Stat 5421 Notes: Sampling Schemes

Charles J. Geyer

September 05, 2023

Contents

Introduction **1**

Three Lemmas About Conditional Probability **1**

Subvectors **2**

Sampling Schemes **4**

Theorems about Sampling Schemes and Conditioning **5**

Maximum Likelihood Estimates **7**

Likelihood Ratio Tests **9**

Pearson Chi-Squared Tests **10**

Wald and Rao Tests **10**

Bibliography **11**

Introduction

These notes back up what is said about sampling schemes in our Chapter Zero and also notes to accompany Agresti Chapter 1. For once, they are not titled "lecture notes". We will not lecture on them, because the details are not important for most applied work. They are just here for reference.

Also, some of what is done here requires theoretical sophistication that goes beyond some of the prerequisites of this course (although not beyond any theory course, even Stat 4101). So this is not required reading for this course, just background.

Three Lemmas About Conditional Probability

Lemma 1. *Repeated marginalization gives consistent results. If X, Y, and Z are random vectors, then calculating the marginal of X and Y and then calculating the marginal of X from that (in two steps) gives the same result as calculating the marginal of X directly (in one step).*

Proof. If X, Y, and Z are discrete having joint PMF f, then what must be shown is

$$\sum_{(y,z) \in S(x)} f(x, y, z) = \sum_{y \in T(x)} \sum_{z \in S(x,y)} f(x, y, z) \tag{1}$$

where S is the domain of f and

$$\begin{aligned} S(x) &= \{\, (y, z) : (x, y, z) \in S \,\} \\ S(x, y) &= \{\, z : (x, y, z) \in S \,\} \\ T(x) &= \{\, y : (x, y, z) \in S \,\} \end{aligned}$$

and the sums on the two sides of (1) are the same because the sets $\{y\} \times S(x, y)$ for $y \in T(x)$ partition $S(x)$.

If any of these variables are continuous, some of the sums are replaced by integrals, but otherwise the proof is the same. □

Lemma 2. *Marginalization and conditionalization can be interchanged. If X, Y, and Z are random vectors, then calculating the marginal of X and Y and then calculating the conditional of X given Y from that (in two steps) gives the same result as calculating the conditional of X and Z given Y and then calculating the marginal of that conditional that is the conditional of X given Y (the same two steps, but in reverse order).*

Proof. Using the notation of the preceding proof, what must be shown is

$$\frac{f_{X,Y}(x, y)}{f_Y(y)} = \sum_{z \in S(x,y)} \frac{f(x, y, z)}{f_Y(y)}$$

where f_Y is the marginal of Y and $f_{X,Y}$ is the marginal of X and Y. But this is obvious because $f_Y(y)$ does not depend on z and hence can be moved outside the sum on the right-hand side.

If z is continuous, the sum is replaced an integral, but otherwise the proof is the same. □

Lemma 3. *Repeated conditionalization gives consistent results. If X, Y, and Z are random vectors, then calculating the conditional of X and Y given Z and then calculating the conditional of X given Y and Z from that (in two steps) gives the same result as calculating the conditional of X given Y and Z directly (in one step).*

Proof. Using the notation of the preceding two proofs, what must be shown is

$$\frac{\frac{f(x,y,z)}{f_Z(z)}}{f_{Y|Z}(y \mid z)} = \frac{f(x, y, z)}{f_{Y,Z}(y, z)}$$

but this is obvious because

$$f_{Y|Z}(y \mid z) = \frac{f_{Y,Z}(y, z)}{f_Z(z)}$$

□

Subvectors

In order to describe the product multinomial sampling scheme, we need the notion of subvectors. If y is a vector having index set I and thus components y_i for $i \in I$, and A is a subset of I, when we say y_A is a *subvector* of y having index set A and components y_i for $i \in A$.

This is rather odd, because convention requires that the index set of a vector be $\{1, 2, \ldots, d\}$ for some positive integer d. But here we are allowing arbitrary index sets. For example, we could have

$$I = \{\, \text{cabbage}, \text{dog food}, \text{kumquats} \,\}$$

and then the components of a vector y having index set I are y_{cabbage}, $y_{\text{dog food}}$, and y_{kumquats}.

R caters to this idea in allowing character string indexing.

```
foo <- rnorm(3)
names(foo) <- c("cabbage", "dog food", "kumquats")
foo["dog food"]
```

```
##  dog food
## 0.1711024
```

This is useful in categorical data analysis because we can have the index sets consist of actual category names rather than arbitrary numbers.

But it is even more important in subvector theory because it allows us to match up components of a vector y and its subvector y_A. If $i \in A$, then y_i means the same thing as a component of y and as a component of y_A.

This trick of using arbitrary index sets is not widely used but leads to much more elegant mathematics in books such as Rockafellar (1984) and Lauritzen (1996).

But what are subvectors if they aren't the usual notion? In advanced math vectors are functions. More precisely if V is a vector space and S is an arbitrary set, then V^S denotes the set of all functions $S \to V$, and these functions can be considered vectors with vector addition $h = f + g$ meaning

$$h(x) = f(x) + g(x), \qquad x \in S,$$

and scalar multiplication $h = af$ meaning

$$h(x) = af(x), \qquad x \in S,$$

where f, g, and h are elements of the vector space V^S and a is a scalar (an element of the field of scalars of V, a real number in vector spaces used in statistics). This is the reason that the study of infinite-dimensional topological vector spaces is usually called *functional analysis*.

In this vectors-are-functions view, a vector y having index set I is a function that is an element of the vector space $\mathbb{R}^I$. And this is a finite-dimensional vector space if and only if I is a finite set.

We continue to write y_i for components of i just to look like conventional notation. But this is really function evaluation: y_i means the same thing as $y(i)$, the value of the function y at the point i in its domain (the index set is the domain of the vector considered as a function).

In this vectors-are-functions view, a vector y is a function and a subvector y_A is the restriction of this function to the subset A of its domain. Both y and y_A have the same rule $i \mapsto y_i$. But they have different domains (index sets); y has domain (index set) I, and y_A has domain (index set) A, and $A \subset I$.

So that takes care of the mathematical formalities, but if you don't bother to think of vectors and subvectors as functions but rather just as vectors with arbitrary sets as index sets, that is OK too.

One oddity. The empty set is a possible index set. This gives us the subvector $y_\emptyset$, which is the one and only element of the vector space $\mathbb{R}^\emptyset$. In the vectors-are-functions view, $y_\emptyset$ is the empty function $\emptyset \to V$, which has no allowed values of its argument and hence no values. Considered as a vector, it has no components. But it is a mathematical object. From linear algebra we know there is only one vector space with a finite number of elements, and that is the zero vector space whose only element is the zero vector. Every vector space must contain a zero vector, and the vector space having only the zero vector does satisfy all the axioms for a vector space. Thus $\mathbb{R}^\emptyset$ must be another notation for the zero vector space (also called trivial vector space). And $y_\emptyset$ must be another notation for the zero vector (regardless of what y is). Hence if Y is a random vector $Y_\emptyset$ is a constant random vector always equal to the zero vector of the trivial vector space $\mathbb{R}^\emptyset$. We usually think a zero vector is one all of whose components are zero. This is also true of $y_\emptyset$ because it does not have any components.

R does have vectors of length zero

```
double(0)
```

```
## numeric(0)
```

```
double(0) |> length()
```

```
## [1] 0
```

You can think of that as the unique element of the vector space $\mathbb{R}^{\emptyset}$.

Sampling Schemes

This repeats what what is said about sampling schemes in our notes to accompany Agresti Chapter 1.

- In *Poisson sampling* the cell counts in a contingency table are assumed to be independent Poisson random variables.
- In *multinomial sampling* the cell counts in a contingency table are are assumed to be components of a multinomial random vector.
- In *product multinomial sampling* the cell counts in a contingency table are components of a random vector Y whose index set has a partition $\mathcal{A}$, and the subvectors Y_A, $A \in \mathcal{A}$ are assumed to be independent multinomial random vectors.

We now comment on these definitions.

In Poisson sampling the cell counts are assumed independent but not identically distributed. The vector of mean values $\mu = E(Y)$ is the mean value parameter vector of the exponential family statistical model which is this sampling scheme. Hence different cells of the contingency table can all have different means, which is the whole point of the model.

In multinomial sampling, the cell counts are not independent because the total number of individuals in all cells (called the *sample size*) is not random but rather specified in the design of the experiment (survey, whatever). Again, this is the whole point of the model. The vector of mean values $\mu = E(Y)$ is the mean value parameter vector of the exponential family statistical model which is this sampling scheme. But now we know that these mean values have the multinomial form $\mu_i = n\pi_i$, where n is the multinomial sample size and π_i is the probability of individuals being classified into cell i of the contingency table. (The probabilities sum to one because the classification is mutually exclusive and exhaustive: every individual goes in exactly one category.)

In product multinomial sampling the elements of the partition $\mathcal{A}$ can be called *strata,* a term taken from the term stratified sampling in sampling theory (this is a Latin word, singular *stratum*, plural *strata*). In many applications the strata are all the same size (same number of cells of the contingency table) but they do not have to be. Our notation applies to arbitrary strata.

In product multinomial sampling, the subvectors Y_A are independent, as the definition says. Because of the multiplication rule for independence the joint distribution of all the cell counts factors as a product of multinomial distributions

$$\begin{aligned} f(y) &= \prod_{A \in \mathcal{A}} f_A(y_A) \\ &= \prod_{A \in \mathcal{A}} \binom{n_A}{y_A} \prod_{i \in A} \pi_i^{y_i} \end{aligned} \tag{2}$$

where π is the vector of cell probabilities for the contingency table (π_i is the probability that an individual in stratum A is classified in cell i, assuming $i \in A$), and

$$n_A = \sum_{i \in A} Y_i \tag{3}$$

is the sample size for the multinomial random vector Y_A, and

$$\binom{n_A}{y_A} = \frac{n_A!}{\prod_{i \in A} y_i!}$$

is a multinomial coefficient.

In this sampling scheme what is not random are the sample sizes n_A, $A \in \mathcal{A}$, which are specified in the design of the experiment (survey, whatever). Again, this is the whole point of the model.

The vector of mean values $\mu = E(Y)$ is the mean value parameter vector of the exponential family statistical model which is this sampling scheme. But now we know that these mean values have the product multinomial form $\mu_i = n_A \pi_i$, where n_A is the multinomial sample size and π_i is the probability of individuals being classified into cell i of the contingency table, where $i \in A$.

Note that the probabilities do not sum to one over the whole table but rather within strata

$$\sum_{i \in A} \pi_i = 1, \qquad A \in \mathcal{A}.$$

Our notation also elegantly applies to all contingency tables of any dimension. If we are working with a three-dimensional contingency table with conventional indices j, k, l (and this word is also Latin, singular *index*, plural *indices*) we can define our index set I to be a set of triples (j, k, l) and then our notation works for three-dimensional tables. Or any-dimensional tables in the same way. And it also works if we put the data in a data frame rather than a contingency table. Is the power of the vectors-are-functions view becoming apparent?

Our notation also has the consequence that we really only have two sampling schemes. The multinomial sampling scheme is a special case of the product multinomial sampling scheme when we have the trivial partition which contains only one element, which must be the original index set, that is, $\mathcal{A} = \{I\}$.

But we have many product multinomial sampling schemes, one for each partition. And in talking about that the following terminology is useful. Consider two partitions $\mathcal{A}$ and $\mathcal{B}$. We say $\mathcal{A}$ is *finer* than $\mathcal{B}$ if every $A \in \mathcal{A}$ is contained in some $B \in \mathcal{B}$. (And by the nature of partitions, each $A \in \mathcal{A}$ is then contained in a unique $B \in \mathcal{B}$.) This same relation can be indicated by saying that $\mathcal{B}$ is *coarser* than $\mathcal{A}$.

Theorems about Sampling Schemes and Conditioning

Theorem 1. *Let Y be the random vector of a Poisson sampling model having mean vector μ and index set I. Let $\mathcal{A}$ be a partition of I. Define product multinomial sample sizes n_A, $A \in \mathcal{A}$. Then the distribution of the product multinomial sampling scheme arises by conditioning the (Poisson sampling scheme) distribution of Y on the events $\sum_{i \in A} Y_i = n_A$, $A \in \mathcal{A}$. And the relationship between the usual parameter vectors of these sampling schemes is $\mu_i = n_A \pi_i$, $i \in A \in \mathcal{A}$.*

Proof. We know from the addition rule for independent Poisson random variables that $\sum_{i \in A} Y_i$ is again Poisson with mean $\sum_{i \in A} \mu_i$. Hence the conditional distribution is joint over marginal

$$\begin{aligned}
f\left(Y \,\middle|\, \sum_{i \in A} Y_i = n_A, A \in \mathcal{A}\right) &= \frac{\prod_{i \in I} \mu_i^{y_i} \exp(-\mu_i) / y_i!}{\prod_{A \in \mathcal{A}} \left(\sum_{j \in A} \mu_j\right)^{\sum_{j \in A} y_j} \exp\left(-\sum_{j \in A} \mu_j\right) \Big/ \left(\sum_{j \in A} y_j\right)!} \\
&= \prod_{A \in \mathcal{A}} \frac{\prod_{i \in A} \mu_i^{y_i} \exp(-\mu_i) / y_i!}{\left(\sum_{i \in A} \mu_j\right)^{\sum_{j \in A} y_j} \exp\left(-\sum_{j \in A} \mu_j\right) \Big/ \left(\sum_{j \in A} y_j\right)!} \\
&= \prod_{A \in \mathcal{A}} \binom{n_A}{y_A} \prod_{i \in A} \left(\frac{\mu_i}{\sum_{i \in A} \mu_i}\right)^{y_i}
\end{aligned}$$

and the last line is the PMF of the product multinomial distribution with success probabilities

$$\pi_i = \frac{\mu_i}{\sum_{i \in A} \mu_i}, \qquad i \in A \in \mathcal{A}.$$

But since $\sum_{i \in A} Y_i = n_A$ implies

$$E\left(\sum_{i \in A} Y_i\right) = n_A = \sum_{i \in A} \mu_i$$

we do have $n_A \pi_i = \mu_i$ as the theorem asserts. □

Corollary 1. *Let Y be the random vector of a Poisson sampling model having mean vector μ and index set I. Then the distribution of the multinomial sampling scheme arises by conditioning the (Poisson sampling scheme) distribution of Y on the event $\sum_{i \in I} Y_i = n$, where n is the multinomial sample size. And the relationship between the usual parameter vectors of these sampling schemes is $\mu_i = n\pi_i$, $i \in I$.*

Proof. This is just the special case of the theorem where $\mathcal{A}$ is the trivial partition $\{I\}$. □

Theorem 2. *Let Y be a random vector having index set I. Let $\mathcal{A}$ and $\mathcal{B}$ be partitions of I with $\mathcal{A}$ finer than $\mathcal{B}$. Define the product multinomial sample sizes n_A, $A \in \mathcal{A}$, and n_B, $B \in \mathcal{B}$, satisfying*

$$\sum_{\substack{A \in \mathcal{A} \\ A \subset B}} n_A = n_B, \qquad B \in \mathcal{B}.$$

Let Y have the product multinomial distribution with partition $\mathcal{B}$ and usual parameter vector β. Then the conditional distribution of Y given the events $\sum_{i \in A} Y_i = n_A$, $A \in \mathcal{A}$ is product multinomial with partition $\mathcal{A}$ and usual parameter vector α, with $n_A \alpha_i = n_B \beta_i$, when $A \in \mathcal{A}$, $B \in \mathcal{B}$, and $i \in A \subset B$.

Proof. Define the random variables $N_A = \sum_{i \in A} Y_i$, so the conditioning in the theorem statement is $N_A = n_A$ for $A \in \mathcal{A}$. It is obvious that collapsing some categories of a multinomial random vector gives another multinomial random vector with fewer categories. Hence the marginal distribution of the N_A is product multinomial with PMF

$$\prod_{B \in \mathcal{B}} \frac{n_B!}{\prod_{\substack{A \in \mathcal{A} \\ A \subset B}} N_A!} \prod_{\substack{A \in \mathcal{A} \\ A \subset B}} \left(\sum_{i \in A} \beta_i\right)^{N_A}$$

Hence the conditional distribution is joint over marginal

$$\begin{aligned} f\left(Y \mid N_A = n_A, A \in \mathcal{A}\right) &= \frac{\prod_{B \in \mathcal{B}} \binom{n_B}{y_B} \prod_{i \in B} \pi_i^{y_i}}{\prod_{B \in \mathcal{B}} \frac{n_B!}{\prod_{\substack{A \in \mathcal{A} \\ A \subset B}} N_A!} \prod_{\substack{A \in \mathcal{A} \\ A \subset B}} \left(\sum_{i \in A} \beta_i\right)^{N_A}} \\ &= \prod_{A \in \mathcal{A}} \binom{n_A}{y_A} \prod_{i \in A} \left(\frac{\beta_i}{\sum_{i \in A} \beta_i}\right)^{y_i} \end{aligned}$$

and the last line is the PMF of the product multinomial distribution with success probabilities

$$\alpha_i = \frac{\beta_i}{\sum_{i \in A} \beta_i}, \qquad i \in A \in \mathcal{A}.$$

By the iterated expectation theorem we get the same unconditional expectation of Y whether we use its unconditional or conditional distribution, hence

$$E\left(\sum_{i \in A} Y_i\right) = n_A = \sum_{i \in A} n_B \beta_i, \qquad A \subset B \in \mathcal{B}$$

we do have $n_A \alpha_i = n_B \beta_i$ as the theorem asserts. □

Corollary 2. *Let Y be the random vector of a multinomial sampling model having sample size n, parameter vector π, and index set I. Let $\mathcal{A}$ be a partition of I. Then the conditional distribution of Y given the events $\sum_{i \in A} Y_i = n_A$, $A \in \mathcal{A}$ is product multinomial having PMF given by (2).*

Proof. This is just the special case of the theorem where $\mathcal{B}$ is the trivial partition $\{I\}$. □

All of the theorems and corollaries in this section have obvious converses where we rearrange

$$\text{conditional} = \frac{\text{joint}}{\text{marginal}}$$

as

$$\text{joint} = \text{conditional} \cdot \text{marginal}$$

and go from conditional to joint rather than the other way. The relevant marginals are found in the proofs of the theorems stated here.

Maximum Likelihood Estimates

In the next theorem we need the following terminology. Consider a conditioning event of the form

$$\sum_{i \in A} Y_i = n_A$$

Then we say the *dummy variable associated with this conditioning event* is the vector u_A having zero-or-one-valued components such that

$$\sum_{i \in A} Y_i = u_A^T Y$$

(clearly the i-th component of u_A is equal to one when $i \in A$ and zero otherwise).

Theorem 3. *Suppose we have two sampling schemes, the one with less conditioning is Poisson, multinomial, or product multinomial, and the one with more conditioning is multinomial or product multinomial with a finer partition than the one with less conditioning if the one with less conditioning is product multinomial. We use canonical affine submodels for both sampling schemes having the same offset vectors and model matrices, and we assume every dummy variable associated with a conditioning event for the model with more conditioning is a column of the model matrix. Then the MLE's of the mean value parameter vector for the two sampling schemes are equal, and any (possibly but not necessarily unique) MLE of the canonical parameter vector for the one with less conditioning is also a (necessarily not unique) MLE of the canonical parameter vector for the one with more conditioning.*

Proof. Suppose the model with less conditioning is Poisson and the one with more conditioning is product multinomial with partition $\mathcal{A}$ and usual parameter vector π. (This includes the possibility that $\mathcal{A} = \{I\}$ so product multinomial is actually multinomial.) Let a denote the offset vector, M the model matrix, and u_A, $A \in \mathcal{A}$ the dummy variables for conditioning events.

By the observed-equals-expected principle the likelihood equations determining the MLE Poisson model are

$$\sum_{i \in I} x_i y_i = \sum_{i \in I} x_i e^{\theta_i} \tag{4}$$

where x is any column of M and $\theta = a + M\beta$.

Similarly, the likelihood equations determining the MLE for the product multinomial model are

$$\sum_{i \in I} x_i y_i = \sum_{A \in \mathcal{A}} n_A \frac{\sum_{i \in A} x_i e^{\theta_i}}{\sum_{j \in A} e^{\theta_j}} \tag{5}$$

where x and θ are as in (4).

What is to be shown is that every β that is a solution to (4) is also a solution to (5). The special case of (4) where $x = u_A$ gives

$$n_A = \sum_{i \in A} y_i = \sum_{i \in A} e^{\theta_i} \tag{6}$$

and together (6) and (4) imply (5). Hence any θ that satisfies (4) also satisfies (5), in particular θ of the form $\theta = a + M\beta$. That proves the assertions about canonical parameters.

Now for any $\theta = a + M\beta$ that satisfies (4) the mean value parameter vector for the Poisson sampling model has i-th component e^{θ_i}. And the mean value parameter vector for the product multinomial sampling model has i-th component

$$\frac{n_A e^{\theta_i}}{\sum_{j \in A} e^{\theta_j}}$$

and (6) shows these are the same. That proves the assertion about mean value parameter vectors.

Now we have to redo the whole proof with the model with less conditioning being product multinomial with a partition $\mathcal{B}$ that is coarser than $\mathcal{A}$. (This includes the possibility that $\mathcal{B} = \{I\}$ so product multinomial is actually multinomial.) The proof is almost the same. Now instead of (4) we need (5) with A and $\mathcal{A}$ replaced by B and $\mathcal{B}$, respectively, that is

$$\sum_{i \in I} x_i y_i = \sum_{B \in \mathcal{B}} n_B \frac{\sum_{i \in B} x_i e^{\theta_i}}{\sum_{j \in B} e^{\theta_j}} \tag{7}$$

Now taking $x = u_A$ in (7) gives

$$n_A = \sum_{i \in A} y_i = \frac{n_B \sum_{i \in A} e^{\theta_i}}{\sum_{j \in B} e^{\theta_j}}, \qquad A \subset B \in \mathcal{B}$$

which we can also write as

$$\frac{n_B}{\sum_{j \in B} e^{\theta_j}} = \frac{n_A}{\sum_{i \in A} e^{\theta_i}}, \qquad A \subset B \in \mathcal{B} \tag{8}$$

And (7) and (8) imply (5). And the rest of this case is the same as before. □

Corollary 3. *Suppose we have models as in the theorem. Then the MLE's of the mean value parameter vector for the two sampling schemes are equal, regardless of the canonical parameterizations used.*

A statistical model is a family of probability distributions. By same model, we mean the same family of probability distributions. Since the mean value parameterization is a parameterization, same model means the same mean value parameter space.

So the assumption of the corollary is that we have two models as described in the theorem, regardless of whether the canonical parameterizations are as described in the theorem. The mean value parameter space of the model with more conditioning is derived from the mean value parameter space with less conditioning by the conditioning. If M is the mean value parameter space of the model with less conditioning and the model with more conditioning is product multinomial with partition $\mathcal{A}$ and sample sizes n_A, then the mean value parameter space of the model with more conditioning is

$$\left\{ \mu \in M : \sum_{i \in A} \mu_i = n_A, \ A \in \mathcal{A} \right\}$$

Proof. This is because mean value parameterizations are unique. $E(Y)$ has the same meaning in both models, even if the canonical parameterizations have nothing to do with each other. □

Likelihood Ratio Tests

Theorem 4. *Suppose we are comparing nested submodels for categorical data. The likelihood ratio test statistic does not depend on either the parameterization of the submodels or the sampling scheme, so long as all models satisfy the conditions of Corollary 3.*

Proof. A version of the log likelihood for the mean value parameter for the Poisson sampling scheme is

$$l_{\text{pois}}(\mu) = \sum_{i \in I} (y_i \log(\mu_i) - \mu_i)$$

(different versions of the log likelihood differ by additive terms that do not depend on the parameter). A version of the log likelihood for the mean value parameter for the product multinomial sampling scheme is

$$\begin{aligned} l_{\text{multi}}(\mu) &= \sum_{A \in \mathcal{A}} \sum_{i \in A} y_i \log\left(\frac{\mu_i}{n_A}\right) \\ &= \left(\sum_{i \in I} y_i \log(\mu_i)\right) - \left(\sum_{A \in \mathcal{A}} \log(n_A) \sum_{i \in A} y_i\right) \end{aligned}$$

and we may drop the term that does not contain parameters giving us a different version

$$l_{\text{multi}}(\mu) = \sum_{i \in I} y_i \log(\mu_i) \tag{9}$$

Define the total sample size

$$n = \sum_{A \in \mathcal{A}} n_A$$

Then the conditions of Corollary 3 guarantee that

$$\sum_{i \in I} \hat{\mu}_i = n$$

for the Poisson sampling scheme. Hence if $\hat{\mu}$ and $\tilde{\mu}$ are maximum likelihood estimators for different models being compared

$$\begin{aligned} l_{\text{pois}}(\hat{\mu}) - l_{\text{pois}}(\tilde{\mu}) &= \sum_{i \in I} y_i \log\left(\frac{\hat{\mu}_i}{\tilde{\mu}_i}\right) \\ &= l_{\text{multi}}(\hat{\mu}) - l_{\text{multi}}(\tilde{\mu}) \end{aligned}$$

(maximum likelihood estimators for the same model but different sampling schemes are equal by Corollary 3, moreover different versions of the log likelihood have the same log likelihood differences because the additive terms not containing parameters by which the versions differ are the same for both terms in a log likelihood difference).

Log likelihoods are invariant under change of parameters, that is, if $\hat{\theta}$ is a different parameter corresponding to $\hat{\mu}$ and $\tilde{\theta}$ is a different parameter corresponding to $\tilde{\mu}$, then

$$l(\hat{\mu}) - l(\tilde{\mu}) = l(\hat{\theta}) - l(\tilde{\theta})$$

and this is true for any log likelihood (any model, any sampling scheme). It is even clear (although we will not fuss about details of the proof) that the same conclusion holds even when MLE do not exist: if Θ_{null} and Θ_{alt} are two parameter spaces of nested models being compared, then

$$\left(\sup_{\theta \in \Theta_{\text{alt}}} l_{\text{pois}}(\theta)\right) - \left(\sup_{\theta \in \Theta_{\text{null}}} l_{\text{pois}}(\theta)\right) = \left(\sup_{\theta \in \Theta_{\text{alt}}} l_{\text{multi}}(\theta)\right) - \left(\sup_{\theta \in \Theta_{\text{null}}} l_{\text{multi}}(\theta)\right)$$

It is also clear (although we will not fuss about details of the proof) that we get a similar conclusion when the sampling schemes being compared are product multinomial with two partitions, one finer than the other (obvious from the fact that the log likelihood (9) does not involve the product multinomial sample sizes). □

Theorem 5. *Suppose we are comparing nested submodels for categorical data. The degrees of freedom for the asymptotic distribution of the likelihood ratio test statistic does not depend on either the parameterization of the submodels or the sampling scheme, so long as all models satisfy the conditions of Corollary 3.*

Proof. Corollary 3 refers back to Theorem 3 so we may assume the conditions of the latter. So first consider the situation as in that theorem. We are using the same offset vector a and model matrix M for both sampling schemes and assuming that every u_A, $A \in \mathcal{A}$ be a column of M, where $\mathcal{A}$ is the partition for the sampling scheme with more conditioning.

Now we also assume M has full column rank. This can always be achieved by dropping some columns that do not include the u_A because the u_A are linearly independent vectors.

Suppose the sampling scheme with less conditioning is Poisson and the sampling scheme with more conditioning is product multinomial with partition $\mathcal{A}$. Then the degrees of freedom (DF) for the former is the number of columns of M, call that d, because the Poisson sampling scheme has no directions of constancy. And the DF for the latter is $d - \operatorname{card}(\mathcal{A})$, where $\operatorname{card}(S)$ denotes the cardinality (number of elements in) a set S, because every u_A is a direction of constancy for this sampling scheme and must be dropped to obtain an identifiable canonical parameterization. So the difference in DF of the two sampling schemes is $\operatorname{card}(\mathcal{A})$.

Now suppose the sampling scheme with less conditioning is product multinomial with partition $\mathcal{B}$, the sampling scheme with more conditioning is product multinomial with partition $\mathcal{A}$, and $\mathcal{B}$ is coarser than $\mathcal{A}$. Then the DF for the former is $d - \operatorname{card}(\mathcal{B})$ and for the latter is $d - \operatorname{card}(\mathcal{A})$ and the difference is $\operatorname{card}(\mathcal{A}) - \operatorname{card}(\mathcal{B})$.

Since this analysis applies to both the null and alternative models, the difference in DF is $d_{\text{alternative}} - d_{\text{null}}$ in all cases, where $d_{\text{alternative}}$ and d_{null} are what d was in our preceding analysis now applied to the alternative and null hypotheses (still assuming their model matrices have full column rank and the conditions of Theorem 3 hold).

Since this is the correct way to count DF regardless of whether or not the model matrices originally had full column rank, we are done. □

Pearson Chi-Squared Tests

Theorem 6. *The Pearson chi-squared test statistic does not depend on either the parameterization of the model or the sampling scheme, so long as all models satisfy the conditions of Corollary 3.*

Proof. This is obvious from the fact that the form of the test statistic

$$\sum_{\text{all cells}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

only depends on the mean value parameter "expected" and Corollary 3 says the MLE of the mean value parameters are the same. □

Wald and Rao Tests

This is as far as we can go. Wald and Rao tests do depend on the sampling scheme. Of course from the asymptotic equivalence of Wald, Wilks, and Rao tests, the differences between these test statistics goes to zero in probability as the sample size goes to infinity. Thus they will be close to the same for large sample sizes but not exactly the same (unlike what we had for the likelihood ratio test).

Bibliography

Lauritzen, S. L. (1996) *Graphical Models*. New York: Oxford University Press.

Rockafellar, R. T. (1984) *Network Flows and Monotropic Optimization*. New York: John Wiley & Sons, Inc.