**Working with MLEs**

There are a lot of nice properties of MLEs!!!

Invariance property of maximum likelihood estimators

Rather than the MLE of θ, suppose we are interested in the MLE of some function of θ, say log(θ). The MLE of log(θ) is simply .

One simple example is with the MLE for the coefficient of variation σ/μ. This is a measure of variability per mean unit. For example, this is used in finance to understand the risk of an investment. If a population can be characterized by a normal PDF, the MLE is .

Normality of maximum likelihood estimators

A maximum likelihood estimate is the actual calculated value based on a sample. Of course, these estimates can vary from one sample to the next. Therefore, we can think of these as random variables. The random variable version is often referred to as a maximum likelihood estimator. Notice that both can use “MLE” as an acronym.

For example, the maximum likelihood estimate for π is  = /n in the field goal kicking example. The maximum likelihood estimator for π is  = /n.

Because a maximum likelihood estimator is a random variable, it has a probability distribution! A very important general property of a MLE is

For a large sample, MLEs can be treated as normal random variables with a mean equal to the parameter it is estimating and a variance computed from the second derivative of the log likelihood function.

Why do you think this property is important?

In general terms, suppose θ is the only parameter of interest for a random sample Y1, …, Yn from a population characterized by the PDF f(y). For a large sample, the random variable version of the MLE of θ, , has an approximate normal distribution with mean θ and variance



where the expectation is taken with respect to Y1, …, Yn. The result here holds under general conditions. Exceptions include instances when θ may be the maximum or minimum possible value of Yi (e.g., uniform distribution).

We can then use this result to find a (1-α)100% CI for θ:



where



is the ESTIMATED variance for . These types of CIs are often referred to as Wald CIs.

The use of “for a large sample” can also be replaced with the word “asymptotically”. You will often hear these results talked about using the phrase “asymptotic normality of maximum likelihood estimators”.

When there is more than one parameter of interest, the variance for each estimator is found in the same manner. There are also covariances to consider in these cases too. For example, the covariance between two estimators  and  is



Example: Field goal kicking (LikelihoodFunction.R, LikelihoodFunction.ipynb)



The estimated variance of  is

.

After recording the video: The negative part of  
has been included in .

To find this, note that



 because only the

Yi’s are random variables



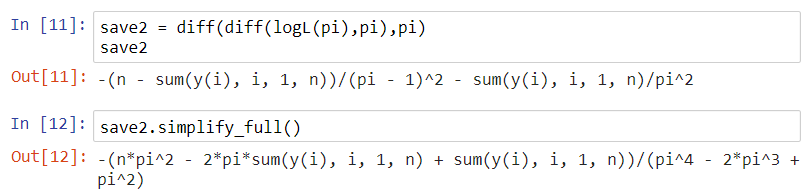




Thus,

= .

Sage:



Wald CI for π



Overall, you can see that properties of maximum likelihood estimation provide a convenient way to find confidence intervals. However, these Wald confidence intervals can have difficulty achieving the (1-α)100% confidence level in small samples.

Students: After making the video for the previous pages of these notes, I realized that Wald tests were not included in the notes. Below is a brief description. This topic will not be on a test.

Suppose we would like to perform a hypothesis test of the form

H0: θ = θ0

Ha: θ ≠ θ0



for a general parameter of interest θ. A Wald test statistic of the form



has an approximate standard normal probability distribution for large samples if the null hypothesis is true. Therefore, we can use



to perform the two-tail test. One-tail tests can be performed as well using the appropriate adjustments.

Notes:

* Similar to CIs, the sample size needs to be large for the standard normal distribution to work well. What does “work well” mean relative to a hypothesis test?
* If θ0 is on the border of possible values for θ (say, 0 < θ < 1 and θ0 = 0), the standard normal distribution approximation will not work well even for large samples. A Wald test of this form should not be used.
* Wald tests can be performed for more than one parameter. For example, suppose the hypothesis test of interest is

Ho: θ1 = θ10, θ2 = θ20

Ha: Not all equal in Ho

The test statistic is best displayed using matrix algebra:



* Computer output from statistical software packages often provide Wald test p-values by default for parameters of interest.