Multinomial models (& goodness of fit)

Multinomial distribution is generalization of the binomial distribution, for categorical variables with more than two response types.

k types

♦ ♦ ★ ♥ ♥ sample n

♦ ♥ ♦ ★ with replacement

★ ♦ ♥ ♥ ♥

π_1 = proportion of type 1 (★) in the urn
π_2 = proportion of type 2 (♥) in the urn
$\vdots$
π_k = proportion of type k (♦) in the urn

The π_j's are **constants** (parameters);

$$\pi_1 + \pi_2 + \cdots + \pi_k = 1.$$

$Y_1, Y_2, \ldots, Y_k$: **random variables**

Y_1 = number of type 1 (★) in the **sample**
Y_2 = number of type 2 (♥) in the **sample**
$\vdots$
Y_k = number of type k (♦) in the **sample**

$$Y_1 + Y_2 + \cdots + Y_k = n.$$

The Y_j's are **dependent**: the value of one affects the others.

$$P(Y_1 = y_1 \text{ and } Y_2 = y_2 \text{ and } \cdots \text{ and } Y_k = y_k)$$

$$= \frac{n!}{y_1! y_2! \cdots y_k!} \pi_1^{y_1} \pi_2^{y_2} \cdots \pi_k^{y_k}$$

where y_1, y_2, ..., y_n are any nonnegative integers that add to n (all possible outcomes).

ex.

$Y_1 =$ # democrats, $Y_2 =$ # republicans, $Y_3 =$ # greens,
$Y_4 =$ # "other", in a random sample of n voters

Multinomial property:

$$\mathrm{E}(Y_j) = \pi_j n$$

Data: y_1, y_2, ..., y_k, n

ML estimates: $\widehat{\pi}_j = \frac{y_j}{n}$

Reduced-parameter models

The ordinary multinomial model has $k-1$ unknown parameters ($\pi_k = 1 -$ sum of the others). An interesting, and scientifically useful, class of multinomial models arise when the π_j's are constructed as functions of fewer underlying parameters (arising out of various scientific mechanisms)

Examples of reduced-parameter models

1. Hardy-Weinberg equilibrium

genes (alleles) A, a
gene proportions: $p =$ proportion of A
$(1-p) =$ proportion of a

bucket of gametes:

A A a
a A a A
a A A a A A

random mating is like drawing two alleles at random (with replacement) from the bucket to make a new individual

$P(\text{AA}) = p^2$
$P(\text{Aa}) = 2p(1-p)$ "Hardy-Weinberg proportions"
$P(\text{aa}) = (1-p)^2$

Draw sample of n individuals in population; determine their genotypes (AA, Aa, or aa). $Y_1 =$ #AA, $Y_2 =$ #Aa, $Y_3 =$ #aa in sample.

Model: $Y_1, Y_2, Y_3 \sim$ multinomial(n, π_1, π_2, π_3)

where
$$\pi_1 = p^2$$
$$\pi_2 = 2p(1-p)$$
$$\pi_3 = (1-p)^2$$

model with *one* unknown parameter, p

2. Bird-banding: band n adult birds

$r =$ prob. of surviving a given yr
$s =$ prob. that band is found & returned (given the bird dies) in the year of death

Study lasts for three years; $Y_1 =$ # band returns in first yr, $Y_2 =$ #band returns in second yr, $Y_3 =$ # band returns in third yr, $Y_4 =$ # birds with unreturned bands.

Model: $Y_1, Y_2, Y_3, Y_4 \sim$ multinomial$(n, \pi_1, \pi_2, \pi_3, \pi_4)$

$$\pi_1 = (1-r)s$$
$$\pi_2 = r(1-r)s$$
$$\pi_3 = r^2(1-r)s$$
$$\pi_4 = 1 - [(1-r)s + r(1-r)s + r^2(1-r)s]$$

Model has *two* unknown parameters r and s

3. Model of independence of two categorical variables in a survey

Draw random sample of n individuals; get opinion from each individual:

for or *against* death penalty (first categorical variable: D. P. opinion)

for or *against* gun registration (second categorical variable: G. R. opinion)

$r =$ proportion in population who favor D. P.
$s =$ proportion in population who favor G. R.

Four different kinds of individuals: $(+\ +)$, $(-\ +)$, $(+\ -)$, $(-\ -)$

If D. P. opinion is *independent* of G. R. opinion, then the proportions of these individuals in the population are the *products* of the marginal proportions

$\pi_1 =$ proportion of $(+\ +) = rs$
$\pi_2 =$ proportion of $(-\ +) = (1-r)s$
$\pi_3 =$ proportion of $(+\ -) = r(1-s)$
$\pi_4 =$ proportion of $(-\ -) = (1-r)(1-s)$

$Y_1 =$ # of $(+\ +)$ in sample, $Y_2 =$ # of $(-\ +)$ in sample, $Y_3 =$ # of $(+\ -)$ in sample, $Y_4 =$ # of $(-\ -)$ in sample

$Y_1, Y_2, Y_3, Y_4 \sim$ multinomial$(n, \pi_1, \pi_2, \pi_3, \pi_4)$

Model has *two* unknown parameters r, s

Counts in tests of independence are often arranged in a **two way table**:

		death penalty opinion	
		$+$	$-$
gun registration opinion	$+$	Y_1	Y_2
	$-$	Y_3	Y_4

n

4. Fitting a probability distribution to grouped data

Group boundaries: $s_1, s_2, ..., s_{k-1}$ (known)

$X_1, X_2, ..., X_n$: a random sample from a normal(μ, σ^2) distribution

$Y_1 =$ # of observations in $(-\infty, s_1)$
$Y_2 =$ # of observations in (s_1, s_2)
$Y_3 =$ # of observations in (s_2, s_3)
$\vdots$
$Y_{k-1} =$ # of observations in (s_{k-2}, s_{k-1})
$Y_k =$ # of observations in $(s_{k-1}, +\infty)$

$\pi_1 =$ area under normal curve from $-\infty$ to s_1
$\pi_2 =$ area under normal curve from s_1 to s_2
$\vdots$
$\pi_k =$ area under normal curve from s_{k-1} to $+\infty$

multinomial model for the Y_j's has *two* unknown parameters: μ and σ^2

Goodness of fit test for a reduced parameter multinomial model

Null hypothesis is a reduced parameter multinomial model for describing the Y_j's, that is, the π_j's can be represented in terms of fewer underlying parameters (let's call them θ_1, θ_2, ..., θ_l)

$$
\begin{aligned}
\text{H}_0\text{:}\ \ \pi_1 &= g_1(\theta_1, \theta_2, ..., \theta_l) \\
\pi_2 &= g_2(\theta_1, \theta_2, ..., \theta_l) \\
&\vdots \\
\pi_k &= g_k(\theta_1, \theta_2, ..., \theta_l)
\end{aligned}
$$

(Note: in goodness of fit, the null hypothesis is often the research model of interest)

Alternative hypothesis is that a more complex multinomial model is necessary to describe the Y_j's, that is, all $k-1$ free parameters π_1, π_2, ..., π_{k-1} are needed

$$
\begin{aligned}
\text{H}_\text{a}\text{:}\ \ \pi_1 &= \pi_1 \\
\pi_2 &= \pi_2 \\
&\vdots \\
\pi_{k-1} &= \pi_{k-1}
\end{aligned}
$$

(Note: H_a is the ordinary multinomial, sometimes called the **saturated model** because it fits perfectly, i.e. estimated expected values equal the observed values)

Data: $y_1, y_2, ..., y_k, \quad n$

Likelihood

H_0: $\quad \widehat{L}_0 = \frac{n!}{y_1! y_2! \cdots y_k!} \widetilde{\pi}_1^{y_1} \widetilde{\pi}_2^{y_2} \cdots \widetilde{\pi}_k^{y_k}$

where $\widetilde{\pi}_j = g_j\left(\widehat{\theta}_1, \widehat{\theta}_2, ..., \widehat{\theta}_l\right)$. $\quad \widehat{L}_0$ is the likelihood evaluated at the ML estimates of the θ_j's.

$\mathrm{H_a}$: $\quad \widehat{L}_\mathrm{a} = \frac{n!}{y_1! y_2! \cdots y_k!} \widehat{\pi}_1^{y_1} \widehat{\pi}_2^{y_2} \cdots \widehat{\pi}_k^{y_k}$

where $\widehat{\pi}_j = \frac{y_j}{n}$

Likelihood ratio test statistic:

$$G^2 = -2 \log_e \left(\frac{\widehat{L}_0}{\widehat{L}_\mathrm{a}} \right)$$

Note, for these multinomial models, the likelihood ratio statistic reduces to the following form:

Let $\widehat{E}_j = n\widetilde{\pi}_j$ (ML-estimated expected value of Y_j under the null hypothesis)

$$G^2 = 2\sum_{j=1}^{k} y_j \log_e\left(\frac{y_j}{\widehat{E}_j}\right)$$

This is in the form $2\Sigma O_j \log_e\left(\frac{O_j}{\widehat{E}_j}\right)$

Also, one can show (by "asymptotic expansion") that

$$G^2 \approx \sum_{j=1}^{k} \frac{\left(y_j - \widehat{E}_j\right)^2}{\widehat{E}_j} \quad \left(= \sum_{j=1}^{k} \frac{\left(O_j - \widehat{E}_j\right)^2}{\widehat{E}_j} = X^2\right)$$

that is, G^2 (likelihood ratio statistic) and Pearson's chi-square statistic are asymptotically similar.

If the null hypothesis (that the reduced model "fits") is true, then G^2 and X^2 both have approximate chi-square distributions with $k - l - 1$ df (# parameters estimated in alternative $-$ # parameters estimated in null)

Rejection region: reject H_0 if $G^2 \geq \chi_\alpha^2$, where the chi-square percentile corresponds to $k - l - 1$ df

Example: are human blood types in H-W proportions?

3 alleles A, B, O with proportions a, b, o in population
(where $a + b + o = 1$; there are *two* unknown parameters)

genotypes	phenotypes	frequencies (H-W)
AA, AO	type A	$\pi_1 = a^2 + 2ao$
BB, BO	type B	$\pi_2 = b^2 + 2bo$
AB	type AB	$\pi_3 = 2ab$
OO	type O	$\pi_4 = o^2$

Data: $y_1 = 182,\ y_2 = 60,\ y_3 = 17,\ y_4 = 176,\ n = 435$

ML estimates under H_0: H-W proportions
(computer maximization of L_0):

$$\widehat{a} = 0.2644431$$
$$\widehat{b} = 0.09316881$$
$$\widehat{o} = 0.64238687$$

	observed
$\widehat{E}_1 = n\widetilde{\pi}_1 = n\left(\widehat{a}^2 + 2\widehat{a}\widehat{o}\right) = 178.21163$	182
$\widehat{E}_2 = n\widetilde{\pi}_2 = n\left(\widehat{b}^2 + 2\widehat{b}\widehat{o}\right) = 55.845851$	60
$\widehat{E}_3 = n\widetilde{\pi}_3 = n\left(2\widehat{a}\widehat{b}\right) = 21.435027$	17
$\widehat{E}_4 = n\widetilde{\pi}_4 = n\left(\widehat{o}^2\right) = 179.50749$	176

$$G^2 = 2\sum_{j=1}^{k} y_j \log_e\left(\frac{y_j}{\widehat{E}_j}\right) = 1.4390029$$

$(X^2 = 1.37570895)$

$\text{df} = 4 - 2 - 1 = 1$

$\chi_{0.05}^2 = 3.843$

$G^2 < \chi_{0.05}^2$ so do not reject H_0