We first download our functions as follows:

```
> source("http://www.math.csi.cuny.edu/st/R/sim.R")
```

There are now two functions. The first is the same as last week. The function is called `sim()` and it is used with the following arguments

```
sim(m = number of simulations,
    n = size of sample,
    family = what family,
    statistic = name of statistic)
```

The default values are $m = 200$, $n = 10$, the normal family and the statistic is the mean.

The second function is called `sim.CI()`. It will simulate confidence intervals for π of for μ. It is called as follows.

For μ

```
sim.CI(m = number of simulations,
       statistic = "mean" (the default),
       alpha = level of confidence (default=0.05),
       n = size of sample,
       mean = mean of sample,
       sd   = standard deviation of sample
       )
```

For π

```
sim.CI(m = number of simulations,
       statistic = "p",
       alpha = level of confidence,
       n = size of sample,
       p = true proportion (pi is reserved word)
       )
```

Most of the time the defaults are reasonable. We only need to specify the arguments we wish to change.

1 how far is the statistic from the parameter value?

The `sim.CI()` returns four values for each simulation, the value of the statistic (p or $\bar{x}$), the standard error, and the lower and upper limits of the confidence interval for confidence level α.

Here are 5 simulations of $\bar{x}$.

```
> sim.CI(m = 5, n = 10, mean = 15)
```

```
   xbar     SE    ll    ul
1 14.83 0.3197 14.21 15.46
2 15.07 0.3923 14.30 15.83
3 15.00 0.1759 14.66 15.35
4 15.20 0.1212 14.96 15.44
5 14.67 0.2899 14.10 15.24
```

Let's first look and see what percent of our values are within cutoffs given by our standard errors.

For example, for $\bar{x}$, with $n = 10$, $\mu = 15$.

```
> tmp = sim.CI(m = 100, n = 10, mean = 15)
> sum(abs(tmp$xbar - 15) < 2 * tmp$SE)/100

[1] 0.94
```

This is the proportion of values within 2 standard errors. It should be roughly 95signs find the variable `xbar` in the variable `tmp`.

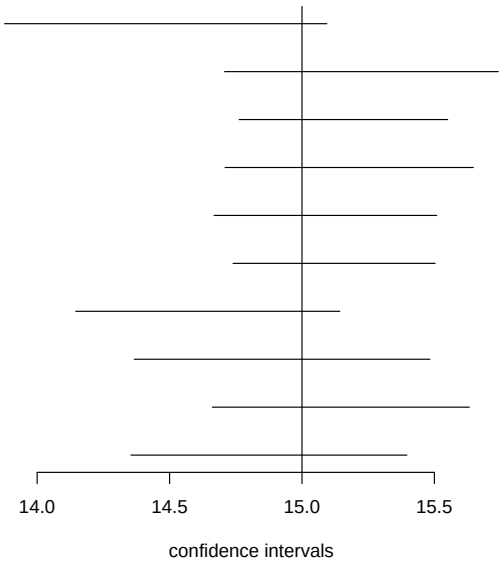
 Question 1: Repeat the above for 1, 1.65 and 3 standard errors. What percentages do you get.

 Question 2: For $m = 100$, $n = 100$, $\pi = 1/2$ repeat the above with 1,1.65, 2 and 3 standard errors.

2 viewing confidence intervals

The `plot()` function will make a graph of the confidence intervals. For example,

```
> tmp = sim.CI(m = 10, alpha = 0.1, n = 10, mean = 15)
> plot(tmp)
> abline(v = 15)
```



The `abline()` function adds the vertical line marking the mean.

Question 3: How many times does the confidence interval cover the vertical line? Should it do it every time? What percentage of the time do you expect?

Question 4: make the same graphic only with $\alpha = .25$. Now what percent of the time is your line covered by the confidence interval?

Question 5: For $m = 25$, $\alpha = .20$, $n = 1000$ and $\pi = .42$ draw the confidence intervals. What percentage of the time does the interval contain the value of π?

3 the *t*-distribution

The distribution of $\bar{x}$ is approximately normal, but the distribution of

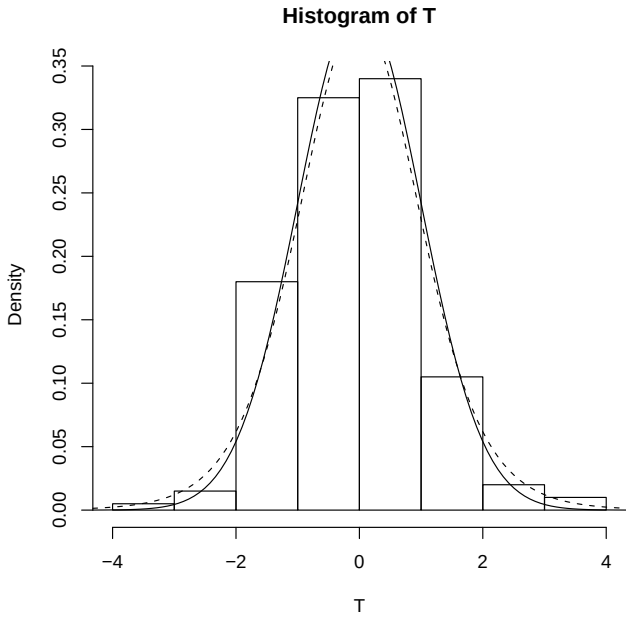
$$t = \frac{\bar{x} - \mu}{\mathbf{SE}}$$

is not. It has the t-distribution with $n-1$ degrees of freedom.

The standard error is $s/\sqrt{n}$. We use this a in the standard deviation, $\sigma/\sqrt{n}$ the value of σ is unknown. If we knew σ, then the distribution of $(\bar{x}-\mu)/\mathbf{SD}$ is normal with mean 0 and variance 1.

We can see what the t-distribution looks like using the simulation

```
> tmp = sim.CI(m = 200, n = 10, mean = 5)
> T = (tmp$xbar - 5)/tmp$SE
> hist(T, probability = TRUE)
> curve(dnorm(x), add = TRUE)
> curve(dt(x, df = 10 - 1), add = TRUE, lty = 2)
```



The two lines are the theoretical densities. The t-distribution looks at first like the standard normal, but it has a larger variance.

Question 6: Why would you think the distribution of t would have a larger variance? If you divided by the standard deviation instead of the standard error you would have the normal instead.

Question 7: Repeat the above graphic with $n = 5$ instead of 10. Does this change the shape of any of the graphics?

Question 8: Find samples of the T statistic for $n = 2, 5, 10, 25$ and 50. Store the values and make side-by-side boxplots. Are the centers similar? The spreads? The tails?

Question 9: We use a normal population above. If we used a exponential – a skewed distribution, the sampling distribution of T is not known. (Only if n is large does it become normal.)

To find a sample from the exponential with mean 1 can be done as follows

```
> tmp = sim.CI(m = 200, n = 6, family = "exp")
```

Does the sampling distribution of T even look symmetric? Compare the tails to that given by `curve(dt(x,df=6-1),add=T)`