STATS305B – Lecture 4

Jonathon Taylor
Scribed by Michael Howes

01/12/22

Contents

1 Inference on contingency tables

1.1 Secondary Analysis

Last time we saw two statistics for testing the hypothesis $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$. Suppose that we do indeed reject H_0, a natural question is why. We will see that both our test statistics can be decomposed and these can highlight which counts broke the independence hypothesis.

1.2 Pearson's test

One of the tests we discussed was Pearson's χ^2-test of independence. This test has statistic,

$$X^2 = \sum_{ij} \frac{(N_{ij} - \widehat{\lambda}_{ij})^2}{\widehat{\lambda}_{ij}},$$

where $\widehat{\lambda}_{ij} = \frac{n_{i+}n_{+j}}{n_{++}} = n_{++}\widehat{\pi}_{i+}\widehat{\pi}_{+j}$ are the MLE estimates under H_0. This is a score statistic with asymptotic distribution $X^2 \sim \chi^2_{(I-1)(J-1)}$. The quantities $e_{ij} = \frac{N_{ij}-\widehat{\lambda}_{ij}}{\sqrt{\widehat{\lambda}_{ij}}}$ are called *Pearson's residuals.* Since,

$$X^2 = \sum_{ij} e_{ij}^2,$$

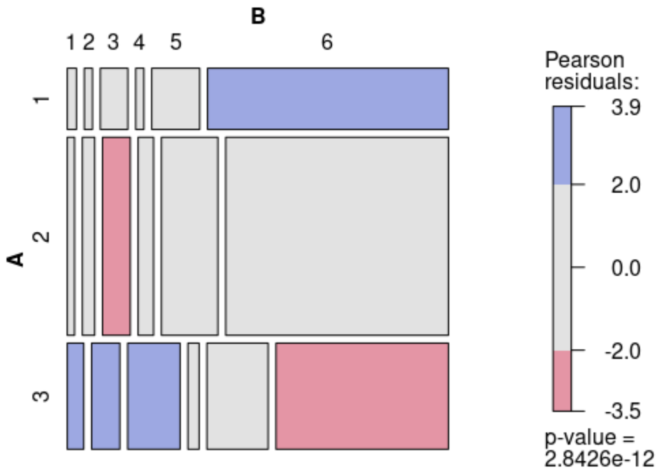
the term e_{ij} can be interpreted as the evidence from cell i, j against H_0. One can ask if $N_{ij} \overset{?}{\sim} \mathsf{N}(0, 1)$ under H_0. This is not true since otherwise we would expect X^2 to have IJ degrees of freedom. The corrected residuals,

$$r_{ij} = \frac{N_{ij} - \widehat{\lambda}_{ij}}{\sqrt{\widehat{\lambda}_{ij}(1-\widehat{\pi}_{i+})(1-\widehat{\pi}_{+j})}},$$

are asymptotically $\mathsf{N}(0,1)$. Consider the following example,

		Belief in god					
		1	2	3	4	5	6
	1	19	8	27	8	47	236
Education level	2	23	39	88	49	179	706
	3	28	48	89	19	104	293

The Pearson statistic can be calculated in R using the function `chisq.test()`. For the above data the statistic is approximately 76 and the p-value is approximately 3×10^{-12}. The package `vcd` has a function `mosaic()` that can be used to visualize the Pearson's residuals and the departure from independence. When applied to the above table, `mosaic()` returns the following



The area of each cell is proportional to the count in that cell. The rows have fixed width and the columns have variable widths. If X and Y were independent, then we would expect the columns to also have approximately the same width. Colored cells indicate cells with large Pearson's residuals with. Blue cells have a higher count than expected under independence and red cells have a lower count than under independence. It can be seen that there are more people with a high belief in god and a low education level than expected. Among people with a high level of education there are more than expected with a low belief in god and less than expected with a high belief in god.

1.3 Likelihood ratio test

Last lecture we also saw that we could test for independence with the statistic,

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

where L_0 is the maximum likelihood under H_0 and L_1 is the maximum likelihood under $H_0 \cup H_1$ (we will sometimes say that L_1 is the maximum likelihood in the *saturated model*). Under the null H_0, G^2 has asymptotic distribution $\chi^2_{(I-1)(J-1)}$. The statistic G^2 can also be calculated in R by using `lr_stat()`. For the education level/belief in god table, we have $G^2 \approx 73 \approx X^2$ and a similar p-value to X^2. We can also decompose G^2. Consider, the following data containing counts from schizophrenia patients.

	biogenic	environmental	combination
eclectic	90	12	78
medical	13	1	6
psychoanalytic	19	13	50

This table has a likelihood ratio statistic of approximately 23. The degrees of freedom is 4 and thus this is a significant result. To decompose G^2 we must first define the sub-tables of an $I \times J$ table. For each i, j such that $2 \leq i \leq I$ and $2 \leq j \leq J$, we can define a 2×2 sub-table $T(i,j)$ as follows. We keep entry (i,j), ignore entries with $h > i, k > j$ and marginalize the entries (h,k) with $h < i$ or $j < k$. More precisely, if our original table had entries n_{hk}, then the entries of T_{ij} are

$T(i,j)$	1	2
1	$\sum_{h<i,k<j} n_{hk}$	$\sum_{h<i} n_{hj}$
2	$\sum_{k<j} n_{ik}$	n_{ij}

Thus $T(2,2)$ is the top-left 2×2 table of our original $I \times J$ table. In general the table $T(i,j)$ contains the counts of $X < i, X = i$ against $Y < j, Y = j$. There are $(I-1)(J-1)$ such sub-tables. Each sub-table has a likelihood ratio statistic $G^2_{ij} \sim \chi^2_1$. It turns out that

$$G^2 = \sum_{2 \leq i \leq I, 2 \leq j \leq J} G^2_{ij}. \tag{1}$$

In the schizophrenia table this decomposition gives

$$G^2 \approx 23 = 0.3 + 1.4 + 13 + 8.3 \approx G^2_{2,2} + G^2_{2,3} + G^2_{3,2} + G^2_{3,3}.$$

So the main contribution to G^2 comes from $G^2_{3,2}$ and $G^2_{3,3}$. The decomposition (1) holds because the terms in the decomposition are the result of comparing a simpler model to a more complex model (this will be explained later). In later terms, the more complicated models become the simpler models and so there is cancellation between the terms. This is because when comparing a model M_r to the simpler model M_s, the corresponding term is of the form

$$\text{DEV}(M_s) - \text{DEV}(M_r),$$

where $\text{DEV}(M)$ is the deviance of the model M and is analogous to the sum of squared errors from the linear model. Thus, when comparing the most complicated model M_k to the simplest model M_0, we can write the comparison as sequence of comparisons M_0 against M_1, M_1 against M_2, …, M_{k-1} against M_k. This is analogous to the ANOVA decomposition for linear models.

1.4 Fisher's exact test

Another way to do inference on 2×2 tables is to use Fisher's exact test. Suppose we use either the Poisson or multinomial model for our cell counts and consider the null $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$. Under H_0, the marginals $N_{1+}, N_{2+}, N_{+1}, N_{+2}$ are sufficient statistics. Furthermore, given the marginals, the individual cell counts are determined by N_{11} and

$$N_{11} | N_{1+}, N_{2+}, N_{+1}, N_{+2} \sim \mathsf{Hypergeometric}(N_{+1}, N_{+2}, N_{1+}),$$

where $\mathsf{Hypergeometric}(a, b, k)$ describes the distribution of the number of red balls when drawing k balls without replacement from an urn with a red balls and b blue balls. The reason why N_{11} follows the above distribution conditioned on the marginals is that we can think of drawing N_{1+} 1's and 2's from an urn with N_{+1} 1's and N_{+2} 2's. The number of 1's in such a draw is the value of N_{11}. Since we know the exact distribution of N_{11} under the null, we can calculate exact p-values by using the hyper-geometric distribution.

This form of exact test can generalize to $I \times J$ tables, but it gets complicated. This is because conditioned on the marginals there are $(I-1)\times(J-1)$ free parameters for the entries N_{ij}. Understanding the distribution of these entries is tricky are requires clever sampling techniques. In particular, naive MCMC does not work.

1.5 Ordinal Association

In the last lecture we discussed Kruskal's γ. Recall that if are variables X and Y are ordinal, then we can define the parameter $\gamma = \frac{\pi_c - \pi_d}{\pi_c + \pi_d}$ where

$$\pi_c = 2\sum_{ij} \pi_{ij} \sum_{h>i,k>j} \pi_{hk}, \quad \text{and} \quad \pi_d = 2\sum_{ij} \pi_{ij} \sum_{h>i,k<j} \pi_{hk}.$$

The parameter $\gamma \in [-1, 1]$ measures the ordinal association between X and Y with $\gamma = 0$ corresponding to no association, $\gamma > 0$ corresponding to positive association and $\gamma < 0$ corresponding to negative association. We also defined the estimator

$$\widehat{\gamma} = \frac{C - D}{C + D},$$

where

$$C = 2\sum_{ij} n_{ij} \sum_{h>i,k>j} n_{hk} \quad \text{and} \quad D = 2\sum_{ij} n_{ij} \sum_{h>i,k<j} n_{hk}.$$

But we are yet to study the variance of $\widehat{\gamma}$. The variance of $\widehat{\gamma}$ can be approximated by using the bootstrap. To do this we need to create bootstrap resamples of our contingency table, and it isn't clear how to do this. One strategy is to "flatten" our contingency table.

1. We first create a new table with n_{++} rows and 2 columns labelled X and Y. Each row corresponds to an individual from our population. In each individual's row we record their X and Y. We can indeed construct such a table from the contingency table since the contingency table records how many individuals have $X = i$ and $Y = j$.

2. For $b = 1, \dots, B$, we then draw n_{++} rows with replacement from the flattened table. From these we can calculate a contingency table T_b, and we can calculate the estimate $\widehat{\gamma}^* b$.

3. We can then use the sample $\widehat{\gamma}_b^*$ to estimate the variance and quantiles of $\widehat{\gamma}$. These can then be used to make confidence intervals.

Another way to measure ordinal association is to assign numerical values to the labels of X and Y and then perform linear regression of Y onto X. This gives us an F-statistic that measures the linear relationship between X and Y. Unfortunately the value of the F-statistic depends on the numerical values assigned to X and Y, and there is no set rule for how to assign such values.

2 Modelling binary data

We now consider modelling binary data. Suppose that we have a model $(X_i, Y_i)_{i=1}^n$ where $Y_i \in \{0, 1\}$, $X_i \in \mathbb{R}^p$ and (X_i, Y_i) are i.i.d. Such data can appear in a variety of applications. For example,

- Medical: presence/absence of a disease.
- Industrial: passes/fails a quality control test.
- Political: voter/non-voter.

Since $Y_i \in \{0, 1\}$, the distribution of $Y|X$ is described by a single parameter

$$\pi(x) = \mathbb{P}(Y = 1|X = x).$$

The negative log-likelihood is thus,

$$-\log L(\pi|Y) = \sum_{i=1}^{n} -Y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) - \log(1-\pi_i).$$

Note that we are only modelling Y in our likelihood. This corresponds to either a fixed design for X or to having conditioned on X. The quantity $\log\left(\frac{\pi_i}{1-\pi_i}\right)$ is the *log-odds* of $Y_i = 1$. The function $\pi \mapsto \log\left(\frac{\pi}{1-\pi}\right)$ is called the logit function. It is an increasing function with domain $(0, 1)$ and range $\mathbb{R}$. The inversion function is $F_{\text{logistic}}(\eta) = \frac{e^\eta}{1+e^\eta}$. The function F_{logistic} is sigmoidal and is the CDF of the sigmoidal distribution.

2.1 Logistic regression

Define $\eta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right)$. The negative log likelihood function in terms of η_i is

$$-\log L(\eta|Y) = \sum_{i=1}^{n} -Y_i\eta_i + \log(1 + e^{\eta_i}).$$

In *logistic regression* we introduce the constraint $\eta_i = X_i^T\beta$. Thus

$$-\log L(\beta|Y) = \sum_{i=1}^{n} -Y_i X_i^T\beta + \sum_{i=1}^{n} \log(1 + \exp(X_i^T\beta)) = -(X\beta)^T Y + \sum_{i=1}^{n} \log(1 + \exp(X_i^T\beta)),$$

where $X \in \mathbb{R}^{n\times p}$ has rows X_i^T. We will see that this function is convex in β and thus optimizing the negative log likelihood is tractable. It also means that we can add convex regularization or convex constraints to β and we will still have a convex problem. This is the default model in most software GLM models for binary data.

To see that the function

$$-\log L(\beta|Y) = -(X\beta)^T Y + \sum_{i=1}^{n} \log(1 + \exp(X_i^T\beta)),$$

is indeed convex, note that it suffices to prove that $-\log L(\eta|Y)$ since the composition of a linear function followed by a convex function is convex. Note that we can write

$$-\log L(\eta|Y) = \sum_{i=1}^{n} \Lambda_B(\eta_i) - \eta_i Y_i,$$

where $\Lambda_B(\eta)$ is the CGF of the Bernoulli distribution. Namely, for an exponential family with natural parameters η and sufficient statistics W, we have

$$\exp(\Lambda(\eta)) = \mathbb{E}_0[\exp(\eta^T W)],$$

where the expectation is taken at $\eta = 0$. The function $\eta \mapsto \Lambda(\eta)$ is convex for any exponential family which can be seen by taking derivatives or by applying Hölder's inequality. Thus, $-\log L(\eta|Y)$ and hence $-\log L(\eta|Y)$ are both convex.

2.2 Interpretation of coefficients

Note that if $\eta = X^T\beta$, then

$$OR_{X_j} = \frac{ODDS(Y=1|\ldots,X_j=x_j+1,\ldots)}{ODDS(Y=1|\ldots,X_j=x_j,\ldots)} = e^{\beta_j}.$$

Thus, β_j determines the change in the odds of success if X_j increases by 1 and all other covariates are fixed. Note that e^{β_j} is the odds ratio when X_j increases by 1. The quantity e^{β_j} does not tell us the relative risk

$$RR_{X_j} = \frac{\mathbb{P}(Y=1|\ldots,X_j=x_j+1,\ldots)}{\mathbb{P}(Y=1|\ldots,X_j=x_j,\ldots)}.$$

However, we have seen that if $\mathbb{P}(Y=1|X=x)$ is small for all values of x, then $OR \approx RR$. Thus, under the rare disease hypothesis,

$$RR_{X_j} \approx OR_{X_j} = e^{\beta_j}.$$

2.3 Other binary models

Suppose there exists a quantity T_i such that $T_i|X_i \sim F$ and

$$Y_i = \begin{cases} 1 & \text{if } T_i \le X_i^T\beta, \\ 0 & \text{if } T_i > X_i^T\beta. \end{cases}$$

In this case $\mathbb{P}(Y_i = 1|X_i) = \mathbb{P}(T_i \le X_i^T\beta|X_i) = F(X_i^T\beta)$. If we set $\eta_i = \text{logit}(F(X_i^T\beta))$ where $\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$, then our log-likelihood is

$$-\log(\beta|Y) = \sum_{i=1}^n Y_i \text{logit} F(X_i^T\beta) - \log(1 - F(X_i^T\beta)).$$

Unfortunately, there is no guarantee that this will be convex in β. However, for the natural choice of $F = \Phi$ the CDF of $\mathsf{N}(0,1)$, then this is convex and this model is called the *probit model.* The parameters β_j are not as interpretable in this type of model.

2.4 Over dispersion

One last comment is that for binary data we have

$$\text{Var}(Y) = \pi(1-\pi) = \mathbb{E}[Y](1 - \mathbb{E}[Y]).$$

Thus, the expectation of Y determines the variance of Y (and indeed all higher moments). This relates to the fact that logistic regression is a generalized linear model, and it relates to a concept called *over-dispersion.*