Statistical and computational preliminaries

Subhash Lele

2024-08-27

Basic ideas

- To write down a likelihood function
- Meaning of the likelihood function
- Meaning of the Maximum likelihood estimator, difference between a parameter, an estimator and an estimate

Frequetist paradigm

- Good properties of an estimator: Consistency, Asymptotic normality, Unbiasedness, Mean squared error
- Frequentist paradigm and quantification of uncertainty
- How to use Fisher information for an approximate quantification of uncertainty

Bayesian paradigm

- Motivation for the Bayesian paradigm
- Meaning of the prior distribution
- Derivation and meaning of the posterior distribution
- Interpretation of a credible interval and a confidence interval

Sophie's choice

- Frequentist answer: Are my inferences replicable?
- Bayesian answer: What are my present beliefs in the light of the data.
- Scope of inference: Should I specify the hypothetical experiment or should I specify the prior distribution? Each one comes with its own scope of inference.

Occupancy studies

Let us start with an occupancy study.

- A site that is being considered for development.
- There is a species of interest that might get affected by this development.
- Hence we need to study what proportion of the area is occupied by the species of interest.
- If this proportion is not very large, we may go ahead with the development or vice versa

Statistical model and assumptions

- We divide the site in several equal area cells.
- Suppose all cells have similar habitats (identical).
- Further we assume that occupancy of one cell does not affect occupancy of other quadrats (independence).
- Let N be the total number of cells.

Bernoulli trials

- Let Y_i be the occupancy status of the i-th quadrat. This is unknown and hence is a random variable.
- It takes values in $0, 1$, 0 meaning unoccupied and 1 meaning occupied. This is a Bernoulli random variable.
- We denote this by $Y_i \sim \mathit{Bernoulli}(\phi)$.
- The random variable Y takes value 1 with probability ϕ. This is the probability of occupancy.
- The value of ϕ is unknown. This is the parameter of the distribution.

Random sample (a realization of the random variable)

- Suppose we visit n, a subset, of these cells. These are selected using simple random sampling without replacement.
- The observations are denoted by $y_1, y_2, y_3, ..., y_n$.
- We can use these to infer about the unknown parameter ϕ.
- The main tool for such inductive inference (data to population and *not* hypothesis to prediction) is the likelihood function.

The likelihood function

Suppose the observed data are ${0,1,1,0,0}. Then we can compute the probabiity of observing these data under various values of the parameter ϕ (assuming independent, identically distributed Bernoulli random variables). It can be written as:

$$L(\phi; y_1, y_2, ..., y_n) = \prod P(y_i; \phi) = \prod \phi^{y_i}(1-\phi)^{1-y_i}$$

Notice that this is a function of the parameter ϕ and the data are fixed. The likelihood function or equivalently the log-likelihood function quantifies the *relative* support for different values of the parameters. Hence only the likelihood ratio function is meaningful.

Maximum likelihood estimator

A natural approach to estimation (inference) of ϕ is to choose the value that is better supported than any other value in the parameter space $(0, 1)$. This is called the maximum likelihood estimator. We can show that this turns out to be:

$$\hat{\phi} = \frac{1}{n} \sum y_i$$

This is called an 'estimate'. This is a fixed quantity because the data are observed and hence not random.

Quantification of uncertainty

As scientists, would you stop at reporting this? I suspect not.

- If this estimate is large,say 0.85, the developer is going to say 'you just got lucky (or, worse, you cheated) with your particular sample'.
- A natural question to ask then is 'how different this estimate would have been if someone else had conducted the experiment?'.
- In this case, the 'experiment to be repeated' is fairly uncontroversial. We take another simple random sample without replacement from the study area. However, that is not always the case as we will see when we deal with the regression model.

Sampling distribution (Neyman-Fisher approach)

Sampling distribution is the distribution of the estimates that one would have obtained had one conducted these replicate experiments. It is possible to get an approximation to this sampling distribution in a very general fashion if we use the method of maximum likelihood estimator. In many situations, it can be shown that the sampling distribution is:

$$\hat{\phi} \sim N(\phi, \frac{1}{n} I^{-1}(\phi))$$

where

$$I(\phi) = -\frac{1}{n} \sum \frac{d^2}{d^2\phi} logL(\phi; y)$$

This is also called the Hessian matrix or the curvature matrix of the log-likelihood function. Higher the curvature, less variable are the estimates from one experiment to other. Hence the resultant 'estimate' is considered highly reliable.

95% Confidence interval

- This is just a set of values that covers the estimates from 95% of the experiments.
- The experiments are not actually replicated and hence this simply tells us what the various outcomes could be.
- Our decisions could be based on this variation *as long as we all agree on the experiment that could be replicated*.
- We are simply covering our bases against the various outcomes and protect ourselves from future challenges.

Approximate confidence intervals using the likelihood function

If we use the maximum likelihood estimator, we can obtain the CI as:

$$\hat{\phi} - \frac{1.96}{n}\sqrt{I^{-1}(\hat{\phi})}, \hat{\phi} + \frac{1.96}{n}\sqrt{I^{-1}(\hat{\phi})}$$

You will notice that as we increase the sample size, the width of this interval converges to zero. That is, as we increase the sample size, the MLE converges to the true parameter value. This is called the 'consistency' of an estimator. This is an essential property of any statistical inferential procedure.

Note: These are extremely loose statements. For proper mathematical statements, you can see our 'future' book.

Replicability crisis: One possible cause

- If we truly replicate the experiment and obtain the sampling distribution, the confidence interval would have the correct coverage.
- Of course, we do not replicate the experiments.
- We can only obtain the 'estimate' of the true sampling distribution.
- The confidence interval based on the estimated sampling distribution *does not* lead to the correct coverage probabilities (e.g. Lele 2020)

Bayesian paradigm

All the above statements seem logical but fake at the same time!

- No one repeats the same experiment (although replication consistency is an essential scientific requirement)!
- What if we have time series? We can never replicate a time series.
- So then should we simply take the estimated value *prima facie*? That also seems incorrect scientifically.
- So where is the uncertainty in our mind coming from?

All in the mind

According to the Bayesian paradigm, it arises because of our 'personal' uncertainty about the parameter values.

Prior distribution: Suppose we have some idea about what values of occupancy are more likely than others *before* any data are collected.
- This can be quantified as a probability distribution on the parameter space $(0, 1)$. - This distribution can be anything, unimodal or bimodal or even multimodal! - Let us denote this by $\pi(\phi)$. - How do we change this *after* we observe the data?

How do I change my belief?

Posterior distribution This is the quantification of uncertainty *after* we observe the data. Usually observing the data decreases our uncertainty, although it is not guaranteed to be the case. The posterior distribution is obtained by:

$$\pi(\phi|y) = \frac{L(\phi; y)\pi(\phi)}{\int L(\phi; y)\pi(\phi)d\phi}$$

Credible interval: This is obtained by using the percentiles of the posterior distribution.

Some key observations

Notice a few things here.

- This involves an integral in the denominator. Depending on how many parameters (unknowns) are in the model, this can be a large dimensional integral. Imagine a regression model with 5 covariates. This integral will be 6 dimensional (add one for the variance).
- Data are fixed. We do not need to replicate the experiment. The uncertainty is completely in the mind of the researcher.
- Different researchers might have different prior uncertainties. This will lead to different posterior uncertainties. Hence this is a subjective or personal quantification of uncertainty. It is not transferable from one researcher to another.

Important result (Lindley 1956, Walker 1969)

An interesting result follows, however. As we increase the samples size, the Bayesian posterior, for ANY prior, converges to the distribution that looks very much like the frequentist sampling distribution of the MLE. That is,

$$\pi(\phi|y) \approx N(\hat{\phi}, \frac{1}{n} I^{-1}(\hat{\phi}))$$

There are subtle differences that we are going to ignore here. Qualitatively, what it says is that for large sample size,

- Posterior mean and the MLE are similar
- Posterior variance is similar to the inverse of the Hessian matrix.

Asymptotically Bayes and Fisher agree

- Hence credible interval and confidence intervals will be indistinguishable for large sample size.
- Effect of the choice of the prior distribution vanishes.
- How large a sample size should be for this to happen? It depends on the number of parameters in the model and how strong the prior distribution is.

But the damn math!!!!

- We can describe even a complex statistical model quite easily.
- But mathematics to compute the likelihood function and the sampling distribution can be daunting
- Similarly mathematics to compute the posterior distribution can be extremely difficult, if not impossible.

NO math please! (Bayesian and ML inference using MCMC and data cloning)

- We now show how one can compute the posterior distribution for any choice of the prior distribution without analytically calculating the integral in the denominator.
- Using the same computing tools, we will show that one can compute the MLE and its asymptotic variance quite easily.

How to write the Occupancy model and get posterior distribution using R and JAGS

How to trick Bayesians into giving frequentist answers?

- In this simple situation, we can write down the likelihood function analytically.
- We can also use calculus and/or numerical optimization such as the 'optim' function in R to get the location of the maximum and its Hessian matrix.
- But suppose we do not want to go through all of that and instead want to use the MCMC algorithm.
- Why? Because it is easy and can be generalized to hierarchical models.

Data cloning in a nutshell

Earlier we noted that as we increase the sample size, the Bayesian posterior converges to the sampling distribution of the MLE. We, obviously, cannot increase the sample size. The data are given to us. Data cloning conducts a computational trick to increase the sample size. We clone the data!

Imagine a sequence of K independent researchers.

Step 1: First researcher has data $y_1, y_2, ..., y_n$. They use their own prior and obtain the posterior distribution.

Step 2: Second researcher goes out and gets their own data. It just so happens that they observed the same exact locations as the first researcher. Being a good Bayesian, they use the posterior of the first researcher as their prior (knowledge accumulation). The posterior for the second researcher is given by:

Step K: The K-th researcher also obtains the same data but uses the posterior at the (K-1) step as their prior.

Posterior distribution for the cloned data

$$\pi(\phi|y) = \frac{L(\phi; y)\pi(\phi)}{\int L(\phi; y)\pi(\phi)d\phi}$$

$$\pi(\phi|y, y) = \frac{L^2(\phi; y)\pi(\phi)}{\int L^2(\phi; y)\pi(\phi)d\phi}$$

$$\pi(\phi|y^{(K)}) = \frac{L^K(\phi; y)\pi(\phi)}{\int L^K(\phi; y)\pi(\phi)d\phi}$$

The posterior distribution is converging to a single point; a degenerate distribution. This is identical to the MLE!

1. As we increase the number of clones, the mean of the posterior distributions converges to the MLE.
2. The variance of the posterior distribution converges to 0.
3. If we scale the posterior distribution with the number of clones (that is, multiply the posterior variance by the number of

Clone the data and apply MCMC

We do not need to implement this procedure sequentially. The matrix of these K datasets is of dimension (n, K) with identical columns.

$$\begin{bmatrix} y_1 & y_1 & y_1 & y_1 & y_1 & y_1 & y_1 & y_1 & y_1 & y_1 \\ y_2 & y_2 & y_2 & y_2 & y_2 & y_2 & y_2 & y_2 & y_2 & y_2 \\ y_3 & y_3 & y_3 & y_3 & y_3 & y_3 & y_3 & y_3 & y_3 & y_3 \\ y_4 & y_4 & y_4 & y_4 & y_4 & y_4 & y_4 & y_4 & y_4 & y_4 \\ y_5 & y_5 & y_5 & y_5 & y_5 & y_5 & y_5 & y_5 & y_5 & y_5 \end{bmatrix}$$

We use the Bayesian procedure to analyze these data. The model function used previously can be used with a minor modification to do this.

BINGO!

- If you can write a Bayesian model in R and JAGS, you can get frequentist inferences.
- You can use any prior you want.
- It reflects your opinion about the starting values for an optimization routine.
- Choice of K determines the Bayes-Fisher compromise. It can lead to robustness of the inference (There is a recent paper on this.)

MORE EXAMPLES

Why use MCMC based Bayesian analysis and data cloning?

1. Writing the model function is much more intuitive than writing the likelihood function, prior etc.
2. Do not need to do numerical integration or numerical optimization
3. Data cloning overcomes multimodality of the likelihood function. Entire prior distribution essentially works as a set of starting values. In the usual optimization, starting values can be quite important when the function is not well behaved. By using data cloning, except for the global maximum, all the local maxima tend to 0.
4. Asymptotic variance of the MLE is simple to obtain. It is also more stable than computing the inverse of the second derivative of the log-likelihood function numerically.