Estimating the mean of a normal population (small sample)

When to use this

- You want to estimate the mean of some characteristic of a population that can be modeled reasonably well by a normal distribution.
- You have a sample that is **not** large enough to treat the sample standard deviation as reliable (say, $n < 30$).

Assumptions

Although the underlying theory assumes a normal distribution for the data, the point estimate is unbiased regardless of whether the data is normal or not.

The theory underlying interval estimates does depend on the data having a normal distribution, but if you have a large sample the mean $\overline{y}$ will be approximately normally distributed by the Central Limit Theorem.

With a small sample, this is a bit more of a stretch. A good approach is rather than thinking "Is my data normal?" think "Is the normal distribution a good model for my data?".

Point estimates

The point estimate for the population mean μ is the sample mean $\overline{y}$:

$$\hat{\mu} = \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Interval estimates

Confidence intervals are constructed using the fact that under the normality assumptions the derived quantity

$$\frac{\overline{y} - \mu}{s}$$

is approximately distributed as a standard normal, that is, $N(0, 1)$

The formula for the upper and lower bounds of a $100(1 - \alpha)\%$ confidence interval for μ is:

$$\overline{y} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$$

Where:

- $\overline{y}$ is the sample mean
- s is the sample standard deviation
- n is the sample size
- $t_{\alpha/2,n-1}$ is the value t such that the proportion of the polulation lying between $-t$ and t is $1 - \alpha$

You will also see this analysis describe as "sigma unknown" rather than "small sample". It just means the sample standard deviation s is not considered reliable enough to replace σ in the formula.

The only difference between "large sample" and "small sample" procedures is that for the "large sample" version you use a normal or z distribution to compute the confidence interval, and the "small sample" version uses a t distribution.

The t distribution converges very rapidly to the z distribution once the sample size reaches about 50, so in practice unless you have a small ($n < 30$) sample, the results of the two versions will be almost identical.