Multinomial sampling

Demo 1, Unit 6B

David Bass

April 13, 2022

Motivation

In this demonstration, we will construct the chi-square test of independence using a random sample $\mathbf{X}$ of multinomial data and Wilks' theorem, which gives that, for a likelihood ratio ρ of an observation $\mathbf{x}$, the distribution of $2 \log \rho(\mathbf{x})$ is approximately chi-square.

Defining a table T

Let T be a table with r rows and c columns. T contains $r \cdot c$ cells, each of which is uniquely identified with a row value j and a column value k.

	1	$\cdots$	c
1	(1, 1)	$\cdots$	(1, c)
$\vdots$	$\vdots$	$\ddots$	$\vdots$
r	(r, 1)	$\cdots$	(r, c)

Motivation for table T

Considering the rows as levels of a factor A and the columns as levels of a factor B, we are interested in testing whether the effects of A and B are independent.

	1	$\cdots$	c
1	(1, 1)	$\cdots$	(1, c)
$\vdots$	$\vdots$	$\ddots$	$\vdots$
r	(r, 1)	$\cdots$	(r, c)

Equivalently, we are interested in testing whether the joint distribution of A and B equals the product of the marginal distributions of A and B.

Defining a multinomial r.v. $\mathbf{X}$ on T

Let $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ be an IID random sample of the cells of T, where

- each $\mathbf{X}_i = (X_{i,1,1}, \ldots, X_{i,1,c}, \ldots, X_{i,r,1}, \ldots, X_{i,r,c})$,
- $X_{i,j,k}$ is the indicator variable that the ith observation equals cell (j, k), and
- the probability of choosing cell (j, k) is $\theta_{j,k}$.

Let $\boldsymbol{\theta} = (\theta_{1,1}, \ldots, \theta_{1,c}, \ldots, \theta_{r,1}, \ldots, \theta_{r,c})$, so $\mathbf{X}$ is a multinomial random variable with parameters n and $\boldsymbol{\theta}$.

Note that, since each $\mathbf{X}_i = (X_{i,1,1}, \ldots, X_{i,1,c}, \ldots, X_{i,r,1}, \ldots, X_{i,r,c})$ encodes one observation in our sample, it is a vector containing one 1 and zeros everywhere else.

θ as a parameter of interest

In order to show that the effects of row j and column k are independent, we will test the null hypothesis

$$H_0 : \theta_{j,k} = \theta_{j\circ}\theta_{\circ k}$$

	1	$\cdots$	c	row total
1	$\theta_{1,1}$	$\cdots$	$\theta_{1,c}$	$\theta_{1\circ}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
r	$\theta_{r,1}$	$\cdots$	$\theta_{r,c}$	$\theta_{r\circ}$
col. total	$\theta_{\circ 1}$	$\cdots$	$\theta_{\circ c}$	1

Note that a dot in the subscript denotes the sum over all values, e.g. $\theta_{1\circ} = \Sigma_{k=1}^{c}\theta_{1,k}$.

Defining a total count variable $\mathbf{Y}$

Let $\mathbf{Y} = (Y_{1,1}, \ldots, Y_{1,c}, \ldots, Y_{r,1}, \ldots, Y_{r,c})$ be the random variable that counts the number of observations of each cell, i.e.

$$Y_{j,k} = \sum_{i=1}^{n} X_{i,j,k}$$

for each $1 \leq j \leq r$ and $1 \leq k \leq c$.

Representing $\mathbf{X}$ in T

We can represent our sample $\mathbf{X}$ in T with components of Y.

	1	$\cdots$	c	row total
1	$Y_{1,1}$	$\cdots$	$Y_{1,c}$	$Y_{1\circ}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
r	$Y_{r,1}$	$\cdots$	$Y_{r,c}$	$Y_{r\circ}$
col. total	$Y_{\circ 1}$	$\cdots$	$Y_{\circ c}$	n

We are again using the dot notation in subscripts to denote the sum over all values, e.g. $Y_{1\circ} = \Sigma_{k=1}^{c} Y_{1,k}$.

Degrees of freedom of $\boldsymbol{\theta}$ in T

Since $\sum_{j=1}^{r}\sum_{k=1}^{c}\theta_{j,k} = 1$, one of the $r \cdot c$ probabilities is determined by the other $r \cdot c - 1$. Thus, the total degrees of freedom for $\boldsymbol{\theta}$ are $r \cdot c - 1$.

	1	$\cdots$	c	row total
1	$\theta_{1,1}$	$\cdots$	$\theta_{1,c}$	$\theta_{1\circ}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
r	$\theta_{r,1}$	$\cdots$	$\theta_{r,c}$	$\theta_{r\circ}$
col. total	$\theta_{\circ 1}$	$\cdots$	$\theta_{\circ c}$	1

Marginal degrees of freedom of θ in T

Similarly, since $\Sigma_{j=1}^{r}\theta_{j,k} = \theta_{\circ k}$ and $\Sigma_{k=1}^{c}\theta_{j,k} = \theta_{j\circ}$, the degrees of freedom within each row and column of T are $c-1$ and $r-1$, respectively.

	1	$\cdots$	c	row total
1	$\theta_{1,1}$	$\cdots$	$\theta_{1,c}$	$\theta_{1\circ}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
r	$\theta_{r,1}$	$\cdots$	$\theta_{r,c}$	$\theta_{r\circ}$
col. total	$\theta_{\circ 1}$	$\cdots$	$\theta_{\circ c}$	1

The log-likelihood function of $\mathbf{X}_i$

For each $\mathbf{X}_i = (X_{i,1,1}, \ldots, X_{i,1,c}, \ldots, X_{i,r,1}, \ldots, X_{i,r,c})$,

The likelihood function of $\mathbf{X}_i$ is

$$f_{\mathbf{X}_i,\boldsymbol{\theta}}(\mathbf{x}) = \prod_{j=1}^{r} \prod_{k=1}^{c} \theta_{j,k}^{X_{i,j,k}}$$

Then the log-likelihood function of $\mathbf{X}_i$ is

$$\ell(\boldsymbol{\theta}; \mathbf{X}_i) = \log f_{\mathbf{X}_i,\boldsymbol{\theta}}(\mathbf{x}) = \sum_{j=1}^{r} \sum_{k=1}^{c} X_{i,j,k} \log \theta_{j,k}$$

Log-likelihood function of $\mathbf{X}$

The likelihood function of $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ is

$$f_{\mathbf{X},\boldsymbol{\theta}}(\mathbf{X}) = \prod_{i=1}^{n} f_{\mathbf{X}_i,\boldsymbol{\theta}}(\mathbf{x})$$

Then the log-likelihood function $\mathbf{X}$ is

$$\ell(\boldsymbol{\theta}; \mathbf{X}) = \log f_{\mathbf{X},\boldsymbol{\theta}}(\mathbf{x}) = \sum_{i=1}^{n} \sum_{j=1}^{r} \sum_{k=1}^{c} X_{i,j,k} \log \theta_{j,k}$$

Recall that $Y_{j,k} = \sum_{i=1}^{n} X_{i,j,k}$, so

$$\ell(\boldsymbol{\theta}; \mathbf{X}) = \sum_{j=1}^{r} \sum_{k=1}^{c} Y_{j,k} \log \theta_{j,k}$$

MLE for $\boldsymbol{\theta}$

The maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ is the component-wise sample proportion $\hat{\boldsymbol{\theta}} = (\hat{\theta}_{1,1}, \ldots, \hat{\theta}_{1,c}, \ldots, \hat{\theta}_{r,1}, \ldots, \hat{\theta}_{r,c})$, where each

$$\hat{\theta}_{j,k} = \frac{1}{n} \sum_{i=1}^{n} X_{i,j,k} = Y_{j,k}/n$$

Maximizing the likelihood under the null hypothesis

Recall the null hypothesis that the effects of the rows and columns are independent

$$H_0 : \theta_{j,k} = \theta_{j\circ}\theta_{\circ k}$$

Under the null hypothesis, the likelihood function $\ell(\boldsymbol{\theta}; \mathbf{X})$ is maximized when $\hat{\theta}_{j,k} = \hat{\theta}_{j\circ}\hat{\theta}_{\circ k} = (Y_{j\circ}/n)(Y_{\circ k}/n) = Y_{j\circ}Y_{\circ k}/n$.

Maximizing the likelihood under the alternative hypothesis

Consider the alternative hypothesis

$$H_1 : \theta_{j,k} \neq \theta_{j\circ}\theta_{\circ k}$$

Under the alternative hypothesis, there is no constraint, so the likelihood function $\ell(\boldsymbol{\theta}; \mathbf{X})$ is maximized when $\hat{\theta}_{j,k} = Y_{j,k}/n^2$.

Log-likelihood ratio

Let ρ be the ratio of the unconstrained likelihood ($\boldsymbol{\theta} \in \Theta_0 \cup \Theta_1$) to the constrained likelihood ($\boldsymbol{\theta} \in \Theta_0$), so

$$\rho(\mathbf{X}) = \frac{\sup\{f_{\mathbf{X},\boldsymbol{\theta}}(\mathbf{X}) \ : \ \boldsymbol{\theta} \in \Theta_0 \cup \Theta_1\}}{\sup\{f_{\mathbf{X},\boldsymbol{\theta}}(\mathbf{X}) \ : \ \boldsymbol{\theta} \in \Theta_0\}}$$

Then

$$\begin{aligned}
\log \rho(\mathbf{X}) &= \sup\{\ell(\boldsymbol{\theta}, \mathbf{X}) \ : \ \boldsymbol{\theta} \in \Theta_0 \cup \Theta_1\} - \sup\{\ell(\boldsymbol{\theta}, \mathbf{X}) \ : \ \boldsymbol{\theta} \in \Theta_0\} \\
&= \sum_{j=1}^{r} \sum_{k=1}^{c} Y_{j,k} \log \frac{Y_{j,k}}{n} - \sum_{j=1}^{r} \sum_{k=1}^{c} Y_{j,k} \log \frac{Y_{j\circ} Y_{\circ k}}{n^2} \\
&= \sum_{j=1}^{r} \sum_{k=1}^{c} Y_{j,k} \log \frac{Y_{j,k}}{Y_{j\circ} Y_{\circ k}/n}
\end{aligned}$$

Distribution of the log-likelihood ratio

Wilks' theorem gives that $2 \log \rho(\mathbf{X})$ approximately follows a chi-square distribution with degrees of freedom equal to the difference in the number of free parameters between H_1 and H_0.

Calculating degrees of freedom of the distribution

Under the alternative hypothesis, there are no constraints, so the degrees of freedom are $\nu_1 = r \cdot c - 1$.

The constraint of the null hypothesis gives we use the marginal degrees of freedom to get that $\nu_0 = (r-1) + (c-1) = r + c - 2$.

Thus, the degrees of freedom of the chi-square distribution followed by $2 \log \rho(\mathbf{X})$ are $\nu = \nu_1 - \nu_0 = r \cdot c - r - c + 1 = (r-1)(c-1)$.

Conclusion

- Described a multinomial random sample $\mathbf{X}$
- Derived $\log \rho(\mathbf{X}) = \sum_{j=1}^{r} \sum_{k=1}^{c} Y_{j,k} \log \frac{Y_{j,k}}{Y_{j\circ} Y_{\circ k}/n}$
- Applied Wilks' theorem to show that, approximately, $2 \log \rho(\mathbf{X}) \sim \chi^2_{(r-1)(c-1)}$
- Constructed the chi-square test of independence