STATS305B – Lecture 2

Jonathon Taylor
Scribed by Michael Howes

01/05/22

Contents

1 Announcements

HW will be up soon (probably today). It will be due in ≈ 1.5 weeks.

2 Inference (Agresti 1.3,1.4)

2.1 Binomial

Consider the null $H_0 : \pi = \pi_0$ where $Y \sim \mathsf{Binomal}(n, \pi)$. The MLE estimate of π is $\widehat{\pi} = \frac{y}{n}$. The *Wald test statistic* is

$$z = \frac{\widehat{\pi} - \pi_0}{\sqrt{\widehat{\pi}(1 - \widehat{\pi})/n}} = \frac{MLE - \theta_0}{\sqrt{\mathrm{Var}_{\widehat{\theta}}(MLE)}}.$$

The statistic z is asymptotically standard normal under H_0. To get a confidence interval we can use $\widehat{\pi} \pm 1.96\sqrt{\widehat{\pi}(1 - \widehat{\pi})/n}$. There is also the *score test statistic*

$$z = \frac{\widehat{\pi} - \pi_0}{\sqrt{\pi_0(1 - \pi_0)/n}} = \frac{MLE - \theta_0}{\sqrt{\mathrm{Var}_{\theta_0}(MLE)}}.$$

This statistic is also asymptotically standard normal under H_0 and it better behaved for values of π close to 0 or 1. The resulting confidence interval is

$$CI = \left\{ \pi : \left| \frac{\widehat{\pi} - \pi}{\sqrt{\pi(1-\pi)/n}} \right| < 1.96 \right\}.$$

The end points can be calculated by solving a quadratic. The *likelihood ratio test statistic* is

$$G^2 = -2 \log L_0 - (-2 \log L_1),$$

where L_0 is the maximum log likelihood under H_0 and L_1 is the maximum log likelihood under $H_0 \cup H_1$. For the binomial model the likelihood ratio test becomes

$$LR(\pi_0) = 2 \left[y \log(\widehat{\pi}/\pi_0) + (n-y) \log((1-\widehat{\pi})/(1-\pi_0)) \right],$$

and, asymptotically, $LR(\pi_0) \sim \chi^2_1$ under $H_0 : \pi = \pi_0$. The resulting confidence interval is

$$CI = \{ \pi : LR(\pi_0) \le 1.96^2 \},$$

which can be calculated numerically.

2.2 Multinomial

Suppose $Y \sim \mathsf{Multinomial}(n, \pi)$ with k classes and $H_0 : \pi = \pi_0$. In this case the likelihood ratio test statistic is

$$G(\pi_0) := 2 \sum_{i=1}^{k} y_i \log \left(\frac{\widehat{\pi}_i}{\pi_{i,0}} \right),$$

under the null $G(\pi_0)$ has an asymptotic χ^2_{k-1} distribution. We also have the Pearson's χ^2 statistic which is essentially a score test

$$\sum_{i=1}^{k} \frac{(y_i - n\pi_{i,0})^2}{n\pi_{i,0}} \overset{n \to \infty}{\sim} \chi^2_{k-1}.$$

The score in this model is $y - n\pi_0$ and a score test statistic would be

$$(y - n\pi_0)^T \operatorname{Var}_{\pi_0}(y - n\pi_0)^{-1} (y - n\pi_0) = (y - n\pi_0)^T \operatorname{Var}_{\pi_0}(y)^{-1} (y - n\pi_0)$$

although in this model the covariance matrix is rank deficient, and so we can't actually invert $\operatorname{Var}_{\pi_0}(y)$. We can instead use the pseudo-inverse or work with only $k-1$ categories.

2.3 Poisson

Suppose $Y_1, \ldots, Y_k \overset{\text{ind}}{\sim} \mathsf{Poisson}(\lambda_i)$ and our null is $H_0 : (\lambda_1, \ldots, \lambda_k) = (\lambda_{1,0}, \ldots, \lambda_{k,0})$. The likelihood ratio test statistic is

$$-2 \sum_{i=1}^{k} y_i \left[\log \left(\frac{y_i}{\lambda_{0,i}} \right) - \left[1 - \frac{\lambda_{0,i}}{y_i} \right] \right] \overset{n \to \infty}{\sim} \chi^2_k,$$

and the score test statistic is

$$\sum_{i=1}^{k} \frac{(y_i - \lambda_{0,i})^2}{\lambda_{0,i}} \overset{n \to \infty}{\sim} \chi^2_k.$$

3 Bayesian inference (Agresti 1.6)

3.1 Beta-binomial

Suppose we have a beta prior $g(\pi) = \pi^{\alpha_1-1}(1-\pi)^{\alpha_2-1}$ and $f(y|\pi) = \binom{n}{y}\pi^y(1-\pi)^{n-y}$. Then the posterior for π is

$$\begin{aligned} g(\pi|y) &\propto g(\pi)f(y|\pi) \\ &\propto \pi^{\alpha_1-1}(1-\pi)^{\alpha_2-1}\pi^y(1-\pi)^{n-y} \\ &\propto \pi^{y+\alpha_1-1}(1-\pi)^{n-y+\alpha_2-1}. \end{aligned}$$

So the posterior is again a beta distribution with parameters $(y+\alpha_1, n-y+\alpha_2)$. Thus, the beta family is conjugate to the binomial family. Note that the prior mean is $\mathbb{E}[\pi] = \frac{\alpha_1}{\alpha_1+\alpha_2}$ and so the posterior mean is $\mathbb{E}[\pi|y] = \frac{\alpha_1+y}{\alpha_1+\alpha_2+n}$. The posterior mean shrinks the MLE towards to the prior mean.

3.2 Multinomial-Dirichlet

Suppose that we have a Dirichlet prior $g(\pi) \propto \prod_{j=1}^k \pi_j^{\alpha_j-1}$ where $\pi_j \geq 0$ and $\sum_{j=1}^k \pi_j = 1$ and our data has a multinomial distribution $f(y|\pi) = \binom{n}{y_1,\dots,y_k}\prod_{j=1}^k \pi_j^{y_j}$. Our posterior is thus

$$\begin{aligned} g(\pi|y) &\propto g(\pi)f(y|\pi) \\ &\propto \prod_{j=1}^k \pi_j^{y_j+\alpha_j-1}. \end{aligned}$$

Thus, the Dirichlet family is conjugate to the multinomial family.

3.3 Poisson-Gamma

Suppose we have a Gamma prior $g(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}$, where $\lambda > 0$. The parameter $\alpha > 0$ is a shape parameter and the parameter $\beta > 0$ is a scale parameter in $\frac{1}{\lambda}$ units. If $f(y|\lambda) = \frac{1}{y!}\lambda^y e^{-\lambda}$, then our posterior is

$$g(\lambda|y) \propto \lambda^{y+\alpha-1}e^{-(\beta+1)\lambda}.$$

So the Gamma family is conjugate to the Poisson family.

4 Contingency tables (Agresti 2)

Consider the following contingency table

	Fatal attack	Non–fatal attack	No attack	
Placebo	$18 = N_{11}$	171	10545	$11034 = N_{1+}$
Aspirin	5	99	10933	11037
	$23 = N_{+1}$	270	21778	$22071 = N_{++}$

We will use Y for the column value and $1, \dots, J$ for the possible values Y can take. We will use X for the row value and $1, \dots, I$ for the possible values X can take. We can describe the distribution of the above counts with a table

	F	NF	NA	
P	π_{11}	π_{12}	π_{13}	π_{1+}
A	π_{21}	π_{22}	π_{23}	π_{2+}
	π_{+1}	π_{+2}	π_{+3}	1

As with $\cdot$, the symbol $i+$ means that the variable replaced with $+$ has been marginalized. There are multiple ways in which the data in the table could have been gathered. Equivalently, there are different descriptions of the sampling process.

1. If both the rows and columns are random, then we could have $N_{ij} \sim \mathsf{Poisson}(\lambda\pi_{ij})$ and hence $N_{++} \sim \mathsf{Poisson}(\lambda)$. In the case we randomly assign either a placebo or aspirin with equal probability we will have $\pi_{ij} = \frac{1}{2} \times \mathbb{P}(Y = j|X = i) = \frac{1}{2}\pi_{j|i}$.
2. For clinical trails/prospective studies we fix the row values and model the column as random. For example, we could select 10 000 people and assign them the placebo and select 10 000 different people and assign them the aspirin. In this case for each row we will have a different distribution for $N_{+|i} \sim \mathsf{Multinomial}(10000, \pi_{\cdot|i})$ where $\pi_{\cdot|i}$ are different parameters for each row.
3. For case control sampling. We fix the column values and treat the row values as random. For example, we could sample 100 people who had fatal heart attacks and 100 people who had non-fatal heart attacks and then compare the frequency of aspirin taking. In this case each column will be a distributed according to $\mathsf{Multinomial}(100, \pi_{\cdot|j})$ for different column parameters $\pi_{\cdot|j}$. This isn't very realistic in this example.

4.1 Case control

Consider the following table where the data was gathered in a case-control experiment:

	lung cancer	
	Yes	No
Smoker	688	650
Non-smoker	21	59
	709	709

Since this is a case-control study the data is given in the form of

$$P(\text{smoker}|\text{cancer}) \text{ and } P(\text{smoker}|\text{no cancer}).$$

What we want is $P(\text{cancer}|\text{smoker})$ and $P(\text{cancer}|\text{non smoker})$ but to calculate these we need $P(\text{cancer})$ which is not available due to the structure of the study. The probabilities we do have are

$$P(\text{smoker}|\text{cancer}) = \frac{688}{709},$$
$$P(\text{smoker}|\text{no cancer}) = \frac{650}{709}.$$

We can use the *odds ratio* to get an idea of the increase in cancer risk due to smoking.

4.2 2 by 2 tables (Agresti 2.2)

Suppose we have the data

	Y	
X	1	2
1	N_{11}	N_{12}
2	N_{21}	N_{22}

Define $\pi_i = \pi_{1|i} = P(Y = 1|X + i) = \frac{\pi_{i1}}{\sum_{j=1}^{J}\pi_{ij}}$. In our example we would have π_i is the probability of cancer given i. The main quantity of interest is the *relative risk* (RR) which is

$$\frac{\pi_1}{\pi_2} = \frac{P(Y = 1|X = 1)}{P(Y = 1|X = 2)}.$$

We cannot directly estimate RR in case control studies since Y is not random. We can estimate the *odds ratio* (OR)

$$\Theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}.$$

Note that $\Theta = 1$ if and only if $\pi_{1|1} = \pi_{1|2}$ (i.e. X and Y are independent). If the roles of X and Y are switched then,

$$\Theta' = \frac{\pi_{11}/\pi_{21}}{\pi_{12}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = \Theta.$$

Thus, we can estimate the OR in both case-control studies and clinical trials. The smoking table is a case-control study and the OR is estimated to be:

$$\Theta = \frac{688/21}{650/59} \approx 3.$$

Note that OR and RR are related in the following way,

$$OR = RR \times \frac{1-\pi_1}{1-\pi_2}.$$

If $1-\pi_1 \approx 1-\pi_2 \approx 1$, then $OR \approx RR$. Thus, if we expect lung cancer to be rare in both smokers and non-smokers, then we would have $RR \approx 3$ and so

$$P(\text{cancer}|\text{smoker}) \approx 3 \times P(\text{cancer}|\text{non smoker}).$$

The assumption that $1-\pi_1 \approx 1-\pi_2 \approx 1$ is called the "rare disease hypothesis".

4.3 Multiway tables

The following example comes from sentencing data for convicted murders. There are three variables of interest

$$\begin{aligned} X &= DR = \text{defendent's race}, \\ Y &= DP = \text{death sentence}, \\ Z &= VR = \text{victim's race}. \end{aligned}$$

We can represent the counts like so.

$VR = W$

	DP		
DR	Y	N	
W	53	414	10%
B	11	37	20%

$VR = B$

	DP		
DR	Y	N	
W	0	16	0%
B	4	139	3%

The percentage are the approximate probability of defendant getting a death sentence given the defendant's and victim's race. We can also create a table where we marginalize over the victim's race

	DP		
DR	Y	N	
W	53	430	11%
B	15	176	8%

We see that marginally, white defendants are more likely to receive the death penalty, but this is reversed if we condition on the victim's race. This is a discrete version of Simpson's paradox. In multiway tables we can calculate the *conditional odds ratio* $\Theta_{XY(k)} = \frac{\pi_{11k}\pi_{22k}}{\pi_{12k}\pi_{21k}}$. Typically, as in the above example, $\Theta_{XY(k)} \neq \Theta_{XY+}$.

Like the odds ratio, the condition OR measures independence conditioned on Z. In particular $X \perp\!\!\!\perp Y | Z$ if and only if $\Theta_{XY(k)} = 1$ for all k. Conditional independence is not the same as marginal independence. For example,

	Y	N	
A	18	12	Clinic 1
B	12	8	
A	2	8	Clinic 2
B	8	32	
A	20	20	Marginal
B	20	40	

In the above example, the columns and rows are conditionally but not marginally independent since

$$\Theta_{XY(1)} = \frac{18 \times 8}{12 \times} = 1, \quad \Theta_{XY(2)} = \frac{2 \times 32}{8 \times 8} = 1, \quad \Theta_{XY+} = \frac{20 \times 40}{20 \times 20} = 2.$$

A weaker property than conditional independence is homogenous association (HA). HA means that the conditional OR is constant. That is,

$$\Theta_{XY(1)} = \Theta_{XY(2)} = \ldots = \Theta_{XY(K)}.$$

As with condition independence, HA does not imply $\Theta_{XY(k)} = \Theta_{XY+}$. HA does mean that are no higher order interactions between X and Y. That is, if we use a log-linear model for our counts $N_{ijk} \sim \mathsf{Poisson}(\lambda_{ijk})$ where

$$\begin{aligned}\log(\lambda_{ijk}) = \alpha_{\cdot\cdot\cdot} & \\ & + \alpha_{i\cdot\cdot} + \alpha_{\cdot j \cdot} + \alpha_{\cdot\cdot k} \\ & + \alpha_{ij\cdot} + \alpha_{i\cdot k} + \alpha_{\cdot jk} \\ & + \alpha_{ijk},\end{aligned}$$

then we would have HA if and only if $\alpha_{ijk} = 0$. We will see this in more details when we study generalized linear models in more detail. The main idea is that HA holds if all interactions beyond second order are zero.

4.4 $I \times J$ tables

For $I \times J$ tables we can define the local odds ratios, these are

$$\Theta_{ij} = \frac{\frac{\pi_{ij}}{\pi_{i,j+1}}}{\frac{\pi_{i+1,j}}{\pi_{i+1,j+1}}} = \frac{\pi_{ij}\pi_{i+1,j+1}}{\pi_{i+1,j}\pi_{j,i+1}},$$

where $1 \leq i \leq I-1$ and $1 \leq j \leq J-1$. The number of local ORs is $(I-1) \times (J-1)$. The local ORs are less interpretable than the OR, but they do determine the associations between X and Y as can be seen by counting the number of free parameters. In the full model there are $IJ-1$ total parameters. For the row marginals there are $I-1$ free parameters and for the column marginals there are $J-1$ free parameters. Thus, the number of free parameters needed to describe the associations is

$$IJ - 1 - (I-1) - (J-1) = IJ - I - J + 1 = (I-1) \times (J-1),$$

which is the number of local ORs. Note that when $I = J = 2$ as before, there is just a single OR and a single parameter is all that is needed to describe the association between X and Y.