directly obtained from a posterior sample of $\boldsymbol{\theta}$ by simply computing the function $f(\boldsymbol{\theta})$ for each posterior draw. Provided the posterior variance of $f(\boldsymbol{\theta})$ exists, a central limit theorem of the form (3.6) exists also in this case, with the expected value and variance replaced by those of $f(\boldsymbol{\theta})$.

To illustrate how simulation immediately provides inference for any function of the parameters, Table 3.1 contains a fourth column named σ/θ with the computed coefficient of variation for each draw. We can now just plot a histogram of this new column to approximate the marginal posterior of the function $f(\theta, \sigma^2) = \sigma/\theta$. The results are presented in the right part of Figure 3.15; the left part of the figure shows the results for the standard deviation $f(\theta, \sigma^2) = \sqrt{\sigma^2}$.

Central limit theorem (CLT)

Let $X_1, X_2, \ldots$ be iid random variables with finite mean μ and variance σ^2. Then

$$\sqrt{n}\,(\bar{X}_n - \mu) \xrightarrow{d} \mathrm{N}(0, 1),$$

as $n \to \infty$ where $\xrightarrow{d}$ denotes convergence in distribution.

The CLT is often informally written as

$$\bar{X}_n \xrightarrow{d} \mathrm{N}(\mu, \sigma^2) \text{ as } n \to \infty.$$

Figure 3.14: The central limit theorem.

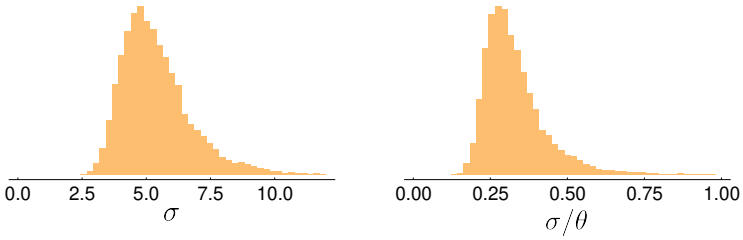

Figure 3.15: Histogram of simulated marginal posteriors for σ (left) and the coefficient of variation σ/θ (right) for the internet speed data.

The final column of Table 3.1 is a binary variable that records if θ was at least 20, i.e. it computes the indicator function $f(\theta, \sigma^2) = I(\theta \geq 20)$. The marginal posterior probability $\Pr(\theta \geq 20|\mathbf{x})$ is then easily approximated by the mean of the final column; the right side of Figure 3.12 illustrates the Monte Carlo convergence of this estimate.

Multinomial data

Categorical data

Categorical data have observations that belong to one of C discrete classes. A computer bug can for example be allocated to C developing teams; an items sold in an auction may reported as: 'defective', 'normal quality', or 'new'; a continuous variable like age can recorded in age intervals: 0-18, 19–28, 29–49, 50–64 and 65+, which would then also be a categorical variable. The categories in the latter two situations are examples of **ordinal data** where the categogies have a natural order. There are special models for ordinal data which we will not cover in this chapter; here we will consider categorical data without natural order. Categorical variables are often called **multi-class** in the machine learning literature.

ordinal data

multi-class

A multi-class random variable X is often written in **one-hot encoding** as $\mathbf{x} = (x_1, \ldots x_C)$ where $X = c$ is encoded as $x_c = 1$ and

one-hot encoding

$x_j = 0$ for $j \neq c$; hence when $C = 3$, $\mathbf{x} = (0, 1, 0)$ means that the observation belongs to the second class. The categorical random variable $X|\boldsymbol{\theta} \sim \text{Cat}(\theta_1, \ldots, \theta_C)$ has probability distribution

$$p(x) = \theta_1^{x_1} \cdots \theta_C^{x_C}, \tag{3.7}$$

where $(x_1, \ldots x_C)$ is the one-hot encoding of x, $0 < \theta_c < 1$ is the probability of class c and $\sum_{c=1}^{C} \theta_c = 1$. Note how Bernoulli data is the special case with $C = 2$ categories 'success' and 'failure', so that the $\text{Cat}(\theta_1, \ldots, \theta_C)$ distribution generalizes the Bernoulli distribution to the case $C > 2$. Figure 3.16 is an example of $\text{Cat}(\theta_1, \ldots, \theta_C)$ for $C = 4$.

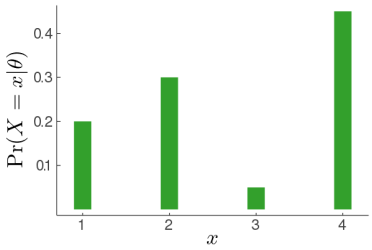

Figure 3.16: Categorical distribution with probabilities $\boldsymbol{\theta} = (0.20, 0.30, 0.05, 0.45)$.

We saw in Section The likelihood function and maximum likelihood estimation that counting the number of successes s in n binary Bernoulli trials gave rise to $S \sim \text{Binomial}(n, \theta)$ data. In the same way we can count the number of observations in category c for $c = 1, \ldots, C$ in multi-class data. This gives data as a count vector $\mathbf{y} = (y_1, \ldots, y_C)$ where y_c is the number of observations in category c in $n = \sum_{c=1}^{C} y_c$ 'trials'. Here is an example:

MOBILE PHONE SURVEY DATA. A survey was conducted among $n = 513$ mobile phone users. Among other questions, the participants were asked: 'What kind of mobile phone do you mainly use?' with the four options:

1. iPhone
2. Android
3. Windows
4. Other/Don't know

The number of responses in the four categories were: $\mathbf{y} = (180, 230, 62, 41)$.

The **multinomial distribution** generalizes the binomial distribution to $C > 2$ categories; its main properties are summarized in Figure 3.17. The Binomial distribution with x successes in n trials with probability θ in Figure 1.4 is the special case with $C = 2$ categories, which is seen by defining $\theta_1 = \theta$, $\theta_2 = 1 - \theta$, $y_1 = x$, $y_2 = n - x$, and noting that

$$\frac{n!}{y_1! y_2!} = \frac{n!}{x!(n-x)!} = \binom{n}{x}. \tag{3.8}$$

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

$$(y_1, y_2, y_3 + y_4) \sim \text{Multinomial}(\theta_1, \theta_2, \theta_3 + \theta_4).$$

multinomial distribution

Multinomial distribution

$(Y_1, \ldots, Y_C) \sim \text{MultiNom}(n, \boldsymbol{\theta})$
where $\sum_{c=1}^{C} Y_c = n$,
$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_C)$ and $\sum_c \theta_c = 1$.

$$p(\mathbf{y}) = \frac{n!}{y_1! \cdots y_C!} \theta_1^{y_1} \cdots \theta_C^{y_C}$$

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

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

Figure 3.17: The multinomial distribution.

In particular, merging to only two categories - for example 'iPhone' and 'not iPhone' - gives a binomial distribution where the probability of success (iPhone) is θ_1 and the probability of failure (not iPhone) is $\theta_2 + \theta_3 + \theta_4$.

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.

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.

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.

unit simplex

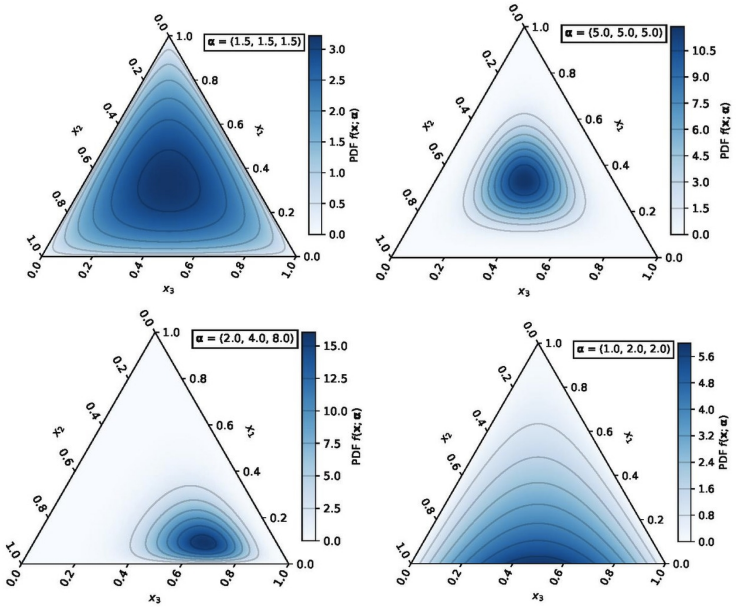

Figure 3.19: Examples of Dirichlet distributions for $\mathbf{x} = (x_1, x_2, x_3)$. Source: Wikipedia.

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

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

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{3.9}$$

$$= \frac{n!}{y_1! \cdots y_C!}\theta_1^{y_1} \cdots \theta_C^{y_C} \cdot \frac{\Gamma(\sum_{c=1}^C \alpha_c)}{\prod_{c=1}^C \Gamma(\alpha_c - 1)}\theta_1^{\alpha_1 - 1} \cdots \theta_C^{\alpha_C - 1} \tag{3.10}$$

$$= \theta_1^{\alpha_1 + y_1 - 1} \cdots \theta_C^{\alpha_C + y_C - 1}, \tag{3.11}$$

which is proportional to the Dirichlet$(\alpha_1 + y_1, \ldots, \alpha_C + y_C)$ density. This is a convenient result: the posterior is simply obtained by adding the data count y_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:	$\mathbf{y}\|\boldsymbol{\theta} \sim$ Multinomial$(\boldsymbol{\theta})$, where
	$\mathbf{y} = (y_1, \ldots, y_C)$ are counts in C categories
	$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_C)$ are category probabilities.
Prior:	$\boldsymbol{\theta} \sim$ Dirichlet$(\boldsymbol{\alpha})$, for $\alpha = (\alpha_1, \ldots, \alpha_C)$
Posterior:	$\boldsymbol{\theta}\|\mathbf{y} \sim$ Dirichlet$(\boldsymbol{\alpha} + \mathbf{y})$

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.

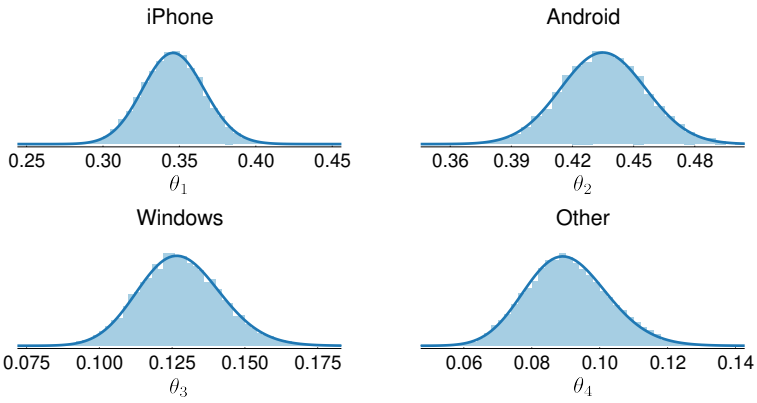

Figure 3.21: Marginal posteriors of the market shares for the mobile phone survey data. Simulated (histogram) draws and analytical density functions (solid curves).

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 $\theta_2 > \max(\theta_1, \theta_3, \theta_4)$.

Posterior simulation - Multinomial data, Dirichlet prior.

Input: data $\mathbf{y} = (y_1, \ldots, y_C)$
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$(\boldsymbol{\alpha} + \mathbf{y})$
end
Output: m posterior draws of $\boldsymbol{\theta} = (\theta_1, ..., \theta_C)$.

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

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 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.

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 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.

Multivariate normal data with known covariance

This section considers the iid **multivariate normal** model for a p-dimensional data vector $\mathbf{x}$:

$$\mathbf{x}_1, \ldots \mathbf{x}_n|\boldsymbol{\theta}, \boldsymbol{\Sigma} \overset{\text{iid}}{\sim} N(\boldsymbol{\theta}, \boldsymbol{\Sigma}), \tag{3.12}$$

where $\boldsymbol{\theta}$ is the p-dimensional mean vector and $\boldsymbol{\Sigma}$ is a $p \times p$ positive definite covariance matrix. We will here take $\boldsymbol{\Sigma}$ to be known and derive the posterior for $\boldsymbol{\theta}$.

Presenting a Bayesian analysis of this model here gives us a chance to meet the important multivariate normal distribution and its properties relatively early in the book; see Figure 3.23 for the density and properties, and Figure 3.24 for contour plots of some example densities.

The likelihood for the multivariate model in (3.12) is the product of the individual densities for each vector observation $\mathbf{x}_i$

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_n|\boldsymbol{\theta}, \boldsymbol{\Sigma}) \propto |\Sigma|^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\theta})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\theta})\right),$$

A vector version of the argument leading up to (2.7) in the univariate case can be used to show that the likelihood can be written as the exponential of a quadratic (form):

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_n|\boldsymbol{\theta}, \boldsymbol{\Sigma}) \propto \exp\left(-\frac{n}{2}(\boldsymbol{\theta} - \bar{\boldsymbol{x}})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \bar{\boldsymbol{x}})\right), \tag{3.13}$$

multivariate normal

Multivariate normal

$\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma} \sim \text{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\mathbf{x} \in \mathbb{R}^p$, $\boldsymbol{\mu} \in \mathbb{R}^p$ and $\boldsymbol{\Sigma}$ is a $p \times p$ positive definite covariance matrix.

$$p(\mathbf{x}) = |2\pi\boldsymbol{\Sigma}|^{-\frac{1}{2}} \times \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

$$\mathbb{E}(\mathbf{x}) = \boldsymbol{\mu}$$
$$\mathbb{V}(\mathbf{x}) = \boldsymbol{\Sigma}$$

Define the decomposition

$$\mathbf{x} = \begin{pmatrix}\mathbf{x}_1 \\ \mathbf{x}_2\end{pmatrix}$$

and similarly for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$

$$\boldsymbol{\mu} = \begin{pmatrix}\boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2\end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{pmatrix}$$

Marginal distributions:

$$x_k \sim \text{N}(\mu_k, \sigma_k^2)$$
$$\mathbf{x}_1 \sim \text{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$$

Conditional distributions:

$$\mathbf{x}_1|\mathbf{x}_2 \sim \text{N}(\tilde{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}}_1)$$

where

$$\tilde{\boldsymbol{\mu}}_1 = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$$
$$\tilde{\boldsymbol{\Sigma}}_1 = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

Figure 3.23: The multivariate normal distribution.