Normal Distributions, t distributions, the Central Limit Theorem and Confidence Intervals

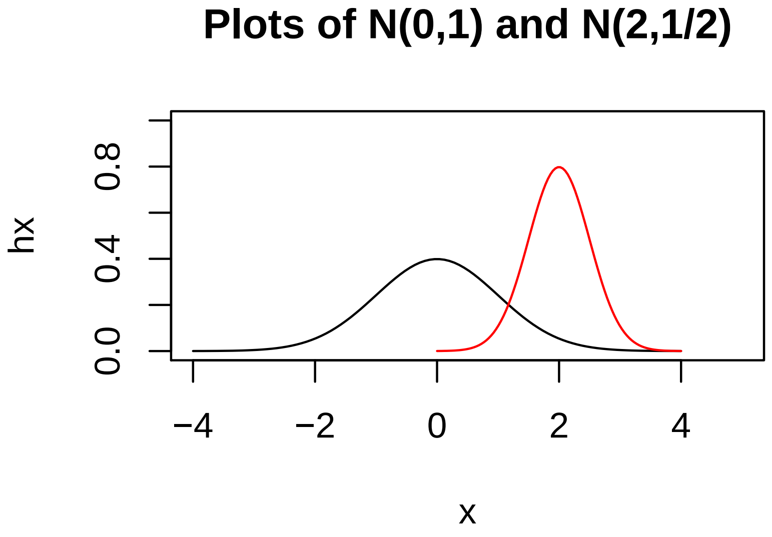
```
> CombineND<-read.csv("combineND.csv",header = TRUE)
> MLB<-read.csv("MLB.csv",header = TRUE)
> NBA<-read.csv("NBA.csv",header = TRUE)
> NFL<-read.csv("NFL.csv",header = TRUE)
> Pass2005<-read.csv("Pass2005.csv",header = TRUE)
> Rush2005<-read.csv("Rush2005.csv",header = TRUE)
```

One important family of distributions for continuous random variables is the family of **Normal Distributions**. The formula for the normal density function is given by

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

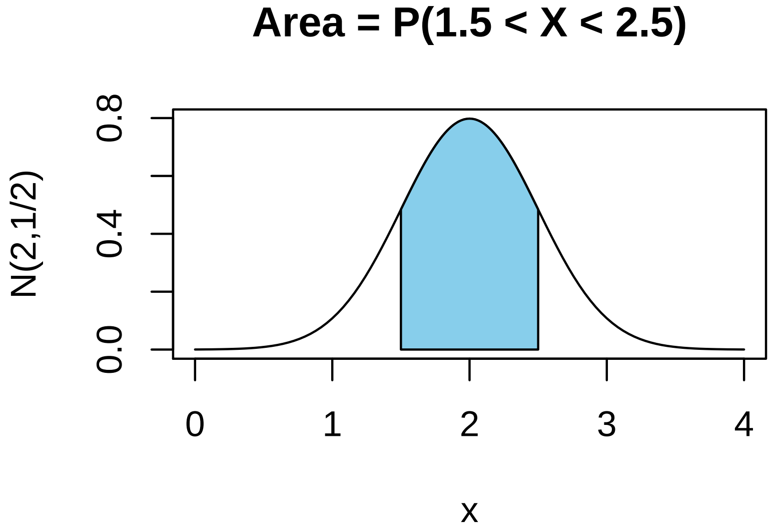
where μ is the mean of the distribution and σ^2 is the variance of the distribution. Since the formula for the distribution depends only on the two unknown parameters μ and σ, one frequently refers to the above density as $N(\mu, \sigma)$ or Normal(μ, σ).

The graph of this distribution is symmetric around the center and bell shaped. The mean, median and mode are the same and the points of inflection are sitiuated one standard deviation from the mean on both sides. The larger the value of σ, the more spread out the distribution is and different values of μ give a different position for the center of the curve. Below we show $N(0, 1)$ and $N(2, 1/2)$(in red).



For data from a population with denisty $N(\mu, \sigma)$, the z-scores of the data $\{\frac{x_i - \mu}{\sigma}\}$ have a standard normal distribution, that is they are normal with mean 0 and standard deviation 1.

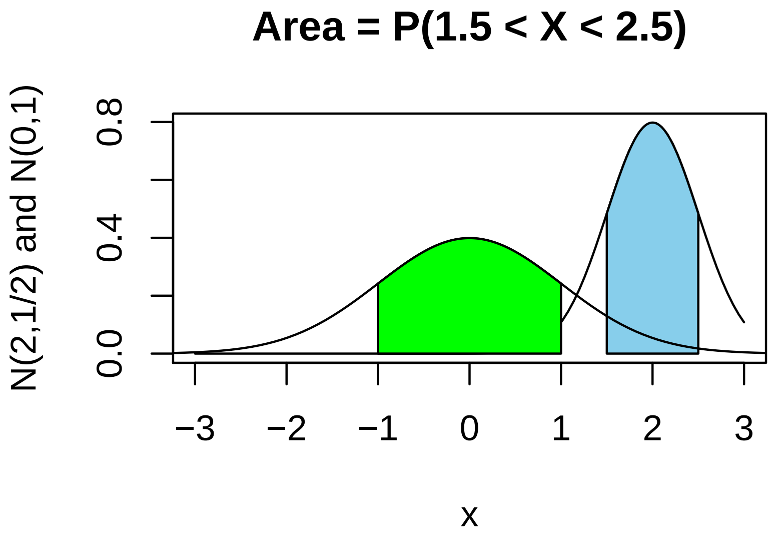
As with all densities, if X is a random variable which is normally distributed, with mean μ and standard deviation σ, we can find the probability that X takes values in the interval (a, b) ($P(a < X < b)$) by calculating the area under the corresponding normal desity above that interval. For example if X is normally distributed with mean 2 and standard deviation 1/2, the probability that X takes a value between 1.5 and 2.5 is shown in the diagram below.



As mentioned above if we scale our observations of the variable X (which is normal with mean 2 and standard deviation 1/2) to their Z-scores, the new variable $Z = \frac{X-2}{1/2}$ is a standard normal random variable. Thus

$$P(1.5 < X < 2.5) = P(\frac{1.5-2}{1/2} < Z < \frac{2.5-2}{1/2}) = P(-1 < Z < 1).$$

Thus the areas highlighted in the garph shown below are the same and the probability that X lies is any interval depends only on how many standard deviations the end points are from the mean.



We can of course calculate this area without drawing the curve. For any family of distributions `R` has four functions for getting information about the density.

The `d`, `p`, `q` and `r` functions.

- The `d` function returns the pdf of the distribution.
- The `p` function returns the cumulative density function.
- The `q` function returns the quantiles.
- the `r` function returns random samples from a distribution.

The normal densities are denoted by `norm` in `R`. They have two parameters which characterize them, the `mean` and the standard deviation `sd`. If left unspecified, R sets the mean to 0 and the standard deviation to 1. We show how the above functions apply to the normal distributions in the following examples.

Example: the `d` function(used to plot the density). To get the height of the normal density curve with mean 2 and standard deviation 1/2 above the point 2 (the highest point on the curve) we can use the command

```
> dnorm(2,mean=2,sd=1/2)
```

```
[1] 0.7978846
```

This is not a calculation we have much interest in other than perhaps to plot normal densities.

The p function($P(X \leq a)$.) This is the function we use to calculate probabilities. The command pnorm(a,mean=μ,sd=σ) will return the probability that a normal random variable X with mean μ and standard deviation σ will take a value less than a.

Example: If I know that the heights of NFL Players is normally distributed with mean 74 inches and standard deviation 2.6 inches then we can find the proportion fo NFL players with height less than or equal to 72 inches with the following command:

```
> pnorm(72,mean=74,sd=2.6)

[1] 0.2208782
```

We see that approximately 22 percent of players have height less than 72 inches.

Example: Find the area under the normal density N(2,1/2) between 1.5 and 2.5 shown above.

```
> pnorm(2.5,mean=2,sd=0.5) - pnorm(1.5,mean=2,sd=0.5)

[1] 0.6826895
```

Empirical rule: The following rule of thumb is good to keep in mind when dealing with data that looks bell shaped:

If X is a normally distributed random variable, then

- roughly 68% of the population values will be within 1 standard deviation of the mean i.e. $P(\mu - \sigma < X < \mu + \sigma) \approx 0.68$.
- roughly 95% of the population values will be within 2 standard deviations of the mean i.e. $P(\mu - 2\sigma < X < \mu + 2\sigma) \approx 0.95$.
- roughly 99.7% of the population values will be within 3 standard deviations of the mean i.e. $P(\mu - 3\sigma < X < \mu + 3\sigma) \approx 0.997$.

Its enough to check this for the distribution of Z-scores:

```
> pnorm(1) - pnorm(-1)

[1] 0.6826895

> pnorm(2) - pnorm(-2)

[1] 0.9544997
```

```
> pnorm(3) - pnorm(-3)

[1] 0.9973002
```

The `q` **function**(quantiles). The command `qnorm(c(p,q,r), mean =` μ`, sd =` σ`)` returns the p th, q th and r th quantiles for a normal random variable X with mean μ and standard deviation σ.

Example: If the height of NFL players is normally distributed with mean 74 inches and standard deviation 2.6 inches, find the 0.1 th, 0.5 th and 0.9th quantiles (10th, 50th and 90th percentiles) for the heights of NFL players.

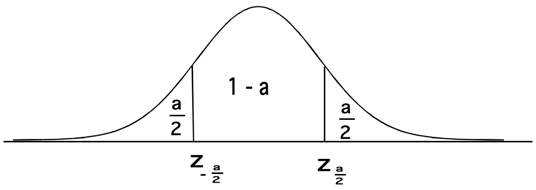
```
> qnorm(c(0.1,0.5,0.9),mean=74,sd=2.6)

[1] 70.66797 74.00000 77.33203
```

Of course it should not surprise us that the 50th percentile is 74 in the above example.

Example Find an interval $(-z_{\alpha/2}, z_{\alpha/2})$ centered at 0 for which

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$$

where Z is a standard normal random variable for the values of α shown below.



$\underline{\alpha = 0.01}$ Here 99% of the area under the curve lies between $-z_{\alpha/2}$ and $z_{\alpha/2}$, thus the tails have area $(1 - 0.99)/2 = .005$. Thus $-z_{\alpha/2}$ and $z_{\alpha/2}$ are the .005 th and .995 th quantiles respectively.

```
> qnorm(.005)

[1] -2.575829

> qnorm(.995)

[1] 2.575829
```

$\underline{\alpha = 0.05}$ Here 95% of the area under the curve lies between $-z_{\alpha/2}$ and $z_{\alpha/2}$, thus the tails have area $(1 - 0.95)/2 = .025$. Thus $-z_{\alpha/2}$ and $z_{\alpha/2}$ are the .025 th and .975 th quantiles respectively.

```
> qnorm(.025)

[1] -1.959964

> qnorm(.975)

[1] 1.959964
```

Note: We can say that for any normal random variable X, with mean μ and standard deviation σ

$$P(\mu - 1.96\sigma < X < \mu + 1.96\sigma) = 0.95$$

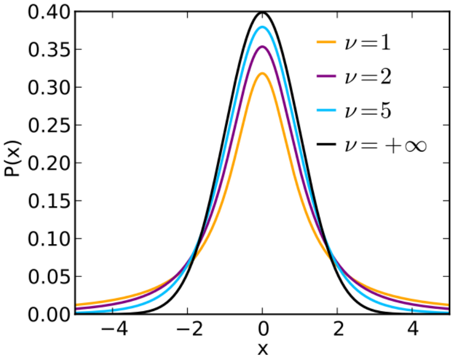
or 95% of the population will have values of X in this interval.

The `r` function(random samples). The `r` function allows us to take a random sample from a population with a particular density.

Example: Testing the empirical rule. Simulate taking a random sample of size 1000 from a normal population with mean 74 and standard deviation 2.6 and check what percentage of the data lies within 1, 2, and 3 standard deviation from the mean.

```
> mu<-74
> sigma<-2.6
> samp<-rnorm(1000,mean=mu,sd=sigma) #take sample
> head(samp)

[1] 77.92306 74.45410 75.25245 75.21480 77.25079 78.12305

> sum((samp> mu-sigma)&(samp<mu+sigma)) #within 1 sd of mean

[1] 682

> sum((samp> mu-2*sigma)&(samp<mu+2*sigma)) #within 2 sds of mean

[1] 952

> sum((samp> mu-3*sigma)&(samp<mu+3*sigma)) #within 3 sds of mean

[1] 999
```

t distributions(used for small samples). This is a family of mound shaped distributions which look similar to the standard normal distribution. They will be used when estimating population parameters or testing hypotheses using small samples. Their shape depends on a parameter called degrees of freedom, $df = \nu$ which is related to sample size. For large values of ν these distributions are very close to the standard normal distribution.



We can of course use the `d, p, q` and `r` functions to explore t distributions. We see that these distributions have fatter tails for smaller degrees of freedom.

```
> pnorm(1.96) - pnorm(-1.96)

[1] 0.9500042

> pt(1.96, df=30) - pt(-1.96, df=30)

[1] 0.9406577

> pt(1.96, df=10) - pt(-1.96, df=10)

[1] 0.9215638

> pt(1.96, df=5) - pt(-1.96, df=5)

[1] 0.892712

> qnorm(c(0.001, 0.01,0.05))

[1] -3.090232 -2.326348 -1.644854

> qt(c(0.001, 0.01,0.05),df=30)

[1] -3.385185 -2.457262 -1.697261
```

```
> qt(c(0.001, 0.01,0.05),df=10)

[1] -4.143700 -2.763769 -1.812461

> qt(c(0.001, 0.01,0.05),df=5)

[1] -5.893430 -3.364930 -2.015048
```

Central Limit Theorem.

Lets consider the following experiment and associated random variable:
Experiment: Take a random sample of size 30 from our data on the population of NFL players in 2014 and record the heights of the players in the sample. A random sample assumes that the players in the sample are chosen independently and hence we say that the heights of the players in the sample are *iid* (independent and identically distributed).
Random Variable $\bar{X}$: $\bar{X}$ is the average height of the players in the sample.

$\bar{X}$ varies from sample to sample and as a random variable it has a probability distribution. As it turns out this distribution is approximately a normal distribution, with mean equal to the population mean $\mu \approx 74$ and standard deviation equal to $\sigma/\sqrt{n} \approx 2.6/\sqrt{30}$ where σ is the population standard deviation and the sample size is 30. (Note that bigger sample sizes give smaller variation in the sample mean). Lets simulate this experiment 1000 times (we choose 10000 random samples of size 30 from the 1699 in the sample) calculate the value of $\bar{X}$ for each sample and plot the density associated to the 1000 means in the data. We will also plot a normal distribution with mean equal to the population mean and standard deviation equal to the population standard deviation divided by $\sqrt{30}$.

First lets calculate the population mean and standard deviation.

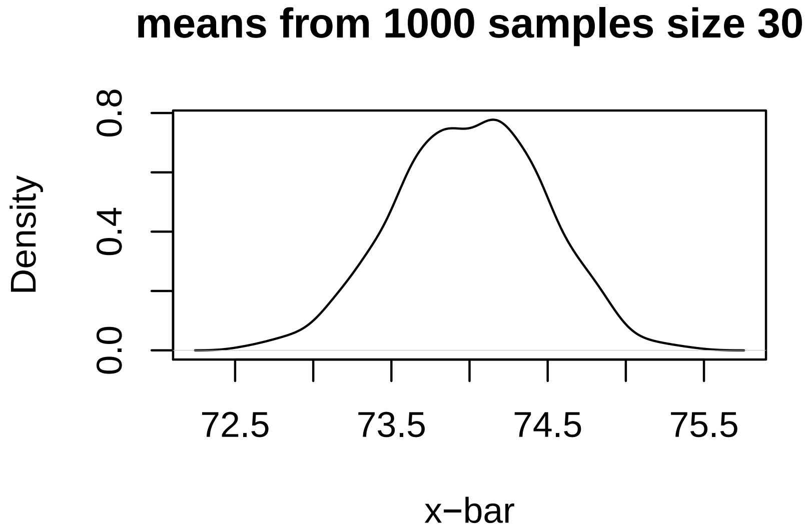
```
>  #the number of different samples of size 30  possible.
> choose(nrow(NFL), 30)

[1] 2.34584e+64

> mean(NFL$HT) #population mean

[1] 74.02001

> sd(NFL$HT)

[1] 2.640105
```

We use our old trick of creating a vector of 0's and filling it up with the sample means using a `for` loop.

```
> smeans<-mat.or.vec(1,1000)
> for(i in 1:1000){
+   smeans[i]<-mean(sample(NFL$HT,size=30))
+ }
```

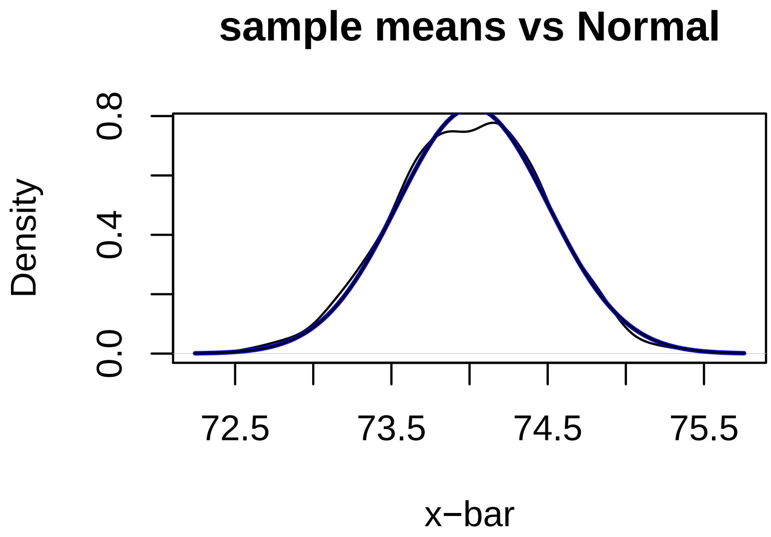
Now lets plot the denity associated to our vector of means.

```
> plot(density(smeans), xlab="x-bar",
+       main = "means from 1000 samples size 30")
```



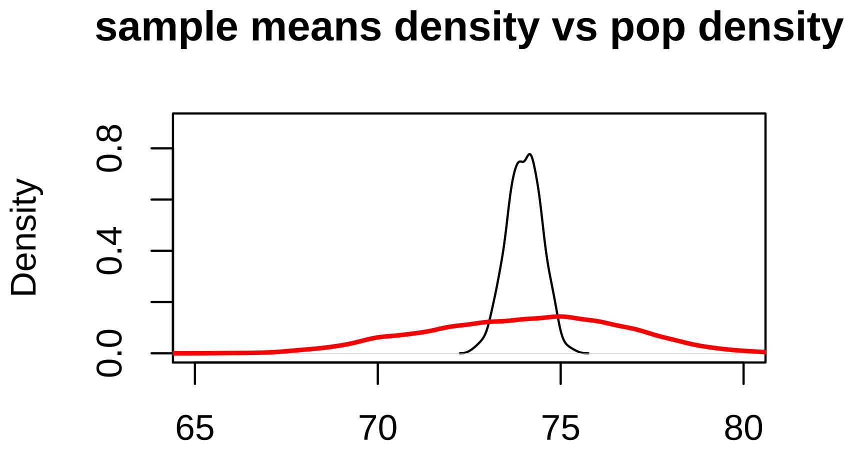
If we add the normal density with the same mean as the population mean $\mu = 74.02001$ and standard deviation equal to the population standard deviation divided by $\sqrt{30}$ in blue, we see that it fits the density of the means very well.

```
> plot(density(smeans), xlab="x-bar",
+       main = "sample means vs Normal")
> lines(curve(dnorm(x,mean=mean(NFL$HT),
+    sd = sd(NFL$HT)/sqrt(30)), lwd=2, col="blue",add=TRUE))
```



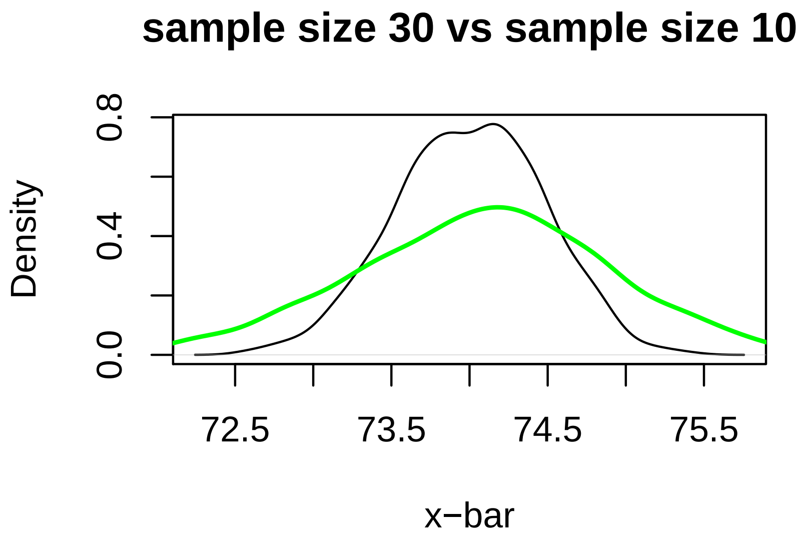
We also compare the density to the original population density and see that the means from the samples are much more concentrated around the mean.

```
> plot(density(smeans), xlab="",
+      xlim=range(c(65,80)),ylim=range(c(0,0.9)),
+      main = "sample means density vs pop density")
> lines(density(NFL$HT, na.rm=TRUE), lwd=2,col="red")
```



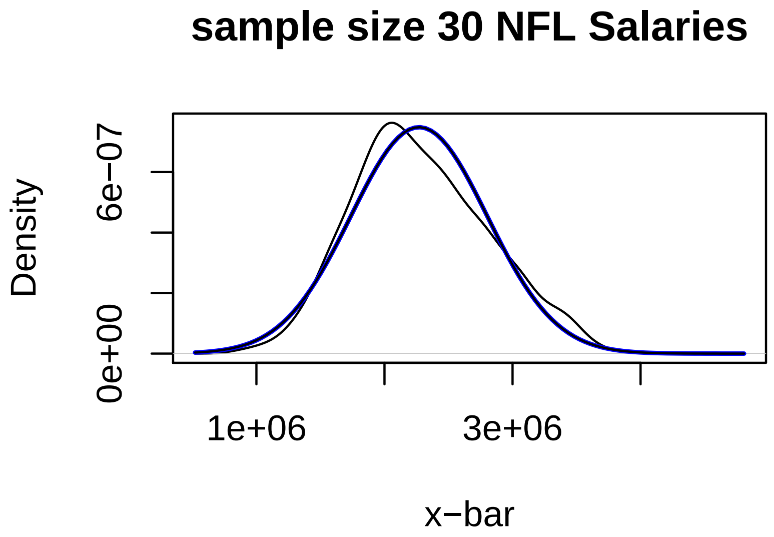
Larger samples with give a smaller variance in the distribution of the sample means and thus less room for error when estimating the population mean using a sample mean. Lets compare to the means of 1000 samples of size 10.

```
> smeans1<-mat.or.vec(1,1000)
> for(i in 1:1000){
+   smeans1[i]<-mean(sample(NFL$HT,size=10))
+ }

> plot(density(smeans), xlab="x-bar",
+      main = "sample size 30 vs sample size 10")
> lines(density(smeans1), lwd=2, col="green")
```



Even if the unerlying distribution is skewed as in player's salaries, the distribution of sample means will still be approximately normal.

```
> smeans2<-mat.or.vec(1,1000)
> for(i in 1:1000){
+   smeans2[i]<-mean(sample(NFL$AVG.Annual.Salary,size=30),na.rm=TRUE)
+ }

> plot(density(smeans2), xlab="x-bar",
+      main = "sample size 30 NFL Salaries")
> lines(curve(dnorm(x,mean=mean(NFL$AVG.Annual.Salary,na.rm=TRUE),
+ sd = sd(NFL$AVG.Annual.Salary,na.rm=TRUE)/sqrt(30)),
+ lwd=2, col="blue",add=TRUE))
```



The Sampling Distribution of $\bar{x}$/ The Central Limit Theorem

For any population, the sample mean, $\bar{x}$, is an unbiased estimator of the population mean, μ.

For a population with any distribution, with mean μ, and standard deviation σ, The sampling distribution of the sample mean, $\bar{x}$, has the following properties:

1. The mean of the sampling distribution of $\bar{x}$ equals the mean of the sampled population, μ. That is :

$$\mu_{\bar{x}} = E(\bar{x}) = \mu.$$

2. The standard deviation of the sampling distribution of $\bar{x}$ is:

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

(This is true if $n < .05N$, where N is the population size).
3. For n large, the sampling distribution of $\bar{x}$ is approximately $N(\mu, \dfrac{\sigma}{\sqrt{n}})$.

Proportions as means. Consider the following experiment:

Experiment: Choose an attempted three point shot at random from the population of Stephen Curry's attempted three point shots in regular season games and record a 1 if he made the shot and a 0 if he did not.

Random Variable The value we record is an example of a Bernoulli random variable. (a variable that takes only two values, in our case 0 or 1.) The **population mean here is** $\mu = \frac{\# \ 1's}{total \ \# \ shots}$ = S. Curry's 3 pt percentage or **the proportion of 3 point shots he made from all he attempted.**

Thus **a population proportion is a population mean** where the variable in question is a **Bernoulli random variable and sample proportions (for a given n) are approximately normally distributed** with mean equal to the population proportion p and standard deviation equal to the population standard deviation $\sqrt{p(1-p)}$ divided by $\sqrt{n}$ where n is the size of the samples. It is not too hard to show that

A Bernoulli random variable (with two values 0 and 1)

has population mean $\mu = p$, and population standard deviation $\sigma = \sqrt{p(1-p)}$.

where p is the proportion of 1's in the population.

Confidence Intervals for the Population Mean

If I take a sample of size n from a population $\{x_1, x_2, \ldots, x_n\}$, I can use the sample mean $\bar{x}$ to estimate the population mean. However, I know that the sample mean varies from sample to sample, so it is likely that my estimate has some error. **Because of the central limit theorem, I can supply an estimate of this error, or a margin of error for my estimate.** In fact what I do is supply an interval around $\bar{x}$ called a confidence interval and say something like "I am 95% confident that the population mean is somewhere in this interval".

We know from the central limit theorem that for large fixed values of n the statistic $\bar{x}$ is approximately normally distributed with mean μ (the unknown population mean) and standard deviation $\sigma/\sqrt{n}$, where σ is the (usually unknown) population standard deviation. Therefore we know that the statistic

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

is approximately normally distributed with mean 0 and standard deviation 1. Thus we know that $P(-1.96 < Z < 1.96) \approx 0.95$. In fact $P(-z_{\alpha/2} < Z < z_{\alpha/2}) \approx (1 - \alpha)$ where the value of $z_{\alpha/2}$ is given in the following table (which can be verified using the

Confidence level $1-\alpha$	$z_{\frac{\alpha}{2}}$
0.60	0.841
0.70	1.036
0.80	1.282
0.90	1.645
0.95	1.96
0.98	2.32
0.99	2.576
0.999	3.291

$$P(\mu - 1.96\frac{\sigma}{\sqrt{n}} < \bar{x} < \mu + 1.96\frac{\sigma}{\sqrt{n}}) \approx 0.95$$

and therefore subtracting μ from both sides of both inequalities, we get

$$P(-1.96\frac{\sigma}{\sqrt{n}} < \bar{x} - \mu < +1.96\frac{\sigma}{\sqrt{n}}).$$

Now subtracting $\bar{x}$ from both sides of both inequlaities and multiplying by -1, we get

$$P(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}) \approx 0.95.$$

Unfortunately we usually do not know the population standard deviation σ. However we can substitute and estimate of σ from the sample, namely the sample standard deviation s. This gives us the following 95% confidence interval for the population mean μ when n is large:

$$P(\bar{x} - 1.96\frac{s}{\sqrt{n}} < \mu < \bar{x} + 1.96\frac{s}{\sqrt{n}}) \approx 0.95$$

or we are 95% confident that the population mean μ is in the interval

$$(\bar{x} - 1.96\frac{s}{\sqrt{n}}, \bar{x} + 1.96\frac{s}{\sqrt{n}}).$$

The statistic SE $= \frac{s}{\sqrt{n}}$ is referred to as the standard error of the mean and in this case $1.96\frac{s}{\sqrt{n}}$ is called the margin of error.

If we want a different level of confidence, we should change 1.96 in the above calculations to an appropriate value of $z_{\frac{\alpha}{2}}$. We get

$$P(\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}) \approx 1 - \alpha$$

where $z_{\frac{\alpha}{2}}$ is given in the table above. In other words, we are $(1-\alpha)100\%$ confident that the population mean μ is in the interval

$$\boxed{\bar{x} \pm z_{\frac{\alpha}{2}} SE.}$$

Note We cheated a little here by estimating the population standard deviation σ by the sample standard deviation s. This changes the distribution of the statistic Z, the statistic

$$T = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

has an approximate t distribution with $n - 1$ degrees of freedom. For large values of n, this is approximately standard normal and our confidence intervals are still fairly accurate. However, for small values of n (less than 20 or so) one should replace the values for $z_{\frac{\alpha}{2}}$ in the table above by the corresponding quantiles for the appropriate t distribution.

Example Suppose we want to estimate the average yards gained for a college football team if they run the ball from anywhere between their 45 yard line to their 50 yard line with less than 3 yards to first down. We might consider all such plays made in the year 2005 as a representative sample (although not entirely random). There were 65 such plays in 2005, the data for which appears in the file `Rush2005.csv`

```
> Rush2005$Yards

 [1]   2   1   0  45   4  -4   0  48   1   0   2   8 -11   1   1   4   4   0   3
[20]  38   2   3   2   8   3  -1   1  -1   0   2   3   8  30   2   3   6   3  -7
[39]   3  13  17   0   0   4   3   1   1   2   1   0   4   5  -1   1  -2  -1   2
[58]   1   0  -1   4   2  -2  -4  -1

> mean(Rush2005$Yards)

[1] 4.092308

> sd(Rush2005$Yards)

[1] 10.2511
```

From this we can make a 95% confidence interval for the average yards gained by such a play in such a position:

```
> xbar<-mean(Rush2005$Yards)
> s<-sd(Rush2005$Yards)
> n<-65
> error <- qnorm(0.975)*s/sqrt(n)
> #left
> xbar-error

[1] 1.600228
```

```
> #right
> xbar+error

[1] 6.584388
```

A Confidence interval for the population proportion

Because the population proportion is an average of a (Bernoulli) random variable (with values 0 or 1) as pointed out above, we can also use the above method to find a confidence interval for the population proportion p, using a sample proportion $\hat{p}$ from a sample of size n. As above we have that $\frac{\hat{p}-p}{\sqrt{p(1-p)/n}}$ is approximately normally distributed with mean 0 and standard deviation 1 for large values of n. We can approximate $\sqrt{p(1-p)/n}$ using the sample proportion as an estimate of p, to get a standard error SE $= \sqrt{\hat{p}(1-\hat{p})/n}$

Therefore, we have

$$\boxed{\hat{p} \pm z_{\frac{\alpha}{2}} SE}$$

gives a confidence interval for the population proportion p with level of confidence $1-\alpha$.

Example Suppose we want to know what percentage of Pass plays on fourth down in College football from the 45-50 yard line ended in a first down. Once again, we take all such plays from the 2005 season as a sample and calculate the proportion of such plays that ended in a first down $\hat{p}$. There were 65 such plays in 2005 and they are stored in a file called `Pass2005.csv`. Lets make a 99% confidence interval for the percentage of successful pass plays in this situation.

```
> Pass2005$X1st.Down

 [1] 0 0 1 0 1 1 0 0 1 0 0 1 0 0 1 1 0 0 1 0 1 0 1 0 0 0 0 0 0 0 1 0 0 1 0 0 0 1
[39] 0 0 1 0 1 0 0 1 0 1 0 1 1 0 1 0 0 1 1 0 0 1 0 1 1 0

> n=length(Pass2005$X1st.Down)
> phat<-mean(Pass2005$X1st.Down)
> phat

[1] 0.390625

> se<-sqrt((phat*(1-phat))/n)
> #left
> phat-qnorm(.995)*se

[1] 0.2335347
```

```
> #right
> phat+qnorm(.995)*se

[1] 0.5477153
```

For smaller samples one needs to change the SE a little.