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 each side of a k-sided die rolled 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.

When k is 2 and n is 1, the multinomial distribution is the Bernoulli distribution. When k is 2 and n is bigger than 1, it is the binomial distribution. When k is bigger than 2 and n is 1, it is the categorical distribution.

The Bernoulli distribution models the outcome of a single Bernoulli trial. In other words, it models whether flipping a (possibly biased) coin one time will result in either a success (obtaining a head) or failure (obtaining a tail). The binomial distribution generalizes this to the number of heads from performing n independent flips (Bernoulli trials) of the same coin. The multinomial distribution models the outcome of n experiments, where the outcome of each trial has a categorical distribution, such as rolling a k-sided die n times.

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. 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 $\boldsymbol{X} = (X_1, \ldots, X_k)$ follows a multinomial distribution with parameters n and $\mathbf{p}$, 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.

In some fields such as natural language processing, categorical and multinomial distributions are synonymous 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.

Multinomial

Parameters	$n > 0$ number of trials (integer) $p_1, \ldots, p_k$ event probabilities ($\sum p_i = 1$)
Support	$x_i \in \{0, \ldots, n\},\ i \in \{1, \ldots, k\}$, with $\sum_i x_i = n$
PMF	$\frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$
Mean	$\mathrm{E}(X_i) = np_i$
Variance	$\mathrm{Var}(X_i) = np_i(1-p_i)$ $\mathrm{Cov}(X_i, X_j) = -np_ip_j \quad (i \neq j)$
Entropy	$-\log(n!) - n\sum_{i=1}^{k} p_i \log(p_i) + \sum_{i=1}^{k}\sum_{x_i=0}^{n} \binom{n}{x_i} p_i^{x_i}(1-p_i)^{n-x_i} \log(x_i!)$
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 $(z_1, \ldots, z_k) \in \mathbb{C}^k$

Contents [hide]

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 from a multinomial distribution
7 Equivalence tests for multinomial distributions
8 Related distributions
9 References
9.1 Citations
9.2 Sources

Specification [edit]

Probability mass function [edit]

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

$$f(x_1, \ldots, x_k; n, p_1, \ldots, p_k) = \Pr(X_1 = x_1 \text{ and } \ldots \text{ and } X_k = x_k)$$
$$= \begin{cases} \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \times \cdots \times p_k^{x_k}, & \text{when } \sum_{i=1}^{k} x_i = n \\ 0 & \text{otherwise,} \end{cases}$$

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

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

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

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

Visualization [edit]

As slices of generalized Pascal's triangle [edit]

Just like one can interpret the binomial distribution as (normalized) one-dimensional (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 equilateral "pyramids" in arbitrary dimension—i.e. a simplex with a grid.[citation needed]

As polynomial coefficients [edit]

Similarly, just like one can interpret the binomial distribution as the polynomial coefficients of $(p + (1-p))^n$ when expanded, one can interpret the multinomial distribution as the coefficients of $(p_1 + p_2 + p_3 + \cdots + p_k)^n$ 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 [edit]

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

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

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

$$\mathrm{Var}(X_i) = np_i(1-p_i).$$

The off-diagonal entries are the covariances:

$$\mathrm{Cov}(X_i, X_j) = -np_ip_j$$

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.

When these expressions are combined into a matrix with i, j element $\mathrm{cov}(X_i, X_j)$, the result is a $k \times k$ positive-semidefinite covariance 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

$$\rho(X_i, X_i) = 1.$$
$$\rho(X_i, X_j) = \frac{\mathrm{Cov}(X_i, X_j)}{\sqrt{\mathrm{Var}(X_i)\,\mathrm{Var}(X_j)}} = \frac{-p_ip_j}{\sqrt{p_i(1-p_i)p_j(1-p_j)}} = -\sqrt{\frac{p_ip_j}{(1-p_i)(1-p_j)}}.$$

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

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

Its number of elements is

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

Matrix notation [edit]

In matrix notation,

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

and

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

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

Example [edit]

Suppose that in a 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 [edit]

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

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

($X_j = 1$, $X_k = 0$ 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 from a multinomial distribution [edit]

Various methods may be used to simulate from a multinomial distribution. A very simple solution is to use a uniform pseudo-random number generator on (0,1). First, we divide the (0,1) interval in k subintervals equal in length to the probabilities of the k categories. Then, we generate n independent pseudo-random numbers to determine in which of the k intervals they occur and count the number of occurrences in each interval.

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 software such as 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. 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.

Since the counts of all categories have to sum to the number of trials, the counts of the categories are always negatively correlated.[1]

Equivalence tests for multinomial distributions [edit]

The goal of equivalence testing is to establish the agreement between a theoretical multinomial distribution and observed counting frequencies. The theoretical distribution may be a fully specified multinomial distribution or a parametric family of multinomial distributions.

Let q denote a theoretical multinomial distribution and let p be a true underlying distribution. The distributions p and q are considered equivalent if $d[p, q] < \varepsilon$ for a distance d and a tolerance parameter $\varepsilon > 0$. The equivalence test problem is $H_0 = \{d[p, q] \geq \varepsilon\}$ versus $H_1 = \{d[p, q] < \varepsilon\}$. The true underlying distribution p is unknown. Instead, the counting frequencies p_n are observed, where n is a sample size. An equivalence test uses p_n to reject H_0. If H_0 can be rejected then the equivalence between p and q is shown at a given significance level. The equivalence test for Euclidean distance can be found in text book of Wellek (2010).[2] The equivalence test for the total variation distance is developed in Ostrovski (2017).[3] The exact equivalence test for the specific cumulative distance proposed in Frey (2009).[4]

The distance between the true underlying distribution p and a family of the multinomial distributions $\mathcal{M}$ is defined by $d[p, \mathcal{M}] = \min_{h \in \mathcal{M}} d[p, h]$. Then the equivalence test problem is given by $H_0 = \{d[p, \mathcal{M}] \geq \varepsilon\}$ and $H_1 = \{d[p, \mathcal{M}] < \varepsilon\}$. The distance $d[p, \mathcal{M}]$ is usually computed using numerical optimization. The tests for this case are developed recently in Ostrovski (2018).[5]

Related distributions [edit]

- 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 [edit]

Citations [edit]

1. ^ "1.7 - The Multinomial Distribution | STAT 504".*onlinecourses.science.psu.edu*. Retrieved 2016-09-11.
2. ^ Wellek, Stefan (2010). *Testing statistical hypotheses of equivalence and noninferiority*. Chapman and Hall/CRC. ISBN 978-1439808184.
3. ^ Ostrovski, Vladimir (May 2017). "Testing equivalence of multinomial distributions". *Statistics & Probability Letters*. **124**: 77–82. doi:10.1016/j.spl.2017.01.004⧀. S2CID 126293429⧀.Official web link (subscription required)⧀. Alternate, free web link⧀.
4. ^ Frey, Jesse (March 2009). "An exact multinomial test for equivalence". *The Canadian Journal of Statistics*. **37**: 47–59. doi:10.1002/cjs.10000⧀.Official web link (subscription required)⧀.
5. ^ Ostrovski, Vladimir (March 2018). "Testing equivalence to families of multinomial distributions with application to the independence model". *Statistics & Probability Letters*. **139**: 61–66. doi:10.1016/j.spl.2018.03.014⧀. S2CID 126261081⧀.Official web link (subscription required)⧀. Alternate, free web link⧀.

Sources [edit]

- Evans, Morton; Hastings, Nicholas; Peacock, Brian (2000). *Statistical Distributions* (3rd ed.). New York: Wiley. pp. 134⧀–136. ISBN 0-471-37124-6.
- Weisstein, Eric W. "Multinomial Distribution"⧀. *MathWorld*. Wolfram Research.

v·t·e **Probability distributions (List)** [hide]

Discrete univariate with finite support	Benford · Bernoulli · beta-binomial · binomial · categorical · hypergeometric · Poisson binomial · Rademacher · soliton · discrete uniform · Zipf · Zipf–Mandelbrot
Discrete univariate with infinite support	beta negative binomial · Borel · Conway–Maxwell–Poisson · discrete phase-type · Delaporte · extended negative binomial · Flory–Schulz · Gauss–Kuzmin · geometric · logarithmic · negative binomial · Panjer · parabolic fractal · Poisson · Skellam · Yule–Simon · zeta
Continuous univariate supported on a bounded interval	arcsine · ARGUS · Balding–Nichols · Bates · beta · beta rectangular · continuous Bernoulli · Irwin–Hall · Kumaraswamy · logit-normal · noncentral beta · PERT · raised cosine · reciprocal · triangular · U-quadratic · uniform · Wigner semicircle
Continuous univariate supported on a semi-infinite interval	Benini · Benktander 1st kind · Benktander 2nd kind · beta prime · Burr · chi-squared · chi · Dagum · Davis · exponential-logarithmic · Erlang · exponential · F · folded normal · Fréchet · gamma · gamma/Gompertz · generalized gamma · generalized inverse Gaussian · Gompertz · half-logistic · half-normal · Hotelling's T-squared · hyper-Erlang · hypoexponential · hypoexponential · inverse chi-squared (scaled inverse chi-squared) · inverse Gaussian · inverse gamma · Kolmogorov · Lévy · log-Cauchy · log-Laplace · log-logistic · log-normal · Lomax · matrix-exponential · Maxwell–Boltzmann · Maxwell–Jüttner · Mittag-Leffler · Nakagami · noncentral chi-squared · noncentral F · Pareto · phase-type · poly-Weibull · Rayleigh · relativistic Breit–Wigner · Rice · shifted Gompertz · truncated normal · type-2 Gumbel · Weibull (discrete Weibull) · Wilks's lambda
Continuous univariate supported on the whole real line	Cauchy · exponential power · Fisher's z · Gaussian q · generalized normal · generalized hyperbolic · geometric stable · Gumbel · Hottsmark · hyperbolic secant · Johnson's S_U · Landau · Laplace · asymmetric Laplace · logistic · noncentral t · normal (Gaussian) · normal-inverse Gaussian · skew normal · slash · stable · Student's t · type-1 Gumbel · Tracy–Widom · variance-gamma · Voigt
Continuous univariate with support whose type varies	generalized chi-squared · generalized extreme value · generalized Pareto · Marchenko–Pastur · q-exponential · q-Gaussian · q-Weibull · shifted log-logistic · Tukey lambda
Mixed continuous-discrete univariate	rectified Gaussian
Multivariate (joint)	*Discrete:* Ewens · **multinomial** · Dirichlet-multinomial · negative multinomial · *Continuous:* Dirichlet · generalized Dirichlet · multivariate Laplace · multivariate normal · multivariate stable · multivariate t · normal-inverse-gamma · normal-gamma · *Matrix-valued:* inverse matrix gamma · inverse-Wishart · matrix normal · matrix t · matrix gamma · normal-inverse-Wishart · normal-Wishart · Wishart
Directional	*Univariate (circular) directional:* Circular uniform · univariate von Mises · wrapped normal · wrapped Cauchy · wrapped exponential · wrapped asymmetric Laplace · wrapped Lévy · *Bivariate (spherical):* Kent · *Bivariate (toroidal):* bivariate von Mises · *Multivariate:* von Mises–Fisher · Bingham
Degenerate and singular	*Degenerate:* Dirac delta function · *Singular:* Cantor
Families	Circular · compound Poisson · elliptical · exponential · natural exponential · location–scale · maximum entropy · mixture · Pearson · Tweedie · wrapped

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

This page was last edited on 4 May 2021, at 18:35 (UTC).

Text is available under the Creative Commons Attribution-ShareAlike License; additional terms may apply. By using this site, you agree to the Terms of Use and Privacy Policy. Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a non-profit organization.

Privacy policy About Wikipedia Disclaimers Contact Wikipedia Mobile view Developers Statistics Cookie statement