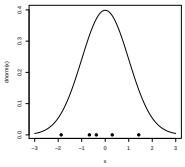

# Confidence Intervals

This project shows how R can be used to find confidence intervals when all the data is available.

## 1 CI for $\mu$ based on $\bar{x}$: `t.test()`

Suppose a random sample of gas prices in New York resulted in

```
2.69 2.72 2.69 2.79 2.65
```

Find a 95% CI for the population mean $\mu$.

We know, that if the population is normally distributed, then an answer is provided by

$$\bar{x} \pm t^* \mathbf{SE}(\bar{x})$$

This can be found in R with these commands:

```
### Use c() to combine numbers into a data vector
> gasPrices = c(2.69, 2.72, 2.69, 2.79, 2.65)
> xbar = mean(gasPrices)
> n = 5                                  # n = length(gasPrices)
> tstar = qt(0.975, df = n - 1)
> SE = sd(gasPrices)/sqrt(n)
> xbar + c(-1, 1) * SE
```

```
[1] 2.684676 2.731324
```

Not too much work, but really the computer should work do this mechanical work for us — and it can.

The function `t.test()` will perform this. The `t` refers to the sampling distribution of $(\bar{x} - \mu)/\mathbf{SE}$, the word "test" to something we haven't discussed yet. To see it work we have:

```
> t.test(gasPrices)
```

```
        One Sample t-test
data:  gasPrices
t = 116.1045, df = 4, p-value = 3.3e-08
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 2.643243 2.772757
sample estimates:
mean of x
    2.708
```

We see in the ouput the line

```
 95 percent confidence interval:
 2.643243 2.772757
```

This is the CI.

A 95% CI is the default, to change this we add the argument `conf.level=`. For instance, a 90% CI is found with

```
> t.test(gasPrices, conf.level = 0.9)

        One Sample t-test
data:  gasPrices
t = 116.1045, df = 4, p-value = 3.3e-08
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
 2.658277 2.757723
sample estimates:
mean of x
    2.708
```

### 1.1 Problems

1. The data set `women` contains height and weight data for women. Suppose it is a random sample of women from a population of interest. Find a 95% CI for the population height. (The data is found using `women$height`.)

2. What assumptions about the population were made to conclude that this CI is valid?

3. Repeat, finding a 95% CI for the population weight (`women$weight`).

4. Again, what assumptions are made about the population of weights? For height and weight, which population assumption seems less likely to actually be true?

5. The `sleep` data set contains measurements on students sleep in the variable `sleep$extra`. Find a 95% CI for the mean amount of extra sleep, if these numbers from a random sample of students.

6. Check as best you can any assumptions necessary to say that your answer is a valid 95% CI.

### 1.2 Proportions: `prop.test()`

A CI for a population proportion is given by

$$\hat{p} \pm z^* \mathbf{SE}(\hat{p}).$$

This formula can actually be improved a bit, but it is better to leave the details to the computer, which we do. The computer provides `prop.test()` for handling tests of proportions. It needs the number of successes (`x`), and the number of trials, `n`.

For instance, US presidential approval ratings are calculated from surveys of size approximately 1,000.

For instance, FOX news had these numbers: (pollingreport.com)

```
  FOX News/Opinion Dynamics Poll. Oct. 11-12, 2005. N=900
registered voters nationwide. LV = likely voters
```

> "Do you approve or disapprove of the job George W. Bush is doing as president?"

only 40% approved.

What is a 95% CI for the population proportion, assuming this is a random sample from likely voters?

Here the number who approve is

```
> 0.4 * 900

[1] 360
```

So we have the CI from

```
> prop.test(0.4 * 900, 900)

        1-sample proportions test with continuity correction
data:  0.4 * 900 out of 900, null probability 0.5
X-squared = 35.6011, df = 1, p-value = 2.421e-09
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.3679384 0.4329264
sample estimates:
  p
0.4
```

(Again, the default confidence level is 0.95, use `conf.level=` to change it.)

The CI has a margin of error of 0.0329.

### 1.3 Problems

1. Another poll had these details:

   ```
    NBC  News/Wall Street Journal Poll conducted by the polling
   organizations of Peter Hart (D) and Bill McInturff (R).
   Oct. 8-10, 2005. N=807 adults nationwide.

   "In general, do you approve or disapprove of the job that
   George W. Bush is doing as president?"
   ```

39% approved.

Find a 95% CI for the population proportion. What is the margin of error? What is the population sampled from?

2. The NBC/Wall Street Journal poll got a lot of press because there were 89 blacks polled and only 2 approved. The media said (Washington Post) "Polling free fall among blacks"

   Assuming this is a random sample from the adult African American population, what is the margin or error for this survey? Is 14% in the 95% CI.
   (What other possible errors can creep in, when the sample is small?)

3. Are gas prices dropping? A random sample of 100 gas stations found that 65% had dropped their prices in the last week. Based on this, find a 90% CI for the proportion of all gas stations that have dropped their prices in the last week.

## 2 CIs for differences of means

A CI for the difference of means is given by

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \mathbf{SE}$$

where $t^*$ comes from the $t$ distribution, but the number of degrees of freedom depends on the assumptions. As well, the formula for **SE** depends on the assumptions.

In class we have if the two samples were randomly sampled from normal populations with means $\mu_i$ and variances $\sigma_i^2$ $(i = 1, 2)$ then $\mathbf{SE} = \sqrt{s_1^2/n_1 + s_2^2/n_2}$, and the degrees of freedom are more than the smaller of $n_1, n_2$ minus 1.

The `t.test()` function will calculate CIs for this problem using a better value for the degrees of freedom. It needs to have two different samples to do its work.

For instance, suppose older gas prices from a random sample were

```
2.75, 2.89, 2.79, 2.75, 2.80, 2.82
```

Find a 95% CI for the difference in the population means (for this sample and the previous one stored in `gasPrices`.

```
> newGasPrices = c(2.75, 2.89, 2.79, 2.75, 2.8, 2.82)
> t.test(gasPrices, newGasPrices)

        Welch Two Sample t-test
data:  gasPrices and newGasPrices
t = -2.9132, df = 8.643, p-value = 0.01796
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.16389279 -0.02010721
sample estimates:
mean of x mean of y
    2.708     2.800
```

The answer

```
95 percent confidence interval:
 -0.16389279 -0.02010721
```

shows that 0 is not in the CI. This indicates in some way that the gas prices have changed. As well, find the degrees of freedom used to calculate $t^*$:

```
data:  gasPrices and newGasPrices
t = -2.9132, df = 8.643, p-value = 0.01796
```

The value of 8.643 is used, and not the minimum of 5 and 6 minus 1 (which is 4).

### 2.1 Problems

1. The data set `chickwts` has two variables `chickwts$casein` and `chickwts$meatmeal`. Both record the weights of a sample of chicks, but for two different diets. Assuming a random sample from two different populations, find a 95% CI for the difference in population means.

2. Repeat the above for the diets given in `chickwts$linseed` and `chickwts$soybean`.

3. Can you verify the assumptions made in saying this is a CI?

4. The data set `warpbreaks` contains data on the number of breaks for two different types of wool. A t.test can be found from variables `warpbreaks$breaks[warpbreaks$wool=="A"]`, but that is a *hassle* to type. Rather, the syntax

   ```
   > t.test(breaks ~ wool, data=warpbreaks)
   ```

   can be used. Read this as the variable `breaks` broken up by the levels of `wool`, where both variables are found in the data set `warp breaks`.

   Find the 95% CI for the difference in population means for the breaks variable.

5. For the `t.test()` function, the argument `var.equal=TRUE` forces an assumption of equal variances for the two populations, which in general will provide smaller confidence intervals.

   For the `warpbreak` data see how it makes a difference. (Look at the CI and the df values.

   Now how to check if the assumption makes sense for our data? A boxplot can be used to compare spreads (which the variance measures). A boxplot for the two variables is also produced with the same syntax

   ```
   > boxplot(breaks ~ wool, data=warpbreaks)
   ```

   Make this boxplot. Do the variances look equal?