Stat 5421 Lecture Notes: To Accompany Agresti Ch 1

Charles J. Geyer

September 16, 2020

Conditional Probability

Conditional probability is just like unconditional probability except it depends on the observed value(s) of some random variable(s).

If there are two random variables X and Y, and you observe X before you observe Y, then what you find out about X may change what you expect to find out about Y when you observe it.

If the (joint) probability mass function (PMF) for X and Y is f, then the conditional PMF of Y given X is proportional to $f(x, y)$ thought of as a function of y for fixed x. So when we are conditioning on x, the variable x is no longer playing the role of a random variable. It is fixed at its observed value throughout the discussion. Thus the PMF of Y given X is

$$f(y \mid x) = c \cdot f(x, y)$$

where the constant c is chosen to make the left-hand side a probability distribution thought of as a function of y for fixed x, that is,

$$\begin{aligned} f(y \mid x) &\geq 0, \qquad \text{for all } y \\ \sum_y f(y \mid x) &= 1 \end{aligned}$$

The PMF $f(\cdot \mid x)$ depends on x, but x is not an argument of the function, so it is really more like a parameter. In fact, there is really no difference between a parametric family of probability distributions f_θ and a conditional distribution. In both case the distribution depends on something that is not considered and argument.

For this reason, Bayesian statisticians always write parametric families as conditional distributions. They write $f(x \mid \theta)$ instead of $f_\theta(x)$, but that is getting a bit ahead of ourselves.

The main point of this section is that conditioning changes the distribution, and the conditioning variable(s) is/are treated as *fixed* in the conditional distribution.

Agresti Section 1.2.5

Poisson Sampling

When we assume *Poisson sampling* we are assuming that the cells of the contingency table, which contain *counts* (nonnegative-integer-valued random variables), are independent Poisson (not necessarily with the same mean, hence not necessarily identically distributed).

Multinomial Sampling

If we start with Poisson sampling and condition on the total we get multinomial sampling. If X_1, ..., X_n are independent Poisson random variables and $N = X_1 + \cdots + X_n$, then the conditional distribution of the random vector $X = (X_1, \ldots, X_n)$ given N is multinomial.

Recall that individuals being sampled form a Poisson process if no individual has any influence on any other. If you take a sample of such individuals over a fixed time interval, then you get Poisson sampling. If you take a sample of such individuals until you get a predetermined number of individuals, then you get multinomial sampling.

Conditioning on N is the same as fixing N.

Product Multinomial Sampling

Here we fix more than one sum of counts. Let $\mathcal{A}$ be a partition of the index set of the counts. This means each index i is an element of exactly one $A \in \mathcal{A}$. Now we fix (condition on)

$$N_A = \sum_{i \in A} X_i, \qquad A \in \mathcal{A}.$$

Now let X_A denote the subvector of X than has just components X_i for $i \in A$. Then these subvectors are independent random vectors conditional on all of the N_A. And the conditional distribution of X_A given all of the N_A is multinomial with sample size N_A.

For a more concrete example, suppose we have a two-way table, in which the data form a matrix. If we condition on the row sums (those are the N_A defined above), then the rows of the table are independent multinomial random vectors. If we condition on the column sums, then the columns of the table are independent multinomial random vectors.

If we start with Poisson sampling, and condition on N_A, $A \in \mathcal{A}$ as defined above, then we get product multinomial sampling.

If we start with multinomial sampling, and condition on N_A, $A \in \mathcal{A}$ as defined above, then we get product multinomial sampling.

Suppose $\mathcal{B}$ is a partition of the index set that is coarser than $\mathcal{A}$, which means that every element of $\mathcal{A}$ is contained in some element of $\mathcal{B}$. If we start with product multinomial sampling for partition $\mathcal{B}$ and condition on N_A, $A \in \mathcal{A}$ as defined above, then we get product multinomial sampling for partition $\mathcal{A}$.

In all three cases, no matter where we start, when we condition on more stuff we get the same thing as if we had decided to fix that stuff in advance.

Summary

Whether you consider this deep or trivial is up to you.

We will find that most of the time it doesn't matter what sampling scheme we "assume" because the same statistical inference results. (But not always.)

Likelihood Inference

Likelihood Function

For parametric family of distributions with PMF f_θ, the *likelihood* for the model when the observed data are x is just PMF

$$L(\theta) = f_\theta(x)$$

with the roles of the parameter and data interchanged. In the PMF the data x is the variable and the parameter θ is fixed. In the likelihood (left-hand side) the data x is fixed at its observed value and the parameter θ is the variable.

You may think this is trivial, but everybody in statistics observes this pedantic distinction. You can really see the difference when it comes to calculus. The derivative of the likelihood function L requires us to differentiate with respect to θ (because that is the variable in that function). The derivative of the PMF f_θ requires us to differentiate with respect to x (because that is the variable in that function).

Log Likelihood Function

For mathematical convenience, the log likelihood is often preferred.

$$l(\theta) = \log L(\theta)$$

Modifications

For reasons that cannot be fully understood until we are done with both frequentist likelihood inference and Bayesian inference (all of which is likelihood-based) it makes no difference whatsoever to statistical inference (frequentist or Bayesian) if

- additive terms that do not contain the parameter(s) are dropped from the log likelihood
- multiplicative terms that do not contain the parameter(s) are dropped from the likelihood

"Principle" (in Scare Quotes) of Maximum Likelihood

You might see in places where they dumb down thing to the point of being wrong. That the maximizer of the likelihood is a good point estimate of the parameter.

This is not a "principle" you should find in any statistics book.

There are statistical models such that the more data you have the worse the maximum likelihood estimator (MLE) is.

What is true is that for statistical models satisfying "suitable regularity conditions" (and neither we nor Agresti go into what that is) we have the following results.

1. the MLE is a consistent and asymptotically normal (CAN) estimator of the parameter, and no other estimator can do better asymptotically than the MLE, except perhaps at a negligible set of parameter values.
2. the MLE is a root-n-consistent, that is,
$$\sqrt{n}(\hat{\theta}_n - \theta)$$
converges in distribution to a mean-zero normal distribution (a multivariate normal distribution if θ is a vector), where $\hat{\theta}_n$ is the MLE for sample size n.

3. (under somewhat weaker regularity conditions than for item 1) if $\tilde{\theta}_n$ is any root-n-consistent estimator, that is,
$$\sqrt{n}(\tilde{\theta}_n - \theta)$$
converges to any distribution whatsoever, and we define $\hat{\theta}_n$ to be the nearest *local* maximum of the likelihood to $\tilde{\theta}_n$, then $\hat{\theta}_n$ is again CAN and best possible (except perhaps for a negligible set of parameter values).

4. The asymptotic variance in the asymptotic (normal) distribution is inverse Fisher information, either observed or expected.
   - Observed Fisher information is minus the Hessian (second derivative) matrix of the log likelihood, evaluated at the MLE.
   - Expected Fisher information is the expectation of observed Fisher information (treating the data as random).

5. Log-linear models for categorical data analysis that are full exponential families (more on that later) *always* satisfy the regularity conditions for item 1.

6. Curved submodels of those in item 4 (for example Agresti Sec. 1.5.4) *always* satisfy the regularity conditions for item 3 (but not necessarily for item 1).

Item 4 means that, so long as we can write down the log likelihood and calculate two derivatives, we know the asymptotic distribution of the MLE (under "suitable regularity conditions"). Item 5 says that for most models used in this class (but not all), the MLE can be defined as the global maximizer of the log likelihood. Item 6 says that for all models used in this class, the MLE can be defined as the the nearest local maximum to a root-n-consistent estimator.

So there is no "principle" that says the global maximizer is best (or even exists). (http://www.stat.umn.edu/geyer/5102/examp/like.html#mix discusses a statistical model for which the global maximizer never exists because the supremum of the log likelihood is infinity. Unfortunately, that model is not categorical data analysis.) But at least for categorical data analysis we do have the estimator of item 6, which we can call the MLE and does have desirable properties.

Except all of this is as sample size goes to infinity. Nothing says the exact sampling distribution of the MLE (for the n we are at) is well approximated by the asymptotic distribution no matter what n is.

But Geyer (2013, IMS Collections, Vol. 10, pp. 1–24) following earlier authors shows that if the log likelihood is well approximated by a quadratic function, then the sampling distribution of the MLE is close to its asymptotic distribution (Agresti also mentions this). So that gives us some purchase on what we need to know about whether asymptotic approximation is good.

We can also use simulation (the so-called *parametric bootstrap*, more on this later) to check how good asymptotic approximation, and also to make the approximation better if it isn't good already.

Agresti Section 1.3.2

One quibble. It so happens that the setting the first derivative of the log likelihood equal to zero and solving for the parameter seems to give the MLE $\hat{\pi} = x/n$. But this is sloppy. The derivative does not exist at the boundary of the parameter space, and even if one uses one-sided derivatives, calculus does not say that the derivative is zero if the maximum occurs on the boundary. Thus to be careful, one needs better analysis (http://www.stat.umn.edu/geyer/5102/slides/s3.pdf slides 20, 21, 25, and 31).

Note that something is fishy about $\pi = 0$ and $\pi = 1$ anyway. In this case the asymptotic variance is zero ($\sqrt{\pi(1-\pi)} = 0$ in either case). So all the asymptotics says is
$$\sqrt{n}(\hat{\pi}_n - \pi) \xrightarrow{\mathcal{D}} 0$$

(the right-hand side is the distribution concentrated at zero). It doesn't tell us much that is useful. (More on this later.)

Agresti Section 1.3.3

Likelihood-based hypothesis tests come in three kinds.

- Likelihood Ratio Tests, also called Wilks tests.
- Wald Tests.
- Score Tests, also called Rao tests, also called Lagrange Multiplier tests (the latter name mostly used by economists).

These tests are all asymptotically equivalent in the sense that (under suitable regularity conditions) they are *asymptotically equivalent* that means that for very large sample sizes they will have nearly equal values of the test statistic and P-value for the same data. (The test statistic and P-value are random quantities when the data are considered random, but for the same data all three tests will have nearly the same test statistics and P-values. The difference between test statistics for any two of these test will be negligible (for large n) compared to either test statistic itself.)

Thus asymptotics gives us no reason to choose one or the other.

Recommendations about which to use are based on mathematical convenience, pedagogical convenience, or simulations. Simulations, of course, must be based on one particular model (or perhaps a few models) and so cannot be general.

Under mathematical convenience we have

- assuming that one can calculate MLE for both the null and alternative hypotheses, the likelihood ratio test statistic is
$$2[l(\hat{\theta}_{\text{big}}) - l(\hat{\theta}_{\text{little}})]$$
and the asymptotic distribution is chi-square with degrees of freedom that is the difference of dimensions of the models. So this is actually easiest if one can fit both big and little models.
- the score test requires only the MLE for the little model, not the big. However the test statistic is complicated to calculate. This it is most useful when the big model is difficult or impossible to fit or when the user does not want to bother with the big model.
- the Wald test requires only the MLE for the big model, not the little. However the test statistic is complicated to calculate. This it is most useful when the little model is difficult or impossible to fit or when the user does not want to bother with the little model.

In particular, R function `summary` computes lots of P-values for lots of tests all based on having fit only one model, the big model (and all of the P-values are for tests of little models that have one of the coefficients set to zero). Thus these must all be Wald tests.

The most famous example of score tests (and one which was invented long before general score tests) is the Pearson Chi-Square test for categorical data analysis. So whenever we use that, we are using a score test.