# Categorical distribution

In probability theory and statistics, a **categorical distribution** (also called a **generalized Bernoulli distribution**, **multinoulli distribution**[1]) is a discrete probability distribution that describes the possible results of a random variable that can take on one of *K* possible categories, with the probability of each category separately specified. There is no innate underlying ordering of these outcomes, but numerical labels are often attached for convenience in describing the distribution, (e.g. 1 to *K*). The *K*-dimensional categorical distribution is the most general distribution over a *K*-way event; any other discrete distribution over a size-*K* sample space is a special case. The parameters specifying the probabilities of each possible outcome are constrained only by the fact that each must be in the range 0 to 1, and all must sum to 1.

**Categorical**

| | |
|---|---|
| **Parameters** | $k > 0$ number of categories (integer)<br>$p_1, \ldots, p_k$ event probabilities<br>$(p_i > 0, \Sigma p_i = 1)$ |
| **Support** | $x \in \{1, \ldots, k\}$ |
| **PMF** | (1) $p(x = i) = p_i$<br>(2) $p(x) = p_1^{[x=1]} \cdots p_k^{[x=k]}$<br>(3) $p(x) = [x = 1] \cdot p_1 + \cdots + [x = k] \cdot p_k$<br>where $[x = i]$ is the Iverson bracket |
| **Mode** | $i$ such that $p_i = \max(p_1, \ldots, p_k)$ |

The categorical distribution is the generalization of the Bernoulli distribution for a categorical random variable, i.e. for a discrete variable with more than two possible outcomes, such as the roll of a die. On the other hand, the categorical distribution is a special case of the multinomial distribution, in that it gives the probabilities of potential outcomes of a single drawing rather than multiple drawings.

## Contents

# Terminology

Occasionally, the categorical distribution is termed the "discrete distribution". However, this properly refers not to one particular family of distributions but to a general class of distributions.

In some fields, such as machine learning and natural language processing, the categorical and multinomial distributions are conflated, and it is common to speak of a "multinomial distribution" when a "categorical distribution" would be more precise.[2] This imprecise usage 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 to *K*; in this form, a categorical distribution is equivalent to a multinomial distribution for a single observation (see below).

However, conflating the categorical and multinomial distributions can lead to problems. For example, in a Dirichlet-multinomial distribution, which arises commonly in natural language processing models (although not usually with this name) as a result of collapsed Gibbs sampling where Dirichlet distributions are collapsed out of a hierarchical Bayesian model, it is very important to distinguish categorical from multinomial. The joint distribution of the same variables with the same Dirichlet-multinomial distribution has two different forms depending on whether it is characterized as a distribution whose domain is over individual categorical nodes or over multinomial-style counts of nodes in each particular category (similar to the distinction between a set of Bernoulli-distributed nodes and a single binomial-distributed node). Both forms have very similar-looking probability mass functions (PMFs), which both make reference to multinomial-style counts of nodes in a category. However, the multinomial-style PMF has an extra factor, a multinomial coefficient, that is a constant equal to 1 in the categorical-style PMF. Confusing the two can easily lead to incorrect results in settings where this extra factor is not constant with respect to the distributions of interest. The factor is frequently constant in the complete conditionals used in Gibbs sampling and the optimal distributions in variational methods.

# Formulating distributions

A categorical distribution is a discrete probability distribution whose sample space is the set of *k* individually identified items. It is the generalization of the Bernoulli distribution for a categorical random variable.

In one formulation of the distribution, the sample space is taken to be a finite sequence of integers. The exact integers used as labels are unimportant; they might be {0, 1, ..., *k* − 1} or {1, 2, ..., *k*} or any other arbitrary set of values. In the following descriptions, we use {1, 2, ..., *k*} for convenience, although this disagrees with the convention for the Bernoulli distribution, which uses {0, 1}. In this case, the probability mass function *f* is:

$$f(x = i \mid \boldsymbol{p}) = p_i,$$

where $\boldsymbol{p} = (p_1, \ldots, p_k)$, $p_i$ represents the probability of seeing element *i* and $\sum_{i=1}^{k} p_i = 1$.

Another formulation that appears more complex but facilitates mathematical manipulations is as follows, using the Iverson bracket:[3]

$$f(x \mid \boldsymbol{p}) = \prod_{i=1}^{k} p_i^{[x=i]},$$

where $[x = i]$ evaluates to 1 if $x = i$, 0 otherwise. There are various advantages of this formulation, e.g.:

- It is easier to write out the likelihood function of a set of independent identically distributed categorical variables.
- It connects the categorical distribution with the related multinomial distribution.

- It shows why the Dirichlet distribution is the conjugate prior of the categorical distribution, and allows the posterior distribution of the parameters to be calculated.

Yet another formulation makes explicit the connection between the categorical and multinomial distributions by treating the categorical distribution as a special case of the multinomial distribution in which the parameter *n* of the multinomial distribution (the number of sampled items) is fixed at 1. In this formulation, the sample space can be considered to be the set of 1-of-*K* encoded[4] random vectors $\mathbf{x}$ of dimension *k* having the property that exactly one element has the value 1 and the others have the value 0. The particular element having the value 1 indicates which category has been chosen. The probability mass function *f* in this formulation is:

$$f(\mathbf{x} \mid \boldsymbol{p}) = \prod_{i=1}^{k} p_i^{x_i},$$

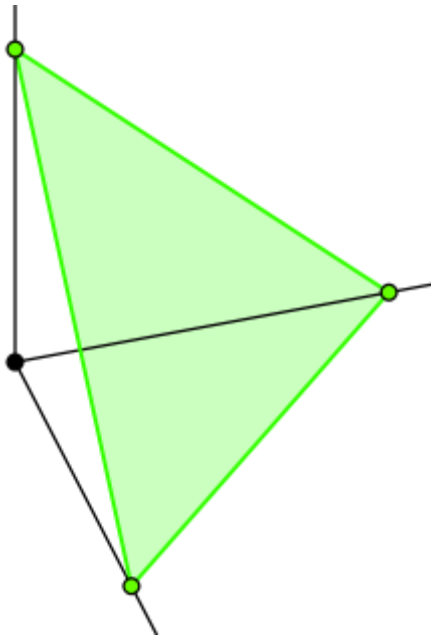
where $p_i$ represents the probability of seeing element *i* and $\sum_i p_i = 1$. This is the formulation adopted by Bishop.[4][note 1]

# Properties

- The distribution is completely given by the probabilities associated with each number *i*: $p_i = P(X = i)$, *i* = 1,...,*k*, where $\sum_i p_i = 1$. The possible sets of probabilities are exactly those in the standard $(k-1)$-dimensional simplex; for *k* = 2 this reduces to the possible probabilities of the Bernoulli distribution being the 1-simplex, $p_1 + p_2 = 1, 0 \leq p_1, p_2 \leq 1$.
- The distribution is a special case of a "multivariate Bernoulli distribution"[5] in which exactly one of the *k* 0-1 variables takes the value one.
- $\mathbf{E}[\mathbf{x}] = \boldsymbol{p}$
- Let $\boldsymbol{X}$ be the realisation from a categorical distribution. Define the random vector *Y* as composed of the elements:

  $$Y_i = I(\boldsymbol{X} = i),$$

  where *I* is the indicator function. Then *Y* has a distribution which is a special case of the multinomial distribution with parameter $n = 1$. The sum of $n$ independent and identically distributed such random variables *Y* constructed from a categorical distribution with parameter $\boldsymbol{p}$ is multinomially distributed with parameters $n$ and $\boldsymbol{p}$.



The possible probabilities for the categorical distribution with $k = 3$ are the 2-simplex $p_1 + p_2 + p_3 = 1$, embedded in 3-space.

- The conjugate prior distribution of a categorical distribution is a Dirichlet distribution.[2] See the section below for more discussion.
- The sufficient statistic from *n* independent observations is the set of counts (or, equivalently, proportion) of observations in each category, where the total number of trials (=*n*) is fixed.
- The indicator function of an observation having a value *i*, equivalent to the Iverson bracket function $[x = i]$ or the Kronecker delta function $\delta_{xi}$, is Bernoulli distributed with parameter $p_i$.

# Bayesian inference using conjugate prior

In Bayesian statistics, the Dirichlet distribution is the conjugate prior distribution of the categorical distribution (and also the multinomial distribution). This means that in a model consisting of a data point having a categorical distribution with unknown parameter vector **p**, and (in standard Bayesian style) we choose to treat this parameter as a random variable and give it a prior distribution defined using a Dirichlet distribution, then the posterior distribution of the parameter, after incorporating the knowledge gained from the observed data, is also a Dirichlet. Intuitively, in such a case, starting from what is known about the parameter prior to observing the data point, knowledge can then be updated based on the data point, yielding a new distribution of the same form as the old one. As such, knowledge of a parameter can be successively updated by incorporating new observations one at a time, without running into mathematical difficulties.

Formally, this can be expressed as follows. Given a model

$$\begin{array}{rcccl} \boldsymbol{\alpha} & = & (\alpha_1, \ldots, \alpha_K) & = & \text{concentration hyperparameter} \\ \mathbf{p} \mid \boldsymbol{\alpha} & = & (p_1, \ldots, p_K) & \sim & \operatorname{Dir}(K, \boldsymbol{\alpha}) \\ \mathbb{X} \mid \mathbf{p} & = & (x_1, \ldots, x_K) & \sim & \operatorname{Cat}(K, \mathbf{p}) \end{array}$$

then the following holds:[2]

$$\begin{array}{rcccl} \mathbf{c} & = & (c_1, \ldots, c_K) & = & \text{number of occurrences of category } i = \sum_{j=1}^{N} [x_j = i] \\ \mathbf{p} \mid \mathbb{X}, \boldsymbol{\alpha} & \sim & \operatorname{Dir}(K, \mathbf{c} + \boldsymbol{\alpha}) & = & \operatorname{Dir}(K, c_1 + \alpha_1, \ldots, c_K + \alpha_K) \end{array}$$

This relationship is used in Bayesian statistics to estimate the underlying parameter **p** of a categorical distribution given a collection of *N* samples. Intuitively, we can view the hyperprior vector **α** as pseudocounts, i.e. as representing the number of observations in each category that we have already seen. Then we simply add in the counts for all the new observations (the vector **c**) in order to derive the posterior distribution.

Further intuition comes from the expected value of the posterior distribution (see the article on the Dirichlet distribution):

$$\operatorname{E}[p_i \mid \mathbb{X}, \boldsymbol{\alpha}] = \frac{c_i + \alpha_i}{N + \sum_k \alpha_k}$$

This says that the expected probability of seeing a category *i* among the various discrete distributions generated by the posterior distribution is simply equal to the proportion of occurrences of that category actually seen in the data, including the pseudocounts in the prior distribution. This makes a great deal of intuitive sense: if, for example, there are three possible categories, and category 1 is seen in the observed data 40% of the time, one would expect on average to see category 1 40% of the time in the posterior distribution as well.

(This intuition is ignoring the effect of the prior distribution. Furthermore, the posterior is a *distribution over distributions*. The posterior distribution in general describes the parameter in question, and in this case the parameter itself is a discrete probability distribution, i.e. the actual categorical distribution that generated the data. For example, if 3 categories in the ratio 40:5:55 are in the observed data, then ignoring the effect of the prior distribution, the true parameter – i.e. the true, underlying distribution that generated our observed data – would be expected to have the average value of (0.40,0.05,0.55), which is indeed what the posterior reveals. However, the true distribution might actually be (0.35,0.07,0.58) or (0.42,0.04,0.54) or various other nearby possibilities. The amount of uncertainty involved here is specified by the variance of the posterior, which is controlled by the total number of observations – the more data observed, the less uncertainty about the true parameter.)

(Technically, the prior parameter $\boldsymbol{\alpha_i}$ should actually be seen as representing $\boldsymbol{\alpha_i - 1}$ prior observations of category $\boldsymbol{i}$. Then, the updated posterior parameter $\boldsymbol{c_i + \alpha_i}$ represents $\boldsymbol{c_i + \alpha_i - 1}$ posterior observations. This reflects the fact that a Dirichlet distribution with $\boldsymbol{\alpha = (1, 1, \ldots)}$ has a completely flat shape — essentially, a uniform distribution over the simplex of possible values of $\mathbf{p}$. Logically, a flat distribution of this sort represents total ignorance, corresponding to no observations of any sort. However, the mathematical updating of the posterior works fine if we ignore the $\boldsymbol{\cdots - 1}$ term and simply think of the $\boldsymbol{\alpha}$ vector as directly representing a set of pseudocounts. Furthermore, doing this avoids the issue of interpreting $\boldsymbol{\alpha_i}$ values less than 1.)

## MAP estimation

The maximum-a-posteriori estimate of the parameter $\mathbf{p}$ in the above model is simply the mode of the posterior Dirichlet distribution, i.e.,[2]

$$\boldsymbol{\arg\max_{\mathbf{p}} p(\mathbf{p} \mid \mathbb{X}) = \frac{\alpha_i + c_i - 1}{\sum_i (\alpha_i + c_i - 1)}, \qquad \forall i \; \alpha_i + c_i > 1}$$

In many practical applications, the only way to guarantee the condition that $\boldsymbol{\forall i \; \alpha_i + c_i > 1}$ is to set $\boldsymbol{\alpha_i > 1}$ for all *i*.

## Marginal likelihood

In the above model, the marginal likelihood of the observations (i.e. the joint distribution of the observations, with the prior parameter marginalized out) is a Dirichlet-multinomial distribution:[2]

$$\begin{aligned} p(\mathbb{X} \mid \boldsymbol{\alpha}) &= \int_{\mathbf{p}} p(\mathbb{X} \mid \mathbf{p}) p(\mathbf{p} \mid \boldsymbol{\alpha}) \mathrm{d}\mathbf{p} \\ &= \frac{\Gamma\left(\sum_k \alpha_k\right)}{\Gamma\left(N + \sum_k \alpha_k\right)} \prod_{k=1}^{K} \frac{\Gamma(c_k + \alpha_k)}{\Gamma(\alpha_k)} \end{aligned}$$

This distribution plays an important role in hierarchical Bayesian models, because when doing inference over such models using methods such as Gibbs sampling or variational Bayes, Dirichlet prior distributions are often marginalized out. See the article on this distribution for more details.

## Posterior predictive distribution

The posterior predictive distribution of a new observation in the above model is the distribution that a new observation $\tilde{\boldsymbol{x}}$ would take given the set $\mathbb{X}$ of *N* categorical observations. As shown in the Dirichlet-multinomial distribution article, it has a very simple form:[2]

$$\begin{aligned} p(\tilde{\boldsymbol{x}} = i \mid \mathbb{X}, \boldsymbol{\alpha}) &= \int_{\mathbf{p}} p(\tilde{\boldsymbol{x}} = i \mid \mathbf{p}) \, p(\mathbf{p} \mid \mathbb{X}, \boldsymbol{\alpha}) \, \mathrm{d}\mathbf{p} \\ &= \frac{c_i + \alpha_i}{N + \sum_k \alpha_k} \\ &= \mathbb{E}[p_i \mid \mathbb{X}, \boldsymbol{\alpha}] \\ &\propto c_i + \alpha_i. \end{aligned}$$

There are various relationships among this formula and the previous ones:

- The posterior predictive probability of seeing a particular category is the same as the relative proportion of previous observations in that category (including the pseudo-observations of the prior). This makes logical sense — intuitively, we would expect to see a particular category according to the frequency already observed of that category.
- The posterior predictive probability is the same as the expected value of the posterior distribution. This is explained more below.
- As a result, this formula can be expressed as simply "the posterior predictive probability of seeing a category is proportional to the total observed count of that category", or as "the *expected count* of a category is the same as the total observed count of the category", where "observed count" is taken to include the pseudo-observations of the prior.

The reason for the equivalence between posterior predictive probability and the expected value of the posterior distribution of $\mathbf{p}$ is evident with re-examination of the above formula. As explained in the posterior predictive distribution article, the formula for the posterior predictive probability has the form of an expected value taken with respect to the posterior distribution:

$$\begin{aligned} p(\tilde{x} = i \mid \mathbb{X}, \boldsymbol{\alpha}) &= \int_{\mathbf{p}} p(\tilde{x} = i \mid \mathbf{p})\, p(\mathbf{p} \mid \mathbb{X}, \boldsymbol{\alpha})\, \mathrm{d}\mathbf{p} \\ &= \mathrm{E}_{\mathbf{p}|\mathbb{X},\boldsymbol{\alpha}}[p(\tilde{x} = i \mid \mathbf{p})] \\ &= \mathrm{E}_{\mathbf{p}|\mathbb{X},\boldsymbol{\alpha}}[p_i] \\ &= \mathrm{E}[p_i \mid \mathbb{X}, \boldsymbol{\alpha}]. \end{aligned}$$

The crucial line above is the third. The second follows directly from the definition of expected value. The third line is particular to the categorical distribution, and follows from the fact that, in the categorical distribution specifically, the expected value of seeing a particular value *i* is directly specified by the associated parameter $p_i$. The fourth line is simply a rewriting of the third in a different notation, using the notation farther up for an expectation taken with respect to the posterior distribution of the parameters.

Observe data points one by one and each time consider their predictive probability before observing the data point and updating the posterior. For any given data point, the probability of that point assuming a given category depends on the number of data points already in that category. In this scenario, if a category has a high frequency of occurrence, then new data points are more likely to join that category — further enriching the same category. This type of scenario is often termed a preferential attachment (or "rich get richer") model. This models many real-world processes, and in such cases the choices made by the first few data points have an outsize influence on the rest of the data points.

## Posterior conditional distribution

In Gibbs sampling, one typically needs to draw from conditional distributions in multi-variable Bayes networks where each variable is conditioned on all the others. In networks that include categorical variables with Dirichlet priors (e.g. mixture models and models including mixture components), the Dirichlet distributions are often "collapsed out" (marginalized out) of the network, which introduces dependencies among the various categorical nodes dependent on a given prior (specifically, their joint distribution is a Dirichlet-multinomial distribution). One of the reasons for doing this is that in such a case, the distribution of one categorical node given the others is exactly the posterior predictive distribution of the remaining nodes.

That is, for a set of nodes $\mathbb{X}$, if the node in question is denoted as $x_n$ and the remainder as $\mathbb{X}^{(-n)}$, then

$$p(x_n = i \mid \mathbb{X}^{(-n)}, \boldsymbol{\alpha}) = \frac{c_i^{(-n)} + \alpha_i}{N - 1 + \sum_i \alpha_i} \qquad \propto c_i^{(-n)} + \alpha_i$$

where $c_i^{(-n)}$ is the number of nodes having category *i* among the nodes other than node *n*.

# Sampling

There are a number of methods, but the most common way to sample from a categorical distribution uses a type of inverse transform sampling:

Assume a distribution is expressed as "proportional to" some expression, with unknown normalizing constant. Before taking any samples, one prepares some values as follows:

1. Compute the unnormalized value of the distribution for each category.
2. Sum them up and divide each value by this sum, in order to normalize them.
3. Impose some sort of order on the categories (e.g. by an index that runs from 1 to *k*, where *k* is the number of categories).
4. Convert the values to a cumulative distribution function (CDF) by replacing each value with the sum of all of the previous values. This can be done in time *O(k)*. The resulting value for the first category will be 0.

Then, each time it is necessary to sample a value:

1. Pick a uniformly distributed number between 0 and 1.
2. Locate the greatest number in the CDF whose value is less than or equal to the number just chosen. This can be done in time *O(log(k))*, by binary search.
3. Return the category corresponding to this CDF value.

If it is necessary to draw many values from the same categorical distribution, the following approach is more efficient. It draws n samples in O(n) time (assuming an O(1) approximation is used to draw values from the binomial distribution[6]).

```
function draw_categorical(n) // where n is the number of samples to draw from the categorical
distribution
  r = 1
  s = 0
  for i from 1 to k // where k is the number of categories
    v = draw from a binomial(n, p[i] / r) distribution // where p[i] is the probability of
category i
    for j from 1 to v
      z[s++] = i // where z is an array in which the results are stored
    n = n - v
    r = r - p[i]
  shuffle (randomly re-order) the elements in z
  return z
```

## Sampling via the Gumbel distribution

In machine learning it is typical to parametrize the categorical distribution, $p_1, \ldots, p_k$ via an unconstrained representation in $\mathbb{R}^k$, whose components are given by:

$$\gamma_i = \log p_i + \alpha$$

where $\alpha$ is any real constant. Given this representation, $p_1, \ldots, p_k$ can be recovered using the softmax function, which can then be sampled using the techniques described above. There is however a more direct sampling method that uses samples from the Gumbel distribution.[7] Let $g_1, \ldots, g_k$ be *k* independent draws from the standard Gumbel distribution, then

$$c = \arg\max_i (\gamma_i + g_i)$$

will be a sample from the desired categorical distribution. (If $u_i$ is a sample from the standard uniform distribution, then $g_i = -\log(-\log u_i)$ is a sample from the standard Gumbel distribution.)

# See also

- Categorical variable

## Related distributions

- Dirichlet distribution
- Multinomial distribution
- Bernoulli distribution
- Dirichlet-multinomial distribution

# Notes

1. However, Bishop does not explicitly use the term categorical distribution.

# References


1. Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*, p. 35. MIT press. ISBN 0262018020.
2. Minka, T. (2003) Bayesian inference, entropy and the multinomial distribution (http://research.microsoft.com/en-us/um/people/minka/papers/multinomial.html). Technical report Microsoft Research.
3. Minka, T. (2003), op. cit. Minka uses the Kronecker delta function, similar to but less general than the Iverson bracket.
4. Bishop, C. (2006) *Pattern Recognition and Machine Learning*, Springer. ISBN 0-387-31073-8.
5. Johnson, N.L., Kotz, S., Balakrishnan, N. (1997) *Discrete Multivariate Distributions*, Wiley. ISBN 0-471-12844-9 (p. 105)
6. Agresti, A., An Introduction to Categorical Data Analysis, Wiley-Interscience, 2007, ISBN 978-0-471-22618-5, pp. 25
7. Adams, Ryan. "The Gumbel–Max Trick for Discrete Distributions" (http://lips.cs.princeton.edu/the-gumbel-max-trick-for-discrete-distributions/).