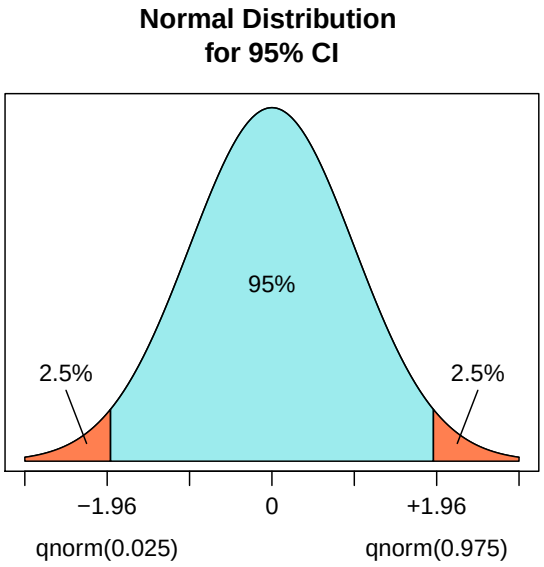
Manually estimating confidence intervals for a mean

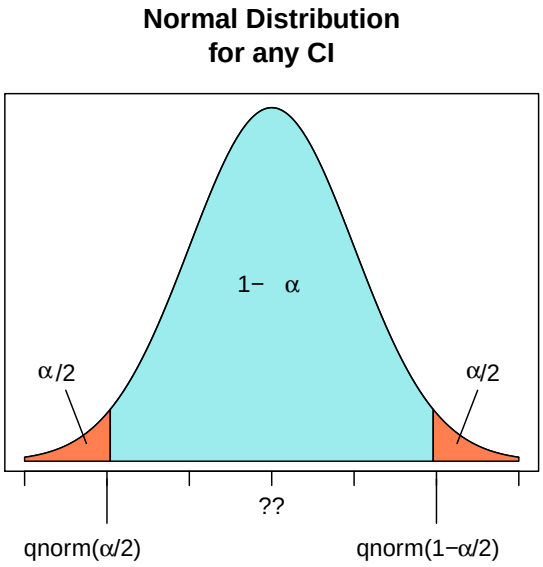
Daloha Rodríguez-Molina

November 5, 2015

I'll try to replicate Riccardo's explanation about how to derivate the way we manually calculate confidence intervals in R, from the standard normal distribution.

All of the formulas we use from now on can be graphically understood by looking at these two graphs:





With known variance

When we know the variance, we can use the standard normal distribution (Gauss distribution) and the `qnorm` function to estimate our confidence intervals.

Let's assume that our point estimate value is called `z`, which in the plots above should be along the x axis somewhere in the turquoise area (1 - α). We want to construct confidence intervals around this point estimate. The confidence intervals will be located at the line between the turquoise and the orange sections.

In the specific case of constructing 95% CI (left plot), we assume that our point estimate `z` will be between the lower and upper bound of the confidence interval, which in the z-scale with `mean = 0` correspond to `-1.96` and `+1.96`, respectively. For the general case (right plot), the formula is:

$$Pr\left[\ qnorm\left(\frac{\alpha}{2}\right)\ <\ z\ <\ qnorm\left(1-\frac{\alpha}{2}\right)\ \right] = 1-\alpha$$

From here, we can start to mathematically derive the whole thing. First, we should remember that `z` can also be expressed as:

$$z = \frac{\overline{x}-\mu}{\sqrt{\frac{\sigma^2}{n}}}$$

Then, we can subtitute `z` in the first equation with the value of the second equation:

$$Pr\left[\ qnorm\left(\frac{\alpha}{2}\right) < \frac{\overline{x}-\mu}{\sqrt{\frac{\sigma^2}{n}}} < \ qnorm\left(1-\frac{\alpha}{2}\right)\ \right] = 1-\alpha$$

Now we want to leave our ?? alone, so we do a little bit of algebra, and we end up with:

$$Pr\left[\overline{x} - qnorm\left(1-\frac{\alpha}{2}\right) \times \sqrt{\frac{\sigma^2}{n}} > \ \mu > \ \overline{x} + qnorm\left(1-\frac{\alpha}{2}\right) \times \sqrt{\frac{\sigma^2}{n}}\right] = 1-\alpha$$

Finally, since the standard normal distribution is symmetric, we can collapse both sides of the inequality using a $\pm$ sign:

$$CI_{1-\alpha/2} = \overline{x} \pm qnorm\left(1-\frac{\alpha}{2}\right) \times \sqrt{\frac{\sigma^2}{n}}$$

Where the lower bound results from the subtraction, and the upper bound results from the sum between the terms:

$$LowerCI_{1-\alpha/2} = \overline{x} - qnorm\left(1-\frac{\alpha}{2}\right) \times \sqrt{\frac{\sigma^2}{n}}$$

$$UpperCI_{1-\alpha/2} = \overline{x} + qnorm\left(1-\frac{\alpha}{2}\right) \times \sqrt{\frac{\sigma^2}{n}}$$

In `R`, we can do this by using the code in Slide 3 - Lecture 4 of the R-course:

```
low <- mean.x - qnorm(0.975)*2/sqrt(n)
 up <- mean.x + qnorm(0.975)*2/sqrt(n)
```

So, if we generate a random sequence from a normal distribution using the function `rnorm()`, which has `mean=1`,`sd=2` and sample size `n=100`, we can do:

```
set.seed(42)                          # set the seed to always generate the same 'random' sequence
x <- rnorm(100,mean=1,sd=2)             # generate sequence and store it in x
mean.x <- mean(x)                       # calculate mean of x and store in mean.x
n.x <- 100                                # sample size is 100
sd.x <- 2                                 # standard deviation is known, and it is 2
low <- mean.x - qnorm(0.975)*sd.x/sqrt(n.x)     # low bound of the CI
up <- mean.x + qnorm(0.975)*sd.x/sqrt(n.x)      # high bound of the CI
c(low,up)                               # look at both lower and upper bound of the CI
```

```
## [1] 0.6730368 1.4570224
```

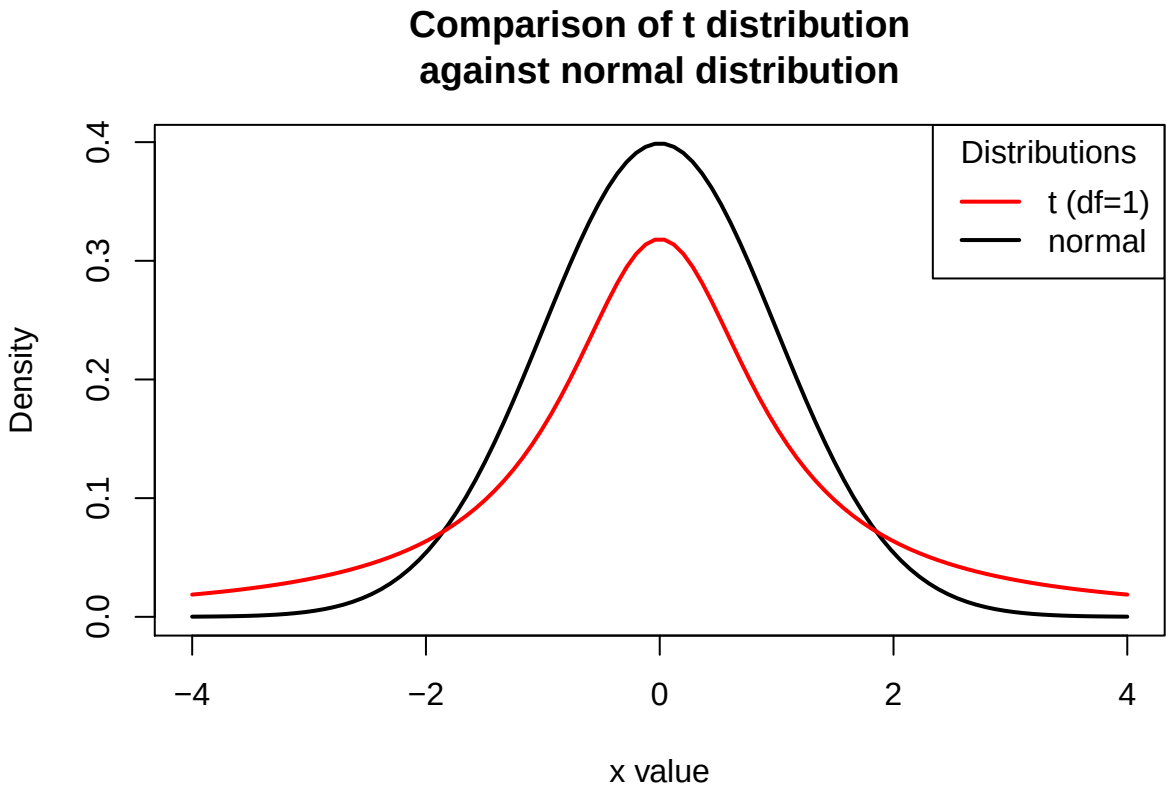
Look how the mean of x is within the boundaries:

```
mean.x
```

```
## [1] 1.06503
```

With unknown variance

For the unknown variance the procedure is pretty much the same, except that now we want to use the Student's t distribution because we need to take into account **extra uncertainty** from the fact that **we don't know** what the variance is:



See how the red line leaves more space near the extremes? This is the t-distribution accounting for uncertainty against the normal distribution (the black line).

In R, we calculate the confidence interval for unknown distribution following the instructions in Slide 3 - Lecture 4 of the R-course. Look how we use the `qt()` function instead of the `qnorm()` function we used before:

```
low <- mean.x - qt(0.975)*2/sqrt(n)
 up <- mean.x + qt(0.975)*2/sqrt(n)
```

So, we can use the same example as before:

```r
set.seed(42)                                  # Always generate the same 'random' sequence
x <- rnorm(100,mean=1,sd=2)                   # generate sequence and store it in x
mean.x <- mean(x)                             # calculate mean of x and store in mean.x
n.x <- 100                                    # sample size is 100
sd.x <- sd(x)                                 # standard deviation is not known
low <- mean.x - qnorm(0.975)*sd.x/sqrt(n.x)   # low bound of the CI
up <- mean.x + qnorm(0.975)*sd.x/sqrt(n.x)    # high bound of the CI
c(low,up)                                     # look at both lower and upper bound of the CI
```

```
## [1] 0.6568252 1.4732341
```

Notice how now the confidence intervals are wider than before, and the mean of x is still within the boundaries:

```r
mean.x
```

```
## [1] 1.06503
```

Conclusions:

- Confidence intervals are interval estimates.
- It is advised to **always** use a point estimate along with a confidence interval.
- The most used confidence intervals are 0.90, 0.95 and 0.99.
- 1 - α is called *confidence level*.

Bonus

- For a more detailed explanation on how to derive the equations, check out this YouTube video.
- You can find the R Scripts for this entry: fig.1, fig.2, Confidence Interval Estimation.

End of script