Multinomial distribution

From Wikipedia, the free encyclopedia

In probability theory, the **multinomial distribution** is a generalization of the binomial distribution. For example it models the probability of counts for rolling a *k* sided die *n* times. For *n* independent trials each of which leads to a success for exactly one of *k* categories, with each category having a given fixed success probability, the multinomial distribution gives the probability of any particular combination of numbers of successes for the various categories.

Multinomial

Parameters	$n > 0$ number of trials (integer) $p_1, \ldots, p_k$ event probabilities ($\Sigma p_i = 1$)
Support	$X_i \in \{0, \ldots, n\}$ $\Sigma X_i = n$
pmf	$\frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$
Mean	$E\{X_i\} = np_i$
Variance	$\mathrm{Var}(X_i) = np_i(1-p_i)$ $\mathrm{Cov}(X_i, X_j) = -np_ip_j \quad (i \neq j)$
MGF	$\left(\sum_{i=1}^k p_i e^{t_i}\right)^n$
CF	$\left(\sum_{j=1}^k p_j e^{it_j}\right)^n$ where $i^2 = -1$
PGF	$\left(\sum_{i=1}^k p_i z_i\right)^n$ for $(z_1, \ldots, z_k) \in \mathbb{C}^k$

When *n* is 1 and *k* is 2 the **multinomial distribution** is the Bernoulli distribution. When *k* is 2 and number of trials are more than 1 it is the Binomial distribution. When *n* is 1 it is the categorical distribution.

The Bernoulli distribution is the probability distribution of whether a Bernoulli trial is a success. In other words it models the number of heads from flipping a coin one time. The binomial distribution generalizes this to the number of heads from doing *n* independent flips of the same coin. For the multinomial distribution the analog to the Bernoulli Distribution is the categorical distribution. Instead of flipping one coin, the categorical distribution models the roll of one *k* sided die. So the multinomial distribution can model *n* independent rolls of a *k* sided die.

Let *k* be a fixed finite number. Mathematically, we have *k* possible mutually exclusive outcomes, with corresponding probabilities $p_1, \ldots, p_k$, and *n* independent trials. Note that since the *k* outcomes are mutually exclusive and one must occur we have $p_i \geq 0$ for $i = 1, \ldots, k$ and $\sum_{i=1}^k p_i = 1$. Then if the random variables X_i indicate the number of times outcome number *i* is observed over the *n* trials, the vector $X = (X_1, \ldots, X_k)$ follows a multinomial distribution with parameters *n* and **p**, where $\mathbf{p} = (p_1, \ldots, p_k)$. While the trials are independent, their outcomes *X* are dependent because they must be summed to n.

Note that, in some fields, such as natural language processing, the categorical and multinomial distributions are conflated, and it is common to speak of a "multinomial distribution" when a categorical distribution is actually meant. This stems from the fact that it is sometimes convenient to express the outcome of a categorical distribution as a "1-of-K" vector (a vector with one element containing a 1 and all other elements containing a 0) rather than as an integer in the range $1 \ldots K$; in this form, a categorical distribution is equivalent to a multinomial distribution over a single trial.

Contents

- 1 Specification
  - 1.1 Probability mass function
- 2 Visualization

- 2.1 As slices of generalized Pascal's triangle
- 2.2 As polynomial coefficients
- 3 Properties
  - 3.1 Matrix notation
- 4 Example
- 5 Sampling from a multinomial distribution
- 6 To simulate a multinomial distribution
- 7 Related distributions
- 8 References

Specification

Probability mass function

Suppose one does an experiment of extracting *n* balls of *k* different colours from a bag, replacing the extracted ball after each draw. Balls from the same colour are equivalent. Denote the variable which is the number of extracted balls of colour *i* ($i = 1, ..., k$) as X_i, and denote as p_i the probability that a given extraction will be in colour *i*. The probability mass function of this multinomial distribution is:

$$f(x_1,\ldots,x_k;n,p_1,\ldots,p_k) = \Pr(X_1 = x_1 \text{ and } \ldots \text{ and } X_k = x_k)$$

$$= \begin{cases} \dfrac{n!}{x_1!\cdots x_k!} p_1^{x1}\cdots p_k^{x_k}, & \text{when } \sum_{i=1}^k x_i = n \\ \\ 0 & \text{otherwise,} \end{cases}$$

for non-negative integers $x_1, ..., x_k$.

The probability mass function can be expressed using the gamma function as:

$$f(x_1,\ldots,x_k;p_1,\ldots,p_k) = \frac{\Gamma(\sum_i x_i + 1)}{\prod_i \Gamma(x_i+1)} \prod_{i=1}^k p_i^{x_i}.$$

This form shows its resemblance to the Dirichlet distribution which is its conjugate prior.

Visualization

As slices of generalized Pascal's triangle

Just like one can interpret the binomial distribution as (normalized) 1D slices of Pascal's triangle, so too can one interpret the multinomial distribution as 2D (triangular) slices of Pascal's pyramid, or 3D/4D/+ (pyramid-shaped) slices of higher-dimensional analogs of Pascal's triangle. This reveals an interpretation of the range of the distribution: discretized equilaterial "pyramids" in arbitrary dimension—i.e. a simplex with a grid.

As polynomial coefficients

Similarly, just like one can interpret the binomial distribution as the polynomial coefficients of $(px_1 + (1-p)x_2)^n$ when expanded, one can interpret the multinomial distribution as the coefficients of $(p_1x_1 + p_2x_2 + p_3x_3 + ... + p_kx_k)^n$ when expanded. (Note that just like the binomial distribution, the coefficients must sum to 1.) This is the origin of the name "*multinomial* distribution".

Properties

The expected number of times the outcome *i* was observed over *n* trials is

$$\operatorname{E}(X_i) = np_i.$$

The covariance matrix is as follows. Each diagonal entry is the variance of a binomially distributed random variable, and is therefore

$$\operatorname{var}(X_i) = np_i(1-p_i).$$

The off-diagonal entries are the covariances:

$$\operatorname{cov}(X_i, X_j) = -np_i p_j$$

for i, j distinct.

All covariances are negative because for fixed n, an increase in one component of a multinomial vector requires a decrease in another component.

This is a $k \times k$ positive-semidefinite matrix of rank $k - 1$. In the special case where $k = n$ and where the p_i are all equal, the covariance matrix is the centering matrix.

The entries of the corresponding correlation matrix are

$$\rho(X_i, X_i) = 1.$$

$$\rho(X_i, X_j) = \frac{\operatorname{cov}(X_i, X_j)}{\sqrt{\operatorname{var}(X_i)\operatorname{var}(X_j)}} = \frac{-p_i p_j}{\sqrt{p_i(1-p_i)p_j(1-p_j)}} = -\sqrt{\frac{p_i p_j}{(1-p_i)(1-p_j)}}.$$

Note that the sample size drops out of this expression.

Each of the k components separately has a binomial distribution with parameters n and p_i, for the appropriate value of the subscript i.

The support of the multinomial distribution is the set

$$\{(n_1, \dots, n_k) \in \mathbb{N}^k | n_1 + \cdots + n_k = n\}.$$

Its number of elements is

$${n+k-1 \choose k-1}.$$

Matrix notation

In matrix notation,

$$\operatorname{E}(\mathbf{X}) = n\mathbf{p},$$

and

$$\operatorname{var}(\mathbf{X}) = n\{\operatorname{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^{\mathrm{T}}\},$$

with $\mathbf{p}^{\mathrm{T}}$ = the row vector transpose of the column vector $\mathbf{p}$.

Example

In a recent three-way election for a large country, candidate A received 20% of the votes, candidate B received 30% of the votes, and candidate C received 50% of the votes. If six voters are selected randomly, what is the probability that there will be exactly one supporter for candidate A, two supporters for candidate B and three supporters for candidate C in the sample?

Note: Since we're assuming that the voting population is large, it is reasonable and permissible to think of the probabilities as unchanging once a voter is selected for the sample. Technically speaking this is sampling without replacement, so the correct distribution is the multivariate hypergeometric distribution, but the distributions converge as the population grows large.

$$\Pr(A = 1, B = 2, C = 3) = \frac{6!}{1!2!3!}(0.2^1)(0.3^2)(0.5^3) = 0.135$$

Sampling from a multinomial distribution

First, reorder the parameters $p_1, \ldots, p_k$ such that they are sorted in descending order (this is only to speed up computation and not strictly necessary). Now, for each trial, draw an auxiliary variable *X* from a uniform (0, 1) distribution. The resulting outcome is the component

$$j = \min\left\{ j' \in \{1, \ldots, k\} : \left(\sum_{i=1}^{j'} p_i\right) - X \geq 0 \right\}.$$

$\{X_j = 1, X_k = 0 \text{ for } k \neq j\}$ is one observation from the multinomial distribution with $p_1, \ldots, p_k$ and $n = 1$. A sum of independent repetitions of this experiment is an observation from a multinomial distribution with n equal to the number of such repetitions.

To simulate a multinomial distribution

Various methods may be used to simulate a multinomial distribution. A very simple one is to use a random number generator to generate numbers between 0 and 1. First, we divide the interval from 0 to 1 in k subintervals equal in size to the probabilities of the k categories. Then, we generate a random number for each of n trials and use a logical test to classify the virtual measure or observation in one of the categories.

Example

If we have :

Categories	1	2	3	4	5	6
Probabilities	0.15	0.20	0.30	0.16	0.12	0.07
Superior limits of subintervals	0.15	0.35	0.65	0.81	0.93	1.00

Then, with a software like Excel, we may use the following recipe:

Cells :	Ai	Bi	Ci	...	Gi
Formulae :	Rand()	=If($Ai<0.15;1;0)	=If(And($Ai>=0.15;$Ai<0.35);1;0)	...	=If($Ai>=0.93;1;0)

After that, we will use functions such as SumIf to accumulate the observed results by category and to calculate the estimated covariance matrix for each simulated sample.

Another way is to use a discrete random number generator. In that case, the categories must be labeled or relabeled with numeric values.

In the two cases, the result is a multinomial distribution with k categories without any correlation. This is equivalent, with a continuous random distribution, to simulate k independent standardized normal distributions, or a multinormal distribution N(0,I) having k components identically distributed and statistically independent.

Related distributions

- When $k = 2$, the multinomial distribution is the binomial distribution.
- Categorical distribution, the distribution of each trial; for $k = 2$, this is the Bernoulli distribution.
- The Dirichlet distribution is the conjugate prior of the multinomial in Bayesian statistics.
- Dirichlet-multinomial distribution.
- Beta-binomial model.
- Negative multinomial distribution
- Hardy–Weinberg principle (it is a trinomial distribution with probabilities $(\theta^2, 2\theta(1-\theta), (1-\theta)^2)$)

References

- Evans, Morton; Hastings, Nicholas; Peacock, Brian (2000). *Statistical Distributions*. New York: Wiley. pp. 134–136. ISBN 0-471-37124-6. 3rd ed.

Categories: Discrete distributions | Multivariate discrete distributions | Factorial and binomial topics | Exponential family distributions | Probability distributions