0.1 The harder way

We've seen that $(1-\alpha)\cdot 100\%$ confidence intervals for an unknown proportion, π, or a population mean, μ, are given by

$$p \pm z^* \mathbf{SE}(p) \quad \text{or} \quad \bar{x} \pm t^* \mathbf{SE}(\bar{x})$$

To find each on the computer is not hard, and follows this pattern. To find for a proportion, you need to use the normal distribution to get $z*$. Suppose 650 out of 1000 are the sample numbers.

```
> x = 650
> n = 1000
> p = x/n
> alpha = 0.05
> zstar = qnorm(1 - alpha/2)
> SE = sqrt(p * (1 - p)/n)
> p + c(-1, 1) * zstar * SE

[1] 0.6204 0.6796
```

For a population mean, say $\bar{x} = 10$, $s = 5$ and $n = 16$. Then the a 90% CI is found with

```
> xbar = 10
> s = 5
> n = 16
> alpha = 0.1
> SE = s/sqrt(n)
> tstar = qt(1 - alpha/2, df = n - 1)
> xbar + c(-1, 1) * tstar * SE

[1]  7.809 12.191
```

 Question 1: Suppose 120 CSI students are surveyed, and 55 support the war in Iraq. Find a 90% confidence interval for the population proportion of CSI students who support the war. Does it include .5?

 Question 2: A newspaper has run an average of 3 articles per day with a standard deviation of 1 for the last 10 days. Find a 95% confidence interval for the mean number of articles it will run if these numbers can be treated as a random sample.

0.2 Using R functions to do the work.

The functions `prop.test()` and `t.test()` will do these calculations for us. The `prop.test()` function uses this format

```
prop.test(x, n, conf.level = 0.95)
```

That is you specify the counts, the size of the sample and the confidence level $(1-\alpha)$. The default confidence level is .95. To do the above example we have

```
> prop.test(650, 1000, conf.level = 0.95)

        1-sample proportions test with continuity correction

data:  650 out of 1000, null probability 0.5
X-squared = 89.4, df = 1, p-value = < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.6194 0.6794
sample estimates:
   p
0.65
```

There is a lot of output, but you should be able to find the confidence interval. Is it the same as what we previously found? (`prop.test()` doesn't use the standard error.)

The `t.test()` function requires the raw data, not the summarized data. If this is stored in `x`, then the format is

```
t.test(x, conf.level = 0.95)
```

Again, 0.95 is the default for the confidence level so it can be left out if that is what is desired.

For example, a baker sells 80, 90, 70, 65, 95, and 100 doughnuts in six days. *If* this can be considered to be a random sample from the population of doughnut sales, and this population is normally distributed, then an 80% confidence interval for the population mean is

```
> x = c(80, 90, 70, 65, 95, 100)
> t.test(x, conf.level = 0.8)

        One Sample t-test

data:  x
t = 14.56, df = 5, p-value = 2.763e-05
alternative hypothesis: true mean is not equal to 0
80 percent confidence interval:
 74.88 91.78
sample estimates:
mean of x
    83.33
```

 Question 3: The data set `nhtemp` can be loaded with the command

```
> data(nhtemp)
```

Is the data appropriate for a t-based confidence interval? (It needs to be normally distributed. Assume it may be treated as a random sample.) How can you tell? If so, find an 80 and 95% confidence interval for the mean.

 Question 4: The data set `UKDriverDeaths` can be loaded with the command

```
> data(UKDriverDeaths)
```

The data contains information about deaths due to driving. Assume it is a random sample for the distribution of deaths. Does the data appear to be normally distributed? How can you tell? If so, find a 80% confidence interval for the population mean.

 Question 5: The data set `trees` can be loaded with the command

```
> data(trees)
```

It contains three variables `Girth`, `Height` and `Volume`. Attach the data frame, then for each variable investigate if a CI based on the t-statistic is appropriate. If not say why, if so, find a 95% CI for the population mean.

0.3 CIs for two samples

Confidence intervals can be constructed for the difference of proportions and the difference of means. These take a familiar form:

$$p_1 - p_2 \pm z^* \mathbf{SE}(p_1 - p_2) \quad \text{or} \quad \bar{x_1} - \bar{x_2} \pm t^* \mathbf{SE}(\bar{x_1} - \bar{x_2}).$$

The standard errors are different though. As far as using the R functions, not much different is required. You need to specify both sets of data. This is done in the following ways

```
    prop.test(c(x1,x2), c(n1,n2), conf.level = 0.95)
t.test(x, y, conf.level = 0.95)
```

The assumptions needed are that the two samples be independent of each other, and for the t-based one, that the populations be normally distributed.

For example, a survey was done two weeks prior on the CSI student opinion about the war in Iraq and of 100 students surveyed, 52 responded favorably. Find a 95% CI for the difference in proportion. DOes it include 0?

The CI is found as follows. Notice how the numbers enter in.

```
> prop.test(c(52, 55), c(100, 120), conf.level = 0.95)

        2-sample test for equality of proportions with
        continuity correction

data:  c(52, 55) out of c(100, 120)
X-squared = 0.6018, df = 1, p-value = 0.4379
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.07992  0.20326
sample estimates:
prop 1 prop 2
0.5200 0.4583
```

For the t-test, here is a simple examle A teacher gives two tests, a random sample from each set of scores yields

```
test 1  65 78 91 87 76 90 60 50
test 2   80 75 80 70 90 95
```

Find an 90% CI for the difference of population means.

We enter the data in, check normality two ways and then apply the function

```
> x = c(65, 78, 91, 87, 76, 90, 60, 50)
> y = c(80, 75, 80, 70, 90, 95)
> plot(density(x))
> lines(density(y))
> qqnorm(x)
> qqnorm(y)
> t.test(x, y, conf.level = 0.9)

        Welch Two Sample t-test

data:  x and y
t = -1.077, df = 11.70, p-value = 0.3031
alternative hypothesis: true difference in means is not equal to 0
90 percent confidence interval:
 -18.717   4.633
sample estimates:
mean of x mean of y
    74.62     81.67
```

0.4 Equal variances for t-based CIs

Do the variances in the data appear to be the same? If so, then the degrees of freedom in calculating t^* can be made greater (so t^* is smaller) by setting the argument `var.equal=TRUE`. By default, this is `FALSE`.

Question 6: The data set `shoes` is loaded as follows

```
> data(shoes, package = "MASS")
```

It contains two variables `A` and `B` which contain measurements of shoe wear. Attach the data frame `shoe` and then find a 90% CI for the difference of means. Is the normality assumption checked? Are the variances equal? Did you use this?

Question 7: The `morley` dataset contains measurements on the speed of light by Morley and Michaelson. There are five Experiments in the data set, we compare the vales for the first two. First attach the data set and form two variables as follows

```
> data(morley)
> attach(morley)
> x = Speed[Run == 1]
> y = Speed[Run == 2]
```

DO the data sets appear to come from normal populations? Are the variances likely the same size? If so find a 90% confidence interval for the difference of means. Does it include 0?

0.5 paired data and CIs

The shoe exercise is actually a little misleading, as the two variables are not independent, but are paired off instead. The experiment was to test shoe wear with 10 children. Rather than randomly assign five kids each to the two types of shoe and measure wear, instead the researchers had each kid wear two types of shoes. The amount of wear for each data point is related then as presumable some kids wear shoes more than others. The difference in shoe wear should be attributable to the difference in shoes though.

This data is paired. In this case, the confidence interval has a different standard error. (it is just the regular standard error for the data set `A - B`). The argument `paired=TRUE` will make the adjustments. FOr example

```
> attach(shoes)
> t.test(A, B, paired = TRUE, conf.level = 0.9)

        Paired t-test

data:  A and B
t = -3.349, df = 9, p-value = 0.008539
alternative hypothesis: true difference in means is not equal to 0
90 percent confidence interval:
 -0.6344 -0.1856
sample estimates:
mean of the differences
                  -0.41
```

Find the confidence interval.

Question 8: Is this confidence interval different from the one you found in the exercise? Is it smaller, large? Does it include 0? Why is this intereting?

Question 9: Why do you think it is better to pair off the shoes, rather than have 5 measurements of shoe wear for one type and 5 for the other? Why might researchers go to the extra step of randomizing which foot the mis-matched shoe goes on?

Question 10: A twin study uses identical twins to test the effect of some drug. By doing so, genetic differences can be controlled. Suppose the measured effects of some drug treatment of pairs of twins are

```
drug              paired data
---------------------------------------
drug 1   70 80 85 90 92 65
drug 2   72 78 87 93 93 67
```

Find a 90% CI for the difference of means. Explain your assumptions (normality? equal variance? paired study?)