



# Confidence intervals

Aren't iPods just so neat — nice design, small, powerful, easy to use. Let's have fun with iPods in a different way — learning about confidence intervals (CIs).<sup>1</sup>

## 1 Background

A confidence interval is a way of making statistical inference about a population parameter based on an observed value of a statistic. The language uses probabilities.

For example, a 95% confidence interval for the population proportion  $p$  based on a sample proportion  $\hat{p}$  is given by

$$\hat{p} - 1.96\text{SE}(\hat{p}) \leq p \leq \hat{p} + 1.96\text{SE}(\hat{p}),$$

where the *standard error* of  $\hat{p}$  is

$$\text{SE}(\hat{p}) = \sqrt{\hat{p}(1 - \hat{p})/n}.$$

To see how this is used, suppose Apple wants to release a new iPod, but would like to have an idea of how reliable the units will be. They let 200 be used internally by employees, and find that in one month only 3 broke. What does this imply about the percentage that would break if they sold 100,000?

The question of interest is the proportion,  $p$ , of broken iPods if they were sold to the public as they are currently constructed. The sample proportion,  $\hat{p} = 3/200$ , gives us a confidence interval for this unknown  $p$ . To calculate we could do this:


```

> x = 3
> n = 200
> phat = x/n
> SE = sqrt(phat * (1 - phat)/n)
> phat + c(-1, 1) * 1.96 * SE

[1] -0.001846311  0.031846311


```

We see that the proportion to break is between 0 and 3.1% with 95% confidence. (The term `c(-1,1)` does the “plus or minus” in an efficient way.)

 **Question 1:** Suppose of these 200 iPods 5 broke in the first 90 days. Find a 95% confidence interval for the proportion of iPods that would break in the first 90 days. (The warranty period.)

---

<sup>1</sup>Of course, our lawyers insist that we say iPod is a registered trademark of Apple Inc., and that this work is an act of fiction, any resemblance to persons living or dead is purely coincidental.

 Question 2: To entice buyers, Apple considers offering a free song download with an iPod purchase. Apple selected 300 registered iPod users at random and offered them this offer. Of these 300, only 75 downloaded a song. (The others presumably didn't know how, had no interest, ...) If they use this sample to infer about the population of all new iPod buyers, find a 95% confidence interval for the proportion of new iPod buyers who would take advantage of the free offer.

### 1.1 Other confidence levels

Of course, the 1.96 comes from the fact that we asked for 95% confidence intervals. In general, for a confidence level of  $1 - \alpha$ , the corresponding number,  $z^*$ , is the quantile of  $1 - \alpha/2$ . To find this on the computer we have for instance

```
> qnorm(1 - 0.05/2)
```


```
[1] 1.959964
```


```
> qnorm(1 - 0.1/2)
```

```
[1] 1.644854
```

The latter finds  $z^*$  for a 90% CI.

 Question 3: What is  $z^*$  when for a 99% confidence interval?

 Question 4: Is the user manual sufficient? Smaller manuals are of course cheaper, but users expect some answers. To find out if a proposed manual is sufficient, 100 people are given a new product with the manual and asked if the manual is sufficient. 12 respond no. Find a 90% CI for the proportion of all users who would find the manual insufficient if it were released as is.

 Question 5: Is the design of a new product too small for some hands? Making smaller and smaller iPods only goes so far, as eventually controls are too small for some hands. To see if a new model is already too small, 125 users are asked to test a prototype. Of these 3 percent complained about the size. If the prototype were released to the public, find a 85% CI for the percentage of buyers who would find it too small.

## 2 Visualizing the randomness involved

Confidence intervals are based on the *randomness* of the statistic, for instance  $\hat{p}$ . But, we have a single value of the statistic from our sample, how is it random? The key lies in realizing that the models assume you could take another sample, and another. For each of these the statistic could likely have a different value, and consequently generate a different CI. To visualize this, we use a function, `plotCI()`, that first must be downloaded:

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

50 95% confidence intervals based on sample proportion

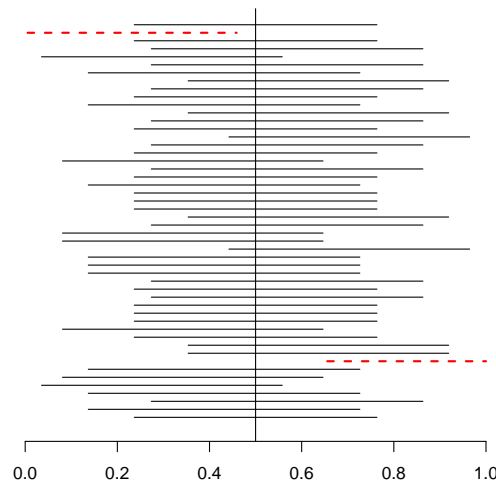





Figure 1: Plot of 50 95% confidence intervals for  $p$  (when true value is known to be  $p = 0.5$ . Those that missed are marked with dashed line and are in red.


```
> plotCI()
```

When this function is run with no arguments, as above, it produces a default graph such as Figure 1 with fifty 95% confidence intervals. For this simulation, unlike in real life, the value of  $p$  is known to be  $p = 0.5$ .

 Question 6: Run the command `plotCI()`. What percent of the CIs missed the true parameter?

 Question 7: The argument `conf.level=0.80` will set the confidence level. Add this argument. Now what percent of the CIs missed the true parameter?

 Question 8: The argument `n=1000` will use  $n = 1000$  and not 10 in the simulation. This should make the confidence intervals smaller. By how much? Either think hard, or compare simulations run with  $n = 1000$  and without specifying  $n$ .

 Question 9: You can plot 100 CIs by specifying `m=100`, rather than the default 50. Repeat with  $n = 500$  and  $m = 200$ . How many missed the true parameter value?

### 3 Simplifying the computing

The computer has a built in function, `prop.test()`, to compute confidence intervals for  $p$  based on the sample proportion. As both  $\hat{p}$  and  $n$  are needed, the function requires values for  $x = n\hat{p}$  and  $n$  in that order.

For instance, if a survey of 500 users finds 15 dissatisfied with the battery life of their iPod, what does this imply about the proportion of all iPod owners?

Assuming a random sample, we can get a 95% confidence interval with

```
> prop.test(x = 15, n = 500)
```

```
1-sample proportions test with continuity correction
```


```
data: 15 out of 500, null probability 0.5
X-squared = 439.922, df = 1, p-value = < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.01750499 0.05012619
sample estimates:
      p
0.03
```

The confidence interval appears in the lines:


```
...
95 percent confidence interval:
 0.01750 0.05013
...
```


We see between 1.75% and 5%.

(The answer is slightly different than had we calculated with  $\hat{p} \pm z^* \text{SE}(\hat{p})$ .)

 Question 10: Will customers pay extra for a sports case? If so, it might be sold separately, otherwise, it might be included in as an incentive. To find out, a test of 225 iPod users was set up, of which 25 bought the sports case.

Find 95% confidence intervals for the proportion of all iPod users who would buy this sports case.

 Question 11: Is an interface intuitive? A new interface is tested by 200 users, and 92% find it easy to learn. Find a 95% confidence interval for all new users, if we can assume these 200 were a random sample from this population.

 Question 12: The extra argument `conf.level=0.80` will return an 80% confidence interval. Repeat the last exercise, only find an 80% CI.

## 4 Confidence intervals for $\mu$ based on $\bar{x}$

The sample mean,  $\bar{x}$ , may be used to form confidence intervals for a population mean. Suppose,  $\bar{x}$  is the sample mean of a random sample of size  $n$  from a population with mean  $\mu$  and standard deviation  $\sigma$ . Let  $s$  be the sample standard deviation.

If the population is assumed to be normal, or  $n$  is large, then a  $(1 - \alpha)100\%$  CI for  $\mu$  based on  $\bar{x}$  is

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

where  $\mathbf{SE}(\bar{x}) = s/\sqrt{n}$  is the *standard error* of  $\bar{x}$  and  $t^*$  satisfies

$$P(-t^* \leq T_{n-1} \leq t^*) = 1 - \alpha$$

with  $T_k$  having a  $t$  distribution with  $k$  degrees of freedom.

Values of  $t^*$  can be found from `qnorm()` after specifying values of  $\alpha$  and  $n$ , for example:

```
> alpha = 0.05
> n = 10
> tstar = qt(1 - alpha/2, df = n - 1)
```

This all follows from the fact that for a random sample from a normal population that the sampling distribution of

$$T = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{\text{observed} - \text{expected}}{\mathbf{SE}},$$

is the  $t$  distribution with  $n - 1$  degrees of freedom.

To illustrate, suppose a new iPod holds 512 megabytes. How many songs is this? Of course it varies as some songs are shorter, some longer. But if we knew the average length of a song, we could divide it into 512 and get an answer. To find out the average length of song, suppose Apple advertising employees looked at their own iPods, and found that of the 1,225 songs, the average size per song was 4.12 megabytes with a standard deviation of 1.25 megabytes. Using this collection as a random sample from the population of all songs people would have on an iPod leads to a 95% CI for the mean of

```
> xbar = 4.12
> s = 1.25
> n = 1225
> SE = s/sqrt(n)
> tstar = qt(1 - 0.05/2, df = n - 1)
> xbar + c(-1, 1) * tstar * SE
```

```
[1] 4.049932 4.190068
```

Using the worst case of 4.20, we can estimate the number of songs as

```
> 512/4.2
```

```
[1] 121.9048
```



Question 13: Why is it safe to assume that the sampling distribution of

$$T = \frac{\bar{x} - \mu}{\text{SE}(\bar{x})}$$

is the  $t$  distribution?



Question 14: At an Apple store, the “genius bar” is responsible for handling repairs and other technical questions. Suppose the average waiting time for 10 randomly chosen customers is 17 minutes with a standard deviation of 5 minutes. If it is assumed that the population of waiting times is normally distributed, use this sample to find a 90% CI for the mean waiting time.

#### 4.1 The `t.test()` function

If the full data set is available, and stored in a data vector `x`, then the command `t.test(x)` will produce a 95% CI based on the mean and standard deviation of `x`.

For instance, suppose the sale price with shipping of 5 iPod minis on eBay were

255.00 257.23 223.00 275.00 249.00

Treating these numbers as a random sample for a normal population find a 95% CI for the mean price of all iPod minis sold on eBay.

We enter the data, and call the function as follows:

```
> x = c(255, 257.23, 223, 275, 249)
> t.test(x)
```

One Sample t-test

```
data: x
t = 29.9389, df = 4, p-value = 7.413e-06
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 228.4905 275.2015
sample estimates:
mean of x
 251.846
```

Again, the CI appears in the output:

```
95 percent confidence interval:
 228.5 275.2
```



Question 15: The battery length of a type of iPod is claimed to be 12 hours, but this varies from player to player. Suppose a random sample of 10 iPods revealed these playback times (in minutes):

672 789 663 725 778 549 506 559 655 709

Find a 90% CI for the mean playback time, assuming this sample comes from a normal population.



Question 16: A person with a 20 gigabyte iPod may not fill it to its entirety. An informal sample of size 8 found this many gigabytes full:

13 20 9 8 17 16 16 1

Find a 90% CI for the mean amount of storage space used by 20-gigabyte iPod users. What assumptions do you make about the data?



Question 17: The `plotCI()` function will plot CIs for the mean if the argument `type="mean"` is given. For instance

```
> plotCI(type = "mean")
```

Run the command above. How many of the CIs miss the true value of  $\mu$ ?



Question 18: Different confidence levels can be specified with `plotCI()`. For instance, 80% CIs are produced with

```
> plotCI(type = "mean", conf.level = 0.8)
```

Run this command. How many of the CIs miss the true value of  $\mu$ ?



Question 19: In 50 samples, a 90% CI is expected to contain the true value of  $\mu$  45 times, but the actual value is random. Specify the distribution of the number of CIs that contain the true value of  $\mu$  in this instance.



Question 20: It is claimed that the assumption of a normally distributed population can be relaxed, as long as the parent distribution is not too skewed. That is, the assumptions are *robust* to small differences in the population. You can verify this by checking if the number of CIs that miss is dramatically different from the expected number when different populations are used.

To use a different population for the simulations is done by specifying the family with `family=` (available types are "norm", the default; "exp", with parameter `rate=`; "t", with parameter `df=`, and `unif`, with parameters `min=` and `max=`).

For instance to use a long-tailed population we could try

```
> plotCI(type = "mean", family = "t", df = 2)
```

A skewed distribution would be

```
> plotCI(type = "mean", mu = 1, family = "exp", rate = 1)
```

Try both and see if the number of intervals missing is much larger or smaller than expected.