STA 114: Statistics

Notes 19. Categorical Data

Category count data

A great deal of social and biological science studies involve category count data. We have seen the simplest example of this in our opinion poll study, where a researcher looks at number of supporters and number of opposers of a federal law among n college students. More generally, one might be interested in counts of k categories where k could be larger than 2. For example, in the opinion poll study, we could consider three categories of response: "support", "oppose" and "do not care".

For n units of observations (e.g., college students) grouped into k categories (e.g., response types), we shall denote the count data as $X = (X_1, \cdots, X_k)$, where $X_l \geq 0$ is the count in the l-th category (e.g., number of supporters), and $X_1 + \cdots + X_k = n$. As usual, observed data will be denoted $x = (x_1, \cdots, x_k)$, with $x_l \geq 0$ and $x_1 + \cdots + x_k = n$. The set of all such vectors x is our samples space S.

The multinomial model

To describe a model for X, we look at the probability vector $p = (p_1, \cdots, p_k)$, with p_l denoting the probability that an observation unit is of type l. For p to be a probability vector, we must have $p_l \geq 0$ and $p_1 + \cdots + p_k = 1$. The set of all k-vectors p with these properties will be denoted Δ_k and called the k-dimensional simplex.

We make the assumption that the category types of the units are independent of each other (same as assuming independent Bernoulli trials in defining a binomial count). Then X can be described by the pmf

$$f(x|p) = \binom{n}{x_1 \ \cdots \ x_k} p_1^{x_1} \cdots p_k^{x_k}, \ \ x \in S$$

and $f(x|p) = 0$ for any other x. This pmf is called the multinomial pmf and is denoted $\mathsf{Multinomial}(n, p)$. In above the multinomial coefficient

$$\binom{n}{x_1 \ \cdots \ x_k} = \frac{n!}{x_1! \cdots x_k!}$$

is an extension of the binomial coefficient and gives the number of ways n units can be split into k distinct groups.

It is easy to see that if $X \sim \mathsf{Multinomial}(n, p)$ then for every $l = 1, \cdots, k$, $X_l \sim \mathsf{Binomial}(n, p_l)$. To see this, simply label being in category l as 'success' and being in any other category as 'failure', which makes X_l the total number of successes in n independent trials each with success probability p_l. Similarly for any two categories $l \neq j$, $X_l + X_j \sim \mathsf{Binomial}(n, p_l + p_j)$ and so on.

Maximum likelihood

Consider the multinomial model $X \sim \mathsf{Multinomial}(n, p)$, $p \in \Delta_k$. For onservation $X = x$, the likelihood function in $p \in \Delta_k$ is given by

$$L_x(p) = \binom{n}{x_1 \ \cdots \ x_k} p_1^{x_1} \cdots p_k^{x_k} = \text{const} \times \prod_{l=1}^{k} p_l^{x_l}$$

and the log-likelihood function is given by

$$\ell_x(p) = \text{const} + \sum_{l=1}^{k} x_l \log p_l.$$

To maximize this over $p \in \Delta_k$, we cannot directly use the standard trick of setting the partial derivatives to zero, because Δ_k imposes the restriction $\sum_{l=1}^{k} p_l = 1$. Instead we use Lagrange multiplies trick and set to zero the partial derivatives of the function

$$g(p, \lambda) = \ell_x(p) + \lambda(\sum_{l=1}^{k} p_l - 1)$$

jointly over (p, λ). That is we solve for (p, λ) in

$$\begin{aligned} \frac{\partial}{\partial p_l} g(p, \lambda) &= \frac{x_l}{p_l} + \lambda = 0, \quad l = 1, 2, \cdots, k \\ \frac{\partial}{\partial \lambda} g(p, \lambda) &= \sum_{l=1}^{k} p_l - 1 = 0. \end{aligned}$$

The first k equations ensure that the solution $(\hat{p}, \hat{\lambda})$ satisfies $\hat{p}_l = -x_l/\hat{\lambda}$, $l = 1, \cdots, k$, which, when plugged into the last equation, gives $\hat{\lambda} = -n$. Therefore $\hat{p}_l = x_l/n$, $l = 1, \cdots, k$. That is, the maximum likelihood estimate of p based on data $X = x$ is given by

$$\hat{p}_{\text{MLE}}(x) = \left(\frac{x_1}{n}, \cdots, \frac{x_k}{n} \right).$$

Hypothesis testing

Categorical data provide a very useful platform for testing various scientific hypotheses. Below are some examples.

Example (Mendel's peas)**.** Mendel, the founder of modern genetics, studied how physical characteristics are inherited in plants. His studies led him to propose the laws of segregation and independent assortment. We'll test this in a simple context. Under Mendel's laws, when pure round-yellow and pure green-wrinkled pea plants are cross-bred, the next generation of plant seeds should exhibit a 9:3:3:1 ratio of round-yellow, round-green, wrinkled-yellow and wrinkled-green combinations of shape and color. In a sample of 556 plants from the next generation the observed counts for these combinations are (315, 108, 101, 32). Does the data support Mendel's laws?

In this case, we have $X = (X_1, X_2, X_3, X_4)$ giving the category counts of the four types of plants with $X \sim \mathsf{Multinomial}(n = 556, p)$, $p \in \Delta_4$. We want to test the point null hypothesis $H_0 : p = (\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16})$ against $H_1 : p \neq (\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16})$. □

Example (Hardy-Weinberg equilibrium)**.** The spotting on the wings of Scarlet tiger moths are controlled by a gene that comes in two varieties (alleles) whose combinations (moths have pairs of chromosomes) produce three varieties of spotting pattern: "white spotted", "little spotted" and "intermediate". If the moth population is in Hardy-Weinberg equilibrium (no current selection drift), then these varieties should be in the ratio $a^2 : (1-a)^2 : 2a(1-a)$, where $a \in (0, 1)$ denotes the abundance of the dominant white spotting allele. In a sample of 1612 moths, the three varieties were counted to be 1469, 5 and 138. Is the moth population in HW equilibrium?

Letting $X = (X_1, X_2, X_3)$ denote the category counts of the three spotting patterns, with model $X \sim \mathsf{Multinomial}(n = 1612, p)$, $p \in \Delta_3$, we want to test whether $H_0 : p \in \Delta_3^{HW}$ against $H_1 : p \neq \Delta_3^{HW}$ where Δ_3^{HW} is a subset of Δ_3 containing all p of the form $(a^2, (1-a)^2, 2a(1-a))$ for some $a \in (0, 1)$. □

A third and widely used type of hypotheses, relating to independence of two or more attributes with categorical outcomes, will be discussed in the next lecture.

ML tests

We'll start with the point null hypothesis. We have $X \sim \mathsf{Multinomial}(n, p)$ and we want to test $H_0 : p = p_0$ against $p \neq p_0$ for some fixed $p_0 = (p_{01}, p_{02}, \cdots, p_{0k}) \in \Delta_k$ of interest. Any ML test is given by

$$\textit{reject } H_0 \text{ if } \frac{L_x(p_0)}{L_x(\hat{p}_{\text{MLE}}(x))} < k$$

for some $k \in (0, 1]$. By using the form of the likelihood function and that of the MLE, we get

$$\frac{L_x(\hat{p}_{\text{MLE}}(x))}{L_x(p_0)} = \prod_{l=1}^{k} \left(\frac{x_l}{np_{0l}}\right)^{x_l} = \prod_{l=1}^{k} \left(\frac{x_l}{e_l}\right)^{x_l}$$

where $e_l = np_{0l}$ is the expected count for category l if H_0 were true. [Under H_0, $X_l \sim \mathsf{Binomial}(n, p_{0l})$ so $\mathrm{E}X_l = np_{0l}$].

It can be shown that if n is fairly large and if none of the coordinates p_{0l} of p_0 is too close to zero, then

$$\frac{L_x(\hat{p}_{\text{MLE}}(x))}{L_x(p_0)} \approx e^{Q(x)}$$

where $Q(x) = \sum_{l=1}^{k} \frac{(x_l - e_l)^2}{e_l}$. Therefore, an ML test is approximately the same as

$$\textit{reject } H_0 \text{ if } Q(x) > c$$

for some positive constant c.

To calculate the size of this test, we need to know the distribution of $Q(X)$ when $X \sim \mathsf{Multinomial}(p_0)$. Karl Pearson showed that $Q \sim \chi^2(k-1)$ approximately. Let F_{k-1} denote

the cdf of the $\chi^2(k-1)$ distribution. Then, an approximately size α ML test is given by

$$\textit{reject } H_0 \text{ if } Q(x) > F_{k-1}^{-1}(1-\alpha).$$

Clearly, the p-value based on $X = x$ for such tests is $1 - F_{k-1}(Q(x))$.

Pearson's chi-square tests

The above test is known as the Pearson's chi-square test. It applies beyond the point null case. In general, suppose we're testing $H_0 : p \in \Delta_k^0$ against $H_1 : p \neq \Delta_k^0$ where Δ_k^0 is determined by r many "free parameters". [In the point null case, Δ_k^0 is a single point, and has $r = 0$ free parameters. In the Hardy-Weinberg equilibrium example above, Δ_k, with $k = 3$ contains all $p = (a^2, (1-a)^2, 2a(1-a))$ with $r = 1$ free parameter $a \in (0, 1)$.]. Given observation $X = x$, the Pearson's chi-square test can be performed by taking the following steps:

1. Find the restricted MLE $\hat{p}_0(x) = \text{argmax}_{p \in \Delta_k^0} L_x(p)$ under the null hypothesis.
2. Calculate "expected" category counts under the estimated null: $\hat{e}_l = n\hat{p}_{0l}$, $l = 1, \cdots, k$.
3. Calculate Pearson's test statistic

$$Q(x) = \sum \frac{(observed - expected)^2}{expected} = \sum_{l=1}^{k} \frac{(x_l - \hat{e}_l)^2}{\hat{e}_l}$$

4. Given a level $\alpha \in (0, 1)$, reject H_0 at level α if $Q(x) > F_{k-1-r}^{-1}(1-\alpha)$.
5. Alternatively, report the p-value $= 1 - F_{k-1-r}(Q(x))$.

Example (Mendel's peas (contd))**.** Here H_0 is a point null consisting of the single point $p_0 = (\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16})$. Therefore the restricted MLE is same as $\hat{p}_0 = p_0$ and so

$$e_1 = 312.75,\ e_2 = 104.25,\ e_3 = 104.25,\ e_4 = 34.75.$$

So fro observed data $x = (315, 108, 101, 32)$ we get $Q(x) = 0.47$. The p-value is $1 - F_3(0.47) =$ `1 - pgamma(0.47, 3/2, 1/2) = 0.92`, because $\chi^2(m) = \mathsf{Gamma}(m/2, 1/2)$.

Example (HW equilibrium)**.** Here H_0 is not a point null, but has a free parameter $a \in (0, 1)$. Writing the likelihood function in terms of a we see,

$$L_{x,H_0}(a) = \text{const} \times \{a^2\}^{x_1} \times \{(1-a)^2\}^{x_2} \times \{2a(1-a)\}^{x_3} = \text{const} \times a^{2x_1+x_3}(1-a)^{2x_2+x_3}$$

and so this is maximized at $\hat{a} = \frac{2x_1+x_3}{2x_1+x_3+2x_2+x_3} = \frac{x_1+x_3/2}{n}$. So for our data, $\hat{a} = \frac{1469+138/2}{1612} = 0.954$. Which leads to $Q(x) = 0.83$ and with p-value $1 - F_1(0.83) = 0.36$.