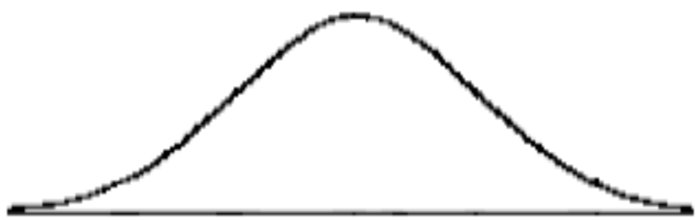
CHAPTER 6.1

The Normal Distribution Curve

Previously we learned that a *normal* histogram is fairly symmetric. Also, the mean, median, and the mode of a normal distribution are just about the same. Let's learn a bit more about this very interesting distribution. First of all, we understood that the shape of a normal histogram was relatively normal. This shape could be compared to a 'Bell Curve' that looks like the one below.



Historical Note of Interest: The Normal Distribution Curve is sometimes referred to as the Gaussian Distribution in honor of the mathematician, Carl Frederick Gauss (1777-1855). He, among several other mathematicians including DeMoivre, contributed to the refinement of the equation used to *generate* the curve above. The equation of this "Normal Distribution Curve" is given below only for ***aesthetic*** purposes. You would only have to learn it if you were in a Calculus class!

$$Y = \frac{1}{\sigma\sqrt{2\pi}} \; e^{-(X-\mu)^2/2\sigma^2}$$

where:

Y = height of curve associated with its corresponding X value
e = logarithmic base
μ = mean of the distribution
σ = standard deviation.

In a **normal distribution** it has been shown that about **68%** of the data will lie within **1** standard deviation of the mean; about **95%** of the data will lie within **2** standard deviations of the mean; and about **99.7%** of the data will lie within 3 standard deviations of the mean. These are called ***Rough and Ready Rules.*** **Summary:**

Rough and Ready rules for a Normal Distribution

1. **68% of the data will lie within one (1) standard deviation of the mean.**
2. **95% of the data will lie within two (2) standard deviations of the mean.**
3. **99.7% of the data will lie within three (3) standard deviations of the mean.**
4. **100% of the data are under the entire curve.**

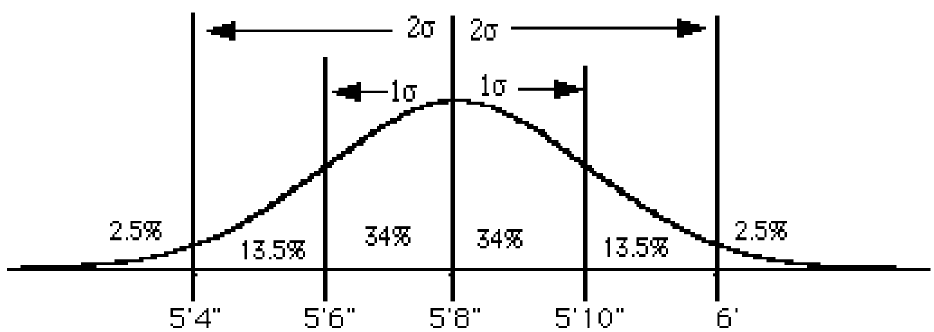
Example: Heights of men (over 25) in USA

What if we wished to know more about the heights of men (say over 25 years old) in the USA. We would need to know information about the relative frequency of the heights of men in the population as a whole. We would take some kind of a survey. (More on this later.)

Let's assume that someone has already taken such a survey and has worked out that the mean height of men in the USA is **5'8"**, and that the standard deviation is **2** inches (don't worry about how he or she arrived at these figures - we'll get to that soon).

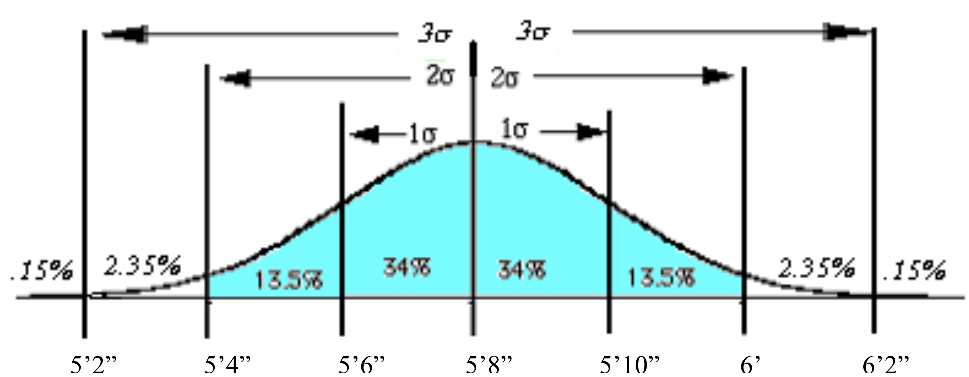
In addition, the distribution of men's heights is ***normal***. We learned that this type of histogram would be fairly symmetric. Also, we learned that the mean, median, and the mode would be just about the same. Thus we have a normal distribution and we can use the normal curve to make calculations.

From our ***rough and ready rules***, we can divide up the curve into sections. Each section is one standard deviation in width. We place the mean of 5'8" on the horizontal axis directly below the middle of the curve. Then we label each standard deviation as shown keeping in mind that we are given the standard deviation as 2". The diagram is only labeled up to two standard deviations here, rather than three. Notice that 68% is within one standard deviation of the mean so each section to the right and left of the mean must be 34%. Similarly, if 95% of the data are within two standard deviations, we must have 47.5% to the right and left of the mean. Thus the section next to 34% must be 13.5%. (34% + 13.5% = 47.5%)



Take, for example, a person who is over 6 feet tall. From the diagram we see that he lies outside **2** standard deviations from the mean. Outside 2 standard deviations is either to the left or right of the mean, that is, in 5% of the data. And since he is in the *upper* half of the **5%**, he is in the top **2.5%** of the data. What does that mean? It says that only 2.5% of males in the US are over 6 feet tall!

We can divide up this last section (2.5%) yet again. This gives us the three standard deviations from the mean. Notice that we now have 99.7% of the data within 3 standard deviations of the mean.

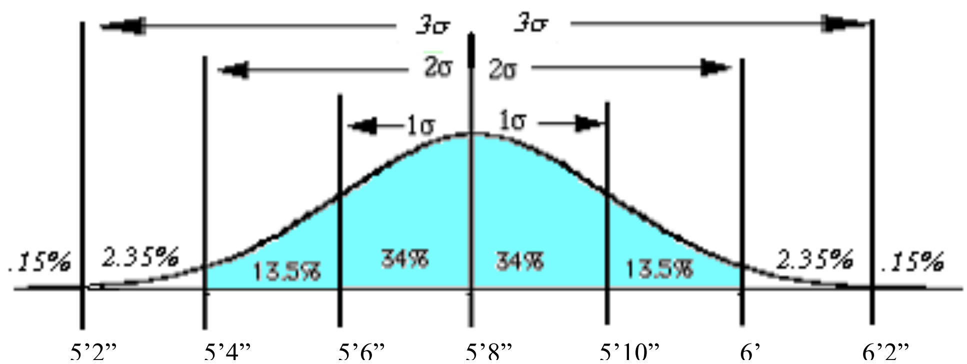


Probability - another type

Interestingly, we can jump from statistics to probability using our normal curve. The normal distribution curve is also known as a *normal probability curve*. With probability, we can predict certain outcomes. For example, if we choose a male (over 25) at random, what is the chance that he will be taller than 6 feet? Looking at the curve we see that there is a **2.5%** chance, or a probability of **0.025**, of finding a male taller than 6 feet, assuming of course that we were looking for him in a *random sample*. (We will talk about random samples a bit later.)

Similarly there would be a **13.5%** chance of finding a man between **5'10"** and **6'** tall. Or a **16%** chance of finding a man who is less than **5'6"** tall.

Question: Find the probability of a male who is relatively short, say less than 5'2". From our rough and ready rules we know that 99.7% of the data are within 3 standard deviations of the mean. Look at a second diagram with three standard deviations marked off.



Answer: Thus, the chance of finding a male over 25 years old who is shorter than 5 feet 2 inches is only **.**15%.

Summary

Because the distribution is **normal**, we can use "rough and ready" percentages to find probabilities.

rough and ready rules:

68% of our data lie within **one** standard deviation of the mean
95% lie within **two** standard deviations of the mean.
99.7% lie within **three** standard deviations of the mean.

Example 2. In a given population of males it is found that the mean weight is 162 lbs. Furthermore, the calculated standard deviation is 19.4 lbs. (We do not need to calculate the standard deviation as it is given to us.) Assume that the distribution of weights among males is normal.

a) What percentage of males have weights between 123.2 lbs and 162 lbs?

b) What percentage of males have weights between 142.6 lbs and 200.8 lbs?

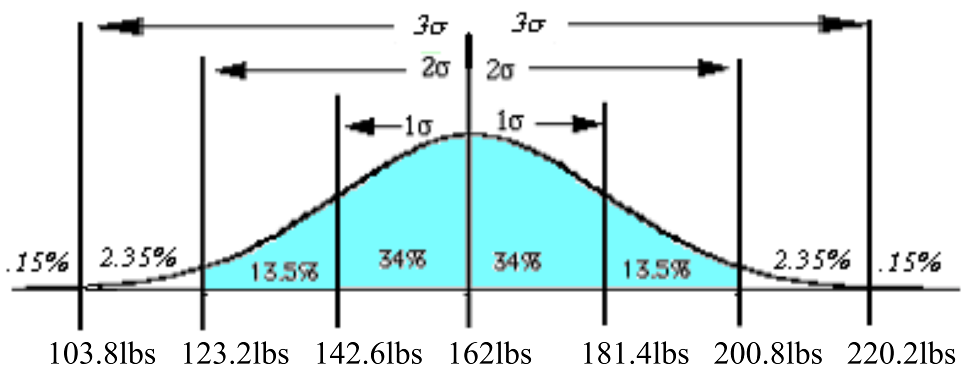
c) What percentage of males have weights over 200.8 lbs?

d) What percentage of males have weights between 103.8 lbs and 142.6 lbs?

Step 1: Draw the normal distribution curve with the rough and ready percentages written INSIDE the curve.

Step 2: On the horizontal axis below the curve, put the mean in the center of the line.

Step 3: Using the mean and the given standard deviation, calculate one standard deviation above the mean simply by adding the standard deviation (in this case 19.4) to the mean. Put this result under the line to the right of the mean. Similarly, subtract the standard deviation from the mean and put that figure on the line to the left of center. Do the same for two standard deviations. Your result looks like the following:



Now we can answer the posed questions:

a) What percentage of males have weights between 123.2 lbs and 162 lbs?

To find the answer, ADD the percentages INSIDE the curve from 123.2 to 162, thus

Answer: 13.5% + 34% = **47.5%**

b) What percentage of males have weights between 142.6 lbs and 200.8 lbs?

Answer: 34% + 34% + 13.5% = **81.5%**

c) What percentage of males have weights over 200.8 lbs?

Answer: 2.5%

d) What percentage of males have weights between 103.8 lbs and 142.6 lbs?

Answer: 15.85%

A New Formula for the Standard Deviation

Now that we have a NORMAL DISTRIBUTION, we can use our rough and ready rules to find percentages and probabilities. But first, we need to calculate the **Standard Deviation**. How would we do that with 'thousands' of sample fractions? Instead of doing it the old way, we now have a simpler formula. This formula can only be used if we are working **with proportions (percents)** rather than with raw data.

The New Formula for Standard Deviation

The formula for calculating the standard deviation of *sample fractions*:

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

p is the mean of the population. (or in this case *hypothetical* or *sample* mean.)

n is the **size** of the sample we are looking at.

Let's calculate the Standard Deviation for the Obama data: I have labeled it as Example 1.

Ex 1: **p = 47**% and our sample size is 100. i.e. **n=100**

Formula: $$\sigma = \sqrt{\frac{p(1-p)}{n}}$$

First turn 47% into a decimal. 47% = 0.47 (Keep at least three decimal places.)

$$\sigma = \sqrt{\frac{.47(1-.47)}{100}}$$

Put (1 – .47) in the calculator first and get (1 – .47) = .53 So.

$$\sigma = \sqrt{\frac{.47(.53)}{100}}$$

$$\sigma = \sqrt{\frac{.2491}{100}} = \sqrt{.002491} = 0.0499099\ldots = \mathbf{4.99099\ldots\ \%}$$

Round: $\sigma = \mathbf{5\%}$

So the standard deviation, $\boldsymbol{\sigma} = \mathbf{5\%}$ In a short time, we will see how to use our standard deviation.

***NOTE:** We always turn our standard deviation into a PERCENT before we do any rounding.

CHAPTER 6.2

Sampling and Statistical Inference

Introduction: The following is an introduction to Sampling, Sample Fractions, and Statistical Inference. This section attempts to give the student an *abbreviated* look at Theoretical Statistics. It also helps to explain some of the more esoteric pieces of statistics and thus help in understanding the next few Chapters. If you wish to learn more about these subjects, think about taking the next level of Statistics, Stats E-50, especially if you will be concentrating in the Social Sciences.

What is a Sample? A sample is a group of observations (or people) chosen from a much larger population; for example, a sample of the heights of male men in the USA. The objective of taking a sample is to utilize the characteristics of that sample and apply those characteristics to the entire population. In the above example, we may want the average height of all male men in the USA. Let's look at another, more timely, example:

> Say you worked for Martha Coakley, the Massachusetts Democratic gubernatorial candidate in 2014. Suppose Coakley wants you to find out her chances of winning the election for Governor. Where would you begin? Would you ask every single voter in Massachusetts if they favored Coakley over Charlie Baker, the Republican candidate? That would be a daunting task to say the least! What about calling all the registered voters in Boston? Would this be a fair group of people to ask? So, how many people would you ask? And whom?
>
> Let's begin by asking say, 100 registered voters randomly chosen from around the state. Then those 100 people are what we call a *random sample*. Should we only ask one group of 100 people? What if we sent out 10 grad students and had each ask 100 people? Then we would have had 10 samples each of size 100. I think you get the idea.

Simple Random Sampling: This is the type of sampling that depends on the type of probability we just studied. For example, we may put names in a hat and choose our sample of names out of the hat. That way, we know that each name in the hat will have an equal chance of being chosen. Often times, researchers will use a computer to generate a list of 'random' numbers. A number is given to every person and those random numbers chosen by the computer become our choices for the sample. Phone books are often used. Simply flipping to a page is not random enough. Again, we might have the computer pick out a random number and that may be the page we use, the column, and the number of names we go down.

Deciding how many people you want in the sample is yet another problem. How big should your sample be? First, it seems reasonable that your sample should be smaller than the population you are studying. Second, it surely depends on your pocketbook. The more people or data that one may wish to have in the study, the more it will cost to collect that data. (Even grad students do not work for free!) Third, the size of the sample will determine how much *error you are willing to accept*. We will see how this works in the next few lessons. There will always be some variation in a sample and errors creep in no matter how good we collect our data. We must always do the best we are able. (Note: Statistical researchers have found that a sample size should be at least 30 observations. This may or may not be enough to give you the results you want, but it is a beginning.)

<u>Statistical Inference:</u> Once we solve this sampling problem, our basic goal is to *infer* the characteristics of the entire population from those of our sample. Thus, we can make *predictions* and *inferences* about the population we are studying as a whole.

Let's go back to Martha Coakley's survey of 100 people and say that 52 out of 100 people said they would vote for her. This is a fraction or a percentage.

$$\frac{52}{100} = 52\%$$

Now let's assume we did have our 10 grad students taking samples of 100 people each. That is, 10 grad students went out and each asked 100 people (randomly selected, of course) whom they would vote for. Here are the results of each of the samples. Notice, we have 10 percentages. These are called the

SAMPLE FRACTIONS

In order to understand this concept better, let's go back a bit and actually look at what happens from the very beginning when we sent out these 10 grad students. Below, you will find the raw data that may be brought back to Coakley's office. Let's follow along what would be done with these data.

Question from Grad Student to a Random passerby: Will you vote for Martha Coakley in the gubernatorial election?
Answers are listed below from each of the 100 people the grad student asked.
Notes: We will assume for the purposes of this exercise that every person asked did give an answer. Hence, the result labeled 'Other.' (e.g. if a respondent answered, "maybe," it was not counted as a yes.)

Grad Student Name:	Results Yes	Other	Percent Yes
Janet	52	48	52%
Jeff	46	54	46%
Steve	56	44	56%
Jane	35	65	35%
Ann	52	48	52%
Dick	47	53	47%
Albert	40	60	40%
Pat	32	68	32%
Joseph	64	36	64%
Mike	46	54	46%

So, as you can see, this very last column is what we call the **"Sample Fractions."**

As we did in Assignment 3, we can find the ***mean* or average *sample fraction*** of people who would vote for Coakley. We add the sample fractions and divide by 10. We get:

Mean = Sum of the data values divided by n. In this case, since we only have 10 fractions, we can easily add these on a calculator and get the average.

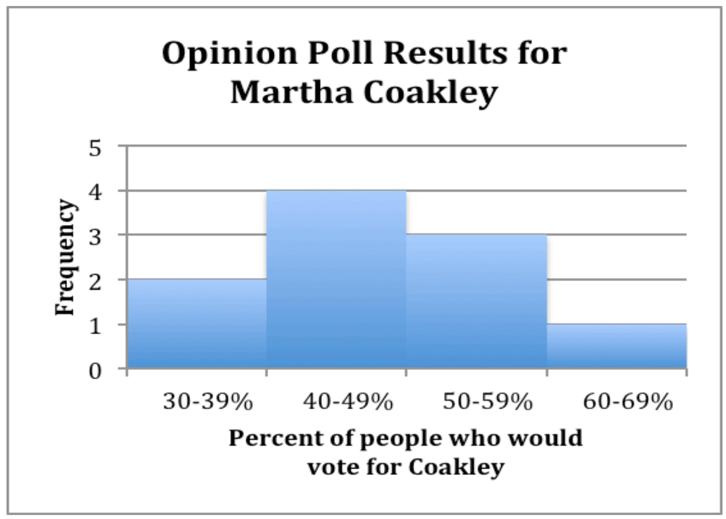
$$\textbf{Mean} = \frac{52 + 46 + 56 + 35 + 52 + 47 + 40 + 32 + 64 + 46}{10} = \mathbf{47\%}$$

If Coakley wished to see a histogram of these Sample Fractions, we can generate one using intervals of 10. We count up the number of times say, 30% - 39% occurs, etc. For example, 2 of the grad students got results of 35% and32% so draw your bar over the 30% up to 2. Here are the data, a frequency table and a histogram similar to one we did in Assignment 3.

Raw Data

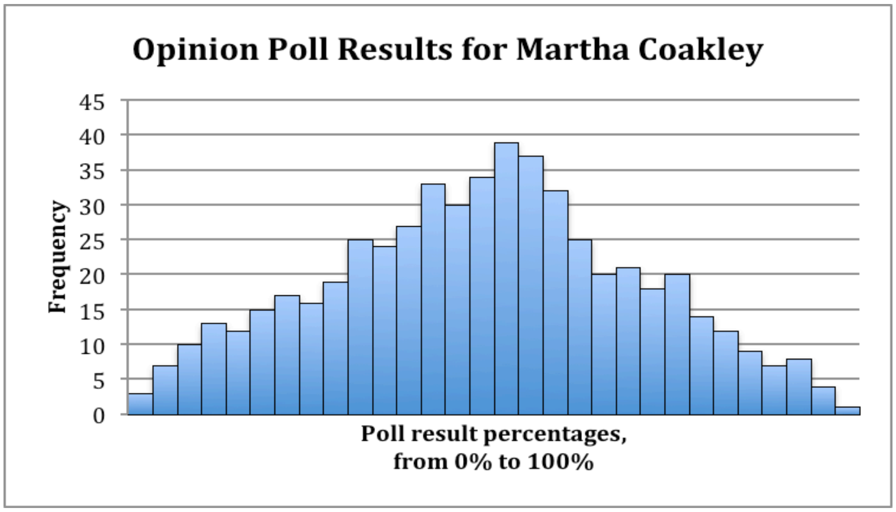
52%
46%
56%
35%
52%
47%
40%
32%
64%
46%

Percent Data	**#Times Appeared Frequency**
30-39%	**2**
40-49%	**4**
50-59%	**3**
60-69%	**1**



THEORETICAL STATISTICS

In theory, we actually *assume* we can take **thousands and thousands** of polls *if* we could. We ask ourselves, "*Hypothetically*, what would happen if we put all these poll results onto a histogram?" <u>*Answer*</u>: The more polls we took would get us very close to the *true mean* of sample percents. Our histogram would look Normal with this mean right in the middle. Thus we say that the *histogram of sample fractions is Normal* with the mean equal to the mean of <u>our sample</u>. It might look something like the histogram on the next page. Notice this is a relatively normal histogram.



NORMAL HISTOGRAM

The Mean = 47%

Now that we have a normal histogram, we can refer to our rough and ready rules.

Recall the New Formula for the Standard Deviation

Now that we have a NORMAL DISTRIBUTION, we can use our rough and ready rules to find percentages and probabilities. But first, we need to calculate the **Standard Deviation**. How would we do that with 'thousands' of sample fractions? Instead of doing it the old way, we now have a simpler formula. This formula <u>can only be used if we are working **with proportions (percents)**</u> rather than with raw data.

The New Formula for Standard Deviation

The formula for calculating the standard deviation of *sample fractions*:

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

p is the <u>mean</u> of the population. (or in this case *hypothetical* or *sample* mean.)

n is the **size** of the sample we are looking at.

Let's calculate the Standard Deviation for the Coakley data: I have labeled it as Example 1.

Ex 1: **p = 47%** and our sample size is 100. i.e. **n = 100**

Formula: $\sigma = \sqrt{\frac{p(1-p)}{n}}$

First turn 47% into a decimal. 47% = 0.47 (Keep at least three decimal places.)

$$\sigma = \sqrt{\frac{.47(1-.47)}{100}}$$

Put (1 – .47) in the calculator first and get (1 – .47) = .53 So.

$$\sigma = \sqrt{\frac{.47(.53)}{100}}$$

$$\sigma = \sqrt{\frac{.2491}{100}} = \sqrt{.002491} = 0.0499099\ldots = \mathbf{4.99099\ldots\ \%}$$

Round: $\sigma = \mathbf{5\%}$

So the standard deviation, $\boldsymbol{\sigma} = \mathbf{5\%}$ In a short time, we will see how to use our standard deviation.

***NOTE:** We always turn our standard deviation into a PERCENT before we do any rounding.

Practice using the new formula for standard deviation:

Ex. 2: Calculate the standard deviation using the new formula with the mean, p = 60% and let's take a sample of size 400. Round your final standard deviation to one decimal place.

Assume p = 60% and our sample size, n = 400. The new standard deviation is thus:

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

First turn 60% into a decimal. 60% = .60 and put it into the formula.

$$\sigma = \sqrt{\frac{.60(1-.60)}{400}}$$

be careful when entering numbers into the calculator. (1 – .60) = .40

$$\sigma = \sqrt{\frac{.60(.40)}{400}}$$

$$\sigma = \sqrt{\frac{.2400}{400}} = \sqrt{.0006} = 0.0244948\ldots = 2.449\% = \mathbf{2.4\%}$$

Notice, I took the decimal 0.0244948 and changed it to a percent **before** I rounded.

Note: the standard deviation *decreases* as the sample size *increases and the reverse is true*. In our case, even though the means were different, our sample size was smaller in Example 1 (n=100) than in Example 2 (n=400). **Since the sample size increased, the standard deviation decreased.**

Review of Rough and Ready Rules and How To Use Them

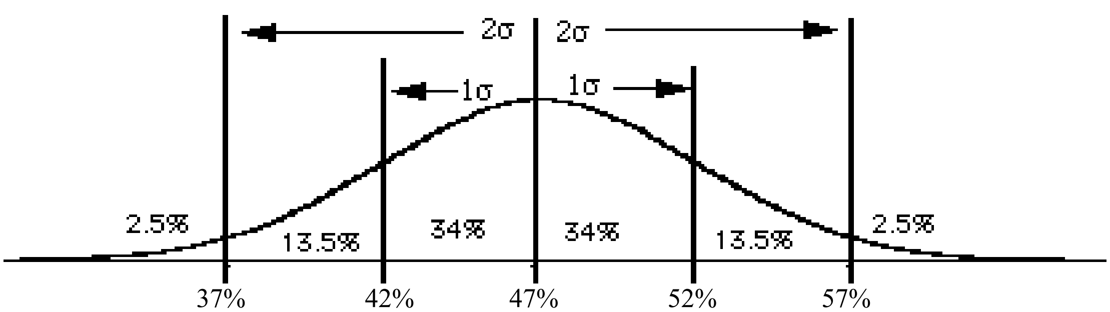
Because the distribution of sample fractions is **normal**, we can use our "rough and ready" percentages from the last Chapter. Remember that our samples must be **randomly** chosen. Here are the *rough and ready rules we lessons back:*

68% of our data lie within **one** standard deviation of the mean
95% lie within **two** standard deviations of the mean.

We also know that:

99.7% lie within **three** standard deviations of the mean. But, for much of our work with election results, we use a 95% interval and don't often need the three standard deviations. We will see how this works a bit later.

Ex. 4: Let's continue with our hypothetical example from the handout on Sampling and Statistical Inference. We looked at those who would vote for Martha Coakley. Our mean was 47% and the standard deviation was 5%. Let's draw the normal distribution curve for these figures. Put 47% as the mean in the center on the horizontal. Add 5% in one direction and subtract in the other, etc.



In this case, we may ask the following questions:

a) What is the probability that Coakley will get between 42% and 57% of the vote?

Answer: Add the inside percentages that are associated with those on the horizontal.

34% + 34% + 13.5% = 81.5% (I think Coakley would be happy with those results.)

b) What is the chance that Coakley will get less than 42% of the vote?

Answer: 13.5% + 2.5% = 16%

We will learn how to calculate Martha Coakley's chances of winning, i.e. of getting more than 50% of the vote!   We will need to calculate what is called z-scores!

First, let's continue with some more definitions of Statistical Inference in the next Chapter. We will get back to Coakley's chance of winning.

CHAPTER 6.3

Confidence Intervals

Let's go back a bit and make believe you worked for a State Candidate before you spread your wings to work for a presidential candidate. (See Sampling and Statistical Inference.) This candidate (say Deval Patrick long before he decided to run for governor) knew you took a statistics course and he asks you to find out his popularity in the State of Massachusetts. How would you answer him? You would need to take your own poll and based on the results of this ONE poll only, you could then say what you thought public opinion was. In this case, we assume no one had any preconceived idea of his popularity and neither did you. We still must bear in mind the following things from our previous study of statistics:

a) was our sample unbiased?
b) how large was our sample?
and thus c) how accurate is our estimate for the population as a whole?

As usual, we can never be 100% certain about anything when dealing with probability. But we will learn how to be 95% sure. Also, we shall see how to use the formula for the standard deviation of sample fractions to find a *range* within which the *true proportion* probably lies (with 95% certainty). So we'll be able to report, for example, that we are 95% certain that our candidate's popularity lies between 37% and 43%. This interval between 37% and 43% is called a **95% Confidence Interval.**

Of course, since we can never be 100% sure of our results, there is a 'margin of error' that creeps into our calculations. This 'margin of error' *depends on how 'large' a sample size we take*. A larger sample size, yields a smaller standard deviation. We will see that the 'margin of error' is directly related to the standard deviation. In fact,

A Margin of Error is equal to "Two times the Standard Deviation," i.e., 2σ

Note that the value for **p** that we use in the calculation of the standard deviation ***is our*** own sample proportion, i.e. our OBSERVED percentage, since we don't have anything else. We will see how this works with some examples.

The *process* of finding the Confidence Interval and the Margin of Error follows.

<u>Steps you need in calculating a Confidence Interval and the Margin of Error</u>

Okay, so Deval Patrick asked you to find out his popularity in Massachusetts and he gave you a small budget to take a survey of about 1000 people. Based on what you know about sampling, you take an <u>unbiased</u> poll of 1000 people and find out that 400 of them say they like Patrick and would vote for him if the election were held then. Now should you go back and tell him that he has a 40% popularity rating??? (400/1000 = 40%) Here's the thinking.

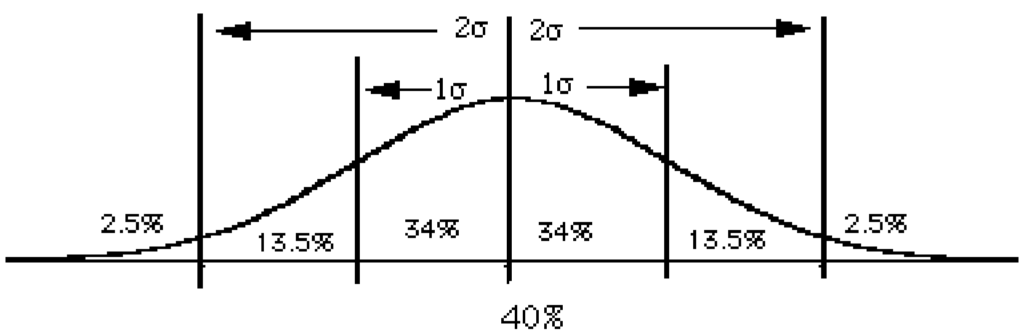
a) You understand that sampling does not necessarily yield the ***exact*** percentage of what the entire population of voters might say. Even if your poll is truly random, there is still a lot of **variation** in sampling results.

b) If you could possibly (which you obviously cannot) take many of these polls of 1000 voters, you would build a histogram of sample fractions. (see handout on Sampling). The true fraction of voters who like your candidate would end up being the mean of all these trials.

c) You know from our study of the normal distribution, that out of all these samples, 95% of them would be within two standard deviations of the true percent. We can turn this around a bit and say that *95% of the time the true percent should be no more than two standard deviations away from our one single sample mean*.

d) However, we do not have very many samples - we only have one. So based on this one and only one sample, we assume it is a *pretty good estimate* of the true fraction of all voters. This is our mean, **p**, that we use in our estimate for the standard deviation.

Looking at the picture:



Our estimate of 40% is the mean and the true proportion is somewhere within two standard deviations of this mean. What do you tell Governor Patrick? Let's actually do the steps and develop our wording.

(This is going to be a CONFIDENCE INTERVAL. It is the percentages between two standard deviations of the mean.)

Example 1: You take an unbiased poll of 1000 people to find out the popularity of a prospective candidate for whom you work. (e.g. Deval Patrick before he became Governor.) You find that out of the 1000 people polled, 400 of them say they like Deval Patrick and would vote for him. ***Find a 95% Confidence Interval for the true proportion of people who approve of your candidate and would vote for him.***

Step 1: Obtain a value for **p** based on our own sample. You may need to calculate 'p' as it could be given in numbers such as in this case. We have 400 out of 1000 (our sample size was 1000), thus

$\mathbf{p} = \frac{400}{1000} = \mathbf{40\%}$ (the 40% came from your survey, i.e. you asked 1000 people and 400 gave a positive answer thus 40%.)

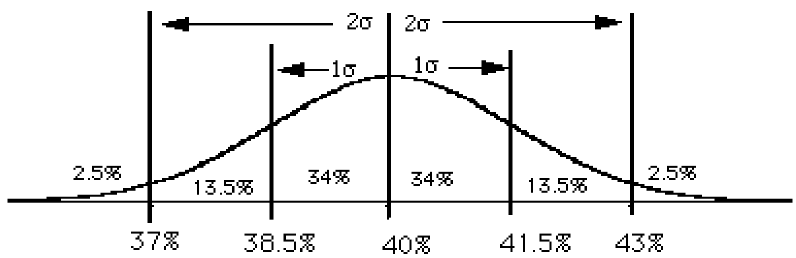
$\mathbf{n = 1000}$

Step 2: Calculate the standard deviation based on this **observed** percentage. Use the new formula.

$$\sigma = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{.40(.60)}{1000}} = 0.01549... = 1.5\%$$

NOTE: 2 times the standard deviation **IS** the Margin of Error. So **Margin of Error = 2(1.5%) = 3%**

Step 3: Calculate **p ± 2σ** Drawing a picture helps here.



p ± 2σ = 40% ± 2(1.5%) = **40% ± 3%** or from **37% to 43%**

Step 4: Construct the 95% confidence interval:

p - 2σ <------> p + 2σ

Answer (1) The 95% confidence interval for the true proportion of voters who would vote for Governor Patrick is between 37% and 43%. Write this in the notation below:

37% <-------------------------> 43%

This means that there is a 95% chance that the real true percent of voters who like Patrick was somewhere between these limits. You can never be 100% sure of this but you can be 95% sure!

Answer (2) (another way of stating a Confidence Interval) The percent of voters who like your candidate is

40% plus or minus a *Margin of Error* of 3%.

Note: You can express your **answer** either by using (1) or (2). If you combine these, it becomes redundant. For example, do not say, "The 95% CI is 37% to 43% with a margin of error of ± 3%."

Finally, we tell our candidate that he has a popularity of 40% plus or minus 3%!!!!

Example 2: Now we'll find the 95% Confidence Interval for the true proportion of people who will vote for Martha Coakley in the Massachusetts gubernatorial election in November.

Step 1: Obtain a value for **p** based on your own sample. You may need to calculate '**p**' as it could be given in numbers. In this case we had p=47%. (See the handout on Sampling and Statistical Inference.)

p = **47%**

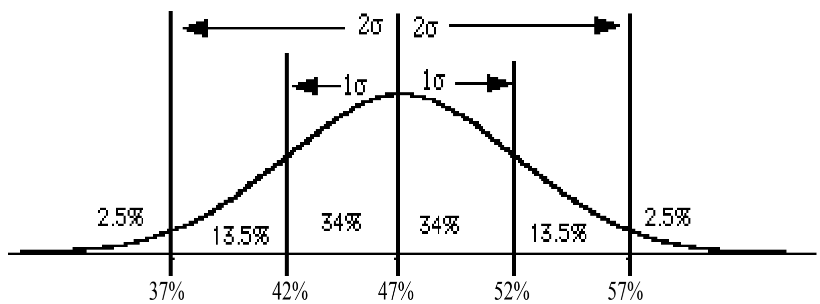
n = 100 This time we had only asked 100 people, a much smaller sample than the 1000.

Step 2: Calculate the standard deviation based on <u>this</u> **observed** percentage. We already calculated the standard deviation as 5% (Page 2).

. **σ = 5%**

NOTE: 2 times the standard deviation **IS** the Margin of Error. So **Margin of Error = 2(5%) = 10%**

Step 3: Calculate **p ± 2σ** We show the curve below.



p ± 2σ = 47% ± 2(5%) = **47% ± 10%** or from **37% to 57%**

Step 4: Construct the 95% confidence interval:

p - 2σ <------> p + 2σ

Use the wording as in ANSWER (1) or ANSWER (2) below.

Answer (1) The 95% confidence interval for the <u>true proportion</u> of people who say they will vote for Martha Coakley is:

37% <-------------> 57%

Note: This means that there is a 95% chance that the real true percent of these voters is somewhere between these limits. You can never be 100% sure of this but you can be 95% sure!

Answer (2) The percent of voters who say they intend to vote for Coakley is

47% plus or minus a *Margin of Error* of 10%.

NB: The Margin of Error is 10%. In example 1, our margin of error was only 3%. We had a much larger sample in example 1. (1000 versus 100 people) The margin of error shrinks as the sample size increases. This is because the standard deviation decreases as the sample size increases. And notice the reverse is true. Since the sample size was much smaller, our margin of error was larger.