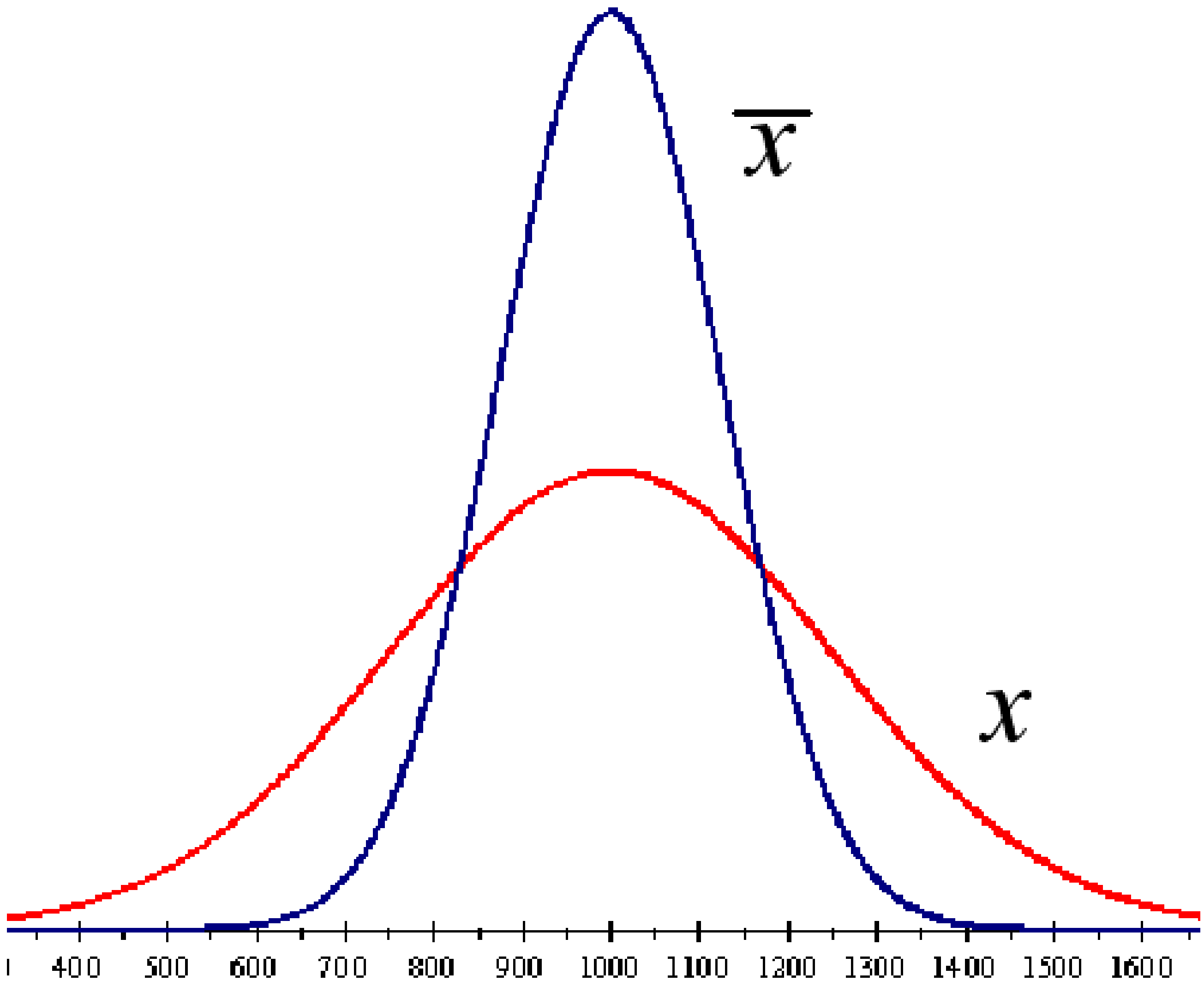
If we look at the distribution of the average life of 4-pack of bulbs from the above population, that is simple random samples of size 4 from the above population, the mean of those $\overline{x}$ has mean 1000 hours st dev is $\frac{250}{\sqrt{4}} = 125$ hours

and the distribution is normal.

$a(x) = \frac{1}{125\sqrt{2\pi}} e^{-\left((x-1000)^2/(2\times 125^2)\right)}$ **(you may ignore this expression, I have used it only to sketch the graph)**

The following picture shows that the distribution of $\overline{x}$ has less spread than x.





red: x
blue: $\overline{x}$

Let us answer the following question:

What is the probability that the average life of 4 randomly selected bulbs from the above popualtion is less than 720 hours ?

that is

Probability that $\overline{x} < 720$

Since $\overline{x}$ **have a normal distribution, we can transform to z as shown below**

$$z < \frac{720 - 1000}{125} = -2.24$$

$$\begin{array}{ccc} & & .04 \\ & & \downarrow \\ -2.2 & \rightarrow & 0.0125 \end{array}$$

$z < -2.24$

area

0.0125

Therefore the probability that average life of randomly selected 4 bulbs will be less than 720 hours is 0.0125.

CENTRAL LIMIT THEOREM:

If x has mean μ **and standard deviation** σ**,**

then for simple random samples of size n from this population,
$\overline{x}$ **shows appromately a normal distribution if the sample size large.**

Many practioners treat 30 as a large enough sample size to meet this criterion.
Again, the mean of $\overline{x}$ **is still** μ **and standard deviation is** $\frac{\sigma}{\sqrt{n}}$

Let us look at an illustration through the following example.

Suppose a population has mean $\mu = 27100$ **and standard deviation** $\sigma = 5200$ **and we take a simple random sample**
of size 30 from this population. We would like to find the probability that the sample mean $\overline{x}$ **is going to be within**
1000 of the population mean 27,100.

this means $-1000 < \overline{x} - 27100 < 1000$
Since, $\overline{x}$ **has approximately a normal distribution**
Note that $z = \dfrac{\overline{x} - 27100}{\left(\dfrac{5200}{\sqrt{30}}\right)}$

Therefore
$-1000 < \overline{x} - 27100 < 1000$
$\Leftrightarrow$
$$\frac{-1000}{\left(\dfrac{5200}{\sqrt{30}}\right)} < \frac{\overline{x} - 27100}{\left(\dfrac{5200}{\sqrt{30}}\right)} < \frac{1000}{\dfrac{5200}{\sqrt{30}}}$$
$\Leftrightarrow$

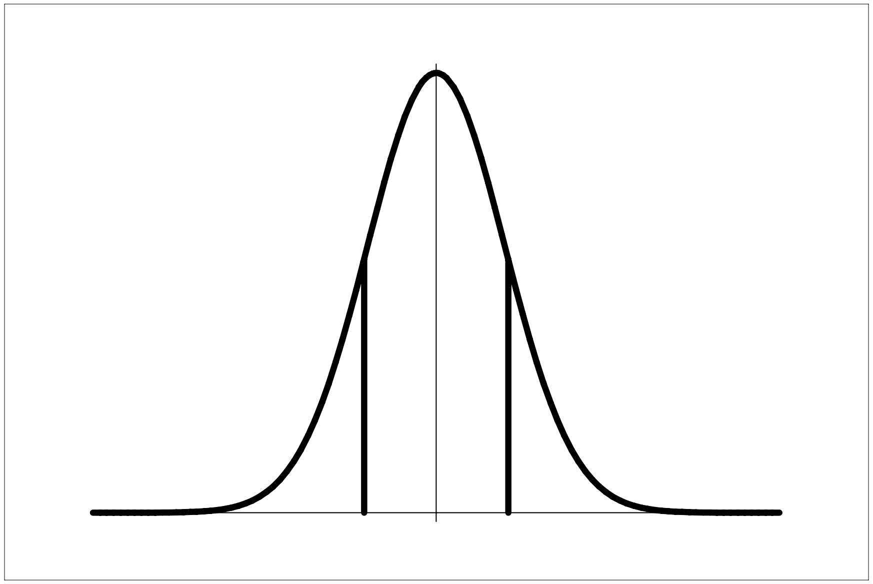
$$\frac{-1000}{\left(\frac{5200}{\sqrt{30}}\right)} < \mathbf{z} < \frac{1000}{\frac{5200}{\sqrt{30}}}$$

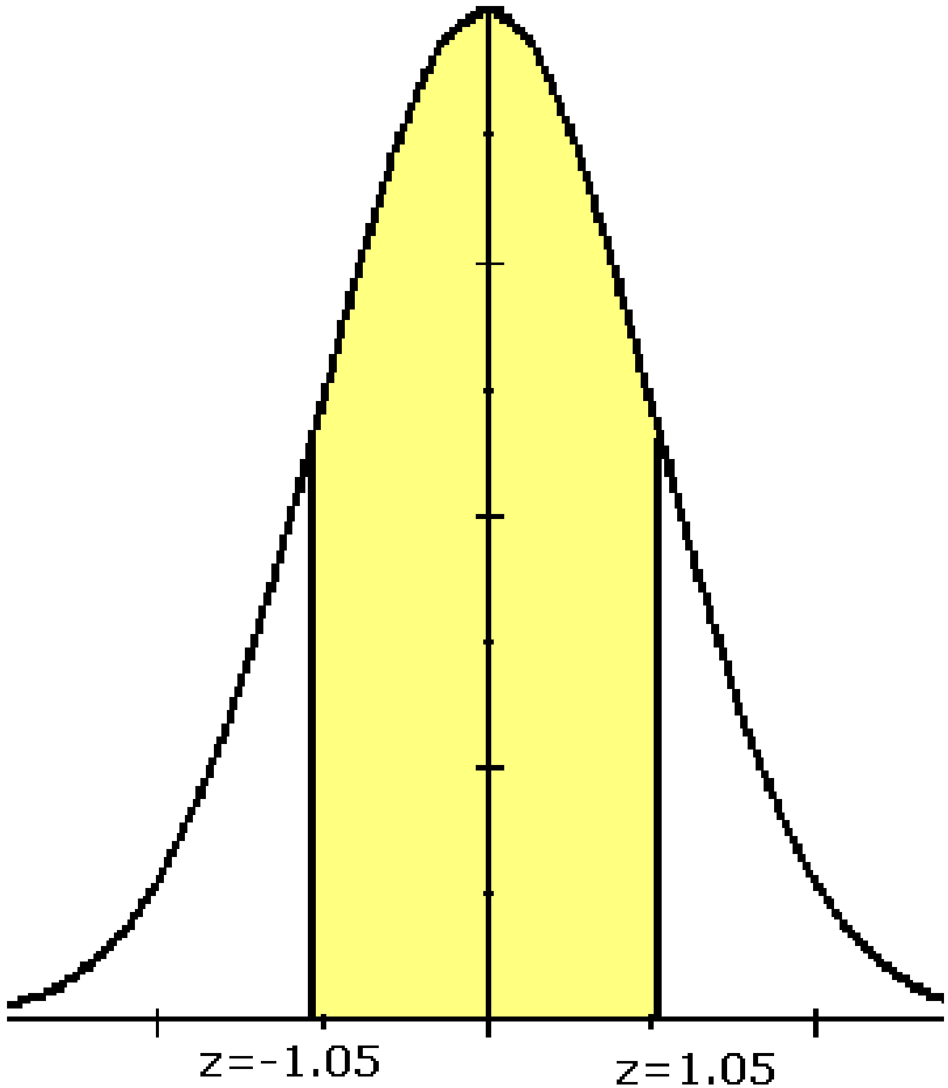
$\Leftrightarrow$

$\mathbf{-1.05 < z < 1.05}$

Therefore the probability of $-1000 < \overline{x} - 27100 < 1000$
is the same as the probability of $-1.05 < z < 1.05$
which is the asme as the shaded area in the two tails in the following picture.
$s(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$ **(you may ignore this expression, used only to sketch the graph)**

$s(z)$





according to the table is

$$
\begin{array}{ccc}
 & & .05 \\
 & & \downarrow \\
\mathbf{1.0} & \rightarrow & \mathbf{0.8531} \\
 & & .05 \\
 & & \downarrow \\
\mathbf{-1.0} & \rightarrow & \mathbf{0.1469}
\end{array}
$$

Therefore, the area between z=-1.05 **and z=1.05 is**
$.8531 - .1469 = 0.7062$

Therefore the probability of $-1000 < \overline{x} - 27100 < 1000$ **is 0.7062.**

Consider another illustration of this idea with the following

Example:

Suppose that the insurance claims for a particular problem, show a mean of $\$987.00$ **and standard deviation $781.00.**
If the same parameters hold, find the probability that the average claim on a random sample of 100 cars will be more than $1100.00.

To find the
Probability that
$\overline{x} >$ **1100**

or

$\rightarrow$ **z** $> \dfrac{1100 - 987}{\left(\dfrac{781}{\sqrt{100}}\right)} =$ **1.4469**

$\rightarrow$ **z** $>$ **1.45**

From the z-table

		.05
		$\downarrow$
1.4	$\rightarrow$	**.9265**

The area is $1 - 0.9265 = 0.0735$

or the probability that the average will exceed $1100 is 0.0735.

Example 4.

The following example is taken from Introduction to Business Statistics by Ronald Weiers, fourth edition, page 290.

From the past experience, an airline has found the luggage weight for individual air travelers on their trans-Atlantic route to have a mean of 80 pounds and standard deviation 20 pounds. The plane is consistently fully booked and holds 100 passengers. The pilot insists on having extra fuel if the TOTAL luggage weight exceeds 8300 pounds. On what percentage of the flights will she end up having extra fuel.

For the weight of the luggage (x) of an individual passenger, we have $\mu = 80$**lbs and** $\sigma = 20$**lbs.**

Note that extra fuel is needed when

total $>$ **8300**
or
average > $\frac{8300}{100} = 83.0$ **(8300 is the total for 100 passengers)**
or
$\overline{x} >$ **83**
or
z $> \frac{83 - 80}{\left(\frac{20}{\sqrt{100}}\right)} =$ **1. 5**
We have to find the area to the right of z=1.5 under the standard normal curve.
From the Normal Curve Areas Table , the area to the left of z=1.5 is 0.9332
therefore the area to the right of z=1.5 is $1 - .9332 = 0.0668$
About 6.68% of the flights will need extra fuelling.

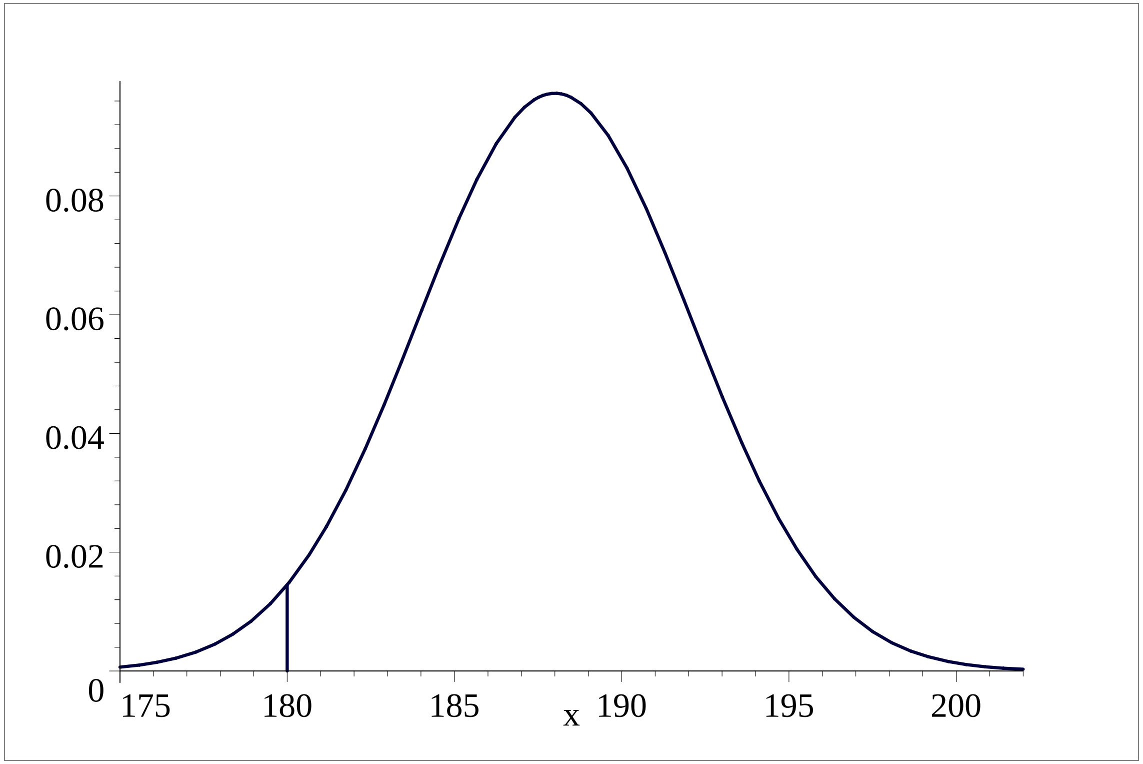
Example 5:

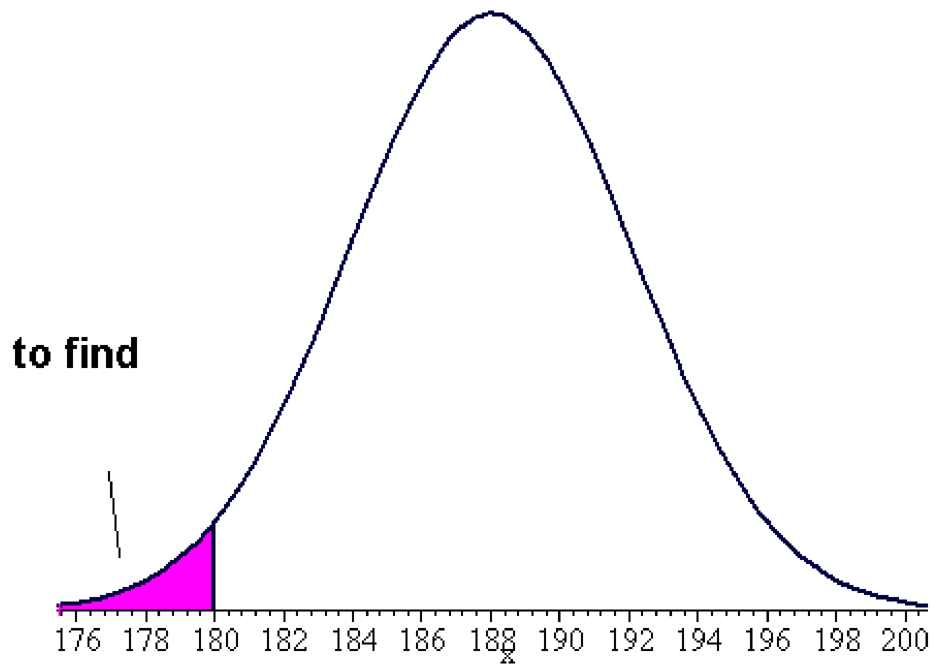
The time that it takes an agent of a certain airline to book the luggage and issue the boarding pass to passengers shows a mean of 188 seconds and standard deviation of 41 seconds. Find the probability that the average time for 100 passengers (treat as a simple random sample from the population of all the passengers) will be less than 180 seconds.

n=100, to state the distribution of $\overline{x}$
For $\overline{x}$**, the mean 188 seconds, st dev** $\frac{41}{\sqrt{100}} = 4.1$ **seconds**
the distribution of

$\overline{x}$ **is approximately normal because 100 passengers is a large sample size**

mean of $\overline{x}$ **s is 188 seconds and st dev 4.1 seconds**





Probability that
$\overline{x} <$ **180**

is the same as that of
z $< \frac{180-188}{4.1} = -1.95121951$

From the z-table

$$\begin{array}{ccc} & & .05 \\ & & \downarrow \\ -1.9 & \rightarrow & .0256 \end{array}$$

The probability is 0.0256

Example 6:

Given that the weights of packages of an expensive compound shows approximately a normal distribution with mean 32 oz and standard deviation 2 oz.

a) Find the probability that the weight of an individual package will be less than 28 oz.

b) Find the probability that the mean weight $\overline{x}$ of randomly selected 10 packeges will be less than 28 oz.

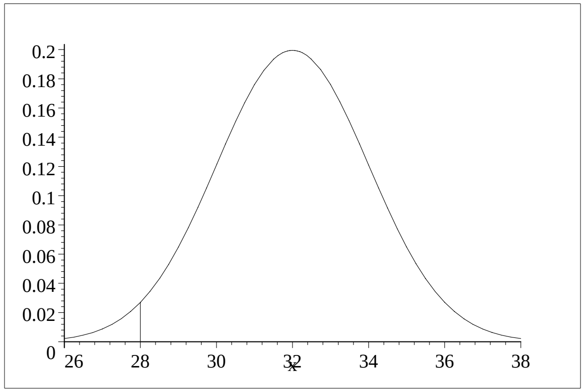
c) Find the level L such that the probability that the mean weight $\overline{x}$ of randomly selected 10 packeges will be less than L is only 0.01.

$$\mathbf{p}(\mathbf{x}) = \frac{1}{2\sqrt{2\pi}}\mathbf{e}^{-\left((x-32)^2/(2\times 2^2)\right)}$$

$$\mathbf{a}(\mathbf{x}) = \frac{1}{\left(2/\sqrt{10}\right)\sqrt{2\pi}}\mathbf{e}^{-\left((x-32)^2/\left(2\times\left(2/\sqrt{10}\right)^2\right)\right)}$$

a)

$p(x)$



The probability that

x $<$ **28**

z $< \frac{28-32}{2} = -2.0$

using the z-table, determine that

The probability is 0.0228.

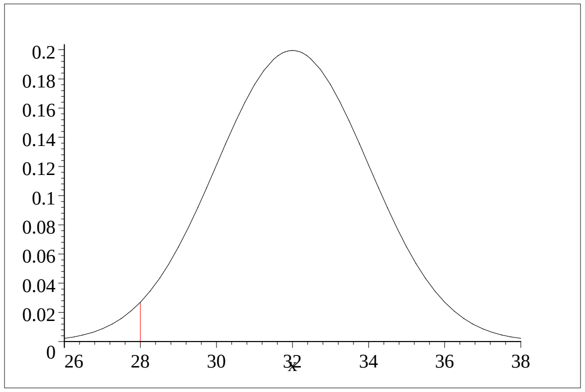
chances are 2.28% that an individual package will weigh less than 28 oz.

b)

Since x is normal,

$\overline{x}$ **has a normal distribution with mean 32 oz and st dev** $\frac{2}{\sqrt{10}} = 0.632455532$ **oz**

Recall that $\frac{2}{\sqrt{10}}$ **is the standard deviation of sample means of simple random samples of size 10**



looks like the probability that $\overline{x} < 28$ oz is approximately 0.

Still, let us see

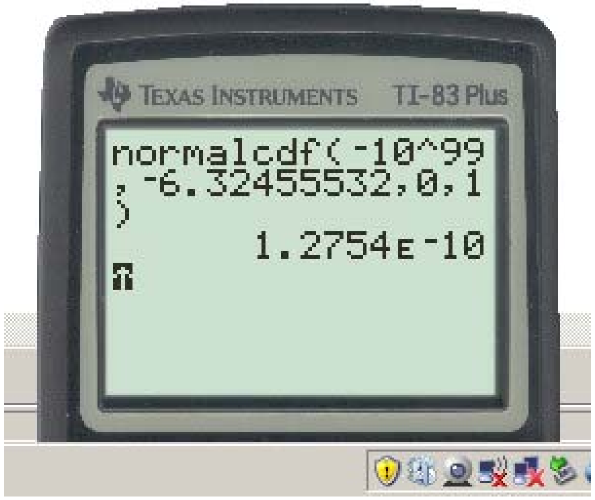
$\overline{x} <$ **28**

z $< \dfrac{28-32}{0.632455532} =$ **-6.32455532**

FYI

$$\int_{-\infty}^{-6.32455532} \frac{1}{\sqrt{2\pi}} \mathbf{e}^{-z^2/2} \mathbf{dz} = \mathbf{1.27} \times \mathbf{10}^{-10}$$

From TI83, you may note that

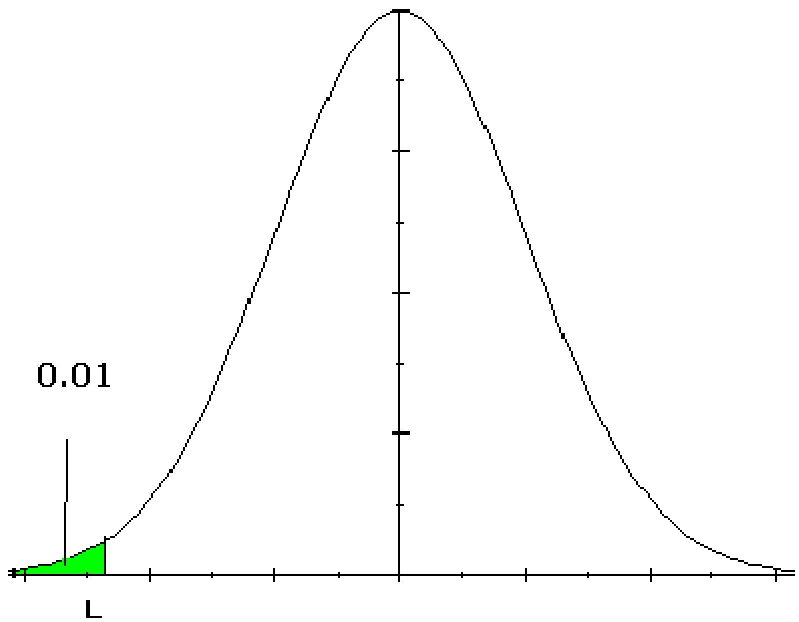


c)

recall:

Find the level L such that the probability that the mean weight $\overline{x}$ of randomly selected 10 packeges will be less than L is only 0.01.

To find z such that the area under the z-curve to the left of z is 0.01



$$\begin{array}{ccc} & & .03 \\ & & \uparrow \\ -2.3 & \leftarrow & .0099 \end{array}$$

since, we have to look at the left tail,

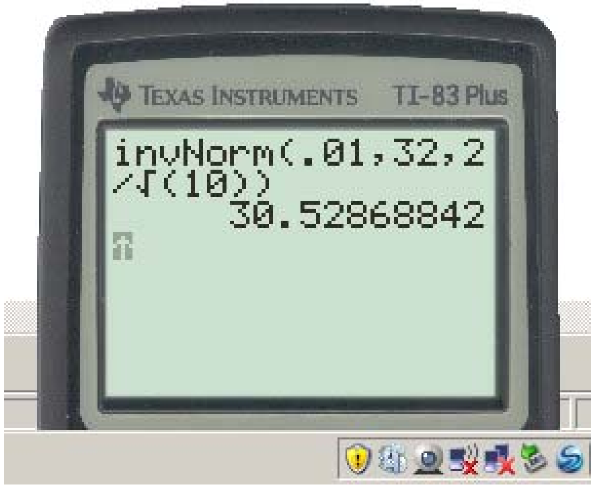
$z = -2.33$

$\dfrac{L-32}{0.632455532} = -2.33$

$\to L - 32 = -2.33 \times 0.632455532$

$\to L = 32 - 2.33 \times 0.632455532 = 30.5263786$ **oz**

OR from TI83 plus



The distribution of Sample Proportions:

Distribution of Sample Proportions:

Let me discuss my Favorite Case for Proportions.

First, remember that when you express the percentage of an attribute as a number, it is called a proportion.

For example:

1.
If 38% of all the households in a region have cable, the proportion of people in the region who have cable is

$\frac{38}{100} = 0.38$

2.

If 217 out of 499 people surveyed say that they would like to see another shopping mall in their city, then the proportion of the people who would like to see another shopping mall is $\frac{217}{499}$

that may be approximated by decimal in the following manner

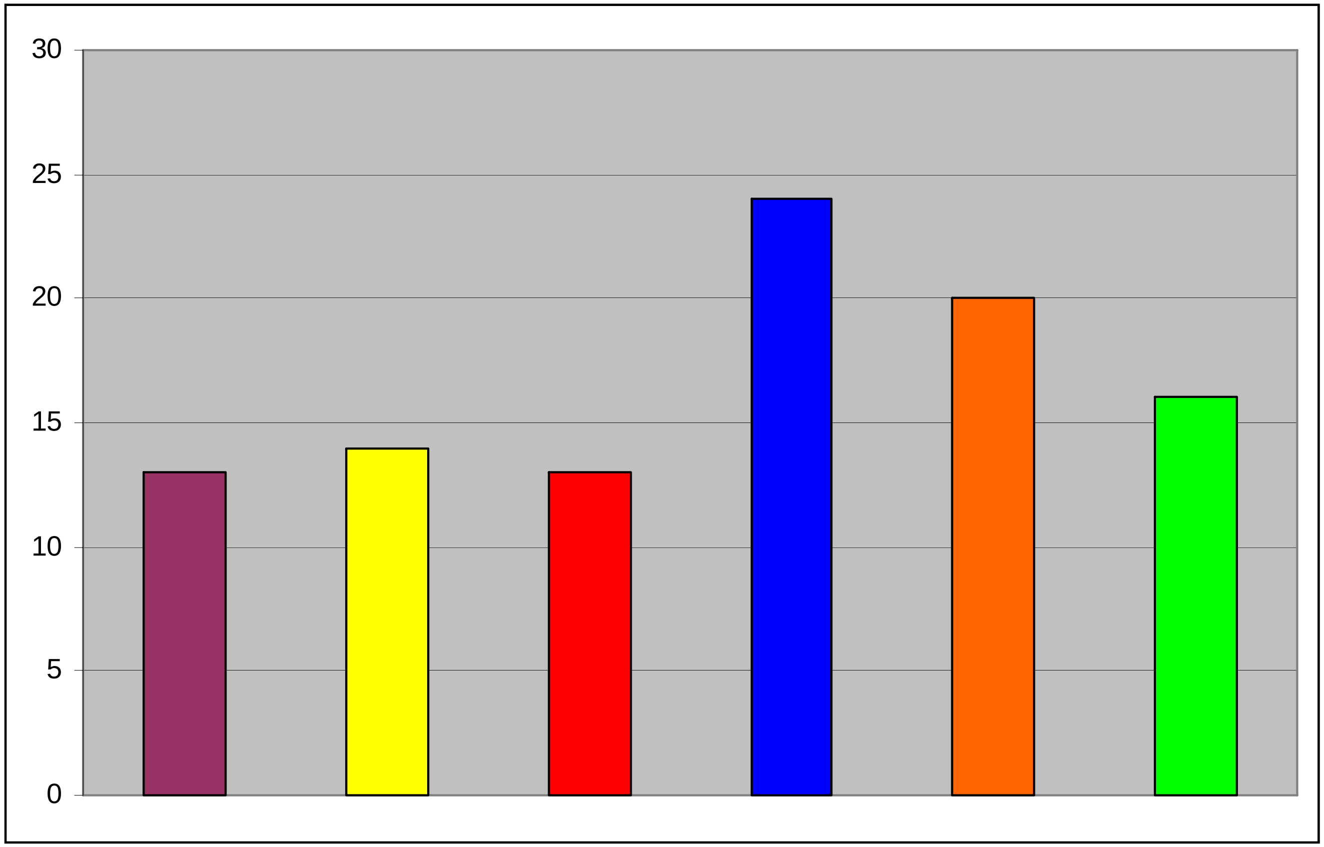
$\frac{217}{499} = \mathbf{0.43486973947895791583}$

Let us work with a basic example which will help us understand bigger cases later.

Consider the following bag of candies (picture from: http://us.mms.com/us/about/products/milkchocolate/)

The companies shows the following picture regarding the percentage of different colors

I used Excel to make the following bar graphs



According to the Company:

COLOR	PROPORTION
Brown	0.13
Yellow	0.14
Red	0.13
Blue	0.24
Orange	0.20
Green	0.16

Look at one of the colors mentioned above, let us say ”orange”

The proportion of the orange color in the population is 0.20.

By population, what we mean here is the collection of all the M&Ms

from which the bags are filled.

Assume that the bags are randomly filled from the above population

so that we can treat these bags are random samples from the above

population.

I used Excel (remember that Excel has a lot of limitation for Statistical work)

to generate 1000 samples of size 100 each from a population with 20% orange

(strictly speaking 20% of the values in this population are 1s and rest 0s.)

and here are a few values of the sample proportions,

The first few values are hown below

proportion of orange

0.24

0.12

0.25

0.19

0.21

0.18

0.21

0.21

0.27

0.23

0.21

0.28

0.23

Notation:

The population proportion is denoted by p:

Example of population proportion:

For the distribution of proportion of orange (may treat ”having an orange color” being ”an attribute” or a ”charactersic of interest” or ”success”) for all the candies, $p = 0.20$

The sample proportion is denoted by $\hat{p}$ **or some times by** p_s

Example of sample proportions:
For the values shown above for the proportion of orange in various bags

proportion of orange in the samples=$\hat{p}$

0.24

0.12

0.25

0.19

0.21

0.18

0.21

0.21

0.27

0.23

0.21

0.28

0.23

Theoretically:

If the population proportion of ”success” is p,
then for sample random samples of size n, from such a population, the sample proprtion of this ”success” $\hat{p}$ varies with mean p and the standard deviation $\sqrt{\frac{p(1-p)}{n}}$
provided that $np \geq 10$ and also $n(1-p) \geq 10$

In addition, for large n,

$\hat{p}$ shows approximately a normal distribution.

Example 1:

An ecommerce company determines that 12% of the people who log on to their website, actually make a purchase.
If 500 people are expected to visit the website on a specific day, find the probbaility that

a) less than 50 of these 500 will actually make a purchase.

b) less than 70 will actually make a purchase.

c) the number of people actually making a purchase will be between 50 and 70.

If we term ”actually making a purchase” as our ”sucess” or ”characterstic of interest” or ”attribute of interest” then the population proportion is $p = 0.12$

If we treat 500 as a simple random sample from a very large population of all the people who could visit this website, then we would like the probability that

$\widehat{p} < \frac{50}{500}$

or

$\widehat{p} < .10$

Note that the mean of the distribution (approximately normal in this case because of large sample size) is

$\mathbf{p} = \mathbf{0.12}$

and standard deviation is

$\sqrt{\frac{p(1-p)}{n}}$

$\sqrt{\frac{.12(1-.12)}{500}} = \mathbf{1.4532721699667959994 \times 10^{-2} \cong 0.014533}$

$\widehat{p} < .10$

is the probability that

$\mathbf{z} < \frac{.10-.12}{\sqrt{\frac{.12(1-.12)}{500}}} = -\mathbf{1.376204706407950757}$

or

$z < -1.38$ **(the standard normal curve areas table goes only upto 2 digits after the decimal for the z-values)**

From the standard normal curve areas table, we have

$$\begin{array}{ccc} & & .08 \\ & & \downarrow \\ -1.3 & \rightarrow & 0.0838 \end{array}$$

therefore the probability is approximately 0.0838 that less than 50 of 500 will actually make a purchase.

b) $\hat{p} < \frac{70}{500}$

$\hat{p} <$ **0.12**

$$z < \frac{.14-.12}{\sqrt{\frac{.12(1-.12)}{500}}} = 1.376204706407950757$$

$$\begin{array}{ccc} & & .08 \\ & & \downarrow \\ 1.3 & \rightarrow & 0.9162 \end{array}$$

therefore the probability that less than 60 of these 500 will actually make a purchase is approximately
0.9162

c)
Now we have to find the probability that

the number of people actually making a purchase will be between

50 and 70

If we include both 50 and 70,

then we have to find the probability of

$$\frac{50}{500} \le \hat{p} \le \frac{70}{500}$$
$$.10 \le \hat{p} \le .14$$
$$\frac{.10-.12}{\sqrt{\frac{.12(1-.12)}{500}}} \le \hat{p} \le \frac{.14-.12}{\sqrt{\frac{.12(1-.12)}{500}}}$$
$$-1.38 \le z \le 1.38$$

Since z varies continuously we can still use the same values from the table as in parts a) and b)

to get the answer:
$0.9162 - 0.0838 = 0.8324$

as the probability that between 50 and 70 will actually make a purchase.* (Please read the note at the end of this lesson)

Example 2:

Given that 57% of the people in a certain region are vegetarians. Jacob has to send lunch packages for 400 people from this region. He has sent 150 vegetarian packages. If we treat these 400 people as a simple random sample, find the probability that

a) less than 62% of these 400 will be vegetarians.
b) more than 62% of these will be vegetarians.

Solution:

a)

$\hat{p} < 0.62$

$z < \frac{.62-.57}{\sqrt{\frac{.57(1-.57)}{400}}}$

$z < 2.02$

$$\begin{array}{lccc} & .00 & .01 & .02 \\ & & & \downarrow \\ 2.0 & & \rightarrow & 0.9783 \end{array}$$

The probability that less than 62% of these 400 will be vegetarians is 0.9783

b) more than 62% of these will be vegetarians will be (because z has continuous distribution)* (Please read the note at the end of this lesson)

is $1 - 0.9783 = 0.0217$

Note 1:

Let us look at the Example 1 above (The ecommerce example)

Now that we have many computing packages, we do not really have to make the approximations

Look at part a) of the example 1

less than 50 is 0 to 49

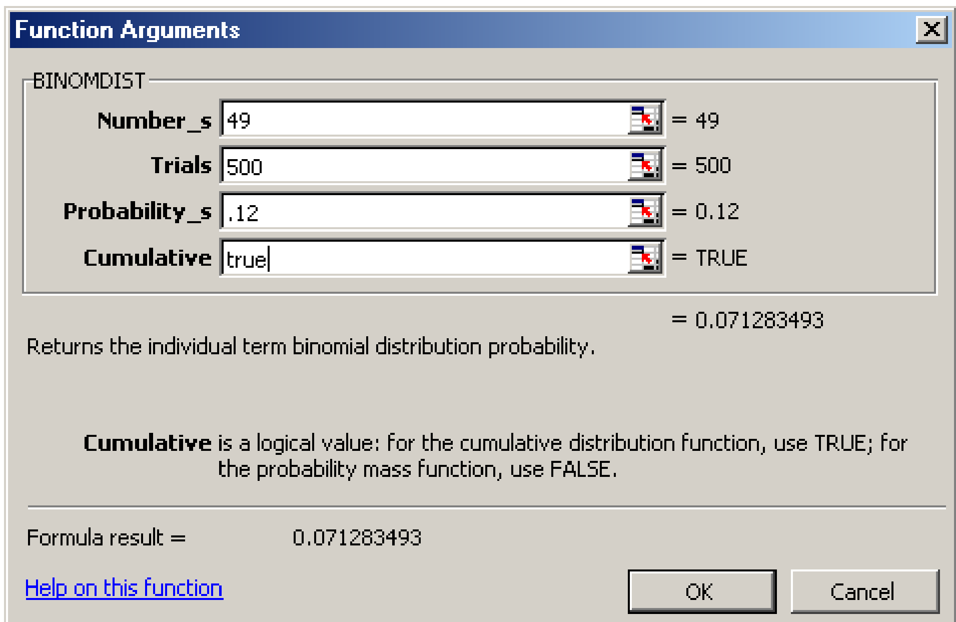
If we use the binomial distribution formula , we can obtain

$$\mathbf{C}(\mathbf{n},\mathbf{x}) = \frac{n!}{x!(n-x)!}$$

$$\sum_{x=0}^{49} \mathbf{C}(\mathbf{500},\mathbf{x})(.12)^x(1-.12)^{500-x} = \mathbf{7.1283492833083139669} \times \mathbf{10}^{-2} \cong \mathbf{0.0713}$$

the answer is quite different beacuse above, we made too many approximations

Excel



TI83plus:

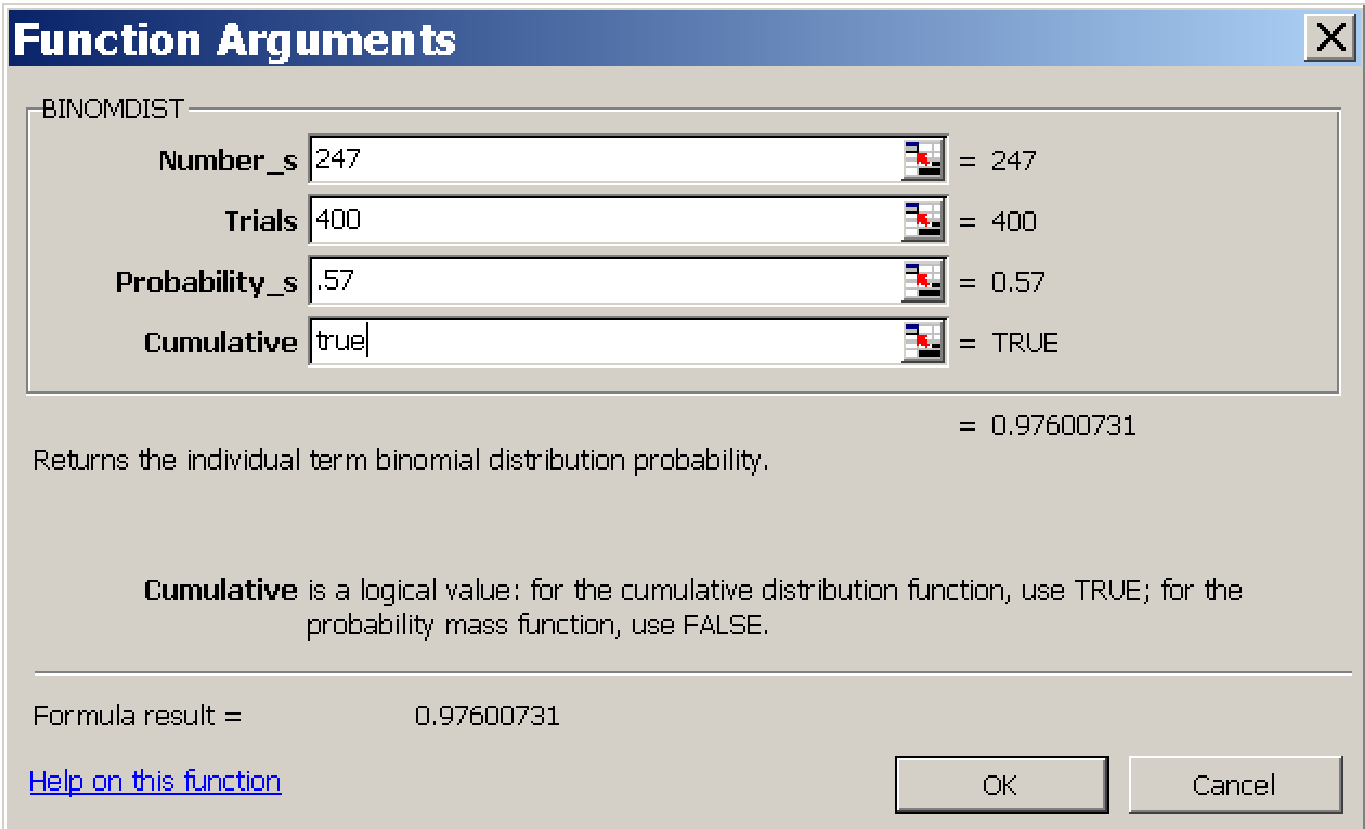
Example 2)

catering

$.62 \times 400 = 248.0$

less than 248 means 247

$$\sum_{x=0}^{247} C(400,x)(.57)^x(1-.57)^{400-x} = 0.97600731049297430672$$



Approximation is better here because p is closer to 0.50