The **multinomial distribution** generalizes the binomial distribution to $C > 2$ categories; its main properties are summarized in Figure 3.17. The Binomial distribution in Figure 1.4 is the special case with $C = 2$ categories, which is seen by defining $\theta = \theta_1$, $\theta_2 = 1 - \theta$, $x = x_1$, $x_2 = n - x$, and noting that

multinomial distribution

$$\frac{n!}{x_1!x_2!} = \frac{n!}{x!(n-x)!} = \binom{n}{x}. \tag{3.8}$$

Multinomial distribution

$(X_1, \ldots, X_C) \sim \text{MultiNom}(n, \boldsymbol{\theta})$
where $\sum_{c=1}^{C} X_c = n$,
$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_C)$ and $\sum_c \theta_c = 1$.

$$p(\mathbf{x}) = \frac{n!}{x_1! \cdots x_C!} \theta_1^{x_1} \cdots \theta_C^{x_C}$$

$\mathbb{E}(X_c) = n\theta_c$

$\mathbb{V}(X_c) = n\theta_c(1 - \theta_c)$

Figure 3.17: The multinomial distribution.

The multinomial distribution is a multivariate distribution with convenient marginalization properties. For example, if we group the counts in one or more categories - for example turning the mobile phone dataset into three categories by merging 'Windows' and 'Other' - the distribution remains multinomial. The probability of a merged category is simply the sum of the probabilities of the merged categories. Hence

$$(x_1, x_2, x_3 + x_4) \sim \text{Multinomial}(\theta_1, \theta_2, \theta_3 + \theta_4).$$

In particular, merging to only two categories - for example 'iPhone' and 'not iPhone' - gives a binomial distribution where the probability of failure (not iPhone) is $\theta_2 + \theta_3 + \theta_4$.

Dirichlet distribution

$\boldsymbol{\theta}|\boldsymbol{\alpha} \sim \text{Dirichlet}(\boldsymbol{\alpha})$ where
$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_C)$, $\sum_c \theta_c = 1$,
$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_C)$ and $\alpha_c > 0$..

$$p(\boldsymbol{\theta}) = k \cdot \theta_1^{\alpha_1 - 1} \cdots \theta_C^{\alpha_C - 1}$$
$$k = \frac{\Gamma(\sum_{c=1}^{C} \alpha_c)}{\prod_{c=1}^{C} \Gamma(\alpha_c - 1)}.$$
$$\mathbb{E}(\theta_c) = \frac{\alpha_c}{\sum_{j=1}^{C} \alpha_j}$$
$$\mathbb{V}(\theta_c) = \frac{\mathbb{E}(\theta_c)(1 - \mathbb{E}(\theta_c))}{1 + \alpha_+}$$
$$\alpha_+ = \sum_{c=1}^{C} \alpha_c.$$

Marginal distributions:

$$\theta_c \sim \text{Beta}\left(\alpha_c, \alpha_+ - \alpha_c\right).$$

Figure 3.18: The Dirichlet distribution.

A Bayesian analysis of multinomial data requires a prior distribution for the model parameters, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_C)$. Since each θ_c is a probability, the first distribution that comes to mind may be a Beta distribution; the Beta distribution is not appropriate here however since it does not enforce the constraint that the probabilities sum to one. Hence, the parameter space of the multinomial distribution is the **unit simplex**, i.e. the set $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_C) : 0 < \theta_c < 1$ and $\sum_c \theta_c = 1$. Luckily, there is a very nice distribution on the unit simplex, the Dirichlet distribution, summarized in Figure 3.18.

unit simplex

The Dirichlet distribution is specified with the prior hyperparameters $\alpha_c > 0$, see Figure 3.19 for some examples. The *relative* sizes of the elements in $\boldsymbol{\alpha}$ determine the prior means for elements of $\boldsymbol{\theta}$. For example, setting $\alpha_1 = \cdots = \alpha_C = 1.5$, as in the upper left graph of Figure 3.19, gives equal prior mean for all categories: $\mathbb{E}(\theta_c) = 1/C$ for all c. The *absolute* size of $\boldsymbol{\alpha}$, measured by $\alpha_+ = \sum_{c=1}^{C} \alpha_c$, is inversely related to the variance, see Figure 3.18; hence, the prior hyperparameters $\boldsymbol{\alpha} = (1.5, \ldots, 1.5)$ and $\boldsymbol{\alpha} = (5, \ldots, 5)$ in the upper part of Figure 3.19 have the same mean, but the latter has smaller variance. Finally, the bottom part of Figure 3.19 shows examples where the prior mean is different over the categories.

The Dirichlet$(1, \ldots, 1)$ has constant density and is therefore the **uniform distribution on the unit simplex**; this generalizes the result that Beta$(1, 1)$ is uniform on the unit interval $[0, 1]$. Finally, when $\alpha_c < 1$, the Dirichlet density becomes 'bathtub shaped with probability mass piling up against the edges of the unit simplex.

uniform distribution on the unit simplex

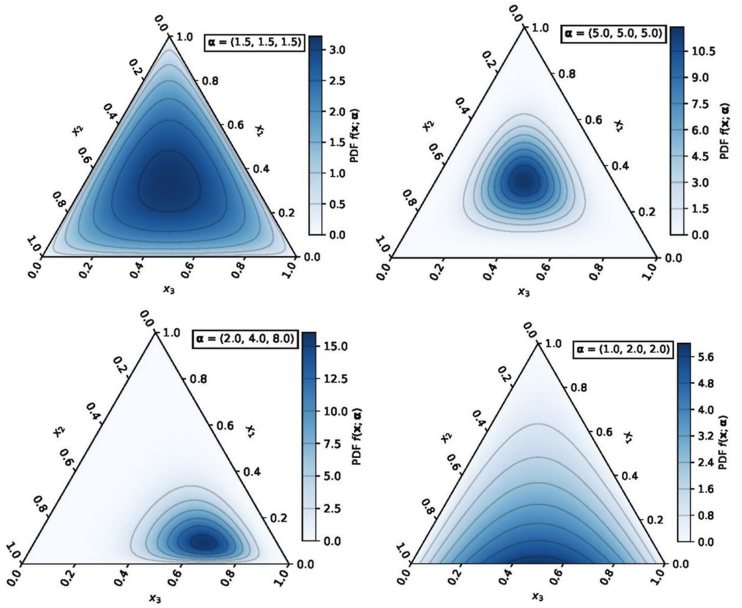

Figure 3.19: Examples of Dirichlet distributions for $\mathbf{x} = (x_1, x_2, x_3)$. Source: Wikipedia.

The Dirichlet distribution is conjugate to the multinomial likelihood which is easily seen by computing the posterior

$$
\begin{aligned}
p(\boldsymbol{\theta}|\mathbf{y}) &\propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) && (3.9)\\
&= \frac{n!}{x_1! \cdots x_C!}\theta_1^{x_1}\cdots\theta_C^{x_C} \cdot \frac{\Gamma(\sum_{c=1}^{C}\alpha_c)}{\prod_{c=1}^{C}\Gamma(\alpha_c - 1)}\theta_1^{\alpha_1-1}\cdots\theta_C^{\alpha_C-1} && (3.10)\\
&= \theta_1^{\alpha_1+x_1-1}\cdots\theta_C^{\alpha_C+x_C-1}, && (3.11)
\end{aligned}
$$

which is proportional to the Dirichlet$(\alpha_1 + x_1, \ldots, \alpha_C + x_C)$ density. This is a convenient result: the posterior is simply obtained by adding the data count x_c to the prior hyperparameter α_c in each category. This parallels and generalizes the binary case where a Beta(α, β) prior was updated to a posterior by adding the number of successes s to α and the number of failures f to β. Figure 3.20 summarizes the prior-to-posterior updating for multinomial data with a Dirichlet prior.

Multinomial data with Dirichlet prior

Model: $\mathbf{n}|\boldsymbol{\theta} \sim \text{Multinomial}(\boldsymbol{\theta})$, where
$\mathbf{n} = (n_1, \ldots, n_C)$ are counts in C categories
$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_C)$ are category probabilities.

Prior: $\boldsymbol{\theta} \sim \text{Dirichlet}(\boldsymbol{\alpha})$, for $\alpha = (\alpha_1, \ldots, \alpha_C)$

Posterior: $\boldsymbol{\theta} \sim \text{Dirichlet}(\boldsymbol{\alpha} + \mathbf{n})$

Figure 3.20: Prior-to-Posterior updating for multinomial data with the Dirichlet prior.

MOBILE PHONE SURVEY DATA We are now ready to analyze the four market shares $\theta_1, \ldots, \theta_4$ in the mobile phone data. We will determine the prior hyperparameters in the Dirichlet prior using data from a similar survey from four year ago. The proportions in the four categories back then were: 30%, 30%, 20% and 20%. This was a large survey, but since time has passed and user patterns most likely have changed, I value the information in this older survey as being equivalent to a survey with only 50 participants. This gives us the prior:

$$(\theta_1, \ldots, \theta_4) \sim \text{Dirichlet}(\alpha_1 = 15, \alpha_2 = 15, \alpha_3 = 10, \alpha_4 = 10)$$

Note that $\mathbb{E}(\theta_1) = 15/50 = 0.3$ and so on, so the prior mean is set equal to the proportions from the older survey. Also, $\sum_{k=1}^{4} \alpha_k = 50$, so the prior information is equivalent to a survey based on 50 respondents, as required.

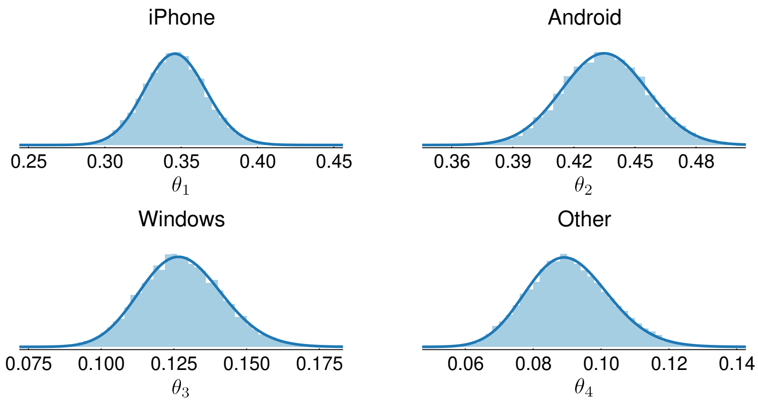

Figure 3.21: Marginal posteriors of the market shares for the mobile phone survey data. Simulated (histogram) draws and analytical density functions (solid curves).

The joint posterior distribution of all four shares is by Figure 3.20 equal to

$$(\theta_1, \ldots, \theta_4)|\mathbf{y} \sim \text{Dirichlet}(15 + 180, 15 + 230, 10 + 62, 10 + 41)$$

The marginal posteriors are plotted in Figure 3.21 as histograms from Monte Carlo simulation (see the algorithm in Figure 3.22); the analytical posteriors from Figure 3.18 are overlayed.

draw	θ_1	θ_2	θ_3	θ_4	θ_2 largest
1	0.338	0.446	0.130	0.086	1
2	0.332	0.457	0.124	0.086	1
3	0.325	0.442	0.136	0.094	1
⋮	⋮	⋮	⋮	⋮	⋮
10,000	0.343	0.443	0.132	0.081	1
Mean	0.346	0.435	0.127	0.090	0.991

Table 3.2: Posterior simulation output for the multinomial model applied to the mobile phone survey data. The last column is a computed binary indicator for the event that Android has the largest market share, i.e. if $\theta_2 > \max(\theta_1, \theta_3, \theta_4)$.

Figure 3.21 indicates that Android may have the largest market share with a posterior mean around 0.44 versus iPhones posterior mean of 0.35. Computing the probability that Android has

Posterior simulation - Multinomial data, Dirichlet prior.

Input: data $\mathbf{n} = (n_1, \ldots, n_C)$
prior hyperparameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_C)$
the number of posterior draws m.

for i in *1:m* **do**
  $\boldsymbol{\theta} \leftarrow$ rDirichlet($\boldsymbol{\alpha} + \mathbf{n}$)
end
Output: m posterior draws of $\boldsymbol{\theta} = (\theta_1, ..., \theta_C)$.

Function rDirichlet(*$\boldsymbol{\alpha}$*)
  for c in *1:C* **do**
    $\mathbf{y}[c] \leftarrow$ rGamma($\boldsymbol{\alpha}[c], 1$)
  end
  return $\mathbf{y}/$Sum$(\mathbf{y})$

Figure 3.22: Algorithm for posterior simulation for the multinomial model with the conjugate Dirichlet prior. The rGamma random number generator is assumed to be part of the standard library.

the largest market share involves integrating the joint posterior $\boldsymbol{\theta}|\mathbf{y} \sim \text{Dirichlet}(\boldsymbol{\alpha} + \mathbf{y})$ over the region $\{\boldsymbol{\theta} : \theta_2 > \max(\theta_1, \theta_3, \theta_4)\}$, a tedious calculation. The probability is however easily computed by simulation by recording for each posterior $\boldsymbol{\theta}$ draw if the condition $\theta_2 > \max(\theta_1, \theta_3, \theta_4)$ is satisfied; see Table 3.2, which shows that

$$\Pr(\text{Andriod has largest market share}|\mathbf{y}) \approx 0.991.$$

We are almost certain that Android is the most popular mobile phone in the population targeted by the survey.