Section 7: The Normal Distribution and Sampling Distributions

The Normal Distribution

The normal distribution (also called Gaussian distribution) is a common type of continuous probability distribution. It is a bell shaped curve.

- The bell is symmetric about the mean of the random variable, μ.
- The standard deviation of the random variable, σ, measures the spread of the bell. The larger σ is, the more spread out the bell.

For the normal distribution the mean, median, and mode are all equal. This should make since because the distribution is symmetric with the point of symmetry being at the peak of the distribution.

The density function for the normal distribution is as follows:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-(x-\mu)^2/(2\sigma^2)}$$

As can be seen in the density function, the values for μ and σ determine which normal distribution we are dealing with. The values μ and σ are the parameters of the normal distribution just like a and b are the parameters of the uniform distribution. There are an infinite number of normal distributions since μ can be any real number and σ can be any positive real number (σ cannot be 0 because all the points would have to be the same which would not give a bell shaped curve). The mathematical ideas of a continuous probability distribution introduced in the last section with the uniform distribution work the same with the normal distribution. However, the density function for the normal distribution is more complicated so we will not deal with all of the math. We will focus on probability and will use a table to help us find the area under the curve. You will not need to know the density function for the normal distribution.

We will start by concentrating on the normal distribution with $\mu = 0$ and $\sigma = 1$. This distribution is called the **standard normal distribution**, denoted by *z* (*z* is used to calculate probabilities for all normal distributions). It is important to distinguish when we are talking about the normal distribution and when we are talking about the standard normal distribution. The normal distribution is any bell shaped curve while the standard normal distribution specifically has a mean of 0 and standard deviation of 1. In addition, when talking about most variables in statistics, they are denoted by *x*. However, some common distributions use different notations. The standard normal is denoted by *z*. So any time you see *z*, you must know that you are dealing with the standard normal distribution.

To calculate probabilities for the standard normal distribution, you will need to use the Standard Normal Distribution table. This table is located on the STA 296 website in the 'tables' link. You should print this table and keep it readily accessible.

When finding probabilities for any normal distribution, I highly recommend that you draw a picture of the desired area!!! You will notice with all the examples I do, there is a picture. Some are simple to see, but this is a good habit to get into because we will continue to draw these pictures and they will get more complex as we move through the material.

Look at your printed Standard Normal Distribution table. The table gives the area between 0 and the *z* value you look up. In other words, the table gives $P(0 \leq Z \leq z)$.

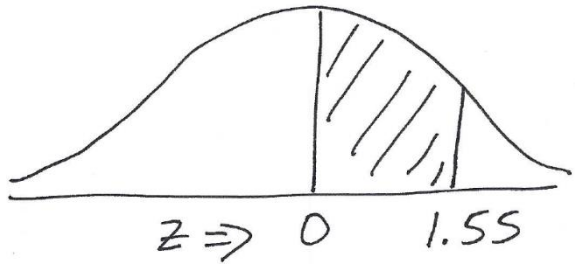
Examples

a. $P(0 \leq Z \leq 1.55)$
b. $P(0 \leq Z \leq 1.96)$

Solutions

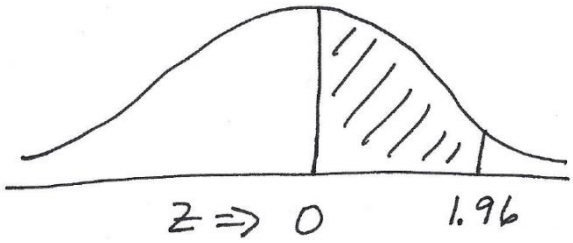
a. $P(0 \leq Z \leq 1.55)$
= 0.4394

All you have to do when finding the area between 0 and a *z* value is look up the *z*. In this case 1.55. So look up 1.5 in the first column and then look up .05 in the first row.



b. $P(0 \leq Z \leq 1.96)$
=0.4750

Again, look up 1.9 in the first column and .06 in the first row. These meet at 0.4750 which is the area/probability.



Combining the facts below with the table allows us to compute any probability for the standard normal distribution.

Facts:

- The total area under the standard normal curve is 1 (this is true for any density function as discussed in the last section of notes)
- The standard normal curve is symmetric about 0 (remember 0 is the mean of the standard normal distribution)
- $P(Z < 0) = 0.5$ and $P(Z > 0) = 0.5$ (This comes directly from the two facts above. If the total area is 1 and the distribution is symmetric about 0, then half of the area must be on each side of 0.)

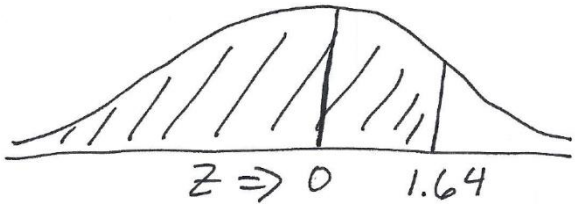
Examples

a. $P(Z \leq 1.64)$
b. $P(Z \geq 1.64)$
c. $P(Z \leq -1.64)$
d. $P(-2.32 \leq Z \leq 0)$
e. $P(-2 \leq Z \leq 2)$
f. $P(1.41 \leq Z \leq 2.18)$
g. $P(Z \geq 12.31)$

Solutions

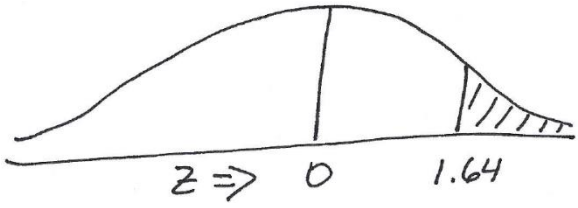
a. $P(Z \leq 1.64)$
= 0. 5 + .4495
= 0.9495

Take the area to the left of 0 (which is 0.5) and add the area between 0 and 1.64 (from table).



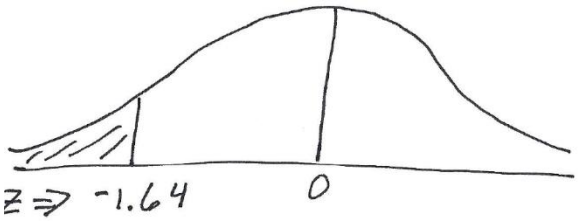
b. $P(Z \geq 1.64)$
= 0.5 - .4495
= 0.0505

Take the area to the right of 0 and subtract the area we don't want between 0 and 1.64.



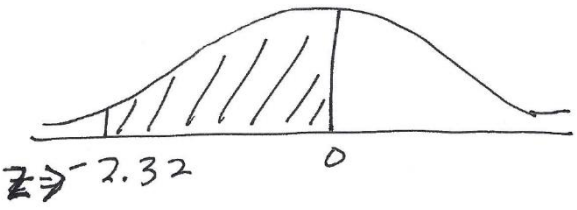
c. $P(Z \leq -1.64)$
= 0.5 - .4495
= 0.0505

Compare this one to part b. By symmetry, you can see that it must be the same.



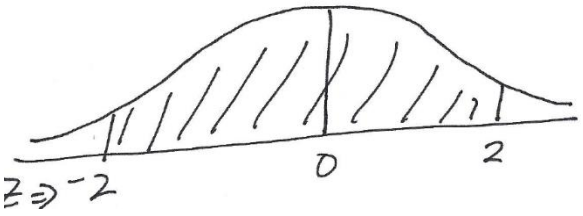
d. $P(-2.32 \leq Z \leq 0)$
= 0.4898

When dealing with values below 0, you must use the fact that the distribution is symmetric. The area between 0 and -2.32 is the same as the area between 0 and 2.32.



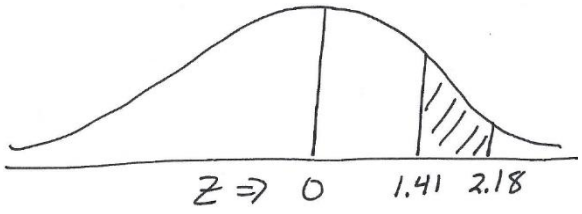
Solutions Continued

e. $P(-2 \leq Z \leq 2)$
= 2(.4772)
= 0.9544



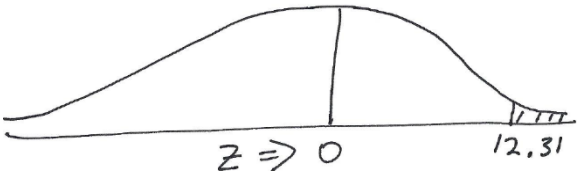
Here the area between 0 and -2 and the area between 0 and 2 is the same. We can just look this up and double it. This solution should make since because the Empirical Rule says approximately 95% of data is within two standard deviations. The Empirical Rule percentages actually come from the normal distribution.

f. $P(1.41 \leq Z \leq 2.18)$
= .4854 - .4207
= 0.0647



Here we want the area between two values above 0. We take the area between 0 and 2.18 and then subtract the area we don't want between 0 and 1.41.

g. $P(Z \geq 12.31)$
= 0



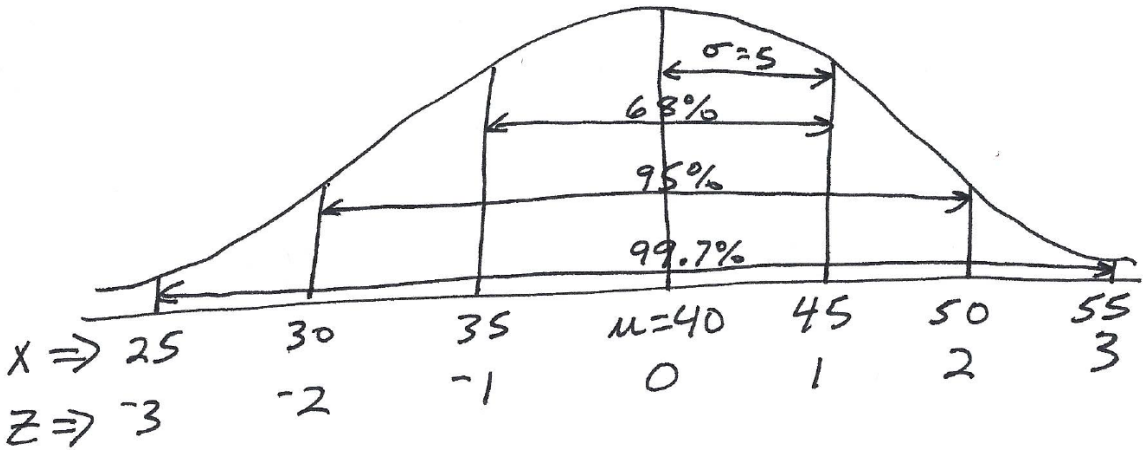
Sometimes we will get Z values that are off the chart and you must know what to do with these. Notice the chart only goes to 3.9, but at this value the area is 0.5. This is not really all of the area to the right of 0, but it does round to 0.5. Therefore, the area between 0 and any value greater than 3.9 is 0.5 when rounded to four decimal places. So here that leaves no area above 12.31. It is almost impossible to get a Z value of 12.31.

We have learned how to compute probabilities for the standard normal distribution. We will now compute probabilities for any normal distribution.

First, we need to look at the z-score. Descriptively, the z-score (standardized score) gives the number of standard deviations an observation is above or below the mean. More importantly, it allows us to transform any normal distribution into a standard normal distribution.

If x is normal with mean μ and standard deviation σ, then $Z = \frac{x-\mu}{\sigma}$ is standard normal.

Illustration of transforming and normal distribution to a standard normal distribution
Suppose we have a normal distribution with $\mu = 40$ and $\sigma = 5$.



The picture above began with a drawing of a normal distribution, *x*. The mean of 40 was identified as the point of symmetry. Next, the standard deviation of 5 was used as our unit of measure (very common in upcoming material). Adding and subtracting a standard deviation we get the 35 and 45 values. By the empirical rule we know about 68% of the *x* values are within this interval. Also, by the empirical rule we know about 95% is between 30 and 50 and about 99.7% is between 25 and 55. This gives a good picture of what the random variable *x* looks like. The final step is to transform the *x* values into *z* values. Notice we are not changing the picture. It is still the same bell shaped curve with the axis being rescaled. You can plug all the *x* values in the *z* formula and you will see they are all correct. You can also just think about the meaning of the *z* values. The *x* value of 40 is the mean so it is 0 standard deviations from the mean, $z = 0$. The value of 45 is 1 standard deviation above the mean, 40 + 5. Each *z* value can be thought of in this same way. The key here is notice when looking at *x*, the mean is 40 and the standard deviation is 5. Now look at *z*. You should see that the mean is 0 and the standard deviation is 1. This means it is a standard normal distribution. Any normal distribution can be transformed to standard normal in this same way. Now, if we want to find the area between 40 and 45 for *x*, it is the same as the area between 0 and 1 for *z*. All normal distributions are the same shape so the area is distributed in the same way.

To calculate probabilities for a normal distribution, *x*

- Write probability statement (and draw picture) for *x*
- Rewrite in terms of *z*

Example

The distribution of IQ scores for the general population is approximately normal with $\mu = 100$ and $\sigma = 10$. The random variable of interest, *x*, is the IQ score of a randomly selected person. Find the following probabilities.

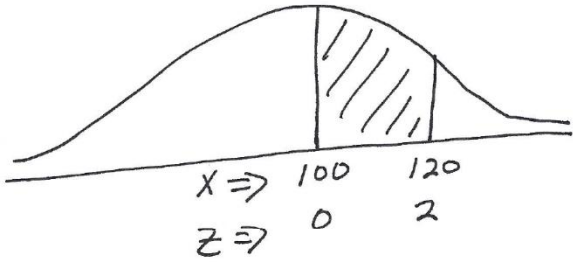
a. $P(100 \leq x \leq 120)$

b. $P(x \geq 130)$

Solution

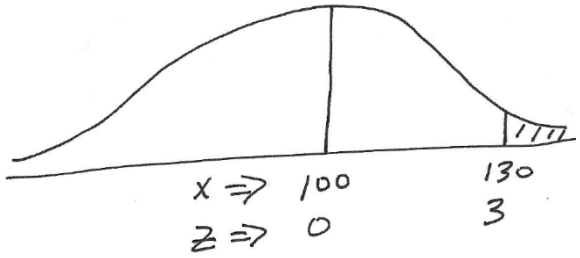
a. $P(100 \leq x \leq 120)$
$= P(0 \leq z \leq 2)$
$= 0.4772$

The first step is to convert the *x* values into *z* values using the *z* transformation formula. Then you use the table just as before.



b. $P(x \geq 130)$
$= P(z \geq 3)$
$= 0.5 - .4987$
$= 0.0013$

Again, transform *x* to *z* and then use the table to calculate the probability.

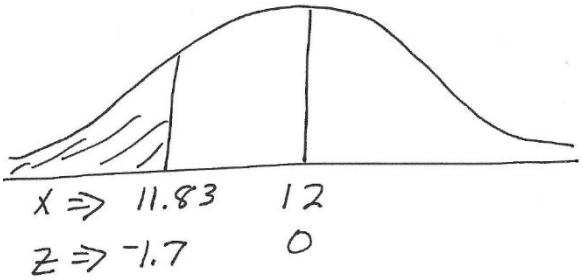


Example

Suppose the amount of Pepsi in a "12 oz" can has a normal distribution with $\mu = 12$ and $\sigma = 0.1$. The random variable of interest, *x*, is the amount of Pepsi in a randomly selected can. Find $P(x \leq 11.83)$.

Solution

$P(x \leq 11.83)$
$= P(z \leq -1.7)$
$= 0.5 - .4554$
$= 0.0446$



The Concept of a Sampling Distribution

Review of important terms

- population – the set of all possible measurements (group of interest)
- sample – a subset of the population
- parameter – a numerical characteristic of a population
- statistic – a numerical characteristic of a sample

Since it is typically unrealistic to do a census, we use a statistic to estimate a parameter. In statistical application, we take a random sample from the population. We compute a statistic, say $\bar{x}$ (the sample mean). The value of the statistic $\bar{x}$ depends on which items are randomly selected for the sample. Different samples yield different values of $\bar{x}$ (we randomly get a value of $\bar{x}$ out of all its possible values). Therefore, $\bar{x}$ is a random variable. Taking a random sample and calculating $\bar{x}$ is equivalent to randomly selecting a value for $\bar{x}$ out of all its possible values.

The sample mean, $\bar{x}$ estimates the population mean μ. We want $\bar{x}$ to be very close to μ (cannot expect $\bar{x}$ to be exactly equal to μ). In order to evaluate how good $\bar{x}$ is at estimating μ, we need to know the probability distribution of $\bar{x}$.

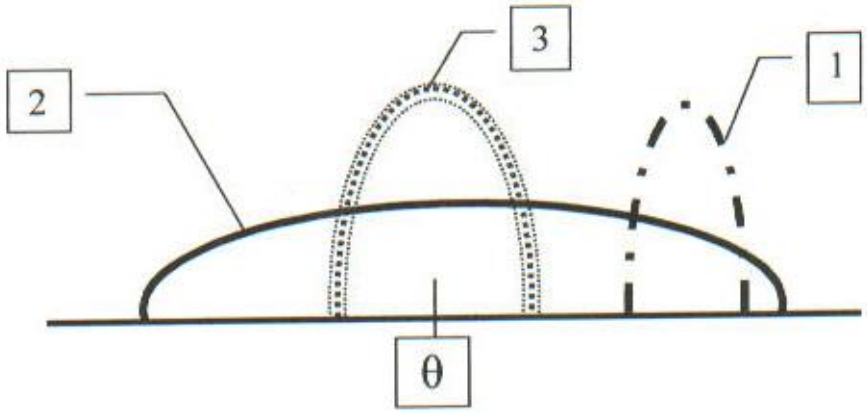
The distance between a statistic and the parameter it is estimating is called **sampling error**.

The probability of a statistic over all possible samples is known as its **sampling distribution**. (This is the same as a probability distribution except the random variable is a statistic rather than an observation.)

In order to be confident that our statistic will give a good estimate of the parameter, we need the sampling distribution to have two characteristics.

a. sampling distribution should be centered at the value of the parameter (means statistic is unbiased)
b. sampling distribution should have a small amount of variation (means there is a small amount of error in the statistic)

We want the sampling distribution to be centered at the value of the parameter and to have little variation. Consider the graph below. In the graph there are three statistics that could be used to estimate the parameter, θ. Which one is best?



Statistic 3 is the best choice in the graph above. The worst choice is statistic 1 which tends to always overestimate the parameter (the center of the distribution is right of the parameter). This is what a sampling distribution could look like for a biased statistic. The way we prevent a statistic from being biased is to take a random sample. A random sample will give a sampling distribution centered over the parameter of interest. When looking at statistic 2 you see there is a lot of variability. This is not good because the spread means there is a reasonable chance of getting a value far from the parameter. Notice with statistic 3 most of the outcomes are very close to the parameter we are estimating. This means that when using the statistic to estimate the parameter you can expect a small amount of error in the estimate. Hopefully, you remember that we can decrease this error by increasing the sample size. How this really works mathematically is that increasing your sample size decreases the variability in the sampling distribution of the statistic.

The statistic is an **unbiased estimator** if it is centered about the parameter of interest. Otherwise, the estimator has bias.

A statistic is a **minimum variance unbiased estimator** if it is an unbiased estimator and has less variance than all other unbiased estimators.

<u>Facts about the Sampling Distribution of $\bar{x}$</u>
The statistic $\bar{x}$ (sample mean) estimates μ (population mean). It should be noted that in this class we will only look at good statistics. This means that with proper techniques the sampling distribution will be centered at the parameter and will have a small amount of variability. With the knowledge that these characteristics hold when looking at the sampling distribution for $\bar{x}$, we can understand some facts.

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

The average value of $\bar{x}$ across all possible samples is μ, the population mean. This should make since because we just said the sampling distribution of $\bar{x}$ is centered about the parameter it estimates which is μ. (The mean of any good statistic is the parameter it estimates)

$\sigma_{\bar{x}}{}^2 = \frac{\sigma^2}{n}$

The variance of the sampling distribution of $\bar{x}$ is the population variance σ^2 divided by the sample size *n*.

$\sigma_{\bar{x}} = \sqrt{\sigma_{\bar{x}}{}^2} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$

The standard deviation of the sampling distribution of $\bar{x}$ is the population standard deviation σ divided by $\sqrt{n}$. Notice that as *n* increases the sample to sample variability in $\bar{x}$ decreases (if you increase the denominator, the overall quantity decreases). Therefore, as has been stated previously, increasing sample size decreases the error when using a statistic to estimate a parameter. Also notice that a larger σ results in a larger $\sigma_{\bar{x}}$. More variability in your data will mean more variability in your statistic.

Standard Error – is the standard deviation of the sampling distribution
($\frac{\sigma}{\sqrt{n}}$ is the standard error of $\bar{x}$)

<u>Illustration of the mean and standard deviation of a sampling distribution</u>
Suppose we have a population *x* that consists of the values: 1, 2, 3, 4. A sample of size 2 is taken. This is not realistic of a true population but illustrates the mathematics.

Population, *x*	Possible Samples When $n = 2$	Probability	$\bar{x}$
1	1, 2	$\frac{1}{6}$	1.5
2	1, 3	$\frac{1}{6}$	2
3	1, 4	$\frac{1}{6}$	2.5
4	2, 3	$\frac{1}{6}$	2.5
	2, 4	$\frac{1}{6}$	3
	3, 4	$\frac{1}{6}$	3.5
$\mu = \frac{10}{4} = 2.5$			$\mu_{\bar{x}} = \frac{15}{6} = 2.5$

Notice that $\mu_{\bar{x}} = \mu = 2.5$.

If we sample without replacement from a finite population containing *N* elements, then we must use the finite population correction factor (FPC) when calculating the variance of the sampling distribution.

$FPC = \left(\frac{N-n}{N-1}\right)$

Therefore, ${\sigma_{\bar{x}}}^2 = \frac{\sigma^2}{n} \cdot \left(\frac{N-n}{N-1}\right)$

You are not required to do this for class, but for this illustration to work we must use the FPC. For class, we will only deal with populations large enough to assume they are infinite. For large populations, you do not need to use the FPC as it will basically be 1.

x	x^2	$P(x)$	$x^2 \cdot P(x)$	$\bar{x}$	$\bar{x}^2$	$P(\bar{x})$	$\bar{x}^2 \cdot P(\bar{x})$
1	1	$\frac{1}{4}$	$\frac{1}{4}$	1.5	2.25	$\frac{1}{6}$	$\frac{2.25}{6}$
2	4	$\frac{1}{4}$	$\frac{4}{4}$	2	4	$\frac{1}{6}$	$\frac{4}{6}$
3	9	$\frac{1}{4}$	$\frac{9}{4}$	2.5	6.25	$\frac{1}{6}$	$\frac{6.25}{6}$
4	16	$\frac{1}{4}$	$\frac{16}{4}$	2.5	6.25	$\frac{1}{6}$	$\frac{6.25}{6}$
			$\sum x^2 \cdot P(x) = \frac{30}{4}$	3	9	$\frac{1}{6}$	$\frac{9}{6}$
$\sigma^2 = \sum x^2 \cdot P(x) - \mu^2$ $= \frac{30}{4} - \left(\frac{10}{4}\right)^2 = \frac{5}{4} = 1.25$				3.5	12.25	$\frac{1}{6}$	$\frac{12.25}{6}$
							$\sum \bar{x}^2 \cdot P(x) = \frac{40}{6}$

${\sigma_{\bar{x}}}^2 = \sum \bar{x}^2 \cdot P(\bar{x}) - \mu^2 = \frac{40}{6} - \left(\frac{10}{4}\right)^2 = \frac{5}{12} = 0.417$

And

${\sigma_{\bar{x}}}^2 = \frac{\sigma^2}{n} \cdot \left(\frac{N-n}{N-1}\right) = \frac{1.25}{2} \cdot \left(\frac{4-2}{4-1}\right) = \frac{5}{12} = 0.417$

$\sigma_{\bar{x}} = \sqrt{0.417} = 0.645$

The sampling distribution of $\bar{x}$ with a sample size of 2 has a mean of 2.5 and a standard deviation of 0.645.

Central Limit Theorem

Remember that if we have a normal distribution and we know the mean and standard deviation, then we can find probability using the transformation $Z = \frac{x-\mu}{\sigma}$. The important part here is that we take an observation and subtract the mean and divide by the standard deviation. We can use the same format for the sampling distribution.

If our sample comes from a <u>normal distribution</u> with mean μ and standard deviation σ then $Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ has a standard normal distribution.

Notice that in the transformation above we know the population is normally distributed. In reality when collecting data it is rare that this would ever be the case. So what can be done when the population is not normally distributed? This has been thoroughly researched by statisticians. The research has led to one of the most important theorems in statistics.

Central Limit Theorem – If we sample from a population with mean μ and standard deviation σ then $Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ is approximately standard normal for large *n*.

Notice that the central limit theorem works when we do not have a normal distribution. Of course the question here becomes how large does *n* have to be in order for the central limit theorem to work? The answer is if $n = 30$ or larger, the central limit theorem will apply in almost all cases. So for class purposes, this is the general rule we will use.

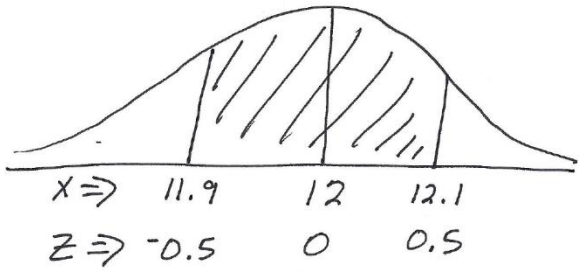
<u>Example</u>
A population of soft drink cans has amounts of liquid following a normal distribution with $\mu = 12$ and $\sigma = 0.2$ oz.

a. What is the probability that a single can, *x*, is between 11.9 and 12.1 oz.?
b. What is the probability that $\bar{x}$ is between 11.9 and 12.1 for $n = 16$ cans?

<u>Solution</u>

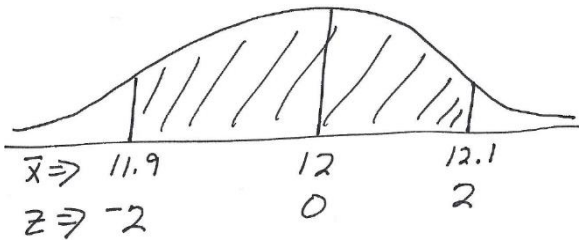
a. $P(11.9 < X < 12.1)$
   $= P(-0.5 < Z < 0.5)$
   $= 2(.1915)$
   $= 0.3830$

   To transform *x* values into *z* value you use the tranformation $Z = \frac{x-\mu}{\sigma}$.



b. $P(11.9 < \bar{x} < 12.1)$
   $= P(-2 < Z < 2)$
   $= 2(.4772)$
   $= 0.9544$

   Notice that in this case we are dealing with $\bar{x}$ instead of *x*. The problem works the same other than our transformation is $Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$.



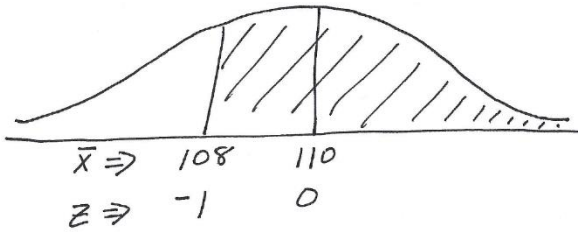
Example

A population of trees have heights that have a mean of 110 feet and a standard deviation of 20 feet. A sample of 100 trees is selected. Find the following.

a. $\mu_{\bar{x}}$
b. $\sigma_{\bar{x}}$
c. $P(\bar{x} > 108)$
d. What about $P(X > 108)$?

Solution

a. $\mu_{\bar{x}} = \mu = 110$
b. $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{100}} = 2$
c. $P(\bar{x} > 108)$
   $= P(Z > -1)$
   $= 0.5 + .3413$
   $= 0.8413$



d. Cannot find the probability because we do not know the distribution of *x*. We know the distribution of $\bar{x}$ is normal because of the central limit theorem which is why we can do part c.

Facts about the Sampling Distribution $\hat{p}$

Population Proportion $= p = \frac{number\ in\ population\ with\ characteristic}{number\ in\ population}$

Sample Proportion $= \hat{p} = \frac{number\ in\ sample\ with\ characteristic}{Number\ in\ Sample}$

$\hat{p}$ is a point estimate of p (just like $\bar{x}$ is a point estimate of μ)

$\mu_{\hat{p}} = p$

The average value of $\hat{p}$ across all possible samples is p, the population proportion. Once again this should make since because the sampling distribution of $\hat{p}$ is centered at the parameter it estimates which is p.

$\sigma_{\hat{p}}^2 = \frac{p \cdot q}{n}$

The variance of $\hat{p}$ is the population variance (which is $p \cdot q$) divided by *n*.

$$\sigma_{\hat{p}} = \sqrt{\frac{p \cdot q}{n}}$$

The standard deviation of $\hat{p}$, also called the standard error, is the square root of the variance of $\hat{p}$.

If we sample from a population with a proportion of p, then $Z = \frac{\hat{p}-p}{\sqrt{\frac{p \cdot q}{n}}}$ is approximately standard normal for large *n*.