Confidence Intervals I

We have seen how we can learn about the expected value of a random variable — The Law of Large Numbers ensures that, whatever the probability distribution, as we take more and more independent observations, we can be more and more confident that the sample average will fall close to the expected value.

The discipline of Statistics is responsible for making this process precise — for a specific probability distribution and sample size, statistics shows how confident we can be that the sample mean will fall within a specified distance of the expected value.

Example

We can observe independent random variables $X_1, X_2, \ldots, X_n$, each of which has the same probability distribution, the normal(μ,σ^2) (assume σ^2 is known).

This is often stated as "$X_1, X_2, \ldots, X_n$ are independent and identically distributed normal(μ,σ^2)."

Or more briefly, "$X_1, X_2, \ldots, X_n$ are iid normal(μ,σ^2)."

We want to learn about the unknown mean, μ.

We know that if n is large, the random variable $\overline{X}_n$ *will probably fall close to μ, so that if we observe the value* $\overline{X}_n = \bar{x}_n$*, we can have some confidence that μ is near our observation,* $\bar{x}_n$*.*

Here is how this intuitive reasoning is made precise:

Because $\overline{X}_n$ has a N(μ, σ^2/n) distribution, we know that the <u>standardized</u> (sample mean) random variable

$$\frac{\overline{X}_n - E[\overline{X}_n]}{\sqrt{Var[\overline{X}_n]}} = \frac{(\overline{X}_n - \mu)}{\sqrt{\sigma^2/n}} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} = Z$$

has a standard normal, N(0,1), distribution.

And we know that the probability that this random variable will exceed 1.96 is only 0.025, so the probability that it will fall inside the interval between -1.96 and 1.96 is 0.95. That is, we can be pretty confident (more precisely, the probability is 0.95) that this will occur:

$$P(-1.96 < \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} < 1.96) = 0.95$$

Now we simply rearrange this expression:

$$0.95 = P(\ -1.96 \ < \ \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \ < \ 1.96 \ )$$

$$= P(\ -1.96\frac{\sigma}{\sqrt{n}} \ < \ (\overline{X}_n - \mu) \ < \ 1.96\frac{\sigma}{\sqrt{n}} \ )$$

$$= P(\ -\overline{X}_n - 1.96\frac{\sigma}{\sqrt{n}} \ < \ -\mu \ < \ -\overline{X}_n + 1.96\frac{\sigma}{\sqrt{n}} \ )$$

$$= P(\ +\overline{X}_n + 1.96\frac{\sigma}{\sqrt{n}} \ > \ +\mu \ > \ +\overline{X}_n - 1.96\frac{\sigma}{\sqrt{n}} \ )$$

$$= P(\ \overline{X}_n - 1.96\frac{\sigma}{\sqrt{n}} \ < \ \mu \ < \ \overline{X}_n + 1.96\frac{\sigma}{\sqrt{n}} \ )$$

This last way of writing the expression shows that the probability that the interval centered at $\overline{X}_n$,

$$\left[\overline{X}_n - 1.96\ \sigma/\sqrt{n}\ ,\ \overline{X}_n + 1.96\ \sigma/\sqrt{n}\right]$$

or

$$\overline{X}_n \pm 1.96\sigma/\sqrt{n}$$

will contain the unknown constant, μ, is 0.95.

The interval, $\overline{X}_n \pm 1.96\sigma/\sqrt{n}$***, is a "95% Confidence Interval."***

It is a <u>random</u> <u>interval</u> whose probability of including the fixed, but unknown constant, μ, is 0.95.

If we use the sample mean as an <u>estimate</u> of the expected value, then the error in that estimate will be $\overline{X}_n - \mu$, and the magnitude of the error will be $\left|\overline{X}_n - \mu\right|$.

Since $\overline{X}_n$ is a random variable, we cannot say exactly how large the error will be. But we can be 95% certain that the error will be less than '1.96 SD($\overline{X}_n$)', or "1.96 standard errors."

There is an important difference between the

random interval $\left[\bar{X}_n - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X}_n + 1.96\frac{\sigma}{\sqrt{n}}\right]$,

which contains the true mean, μ , 95% of the time

$$P\left(\bar{X}_n - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X}_n + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

and the **fixed interval** $\left[\mu - 1.96\frac{\sigma}{\sqrt{n}}, \mu + 1.96\frac{\sigma}{\sqrt{n}}\right]$,

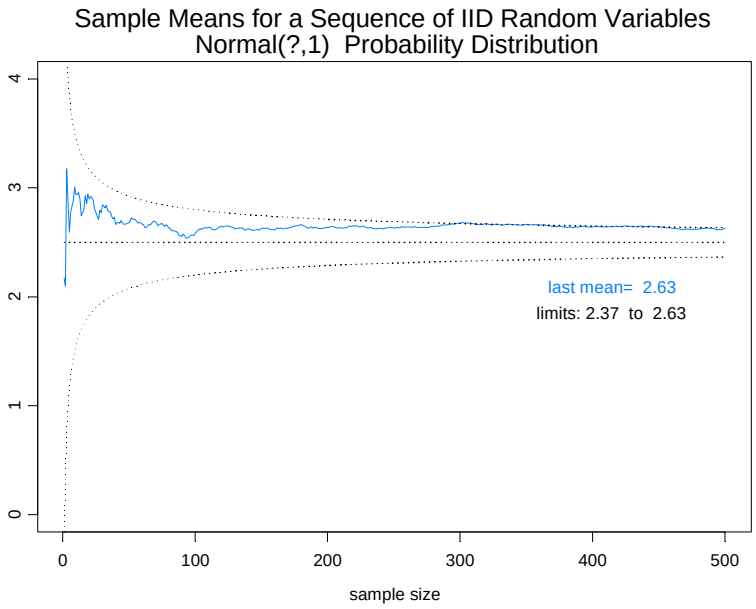
which contains the sample mean, $\bar{X}_n$, 95% of the time

$$P\left(\mu - 1.96\frac{\sigma}{\sqrt{n}} < \bar{X}_n < \mu + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

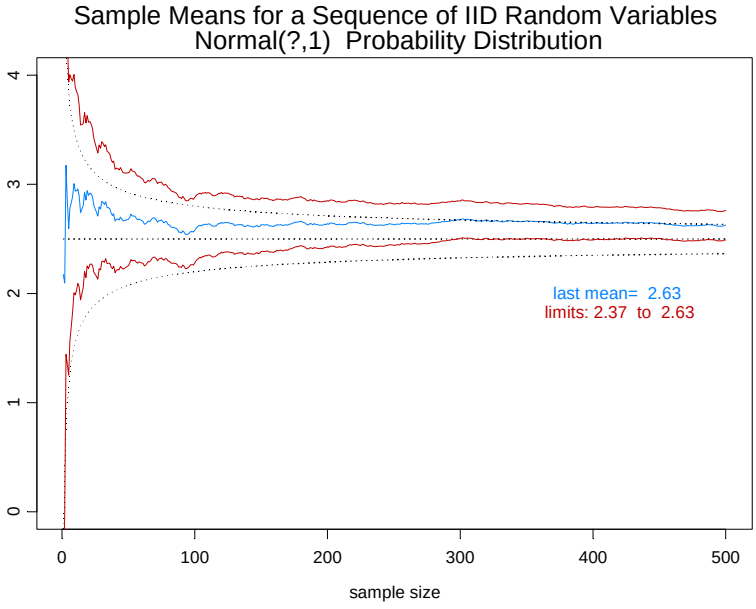
Even though both intervals are derived from the same probability statement:

$$P\left(-1.96 < \frac{\sqrt{n}\left(\bar{X}_n - \mu\right)}{\sigma} < 1.96\right) = 0.95$$

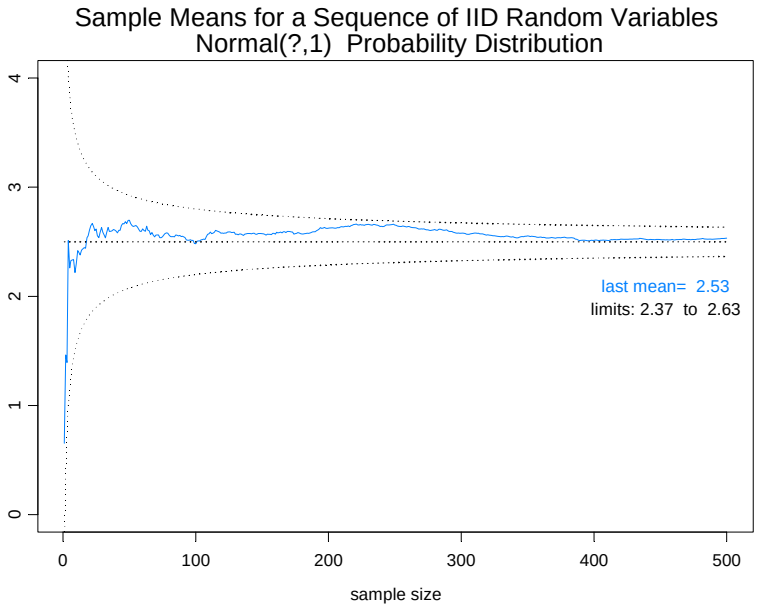
Fixed interval



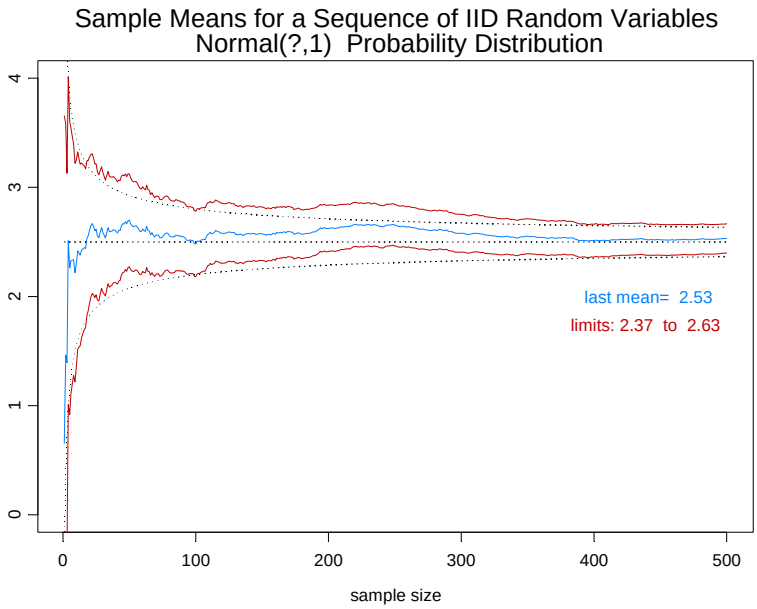
Random interval



Fixed interval



Random Interval



Let's illustrate this procedure in another way, using the Pagano and Gauvreau example of serum cholesterol levels. Suppose men's cholesterol levels are normally distributed with some mean, μ, and standard deviation, σ = 46. (The mean happens to be 211, but let's pretend we don't know that.) We observe the levels in 12 men, and here are the results (mg/100ml) :

277, 103, 202, 160, 280, 214, 259, 141, 221, 260, 237, 198

The mean of this sample is 212.56, so the 95% CI for μ, the expected serum cholesterol level in men like these, is

$$212.56 \pm 1.96\ (46)/\sqrt{12}$$

$$212.56 \pm 26.03$$

$$(\ 186.53\ ,\ 238.59\ )$$

I repeated this process, this time observing a *different* sample with mean 211.32. The 95% CI for μ based on this sample is

$$211.32 \pm 1.96\ (46)/\sqrt{12}$$

$$211.32 \pm 26.03$$

$$(185.29\ ,\ 237.35\ )$$

When I repeated the process a third time, I observed yet another sample with a mean of 223.74, so the 95% CI for μ is

$$223.74 \pm 1.96\ (46)/\sqrt{12}$$

$$223.74 \pm 26.03$$

$$(197.71\ ,\ 249.77\ )$$

Each time I draw a sample and construct a CI, the probability is 0.95 that my procedure will yield an interval that includes μ, so if I construct many of these intervals, about 95% of them will contain μ.

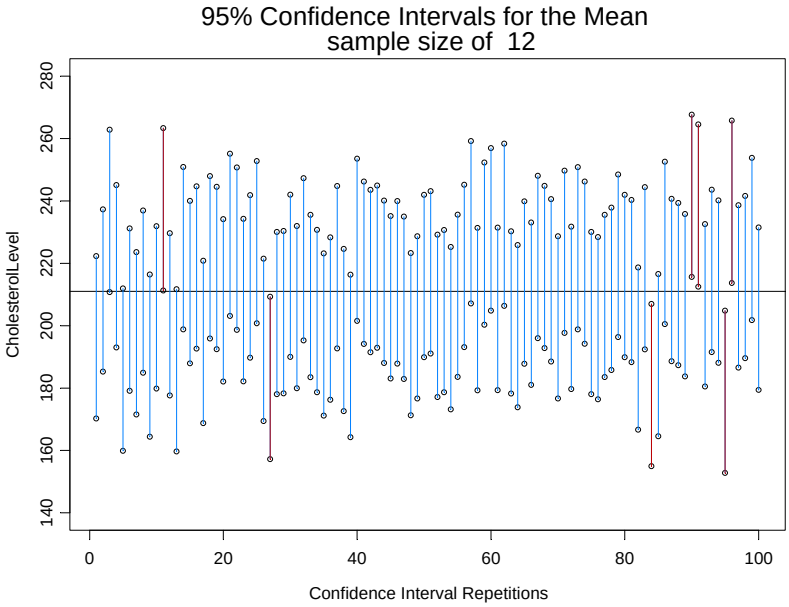
In a real application I will usually observe only one sample, and construct one CI. Therefore I will have no way of knowing whether that interval, e.g. the last one above, (197.71 , 249.77), does or does not contain μ.

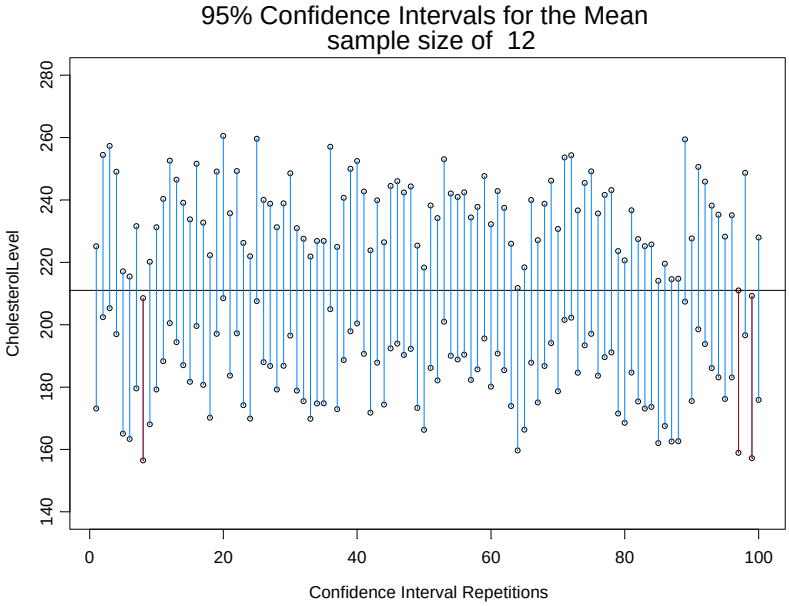
> *My* **confidence** *that it does is justified by the fact that it was produced by a procedure that is successful 95% of the time.*

I generated, not 3, but 100 samples of size 12. The next page shows all 100 of the 95% confidence intervals.

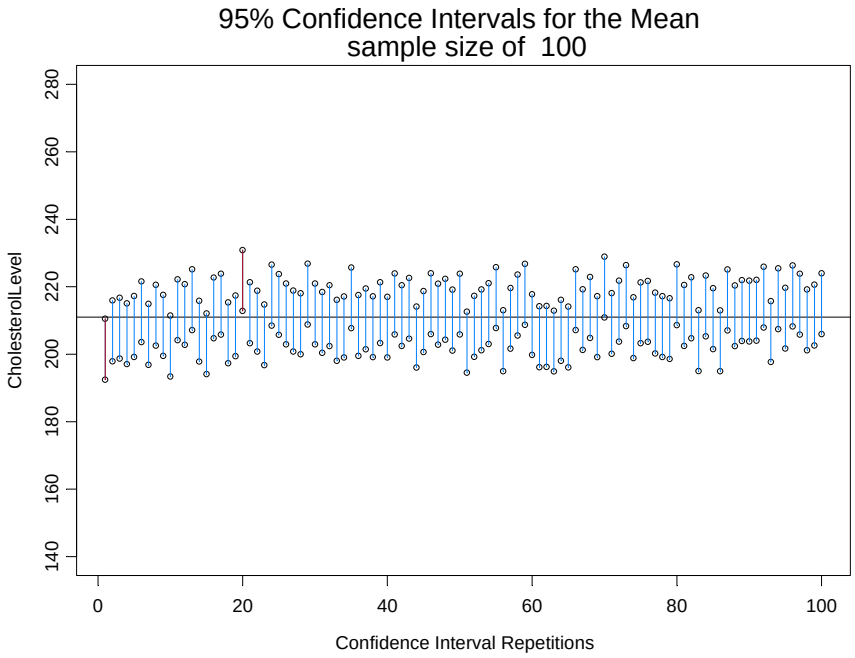
Since we know the value of the mean, (μ=211), used to generate these observations, we can see whether each interval contains that value or not.

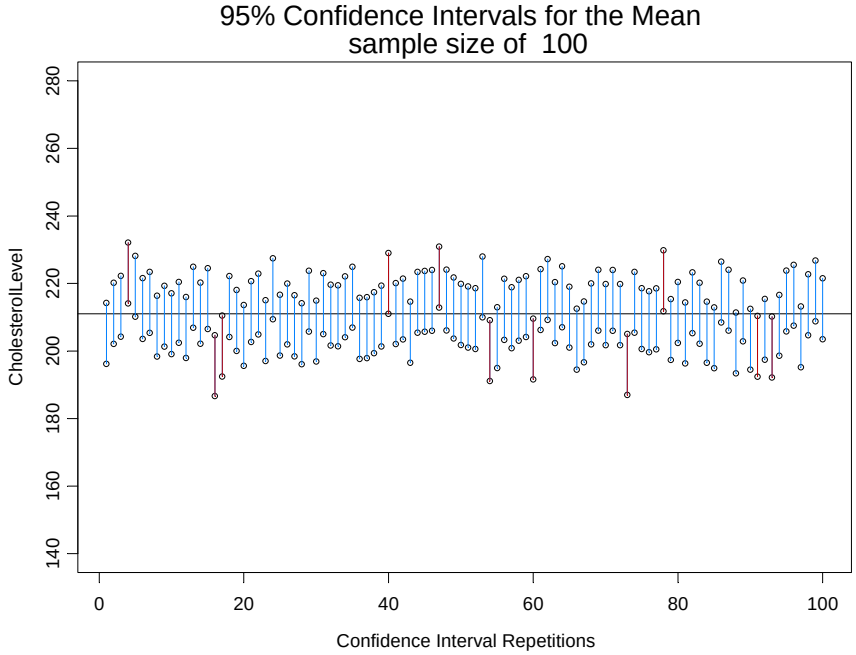
Repetitions with sample size of 12:





Repetitions with sample size of 100:





If 95% is not a sufficiently high degree of confidence, then, instead of the value 1.96, for which P(-1.96 < Z < 1.96) = 0.95, we can find a value in for which the probability is greater.

If we want 99% confidence, this value is 2.576 (P(-2.576 < Z < 2.576) = 0.99), so that

$$\overline{X}_n \pm 2.576 \frac{\sigma}{\sqrt{n}}$$

is a 99% confidence interval.

Similarly, since P(-1.645 < Z < 1.645) = 0.90, the interval

$$\overline{X}_n \pm 1.645 \frac{\sigma}{\sqrt{n}}$$

is a 90% confidence interval.

For the last of our the three samples of n=12 cholesterol levels, where the sample mean was 223.74, the 90% confidence interval is

$$223.74 \pm 1.645(46)/\sqrt{12}$$

$$223.74 \pm 21.84$$

$$(201.90 , 245.58)$$

This interval was produced by a procedure that is successful in covering μ only 90% of the time, so it deserves somewhat less confidence than the 95% CI, which was

$$(197.71 , 249.77).$$

And the 99% CI,

$$223.74 \pm 2.576 (46)/\sqrt{12}$$

$$223.74 \pm 34.21$$

$$(189.53 , 257.95)$$

is produced by a procedure that misses μ only 1% of the time, deserves more confidence.

Note on Terminology and Symbols

The coverage probability of a confidence interval is usually denoted by $1-\alpha$. It is called the confidence level, or the confidence coefficient. The probability that the CI misses the true value of the parameter is α.

A 95% CI has confidence coefficient $1-\alpha = 0.95$, so

$$\alpha = 0.05.$$

A 99% CI has confidence coefficient $1-\alpha = 0.99$, so

$$\alpha = 0.01.$$

To construct an approximate two-sided $100(1-\alpha)\%$ confidence interval, you often need to find the value, z, that cuts off probability $\alpha/2$ in the upper tail of the standard normal distribution. Most writers use the notation $z_{1-\alpha/2}$ to represent this value.

Others use a different symbol, e.g. $z_{\alpha/2}$, for the same thing, while still others, like Pagano and Gauvreau simply call it z.

Confidence Intervals for the Normal Mean When the Standard Deviation is Unknown

Recall that in order to get the confidence interval that we have been looking at, we used the fact that

$$Z = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}$$

has a probability distribution (the standard normal) that does not depend on μ or σ.

This meant that we could find numbers a and b so that P(a<Z<b) = 0.95, then rearrange this expression to find a 95% confidence interval for μ.

When σ is unknown, we need another expression, similar to the above Z, whose probability distribution can be tabulated, but which, unlike Z, does not depend on the unknown standard deviation, σ.

This problem was solved by an English statistician named William Gosset, who showed in 1908 that the simple solution of substituting the sample variance, S_n^2, for σ^2 "works."

Gosset, who published his findings under the pseudonym "Student," discovered that when the *X*'s are i.i.d. $N(\mu,\sigma^2)$, the random variable,

$$T_{n-1} = \frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{S_n}$$

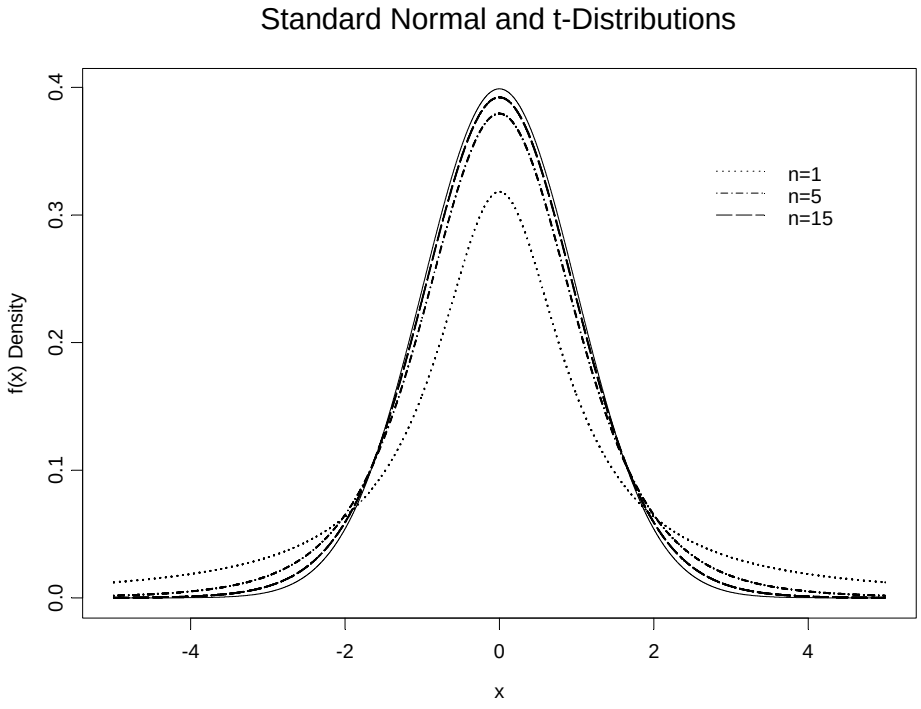
has a probability distribution with the key property that we need — it does not depend on either of the parameters, μ and σ^2.

It is known as "Student's t distribution." (Like the standard normal, it has a bell-shaped density that is symmetric about zero. The values that cut off certain upper tail areas of this distribution (0.10, 0.05, 0.025, etc.) are given in Table A.4.)

What we're doing is simply estimating the unknown variance, σ^2, by the sample variance,

$$S_n^2 = \frac{\sum_1^n (X_i - \overline{X}_n)^2}{n-1}.$$

We call S_n^2 the "sample variance" to distinguish it from σ^2, the variance of the probability distribution (σ^2 is often called the "population variance").


Standard Normal and t-Distributions
f(x) Density
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
x
n=1
n=5
n=15

Because the *X*'s are random, S_n^2 is random, too, and skilled mathematical statisticians have proven that the expected value of this random variable is precisely the thing that we're using it to estimate: $E(S_n^2) = \sigma^2$.

These diligent members of that most admirable of professions have also proved (using the Law of Large Numbers) that, just as $\overline{X}_n$ converges to μ, S_n^2 converges to σ^2 as n grows.

We use the "t table" (Table A.4) in the same way that we used the standard normal table before: We find in the table the value,
t_{n-1} , for which

$$P(-t_{n-1} < T_{n-1} < t_{n-1}) = 0.95,$$

then rearrange the inequalities to get a 95% confidence interval for μ.

This interval, is $\left[\overline{X}_n - t_{n-1}\frac{S_n}{\sqrt{n}}, \overline{X}_n + t_{n-1}\frac{S_n}{\sqrt{n}}\right]$, depends only on the observable quantities, $\overline{X}_n$ and S_n , plus the number , t_{n-1} , that we find in the "t table."

The rearrangement goes just like it did when the variance, σ^2, was known: Dropping the subscripts on $\overline{X}_n$, S_n, and t_{n-1} to make the expressions less cluttered,

$$0.95 = P(\; -t \; < \; T_{n-1} \; < \; t \;)$$

$$= P(\; -t_{n-1} \; < \; \frac{\sqrt{n}(\overline{X} - \mu)}{S_n} \; < \; t_{n-1} \;)$$

$$= P(\; -t_{n-1}\frac{S_n}{\sqrt{n}} \; < \; (\overline{X} - \mu) \; < \; t_{n-1}\frac{S_n}{\sqrt{n}} \;)$$

$$= P(\; -\overline{X} - t_{n-1}\frac{S_n}{\sqrt{n}} \; < \; -\mu \; < \; -\overline{X} + t_{n-1}\frac{S_n}{\sqrt{n}} \;)$$

$$= P(\; +\overline{X} + t_{n-1}\frac{S_n}{\sqrt{n}} \; > \; +\mu \; > \; +\overline{X} - t_{n-1}\frac{S_n}{\sqrt{n}} \;)$$

$$= P(\; \overline{X} - t_{n-1}\frac{S_n}{\sqrt{n}} \; < \; \mu \; < \; \overline{X} + t_{n-1}\frac{S_n}{\sqrt{n}} \;)$$

This last way of writing the expression shows that the random interval,

$$\left[\overline{X}_n - t_{n-1}\ \frac{S_n}{\sqrt{n}}, \overline{X}_n + t_{n-1}\ \frac{S_n}{\sqrt{n}}\right],$$

or

$$\overline{X}_n \pm t_{n-1}\frac{S_n}{\sqrt{n}}\ ,$$

will contain μ (the expected value of $\overline{X}_n$) with probability 0.95.

The comparable interval, when the variance is known, is

$$\overline{X}_n \pm 1.96\frac{\sigma}{\sqrt{n}}\ .$$

We can't write a specific number, like 1.96, for the 95% t-interval because there is a different value, t_{n-1}, for every sample size n. The quantity n-1 is called the "degrees of freedom" (df).

When n=6, for example, if we look in Table A.4, under the column "Area in One Tail " = 0.025, we find, for n-1 = 5 degrees of freedom, the value, t_5 =2.571.

This means that for n=6, (5 df) the 95% confidence interval (CI) is

$$\overline{X}_6 \pm 2.571 \frac{S_6}{\sqrt{6}},$$

For n=81, (80 df) we find that t_{80} =1.990, so the 95% CI is

$$\overline{X}_{81} \pm 1.990 \frac{S_{81}}{\sqrt{81}},$$

In the previous section we used three samples of n = 12 cholesterol levels to illustrate the procedure for setting confidence intervals for the mean of a normal distribution when standard deviation is known (σ was 46, in that example).

Let's pretend that we don't know the standard deviation, and see how the confidence intervals for μ look in this case.

From Table A.4 we find that with n-1 = 11 df, $P(T_{11} > 2.201) = 0.025$. Therefore the 95% "Student's t" confidence interval is $\overline{X}_{12} \pm 2.201 S_{12}/\sqrt{12}$.

The first of our three samples had a mean of 212.56 and a sample standard deviation of 55.63. For this sample the 95% confidence interval for the expected serum cholesterol level is

$$212.56 \pm 2.201(55.63)/\sqrt{12}$$

$$212.56 \pm 35.35$$

$$(177.21 , 247.91)$$

The second sample, whose mean was 211.32, had a standard deviation of 47.08, so the 95% CI is

$$211.32 \pm 2.201(47.08)/\sqrt{12}$$

$$211.32 \pm 29.91$$

$$(181.40 , 241.23)$$

For the third sample, with mean 223.74 and standard deviation 55.98, the 95% Student's t CI is

$$223.74 \pm 2.201(55.98)/\sqrt{12}$$

$$223.74 \pm 35.57$$

$$(188.17 , 259.31)$$

The last line in the t-table shows that, as the df grows, the value in each column approaches the corresponding value in the standard normal table. This means that when n is large, the t distribution (with n-1 df) and the standard normal distribution are approximately the same.

For example, the value that cuts off the upper 2.5% of the t distribution approaches 1.96, the value that cuts off the upper 2.5% of the standard normal distribution. And the value that cuts off 5% of the t distribution approaches 1.645, which is the 5% cutoff value for the normal.

The numbers in the t table are always greater than the corresponding ones in the normal table. This is because of the additional variability that is introduced when we replace σ by its estimate, S_n. To allow for the additional variability, the Student's t CI must tend to be wider than the corresponding (normal) CI.