Discrete Categorical Distribution

Ayush Tewari

November 24th, 2025

Adapted from Carl Edward Rasmussen

Key concepts

We generalize the concepts from binary variables to multiple discrete outcomes.

- discrete and multinomial distributions

Key concepts

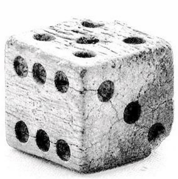
We generalize the concepts from binary variables to multiple discrete outcomes.

- discrete and multinomial distributions
- the Dirichlet distribution

The multinomial distribution (1)

- Generalisation of the binomial distribution from 2 outcomes to m outcomes.
- Useful for random variables that take one of a finite set of possible outcomes.

The multinomial distribution (1)



- Generalisation of the binomial distribution from 2 outcomes to m outcomes.
- Useful for random variables that take one of a finite set of possible outcomes.

Throw a die $n = 60$ times, and count the observed (6 possible) outcomes.

Outcome	Count
$X = x_1 = 1$	$k_1 = 12$
$X = x_2 = 2$	$k_2 = 7$
$X = x_3 = 3$	$k_3 = 11$
$X = x_4 = 4$	$k_4 = 8$
$X = x_5 = 5$	$k_5 = 9$
$X = x_6 = 6$	$k_6 = 13$

The multinomial distribution (1)

- Generalisation of the binomial distribution from 2 outcomes to m outcomes.
- Useful for random variables that take one of a finite set of possible outcomes.

Throw a die $n = 60$ times, and count the observed (6 possible) outcomes.

Outcome	Count
$X = x_1 = 1$	$k_1 = 12$
$X = x_2 = 2$	$k_2 = 7$
$X = x_3 = 3$	$k_3 = 11$
$X = x_4 = 4$	$k_4 = 8$
$X = x_5 = 5$	$k_5 = 9$
$X = x_6 = 6$	$k_6 = 13$

Note: We have one parameter too many.
We don't need to know all the k_i and n, because $\sum_{i=1}^{6} k_i = n$.

The multinomial distribution (2)

Consider a discrete random variable X that can take one of m values $x_1, \ldots, x_m$.

- Out of n independent trials, let k_i be the number of times $X = x_i$ was observed.
  It follows that $\sum_{i=1}^{m} k_i = n$.

The multinomial distribution (2)

Consider a discrete random variable X that can take one of m values $x_1, \ldots, x_m$.

- Out of n independent trials, let k_i be the number of times $X = x_i$ was observed.
  It follows that $\sum_{i=1}^{m} k_i = n$.
- Denote by π_i the probability that $X = x_i$, with $\sum_{i=1}^{m} \pi_i = 1$.

The multinomial distribution (2)

Consider a discrete random variable X that can take one of m values $x_1, \dots, x_m$.

- Out of n independent trials, let k_i be the number of times $X = x_i$ was observed.
  It follows that $\sum_{i=1}^{m} k_i = n$.
- Denote by π_i the probability that $X = x_i$, with $\sum_{i=1}^{m} \pi_i = 1$.

The probability of observing a vector of occurrences $\mathbf{k} = [k_1, \dots, k_m]^\top$ is given by the *multinomial distribution* parametrised by $\boldsymbol{\pi} = [\pi_1, \dots, \pi_m]^\top$:

$$p(\mathbf{k}|\boldsymbol{\pi}, n) \;=\; p(k_1, \dots, k_m|\pi_1, \dots, \pi_m, n) \;=\; \frac{n!}{k_1!k_2!\dots k_m!} \prod_{i=1}^{m} \pi_i^{k_i}$$

The multinomial distribution (2)

Consider a discrete random variable X that can take one of m values $x_1, \dots, x_m$.

- Out of n independent trials, let k_i be the number of times $X = x_i$ was observed.
  It follows that $\sum_{i=1}^{m} k_i = n$.
- Denote by π_i the probability that $X = x_i$, with $\sum_{i=1}^{m} \pi_i = 1$.

The probability of observing a vector of occurrences $\mathbf{k} = [k_1, \dots, k_m]^\top$ is given by the *multinomial distribution* parametrised by $\boldsymbol{\pi} = [\pi_1, \dots, \pi_m]^\top$:

$$p(\mathbf{k}|\boldsymbol{\pi}, n) = p(k_1, \dots, k_m|\pi_1, \dots, \pi_m, n) = \frac{n!}{k_1!k_2!\dots k_m!} \prod_{i=1}^{m} \pi_i^{k_i}$$

- Note that we can write $p(\mathbf{k}|\boldsymbol{\pi})$ since n is redundant.
- The multinomial coefficient $\frac{n!}{k_1!k_2!\dots k_m!}$ is a generalisation of $\binom{n}{k}$.

The multinomial distribution (2)

Consider a discrete random variable X that can take one of m values $x_1, \ldots, x_m$.

- Out of n independent trials, let k_i be the number of times $X = x_i$ was observed.
  It follows that $\sum_{i=1}^{m} k_i = n$.
- Denote by π_i the probability that $X = x_i$, with $\sum_{i=1}^{m} \pi_i = 1$.

The probability of observing a vector of occurrences $\mathbf{k} = [k_1, \ldots, k_m]^\top$ is given by the *multinomial distribution* parametrised by $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_m]^\top$:

$$p(\mathbf{k}|\boldsymbol{\pi}, n) = p(k_1, \ldots, k_m|\pi_1, \ldots, \pi_m, n) = \frac{n!}{k_1!k_2!\ldots k_m!} \prod_{i=1}^{m} \pi_i^{k_i}$$

- Note that we can write $p(\mathbf{k}|\boldsymbol{\pi})$ since n is redundant.
- The multinomial coefficient $\frac{n!}{k_1!k_2!\ldots k_m!}$ is a generalisation of $\binom{n}{k}$.

The discrete or *categorical distribution* is the generalisation of the Bernoulli to m outcomes, and the special case of the multinomial with one trial:

$$p(X = x_i|\boldsymbol{\pi}) = \pi_i.$$

Example: word counts in text

Consider describing a text document by the frequency of occurrence of every distinct word.

The UCI *Bag of Words* dataset from the University of California, Irvine. [1]

[1] http://archive.ics.uci.edu/ml/machine-learning-databases/bag-of-words/

Example: word counts in text

Consider describing a text document by the frequency of occurrence of every distinct word.

The UCI *Bag of Words* dataset from the University of California, Irvine. [1]

[1] http://archive.ics.uci.edu/ml/machine-learning-databases/bag-of-words/

Priors on multinomials: The Dirichlet distribution

The Dirichlet distribution is to the categorical/multinomial what the Beta is to the Bernoulli/binomial.

Priors on multinomials: The Dirichlet distribution

The Dirichlet distribution is to the categorical/multinomial what the Beta is to the Bernoulli/binomial.
It is a generalisation of the Beta defined on the $m - 1$ dimensional simplex.

Priors on multinomials: The Dirichlet distribution

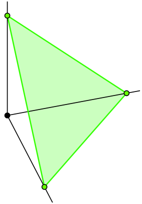
The Dirichlet distribution is to the categorical/multinomial what the Beta is to the Bernoulli/binomial.
It is a generalisation of the Beta defined on the $m - 1$ dimensional simplex.

- Consider the vector $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_m]^\top$, with $\sum_{i=1}^{m} \pi_i = 1$ and $\pi_i \in [0, 1] \;\; \forall i$.

Priors on multinomials: The Dirichlet distribution

The Dirichlet distribution is to the categorical/multinomial what the Beta is to the Bernoulli/binomial.
It is a generalisation of the Beta defined on the $m-1$ dimensional simplex.

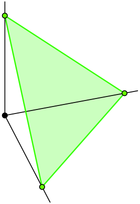
- Consider the vector $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_m]^\top$, with $\sum_{i=1}^{m} \pi_i = 1$ and $\pi_i \in [0, 1] \ \ \forall i$.
- Vector $\boldsymbol{\pi}$ lives in the open standard $m-1$ simplex.
- $\boldsymbol{\pi}$ could for example be the parameter vector of a multinomial. [Figure on the right $m = 3$.]



Priors on multinomials: The Dirichlet distribution

The Dirichlet distribution is to the categorical/multinomial what the Beta is to the Bernoulli/binomial.
It is a generalisation of the Beta defined on the $m-1$ dimensional simplex.

- Consider the vector $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_m]^\top$, with $\sum_{i=1}^m \pi_i = 1$ and $\pi_i \in [0, 1] \ \ \forall i$.
- Vector $\boldsymbol{\pi}$ lives in the open standard $m-1$ simplex.
- $\boldsymbol{\pi}$ could for example be the parameter vector of a multinomial. [Figure on the right $m = 3$.]



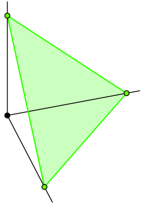
The Dirichlet distribution is given by

$$\mathrm{Dir}(\boldsymbol{\pi}|\alpha_1, \ldots, \alpha_m) \;=\; \frac{\Gamma(\sum_{i=1}^m \alpha_i)}{\prod_{i=1}^m \Gamma(\alpha_i)} \prod_{i=1}^m \pi_i^{\alpha_i - 1} \;=\; \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^m \pi_i^{\alpha_i - 1}$$

Priors on multinomials: The Dirichlet distribution

The Dirichlet distribution is to the categorical/multinomial what the Beta is to the Bernoulli/binomial.
It is a generalisation of the Beta defined on the $m-1$ dimensional simplex.

- Consider the vector $\boldsymbol{\pi} = [\pi_1, \dots, \pi_m]^\top$, with $\sum_{i=1}^m \pi_i = 1$ and $\pi_i \in [0,1] \ \ \forall i$.
- Vector $\boldsymbol{\pi}$ lives in the open standard $m-1$ simplex.
- $\boldsymbol{\pi}$ could for example be the parameter vector of a multinomial. [Figure on the right $m = 3$.]

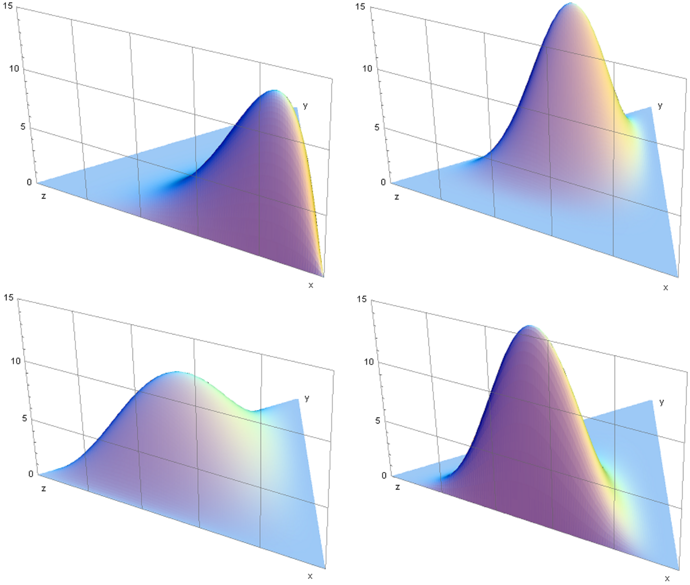


The Dirichlet distribution is given by

$$\mathrm{Dir}(\boldsymbol{\pi}|\alpha_1, \dots, \alpha_m) \ = \ \frac{\Gamma(\sum_{i=1}^m \alpha_i)}{\prod_{i=1}^m \Gamma(\alpha_i)} \prod_{i=1}^m \pi_i^{\alpha_i - 1} \ = \ \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^m \pi_i^{\alpha_i - 1}$$

- $\boldsymbol{\alpha} = [\alpha_1, \dots, \alpha_m]^\top$ are the shape parameters.
- $B(\boldsymbol{\alpha})$ is the multivariate beta function.
- $E(\pi_j) = \frac{\alpha_j}{\sum_{i=1}^m \alpha_i}$ is the mean for the j-th element.

Dirichlet Distributions from Wikipedia



The symmetric Dirichlet distribution

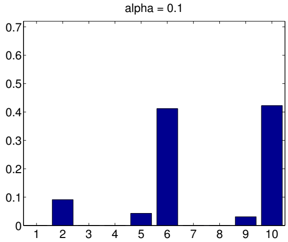
In the symmetric Dirichlet distribution all parameters are identical: $\alpha_i = \alpha, \ \ \forall i$.

`en.wikipedia.org/wiki/File:LogDirichletDensity-alpha_0.3_to_alpha_2.0.gif`

To sample from a symmetric Dirichlet in D dimensions with concentration α use: `w = randg(alpha,D,1); bar(w/sum(w));`

The symmetric Dirichlet distribution

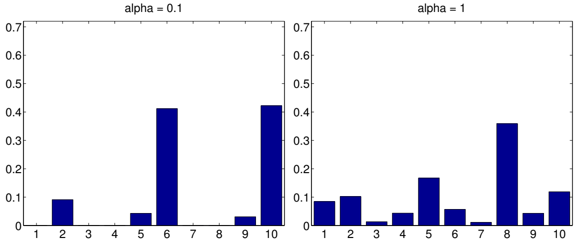
In the symmetric Dirichlet distribution all parameters are identical: $\alpha_i = \alpha, \;\; \forall i$.
`en.wikipedia.org/wiki/File:LogDirichletDensity-alpha_0.3_to_alpha_2.0.gif`
To sample from a symmetric Dirichlet in D dimensions with concentration α use: `w = randg(alpha,D,1); bar(w/sum(w));`



- Left: $\alpha = 0.1$ (Sparse / Spiky)

The symmetric Dirichlet distribution

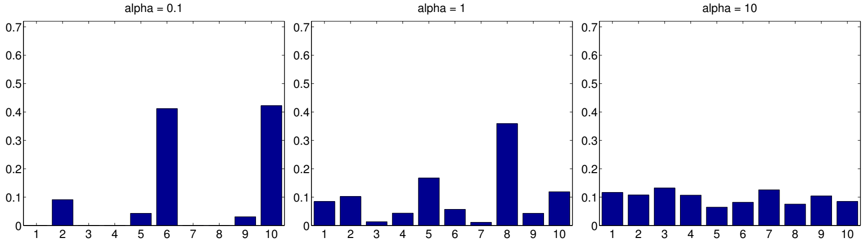
In the symmetric Dirichlet distribution all parameters are identical: $\alpha_i = \alpha, \ \ \forall i$.
`en.wikipedia.org/wiki/File:LogDirichletDensity-alpha_0.3_to_alpha_2.0.gif`
To sample from a symmetric Dirichlet in D dimensions with concentration α use: `w = randg(alpha,D,1); bar(w/sum(w));`



- Left: $\alpha = 0.1$ (Sparse / Spiky)
- Middle: $\alpha = 1.0$ (Uniform)

The symmetric Dirichlet distribution

In the symmetric Dirichlet distribution all parameters are identical: $\alpha_i = \alpha, \ \ \forall i$.
`en.wikipedia.org/wiki/File:LogDirichletDensity-alpha_0.3_to_alpha_2.0.gif`
To sample from a symmetric Dirichlet in D dimensions with concentration α use: `w = randg(alpha,D,1); bar(w/sum(w));`



- Left: $\alpha = 0.1$ (Sparse / Spiky)
- Middle: $\alpha = 1.0$ (Uniform)
- Right: $\alpha = 10$ (Peaked at center)