Confidence Intervals for the Sample Mean

As we saw before, parameter estimators are themselves random variables. If we are going to make decisions based on these uncertain estimators, we would benefit from having a way to quantify how **confident** we are in those estimates. A estimate that is just composed of a single number can't convey information about the reliability too. To quantify the degree of (un)certainty in an estimate, it is useful to construct an **interval** in to which the parameter falls with some certainty.

We will explore this idea in the context of using the sample mean to estimate the expected value, but the concept can apply to any parameter estimate.

The scenario is the same as in the last section: we are given a series of independent and identically distributed data $X_1, X_2, \ldots, X_N$ whose mean $\mathrm{E}[X_i] = \mu$ is unknown. We specify a desired **confidence level** $1 - \alpha$, for some $0 \leq \alpha \leq 1$, and then use the data to construct a *lower estimator* M_N^- and *upper estimator* M_N^+ for μ designed so that

$$\mathrm{P}\left(M_N^- \leq \mu \leq M_N^+\right) \;\geq\; 1 - \alpha.$$

We call $[M_N^-, M_N^+]$ a $1 - \alpha$ *confidence interval.*

As before, the M_N^- and M_N^+ are random variables that depend on the observations $X_1, X_2, \ldots, X_N$. In the probability expression above, it is not μ which is random, but rather the end points of the interval.

Example: Suppose that $X_1, X_2, \ldots, X_N$ are independent and identically distributed (iid) normal random variables with *unknown mean* μ and *known variance* σ^2. We use the sample mean to estimate μ:

$$M_N = \frac{X_1 + X_2 + \cdots + X_N}{N}.$$

In this case M_N is also normal with mean μ and variance σ^2/N:

$$M_N \ \sim\ \text{Normal}\left(\mu, \frac{\sigma^2}{N}\right).$$

We set $\alpha = 0.05$, and try to form a 95% confidence interval.

In other words, we would like to find an interval around M_N that contains μ with probability 0.95. We will use M_N as the center point in this interval, and then take the endpoints to be

$$M_N^- = M_N - \ell, \quad \text{and} \quad M_N^+ = M_N + \ell,$$

for some appropriately chosen ℓ. We may not always want symmetric intervals, but it is a fine place to start.

We know that

$$\mathrm{P}\left(M_N - \ell \le \mu \le M_N + \ell\right) = \mathrm{P}\left(-\ell \le M_N - \mu \le \ell\right)$$

and since $M_N - \mu$ is a normal random variable with mean 0 and variance σ^2/N,

$$\begin{aligned}\mathrm{P}\left(-\ell \le M_N - \mu \le \ell\right) &= \Phi\left(\frac{\ell}{\sqrt{\sigma^2/N}}\right) - \Phi\left(\frac{-\ell}{\sqrt{\sigma^2/N}}\right) \\ &= 2\Phi\left(\frac{\ell}{\sqrt{\sigma^2/N}}\right) - 1.\end{aligned}$$

Where last last step comes from recalling that $\Phi(-y) = 1 - \Phi(y)$. Thus we want to choose ℓ such that

$$1 - \alpha = 2\Phi\left(\frac{\ell}{\sqrt{\sigma^2/N}}\right) - 1$$

which means

$$\ell = \Phi^{-1}(1 - \alpha/2) \cdot \sqrt{\frac{\sigma^2}{N}}.$$

For $\alpha = 0.05$, we look at a table for $\Phi(\cdot)$, we find that

$$\Phi(1.96) = 0.975 = 1 - \alpha/2,$$

and so

$$\left[\underbrace{M_N - 1.96\sqrt{\frac{\sigma^2}{N}}}_{M_N^-}, \quad \underbrace{M_N + 1.96\sqrt{\frac{\sigma^2}{N}}}_{M_N^+} \right]$$

is a 95% confidence interval.

Normality. The above relies on the fact that we know the distribution of $M_N - \mu$ to construct the confidence interval — $M_N - \mu$ was normal, so we used the corresponding (inverse) cdf for our calculations.

In general, the distribution of the data $X_1, \ldots, X_N$ may not be known. However, it is often the case that $M_N - \mu$ is very well approximated by a normal random variable. Recall the the *central limit theorem*, which says that if the $X_1, \ldots, X_N$ are independent and have finite variance, then their sum becomes a normal distribution as N gets large. In turn, this means that

$$\frac{M_N - \mu}{\sqrt{\mathrm{var}(M_N)}} \ \rightarrow \ \mathrm{Normal}(0,1)$$

as $N \rightarrow \infty$.

If we can calculate $\mathrm{var}(M_N)$, and this quantity is independent of μ, we can construct the confidence intervals in exactly the same way as above no matter what the distribution of the X_i are, provided, of course, that N is large enough.

If not ...

Intervals based on estimator variance approximations

Let's see how we might construct confidence intervals for the sample mean when the **variance is unknown**. That is, the variance is another parameter ν which needs to be estimated.

Given $X_1, \ldots, X_N$ we form

$$M_N = \frac{X_1 + \cdots + X_N}{N}$$

and *estimate* the variance ν with

$$\widehat{S}_N^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(X_i - M_N\right)^2.$$

Recall from the previous section that $\mathrm{E}[\widehat{S}_N^2] = \nu$. We can use this expression to estimate the variance of the sample mean by $\widehat{S}_N^2/N$:

$$\mathrm{var}(M_N) \ \approx \ \widehat{S}_N^2/N$$

Then provided that N is large enough so that this approximation is acceptable, we can form the $1-\alpha$ confidence interval

$$\left[\ M_N - z\,\frac{\widehat{S}_N}{\sqrt{N}}\ \ ,\ \ M_N + z\,\frac{\widehat{S}_N}{\sqrt{N}}\ \right]$$

where z satisfies

$$\Phi(z) = 1-\alpha/2 \quad \Leftrightarrow \quad z = \Phi^{-1}(1-\alpha/2).$$

Example: Polling

You are polling a population to get an idea of who will win the next election. A certain proportion p of voters will vote for candidate R while the proportion $(1-p)$ will vote for candidate D. If we select N people from the population at random, and ask them who they will vote for, we can model the responses as independent Bernoulli random variables:

$$X_i = \begin{cases} 1, & \text{(vote for } R\text{), probability } p \\ 0, & \text{(vote for } D\text{), probability } 1-p \end{cases}.$$

The X_i have mean $\mathrm{E}[X_i] = p$; this mean is unknown and we are trying to estimate it (this is the whole point of the poll).

We estimate p using the sample mean:

$$M_N = \frac{X_1 + \cdots + X_N}{N}.$$

Suppose we poll $N = 100$ people, and 58 of them reply that they will support R in the upcoming election. What is a 95% confidence interval for p?

From the information above, we can quickly see that the sample mean is $M_N = 58/100 = 0.58$. We can also compute the sample variance:

$$\begin{aligned}\hat{S}_N^2 &= \frac{1}{99}\sum_{i=1}^{N}(X_i - 0.58)^2 \\ &= \frac{1}{99}\sum_{i=1}^{N} X_i^2 - 1.16X_i + 0.3364.\end{aligned}$$

Now since in this case the X_i are 0 or 1, $X_i^2 = X_i$, and so

$$\hat{S}_N^2 = \frac{(1-1.16)58 + 33.64}{99} = 0.2461.$$

A 95% percent confidence interval corresponds to $\alpha = 0.05$, and so with

$$z = \Phi^{-1}(1-\alpha/2) = \Phi^{-1}(0.975) = 1.96,$$

our interval is

$$\left[\ M_N - z\frac{\widehat{S}_N}{\sqrt{N}}\ \ ,\ \ M_N + z\frac{\widehat{S}_N}{\sqrt{N}}\ \right]\ =\ [0.58{\pm}0.0972]\ =\ [0.4828,\ 0.6772].$$

So while data you have collected indicate that R is a favorite, the 95% confidence interval still contains values of p that are less than 0.5.

We can generalize the calculation above. If I ask N people, and K say they are voting for R, then

$$M_N = \frac{K}{N}, \quad \text{and} \quad \hat{S}_N^2 = \frac{K - K^2/N}{N-1},$$

and so the 95% confidence interval is

$$\left[\frac{K}{N} - 1.96\sqrt{\frac{K - K^2/N}{N(N-1)}}\ ,\ \ \frac{K}{N} + 1.96\sqrt{\frac{K - K^2/N}{N(N-1)}}\right].$$

Now if I ask $N = 5000$ people and $K = 2900$ say they are voting for R, the 95% confidence interval for p becomes much tighter:

$$[0.58 \pm 0.0137] \ = \ [0.5663,\ 0.5937].$$

Worst-case variance. We know that the variance of a Bernoulli random variable X_i is $p(1-p)$. Since $0 \leq p \leq 1$, we know that $\mathrm{var}(X_i) \leq 1/4$, and that this upper bound is achieved when $p = 1/2$. Also note that polls are most informative when p is near 0.5, that is, it is not obvious who is going to win.

The moral is that we lose almost nothing by simply replacing $\hat{S}_N^2$ with $1/4$ in the calculations above. The intervals would be slightly larger, but not by much — in the $N = 100, K = 58$ example, we would have found a 95% confidence interval of $[0.4820, 0.6780]$; in the $N = 5000, K = 2900$ example, the interval would have been $[0.5661, 0.5939]$.

Student's t-distribution

The methodology in the previous section is based on two approximations:

1. the estimation error $M_N - \mu$ is (at least close to) normal,
2. $\widehat{S}_N^2/N$ is (at least close to) the variance of M_N.

There are important situations where item 1 is reasonable, but item 2 is not ... one such scenario is if the X_i are indeed normal (so item 1 is exactly true) with unknown mean and variance, and N is small.

In this section, we consider the following problem:

We observe $X_1, X_2, \ldots, X_N$ that are normally distributed with unknown mean μ and variance v, and we wish to construct a $1-\alpha$ confidence interval for the mean μ.

As before, we estimate the mean and variance using

$$M_N = \frac{1}{N}\sum_{i=1}^{N} X_i, \qquad \widehat{S}_N^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - M_N)^2.$$

The normalized error is

$$T_N \quad = \quad \frac{M_N - \mu}{\widehat{S}_N/\sqrt{N}} \quad = \quad \frac{\sqrt{N}(M_N - \mu)}{\widehat{S}_N}.$$

Even when $M_N - \mu$ is exactly normal, T_N is not normal, since $\widehat{S}_N$ is itself a random variable (as it is computed from the data). Thus using the normal cdf Φ to construct confidence intervals might not be a good idea.

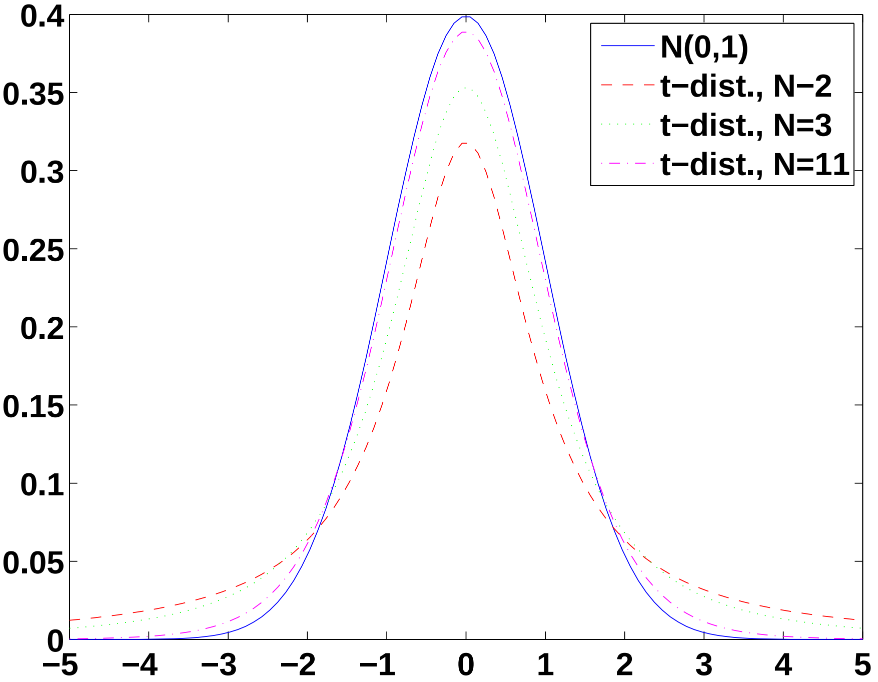
However, the distribution for T_N is known; it is called **Student's t-distribution with $N-1$ degrees of freedom**. The pdf is

$$f_{T_N}(t) = \frac{\Gamma(N/2)}{\sqrt{(N-1)\pi}\cdot\Gamma((N-1)/2)}\cdot\left(1+\frac{t^2}{N-1}\right)^{-N/2},$$

where $\Gamma(n)$ is the gamma function

$$\Gamma(n) = \int_0^\infty t^{n-1}\mathrm{e}^{-t}\,\mathrm{d}t \qquad (= (n-1)!\text{ when } n \text{ is an integer}).$$

As N gets large, the t-distribution becomes normal, but for N small it has much "heavier tails":



Thus for small samples sizes, it is usually better to use the cdf $\Psi_{N-1}(z)$ of the t-distribution to construct the confidence intervals in place of the normal cdf Φ.

Like the normal distribution, the cdf $\Psi_{N-1}(z)$ for the t-distribution

$$\Psi_{N-1}(z) = F_{T_N}(z) = \frac{\Gamma(N/2)}{\sqrt{(N-1)\pi}\cdot\Gamma((N-1)/2)} \int_{-\infty}^{z} \left(1 + \frac{t^2}{N-1}\right)^{-N/2} \,\mathrm{d}t$$

does not have a nice closed-form expression. We will again use a chart when we need to know actual values. Since we will be only be using $\Psi_{N-1}(z)$ to construct confidence intervals, we will tabulate the **inverse** values, which are more convenient.

$\beta \rightarrow$

		0.1000	0.0500	0.0250	0.0100	0.0050	0.0025	0.0010
$N-1 \rightarrow$	1	3.0777	6.3138	12.7062	31.8205	63.6567	127.3213	318.3088
	2	1.8856	2.9200	4.3027	6.9646	9.9248	14.0890	22.3271
	3	1.6377	2.3534	3.1824	4.5407	5.8409	7.4533	10.2145
	4	1.5332	2.1318	2.7764	3.7469	4.6041	5.5976	7.1732
	5	1.4759	2.0150	2.5706	3.3649	4.0321	4.7733	5.8934
	6	1.4398	1.9432	2.4469	3.1427	3.7074	4.3168	5.2076
	7	1.4149	1.8946	2.3646	2.9980	3.4995	4.0293	4.7853
	8	1.3968	1.8595	2.3060	2.8965	3.3554	3.8325	4.5008
	9	1.3830	1.8331	2.2622	2.8214	3.2498	3.6897	4.2968
	10	1.3722	1.8125	2.2281	2.7638	3.1693	3.5814	4.1437
	11	1.3634	1.7959	2.2010	2.7181	3.1058	3.4966	4.0247
	12	1.3562	1.7823	2.1788	2.6810	3.0545	3.4284	3.9296
	13	1.3502	1.7709	2.1604	2.6503	3.0123	3.3725	3.8520
	14	1.3450	1.7613	2.1448	2.6245	2.9768	3.3257	3.7874
	15	1.3406	1.7531	2.1314	2.6025	2.9467	3.2860	3.7328
	16	1.3368	1.7459	2.1199	2.5835	2.9208	3.2520	3.6862
	17	1.3334	1.7396	2.1098	2.5669	2.8982	3.2224	3.6458
	18	1.3304	1.7341	2.1009	2.5524	2.8784	3.1966	3.6105
	19	1.3277	1.7291	2.0930	2.5395	2.8609	3.1737	3.5794
	20	1.3253	1.7247	2.0860	2.5280	2.8453	3.1534	3.5518
	50	1.2987	1.6759	2.0086	2.4033	2.6778	2.9370	3.2614
	100	1.2901	1.6602	1.9840	2.3642	2.6259	2.8707	3.1737
	150	1.2872	1.6551	1.9759	2.3515	2.6090	2.8492	3.1455
	∞	1.2816	1.6449	1.9600	2.3263	2.5758	2.8070	3.0902

Inverse of $\Psi_{N-1}(z)$.
Top row: Tail probability β.
Left column: Number of degrees of freedom $N-1$.
Entries: A value z such that $\Psi_{N-1}(z) = 1 - \beta$.

For example, by looking in the first column of the second row, we can say

$$\mathrm{P}\left(T_3 > 1.886\right) = 0.1$$

and by looking in the 5th column of the 10th row,

$$\mathrm{P}\left(T_{11} > 3.169\right) = 0.005.$$

Example: You are brewing a large batch of beer, and are trying to measure the pH level of the mash. You make five measurements; each measurement is the true pH level plus a random error that is normally distributed with zero mean and unknown variance. Assume that the errors in the observations are independent. The following measurement are made:

$$4.53, \quad 6.72, \quad 5.56, \quad 5.17, \quad 5.56$$

Compute a 95% confidence interval ($\alpha = 0.05$) using the t-distribution.

With $N = 5$, we can quickly compute the sample mean and variance

$$\begin{aligned} M_N &= 5.5080, \\ \widehat{S}_N^2 &= 0.6363, \end{aligned}$$

and $\widehat{S}_N/\sqrt{N} = 0.3567$.

We want an ℓ such that for the unknown mean μ,

$$\mathrm{P}\left(M_N - \ell \ \leq \ \mu \ \leq \ M_N + \ell\right) = 0.95,$$

which is the same thing as asking for an ℓ which obeys

$$\mathrm{P}\left(\frac{-\ell}{\widehat{S}_N/\sqrt{N}} \ \leq \ \frac{M_N - \mu}{\widehat{S}_N/\sqrt{N}} \ \leq \ \frac{\ell}{\widehat{S}_N/\sqrt{N}}\right) = 0.95.$$

Since $(M_N - \mu)\sqrt{N}/\widehat{S}_N$ has a t-distribution with $N-1=4$ degrees of freedom, this is the same as asking for an ℓ which obeys

$$\Psi_4\left(\frac{\ell}{0.3567}\right) - \Psi_4\left(\frac{-\ell}{0.3567}\right) = 0.95.$$

Since the Student's t-distribution is symmetric about the origin, we can also use the fact that $\Psi_{N-1}(-z) = 1 - \Psi_{N-1}(z)$. This leads us to an equivalent condition

$$2\Psi_4\left(\frac{\ell}{0.3567}\right) - 1 = 0.95,$$

or

$$\Psi_4\left(\frac{\ell}{0.3567}\right) = 0.975 = 1 - 0.025.$$

Using the 4th row of the third column in the t-table yields $\Psi_4(2.776) = 1 - 0.025 = 0.975$, so we choose

$$\ell = 0.3567 \times 2.776 = 0.9903.$$

Thus our 95% confidence interval is

$$\left[M_N - 2.776\tfrac{\widehat{S}_N}{\sqrt{N}}\ ,\ M_N + 2.776\tfrac{\widehat{S}_N}{\sqrt{N}}\right] = [M_N - \ell\ ,\ M_N + \ell] = [4.512,\ 6.498]\,.$$

Note that if we had used the normal tables, we would have had the more optimistic (i.e. tighter) interval

$$\left[M_N - 1.96\frac{\widehat{S}_N}{\sqrt{N}}\ ,\ M_N + 1.96\frac{\widehat{S}_N}{\sqrt{N}}\right] = [4.819,\ 6.207]\,.$$