THE MULTINOMIAL DISTRIBUTION

Discrete distribution -- The Outcomes Are Discrete. A generalization of the binomial distribution from only 2 outcomes to k outcomes.

Typical Multinomial Outcomes:

red	A	area1	year1
white	B	area2	year2
blue	C	area3	year3
	D	area4	year4
	F	area5	**never**

Individual trials are **independent.**
Outcomes are **mutually exclusive and all inclusive.**

Throwing Dice and the Multinomial Distribution

Assume that a die is thrown 60 times (n=60) and a record is kept of the number of times a 1, 2, 3, 4, 5, or 6 is observed. The outcomes of these 60 *independent* trials are shown below:

"Face"	Number	Notation
1	13	y_1
2	10	y_2
3	8	y_3
4	10	y_4
5	12	y_5
6	7	y_6
	60	n

Each trial (e.g., throw of a die) has a Mutually Exclusive Outcome (1 or 2 or 3 or . . . or 6). Note that there is a type of **dependency** in the cell counts in that once

$$n \text{ and } y_1, y_2, y_3, y_4, \text{ and } y_5$$

are known, then y_6 can be gotten by subtraction, because the total (n) is known. Of course, the dependency applied to any count, not just y_6.

This dependency is seen in the binomial as it is not necessary to know the number of tails, if the number of heads and the total (n) are known. The “last" cell is redundant.

The multinomial distribution is useful in a large number of applications in ecology. Its **probability function** for k = 6 is

$$\boldsymbol{f(y_i \mid n, p_i) = \binom{n}{y_i} p_1^{y_1}\, p_2^{y_2}\, p_3^{y_3}\, p_4^{y_4}\, p_5^{y_5}\, p_6^{y_6}}$$

This allows one to compute the probability of various combinations of outcomes, given the number of trials and the parameters. That is, the parameters must be known.

The **multinomial coefficient** $\binom{n}{y_i}$ is shorthand for

$$n!/\Big((y_1)!\,(y_2)!\ \cdot\cdot\cdot\ (y_k)!\Big),$$

where ! is the factorial operator (5! = 5 · 4 · 3 · 2 · 1 = 120). This term does not involve any of the unknown parameters and is ignored for many estimation issues.

In the die tossing data, k = 6 and the multinomial coefficient is

$$60!\Big/\Big(13!\ 10!\ 8!\ 10!\ 12!\ 7!\Big),$$

which is a very large number.

Some examples: Suppose you roll a fair die 6 times (6 trials), First, assume (y_1, y_2, y_3, y_4, y_5, y_6) is a multinomial random variable with parameters

$$p_1 = p_2 = \cdot\ \cdot\ \cdot = p_6 = 1/6 \text{ and } n = 6.$$

What is the probability of that each face is seen exactly once? This is simply

$$f(1, 1, 1, 1, 1, 1 \mid 6, 1/6, 1/6, 1/6, 1/6, 1/6, 1/6) = \frac{6!}{(1!)^6}\left(\frac{1}{6}\right)^6 = \frac{5}{324}\,.$$

What is the probability that exactly four 1's and two 2's occur? Then,

$$f(4, 2, 0, 0, 0, 0 \mid 6, 1/6, 1/6, 1/6, 1/6, 1/6, 1/6) = \frac{6!}{4!\,2!\,(0!)^4} \left(\frac{1}{6}\right)^4 \left(\frac{1}{6}\right)^2 = \frac{5}{15552},$$

hardly a high probability.

What is the probability of getting exactly two 3's two 4's and two 5's? Try this and get familiar with the notation and use of the probability function. You can see why such a tool might be useful if you were a gambler and wanted to know something quantitative about “the odds" of various outcomes. Hopefully, your answer will be about 5/2592.

Biologists have the reverse problem in their research. They do not know the parameters – they want to estimate parameters from data, using a model. These issues are the domain of the likelihood and log-likelihood functions.

If the die is “fair" we know that the probability of any of the 6 outcomes is 1/6. But, if the die is known to be unfair, how might we estimate the probabilities (p_i), i= 1, 2, ..., 6, that underlie the data observed?

The key to this estimation issue is the **multinomial distribution** and, particularly the **likelihood and log-likelihood functions.**

$$\mathcal{L}(\theta \mid \text{data}) \quad \text{or} \quad \mathcal{L}(p_i \mid n, y_i)$$

“the likelihood of the parameters, *given* the data."

At first, the likelihood function looks messy but it is only a different view of the probability function. Both functions assume n is given; the probability function assumes the parameters are given, while the likelihood function assumes the data are given. The likelihood function for the multinomial distribution is

$$\mathcal{L}(p_i \mid n, y_i) = \binom{n}{y_i} p_1^{y_1}\, p_2^{y_2}\, p_3^{y_3}\, p_4^{y_4}\, p_5^{y_5}\, p_6^{y_6}$$

The first term (multinomial coefficient--more on this below) is a constant and does not involve any of the unknown parameters, thus we often ignore it.

Note, $\sum p_i = 1$, does this make sense to you? Why?

Because of the dependency, there are only 5 “free" parameters, the 6^{th} one is defined by the other 5 and the total, n. We will use the symbol K to denote the total number of estimable parameters in a model; here $K = 5$. Then, the likelihood function could be written as

$$\boldsymbol{\mathcal{L}(p_i \mid n, y_i) = \binom{n}{y_i} p_1^{y_1}\, p_2^{y_2}\, p_3^{y_3}\, p_4^{y_4}\, p_5^{y_5} \left(1-\sum_{i=1}^{5} p_i\right)^{n-\sum_{i=1}^{5} y_i}}$$

This gets a bit awkward, but necessary to keep the concept clearly in mind.

If the die had 10-20 faces, the likelihood would be messy to write out. Thus, a shorthand notation is merely,

$$\mathcal{L}(p_i \mid n, y_i) = \mathrm{C} \prod_{\mathrm{i}=1}^{\mathrm{k}} p_i^{y_i}$$

where C is the multinomial coefficient and the symbol $\prod$ is the product operator. Here, one must remember that the final term (k^{th}) is actually

$$\left(1-\sum_{i=1}^{k-1} p_i\right)^{n-\sum_{i=1}^{k-1} y_i} .$$

Products are often difficult to work with, thus the **log-likelihood function** is of primary interest,

$$\log_e\left(\mathcal{L}(p_i \mid n, y_i)\right)$$

or, often

$$\log_e(\mathcal{L})$$

for short, or more generally

$$\log_e \mathcal{L}(\theta \mid \text{data}),$$

where θ = parameters (e.g., the p_i) and the data are given. Not only is this convenient, but it is the basis for many procedures in statistics. For multinomial random variables, the log-likelihood is

$$\log_e\Big(\mathcal{L}(p_i \mid n, y_i)\Big) = \log_e(\mathrm{C}) + y_1\log_e(p_1) + y_2\log_e(p_2) + \cdot \cdot \cdot + y_k\log_e(p_k).$$

Taking natural logarithms makes products into sums. A shorthand notation is

$$\log_e\Big(\mathcal{L}(p_i \mid n, y_i)\Big) = \log_e(\mathrm{C}) + \sum_{\mathrm{i}=1}^{k} y_i \log_e(p_i).$$

The log-likelihood function links the *DATA* (*n*, y_i) with the unknown *PARAMETERS* (p_i) through a *MODEL* and makes implicit the *ASSUMPTIONS*. This is the basis for rigorous inference.

The log-likelihood function is the (optimal) basis for estimation of parameters and their precision (variance, standard errors, coefficients of variation and confidence intervals), in addition to other important quantities.

Note, each term that includes the parameters in the log-likelihood function is of the form

DATA * LOG(PROBABILITY).

In the example, just above, the DATA are y_i and PROBABILITY is p_i, thus

$$\mathbf{y}_i \cdot \log_e(p_i)\ .$$

The typical log-likelihood function is the sum of such terms (plus, sometimes, the binomial or multinomial coefficient, which does not involve the parameters). Get used to seeing log-likelihood functions in this form,

$$\sum_{i=1}^{k} y_i \cdot \log_e(p_i) \ .$$

The Fisher Information Matrix and the Variance-Covariance Matrix

Measures of precision of the parameter estimator or notion of repeatability.

Reference: Section 1.2.1.2 (pages 12-14) in Burnham et al. (1987).