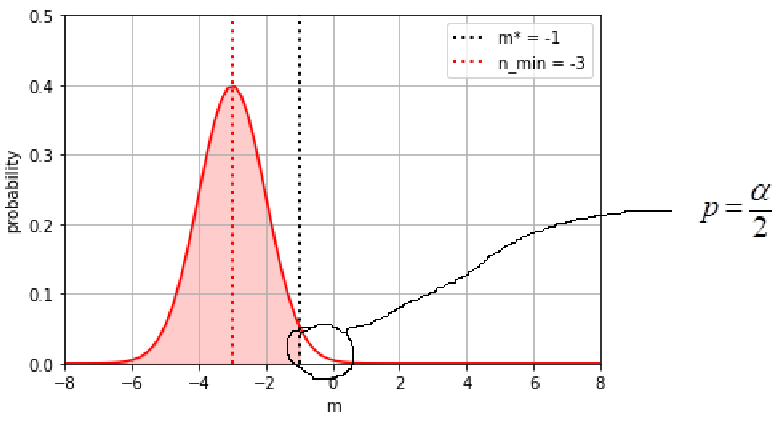
**Confidence Intervals**

Determining confidence intervals is kind of, but not simply, the reverse of doing a hypothesis test. In this case consider a set of possible statistics {η} for the population. And then the commensurate set of probability distributions for some sample measurement {Pη(m)}. And say we measure m = m\*. We’d like to know what this tells us about η, specifically, what range of values η is likely to span, at the α level of significance, where α can range between 0 and 1. We can frame this as a Hypothesis test. Apropos the minimum value, ηmin, we may frame a ‘Null Hypothesis’ such as, ‘the true value of η is less than ηmin’. And we may ask for what value of ηmin would the measurement of m = m\* cause us to reject this Null Hypothesis at the p = α/2 significance level. Mathematically, we’d be looking for the ηmin such that,



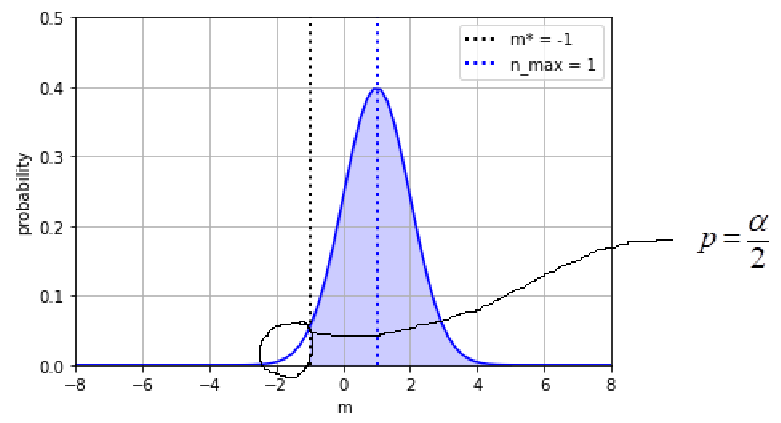
Graphically, it would look like this (integrated region is the circled, white region):



We can reason similarly apropos the max value, ηmax. We may frame a ‘Null Hypothesis’ such as, ‘the true value of η is greater than ηmax’. And we may ask for what value of ηmax would the measurement of m = m\* cause us to reject this Null Hypothesis at the p = α/2 significance level. Mathematically, we’d be looking for,



Graphically, it would look like this (integrated region is the circled, white region):



Its worth simplifying these formulas for the special, and ubiquitous, case of the measurement following a symmetric distribution, Pη(m), like a normal distribution, or Student’s T distribution. Let’s say the probability distribution is translationally invariant w/r to η, so that Pη(m) = P0(m-η). And we’ll also, eventually, presume it’s symmetric such that P0(-m) = P0(m). Then our formulas simplify to:



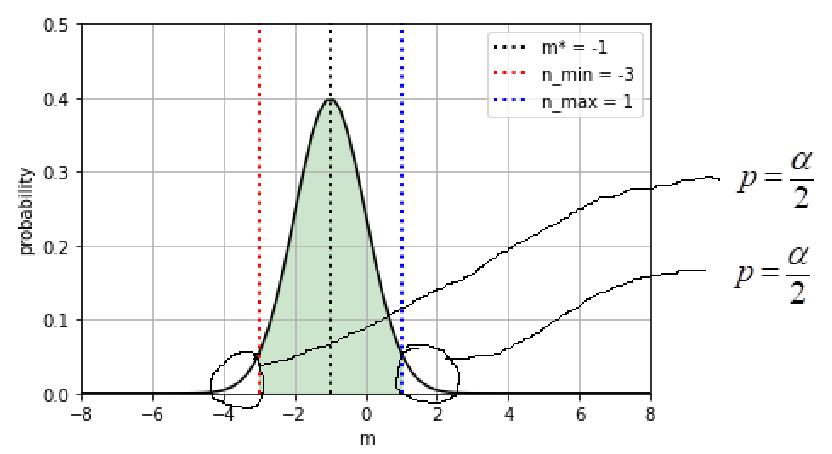
Adding the two together gives us:



So finally we have:



So given our symmetric, translationally invariant assumptions, we can take the statistic η to be given by m\*, calculate the sample measurement probability distribution accordingly, and then work out the left and right bounds as the min and max values of η. Graphically it looks like this:



ηmin and ηmax would symmetrically situated about the point estimator, η = m\*. So we can go a little bit further. Like with the normal distribution, let’s introduce a variable



where σm\* is the standard deviation of Pη=m\*(m), and m\* is the average (necessarily, since Pη=m\*(m) is symmetric about m\*). Then we can say Pη=m\*(m) is a function of z, which we’ll just call Pη=m\*(z). Let zα designate the point to the right of which lies an area α under the sample distribution curve Pη=m\*(z). Then



And also,



So our range is:



**Example**

Say we have an exponential distribution, P = (1/λ)e-x/λ. Suppose we measure a value of x and get x\*. What is a 1-α confidence interval for λ? The exponential distribution is not a symmetric distribution, so we have to use the more general formulation. So we’d do:



and we’d also say,



So our interval would be:



**Example**

Let’s revisit the prior example. Say we have an exponential distribution, P = (1/λ)e-x/λ. Suppose we measure n (a large number) values of x and get an average x\*. What is a 1-α confidence interval for λ? There’s a couple different approaches. So our measurement is:



And our measurement would be m\* = (1/n)Σi=1..n(xi). And as we saw in the Multiple Variables PD file, m would follow a Gamma distribution,



with mean and variance λ, λ2/n, respectively. So m\* is a point predictor for λ. So the first approach we could take would be like above,



It might be onerous to integrate the Γ function. Another option is presume n is large (enough). Then we can approximate Pλ(m) with a normal distribution,



It is the case, in this case, that Pλ(m) = P(m-λ), and that Pλ(m) is symmetric. So it follows that an approximate 1-α confidence interval for λ would be:



Say we measure the decay times of 50 nuclei and find an average value of 32ns. Then a 95% confidence interval for the decay time would be:



**Example**

In a previous example it seemed somewhat unlikely that male/female births are truly random events, but that there is probably some biological mechanism behind the plurality. Now we’d like to find a say 95% confidence interval for the actual probability of male births. Specifically, this means we want to find a set of p-values, {p} (that’s our statistic η), for which the associated probability distributions Pp(x) (that’s the Pη(m) obviously) would encompass the measurement x = 530 (that’s m\*) within at least 95% of their area (which is two std’s from mean of course, as we’re going to approximate the binomial distribution as a normal one for simplicity).

For instance, if p = 0.53, then mean = 530, std = (npq)0.5 = 15.8 and getting getting 530 male births would be exactly 0 std away, which is less than 2. So p = 0.53 is possibility. If p = 0.52, then mean = 520, std = (npq)0.5 = 15.8 still, and getting 530 births would be 10/15.8 std’s away from mean and so still a possibility. Similarly for p = 0.51. Say p = 0.5 though. Then mean = 500, std = (npq)0.5 = 15.8-ish still. And so getting 530 births would be 30/15.8 std’s away, which is still barely less than 2, and so acceptable. Anything much less than this won’t work. So we expect minimum p is about pmin = 0.50. And by symmetry we’d expect maximum p can be is about pmax = 0.56. Anyway, working this out for real….So,



Solving for p…



Guess we’ll call = x/n. Then, we can say,



So this is the exact answer, but if n is large (probably had to be anyway, to use normal distribution approximation) then this is:



Where we recognize = √(pq) = √(p(1-p))as the sample estimate of the variance of the population proportion. This formula compares well with what we’ll get in the next example using the normal distribution, and even better with what we’ll get when we do same with Student’s T distribution. Anyway, plugging in x = 530, n = 1000, and α = 0.05, we get:



**Example**

How many births should we record, if we want to be able to estimate the true probability/ratio of births to within 0.001 with a 95% certainy?



From formula above, we need,



So our result depends on what we think p is. Say we estimate p = 0.5. This is also the largest n would be:



So yeah.

**Example**

Say we have a population of 10,000 wolves and we want to know their average height, μ, (that’s the statistic η). So we sample 100 of them and find for our sample that the mean is = 82cm (that’s the measurement m\*). So what is a 95% confidence interval for the true population mean? Let’s say the sample variance is 8.

So we have to find the set of population means, {μ}, that would produce sample probability distributions, Pμ(), for which = 82 would be within their 95% area range. There’s a problem though. We need to presuppose more than just the population average, μ, in order to ascertain the probability distribution Pμ(). We need to know the population variance, σ2, too. An approximate way to deal with this is to take the variance of our sample and naively equate it to the population variance (should use the 1/√(n-1) formula here, though, rather than 1/√n formula right?).

Okay, well if we do, then Pμ() would be, in the large n limit, normally distributed with mean μ , and var = σ2/n. So we want,



And we get:



Observe similarity with previous example’s answer. Filling in numbers,



This method works pretty well for n > 30 or so. And it doesn’t require the underlying population heights to be normally distributed for the average height to be. And note how if we wanted to tighten the interval, we could do so by putting more people in our sample. Another interesting thing that makes clear our analysis is only approximate, is that if we made n = 10,000, we would still have error bars on our estimate, even though it would be exact.

**Example**

Say we have a population of 10,000 wolves and we want to know their average height, μ, (that’s the statistic η). So we sample 100 of them and find for our sample that the mean is = 82cm (that’s basically the measurement, m\*). We’ll also presume to find a sample std dev s = 8. So what is a 95% confidence interval for the true population mean?

So we have to find the set of population means, {μ}, that would produce sample probability distributions, Pμ(), for which = 82 would be within their 95% area range. Last time we had to presume by the CLT that the sample mean was normally distributed, since we didn’t know anything about the underlying population. But if we presume the underlying population heights to be normally distributed, then there is ameasurement whose probability distribution depends only on the presumption of the average, μ, of that underlying population, and that’s the *sample* z measurement.



So z will be our measurement m, and as can see, z only presumes the population average. And it is known that z is Student’s T distributed, where n = size of sample.



We do need to know within what interval z needs to be to encompass 95% of the area under a Student’s T distribution function.



So we have,



So our answer is basically the same. Only thing that’s different is zα/2. It’s value is different than that of the normal distribution. Turns out for ν = n – 1 = 99 d.o.f., this is practically 2 still. So our calculation looks like before:



But for ν = n – 1 = 9, Zα/2 would be more like 2.23. Say our sample size were n = 20, and we wanted a 90% confidence interval. Then ν = 19, and zα/2 = 1.73. And we’d have:



**Example**

Suppose we’re manufactoring a drug, which is supposed to have an active ingredient with amount 100μg ± 1μg. We take a random sample of 50 tablets, and find a sample standard deviation of 1.5μg. What’s a 95% confidence interval for the variance?

We can do a χ2 test. So recall from the Multiple Variable PDF file that the sample variance, S2, is χ2 distributed with n-1 = 49 d.o.f., according to,



Let our measurement m be W. Since the χ2 distribution is not symmetric (though with 49 d.o.f., it is probably pretty close), we’ll proceed with the more general formulation of the confidence interval. So we want a σmax such that the probability that we’d find m < m\* = (50-1)(1.5)2/σmax2 = 110.25/σmax2 is α/2. So,



where cdfχ2(x) is the χ2 cumulative distribution function, and cdfχ2-1(x) is its inverse (we can access these guys from scipy). And the lower bound for σ2 is:



So we have: σ2 ∈ (1.57, 3.49), or σ ∈ (1.25, 1.87). Reassuringly, this interval covers our sample standard deviation of 1.5.