2-17-09 Math 152

Multinomial Cell probabilities

Suppose there are m possible outcomes in an experiment which occur with probabilities $p_1, \ldots, p_m$ respectively ($p_1 + \ldots + p_m = 1$).

Repeat the experiment n times and let X_i be the number of times the i^{th} outcome occurs.

Marginal distribution of X_i:

$$P(X_i = k) = \binom{n}{k} p_i^k (1-p_i)^{n-k}$$

$$0 \leq k \leq n.$$

Joint distribution of $(X_1, \ldots, X_m)$:

The X_i are not independent since

$$X_1 + \ldots + X_m = n.$$

$$\text{Prob}\left(X_1 = x_1, \ldots X_m = x_m\right)$$

$$= \frac{n!}{x_1!\, x_2! \ldots x_m!}\, p_1^{x_1} p_2^{x_2} \cdots p_m^{x_m}$$

Given data $x_1, \ldots, x_m$ on "cell counts"

What is the maximum likelihood estimate of the probabilities $p_1, \ldots p_m$?.

Claim: $\hat{p}_j = \frac{x_j}{n}$

Perhaps this seems intuitively clear.

To verify the claim we maximize

$$l(p_1, \ldots, p_m) = \log n! - \sum_{i=1}^{m} \log(x_i!) + \sum_{i=1}^{m} x_i \log p_i$$

subject to the constraint

$$g(p_1 \ldots, p_m) = p_1 + \ldots + p_m = 1.$$

$$\frac{\partial \ell}{\partial p_j} = \frac{x_j}{p_j}$$

$$\frac{\partial g}{\partial p_j} = 1.$$

At the critical point, we must have.

$$\frac{x_j}{p_j} = \lambda \qquad j = 1, \ldots, m$$

for some λ.

Then $\sum_{j=1}^{m} \frac{x_j}{\lambda} = 1.$

so $\lambda = n.$ and $p_j = \frac{x_j}{n}.$

Note that $\log p_i \leq 0 \quad \forall\, i = 1, \ldots m$

so $\ell(p_1, \ldots, p_m) \to -\infty$ if $x_i \to +\infty$ for any i.

$\therefore \hat{p}_j = \frac{x_j}{n}$ gives a maximum.

If the p_i $i = 1, \ldots m$ are functions of a parameter

$p_i = p_i(\theta)$ then we can write

$$l(\theta) = \log n! - \sum_{i=1}^{m} \log(x_i)! + \sum_{i=1}^{m} x_i \log p_i(\theta)$$

and maximize w.r.t. θ. by setting

$$l'(\theta) = \sum_{i=1}^{m} \frac{p_i'(\theta)}{p_i(\theta)} x_i = 0.$$

and solving for θ.

e.g. 1029 blood types, Hong Kong 1937

	M	MN	N	Total
Frequency	342	500	187	1029

Hardy-Weinberg:

genotypes AA, Aa, aa

occur with frequencies $(1-\theta)^2$, $2\theta(1-\theta)$, θ^2

$$-\frac{2(1-\theta)}{(1-\theta)^2}X_1 + \frac{2(1-\theta)-2\theta}{2\theta(1-\theta)}X_2 + \frac{2\theta}{\theta^2}X_3 = 0.$$

$$= -\frac{2}{1-\theta}X_1 + \frac{X_2}{\theta} - \frac{X_2}{1-\theta} + \frac{2X_3}{\theta} = 0.$$

$$\Rightarrow -2\theta X_1 + X_2(1-\theta) - X_2\cdot\theta + 2X_3(1-\theta) = 0.$$

$$-2\theta X_1 + X_2 - X_2\theta - X_2\theta + 2X_3 - 2X_3\theta = 0.$$

$$\theta(-2X_1 - 2X_2 - 2X_3) + X_2 + 2X_3 = 0.$$

$$\frac{X_2 + 2X_3}{2X_1 + 2X_2 + 2X_3} = \theta.$$

$$\hat{\theta} \approx .4247.$$

and the estimated cell probabilities are

$$p_1 \simeq .331$$
$$p_2 \simeq .489$$
$$p_3 \simeq .180.$$

To use the Boobstrap method to estimate the variability in $\hat{\theta}$,

generate 1000 random. multinomial

counts with probabilities $\hat{P_1}, \hat{P_2}, \hat{P_3}$ and $n = 1029$ trials.

$X_{1,i}^*$

$X_{2,i}^*$ $\quad i = 1, \ldots 1000.$

$X_{3,i}^*$

and then use a histogram of

$$\hat{\theta}_i^* = \frac{X_{2,i}^* + 2X_{3,i}^*}{2X_1^* + 2X_2^* + 2X_3^*} = \frac{X_{2,i}^* + 2X_{3,i}^*}{2(1029)}.$$

to estimate the variability of $\hat{\theta}$.

Confidence Intervals using the Bootstrap.

If $\hat{\theta}$ is an estimate of a parameter θ whose true value is $\theta = \theta_0$ and if we knew the distribution of $\hat{\theta} - \theta_0$, we could find for any given $\alpha > 0$.

$\underline{\delta}$ and $\overline{\delta}$ such that

$$P(\hat{\theta} - \theta_0 \leq \underline{\delta}) = \frac{\alpha}{2}$$

and $P(\hat{\theta} - \theta_0 \leq \overline{\delta}) = 1 - \frac{\alpha}{2}$

so that

$$P(\underline{\delta} \leq \hat{\theta} - \theta_0 \leq \overline{\delta}) = 1 - \alpha.$$

and

$$P(\hat{\theta} - \overline{\delta} \leq \theta_0 \leq \hat{\theta} - \underline{\delta}) = 1 - \alpha.$$

Since we don't know the distribution of $\hat{\theta} - \theta_0$, the next best thing would be to simulate it assuming we know the true value θ_0.

We would generate many samples of observations from the distribution with parameter θ_0, record the values $\hat{\theta}_i - \theta_0$ (say for $i = 1, \ldots, 1000$) and compute the simulated values $\underline{\delta}$ and $\bar{\delta}$.

Since we don't have θ_0 we use our original determination of $\hat{\theta}$ in its place.

- generate samples from dist. w/ param $\hat{\theta}$
- record $\hat{\theta}_i^* - \hat{\theta}$
- find $\underline{\delta}$, $\bar{\delta}$ for these data

- The approximate $(1-\alpha)\%$ confidence interval is given by.

$$(\hat{\theta} - \bar{\delta}, \hat{\theta} - \underline{\delta})$$

If $\underline{\theta}$ and $\bar{\theta}$ are the upper and lower quantiles of the distribution of θ^* then

$$\underline{\delta} = \underline{\theta} - \hat{\theta}$$

$$\bar{\delta} = \bar{\theta} - \hat{\theta}$$

and the confidence interval can be written as.

$$(2\hat{\theta} - \bar{\theta}, 2\hat{\theta} - \underline{\theta}).$$

If the distribution of θ^* is symmetric about $\hat{\theta}$ then

$$\bar{\theta} - \hat{\theta} = \hat{\theta} - \underline{\theta}$$

and $\therefore \quad 2\hat{\theta} = \bar{\theta} + \underline{\theta}$.

In this case our interval is $(\underline{\theta}, \bar{\theta})$.