STA 360/602L: Module 8.1

The multinomial model

Dr. Olanrewaju Michael Akande

Categorical data (univariate)

- Suppose
  - $Y \in \{1, \ldots, D\}$;
  - $\Pr(y = d) = \theta_d$ for each $d = 1, \ldots, D$; and
  - $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_D)$.
- Then the pmf of Y is

$$\Pr[y = d|\boldsymbol{\theta}] = \prod_{d=1}^{D} \theta_d^{1[y=d]}.$$

- We say Y has a multinomial distribution with sample size 1, or a categorical distribution.
- Write as $Y|\boldsymbol{\theta} \sim \text{Multinomial}(1, \boldsymbol{\theta})$ or $Y|\boldsymbol{\theta} \sim \text{Categorical}(\boldsymbol{\theta})$.
- Clearly, this is just an extension of the Bernoulli distribution.

Dirichlet distribution

- Since the elements of the probability vector $\boldsymbol{\theta}$ must always sum to one, the support is often called a simplex.
- A conjugate prior for categorical/multinomial data is the Dirichlet distribution.
- A random variable $\boldsymbol{\theta}$ has a Dirichlet distribution with parameter $\boldsymbol{\alpha}$, if

$$p[\boldsymbol{\theta}|\boldsymbol{\alpha}] = \frac{\Gamma\left(\sum_{d=1}^{D} \alpha_d\right)}{\prod_{d=1}^{D} \Gamma(\alpha_d)} \prod_{d=1}^{D} \theta_d^{\alpha_d - 1}, \quad \alpha_d > 0 \text{ for all } d = 1, \ldots, D.$$

  where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_D)$, and

$$\sum_{d=1}^{D} \theta_d = 1, \;\; \theta_d \geq 0 \text{ for all } d = 1, \ldots, D.$$

- We write this as $\boldsymbol{\theta} \sim \text{Dirichlet}(\boldsymbol{\alpha}) = \text{Dirichlet}(\alpha_1, \ldots, \alpha_D)$.
- The Dirichlet distribution is a multivariate generalization of the beta distribution.

Dirichlet distribution

- Write

$$\alpha_0 = \sum_{d=1}^{D} \alpha_d \quad \text{and} \quad \alpha_d^{\star} = \frac{\alpha_d}{\alpha_0}.$$

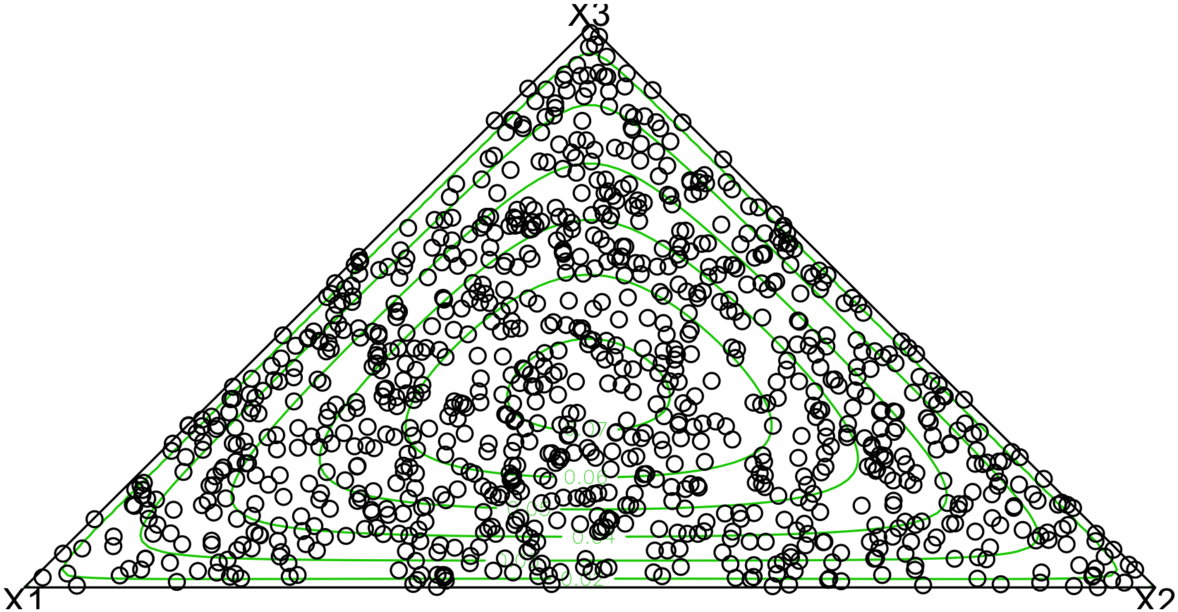
- Then we can re-write the pdf slightly as

$$p[\boldsymbol{\theta}|\boldsymbol{\alpha}] = \frac{\Gamma(\alpha_0)}{\prod_{d=1}^{D} \Gamma(\alpha_d)} \prod_{d=1}^{D} \theta_d^{\alpha_d - 1}, \quad \alpha_d > 0 \text{ for all } d = 1, \ldots, D.$$

- Properties:
  - $$\mathbb{E}[\theta_d] = \alpha_d^{\star};$$
  - $$\text{Mode}[\theta_d] = \frac{\alpha_d - 1}{\alpha_0 - d};$$
  - $$\mathbb{V}\text{ar}[\theta_d] = \frac{\alpha_d^{\star}(1 - \alpha_d^{\star})}{\alpha_0 + 1} = \frac{\mathbb{E}[\theta_d](1 - \mathbb{E}[\theta_d])}{\alpha_0 + 1};$$
  - $$\mathbb{C}\text{ov}[\theta_d, \theta_k] = \frac{\alpha_d^{\star}\alpha_k^{\star}}{\alpha_0 + 1} = \frac{\mathbb{E}[\theta_d]\mathbb{E}[\theta_k]}{\alpha_0 + 1}.$$

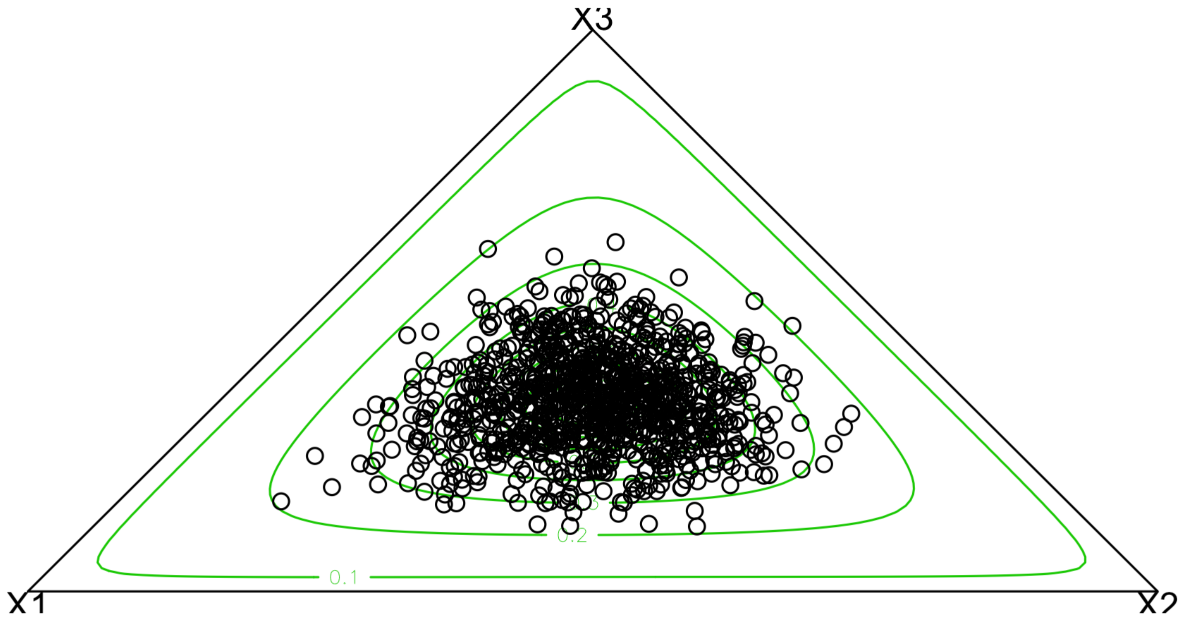
Dirichlet examples

$\text{Dirichlet}(1, 1, 1)$



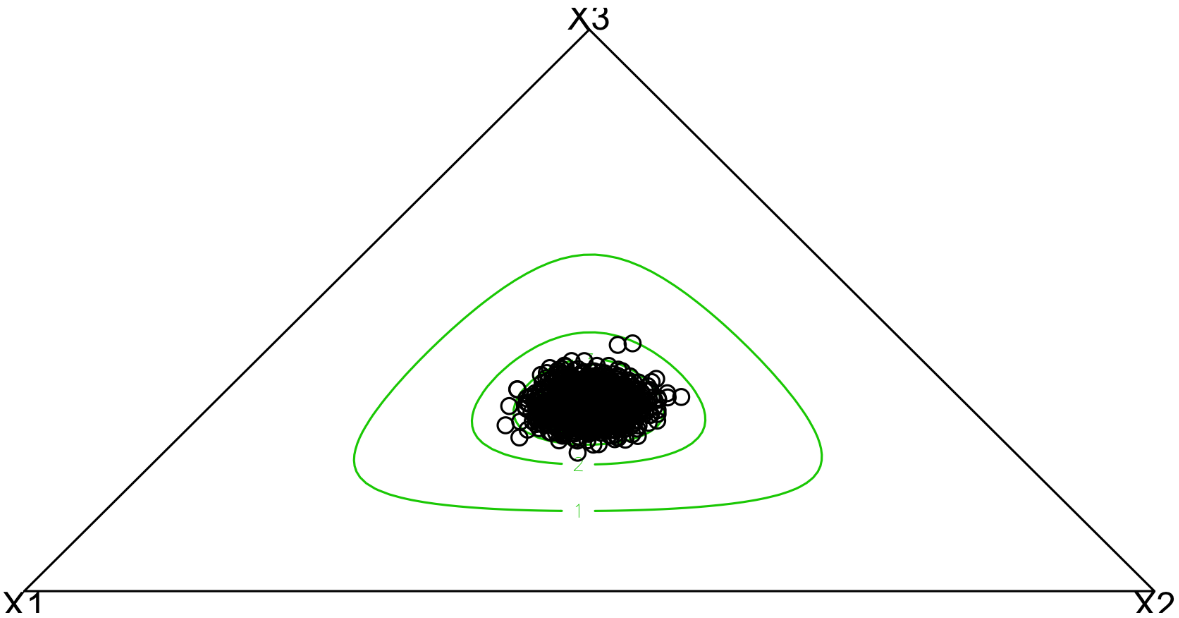
Dirichlet examples

$\text{Dirichlet}(10, 10, 10)$



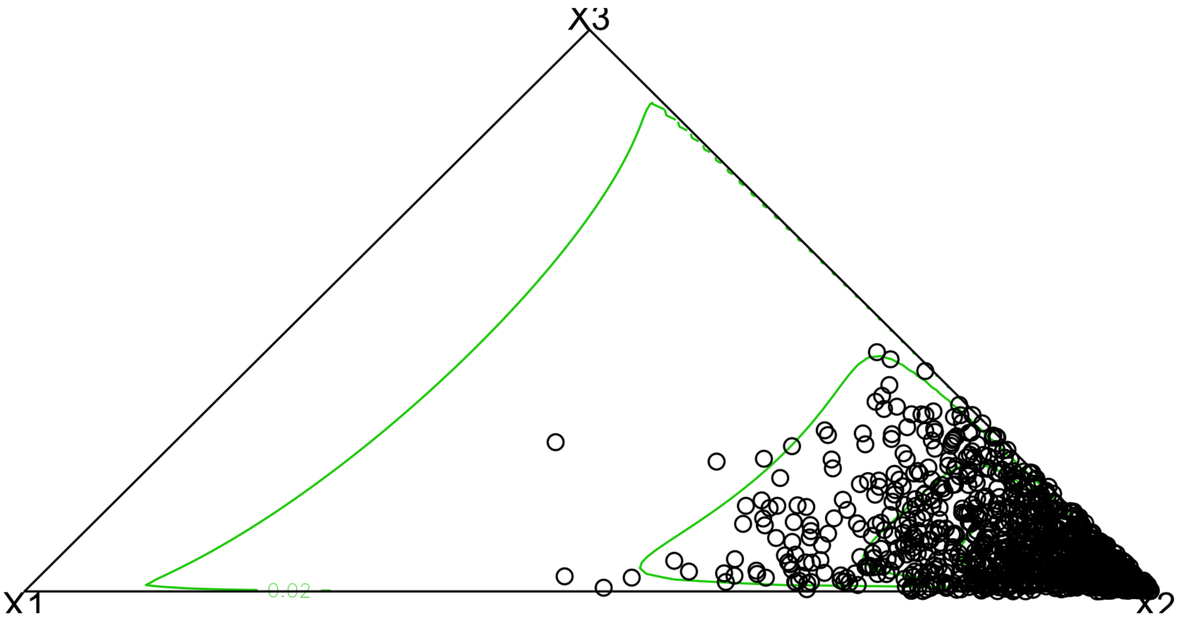
Dirichlet examples

$\text{Dirichlet}(100, 100, 100)$



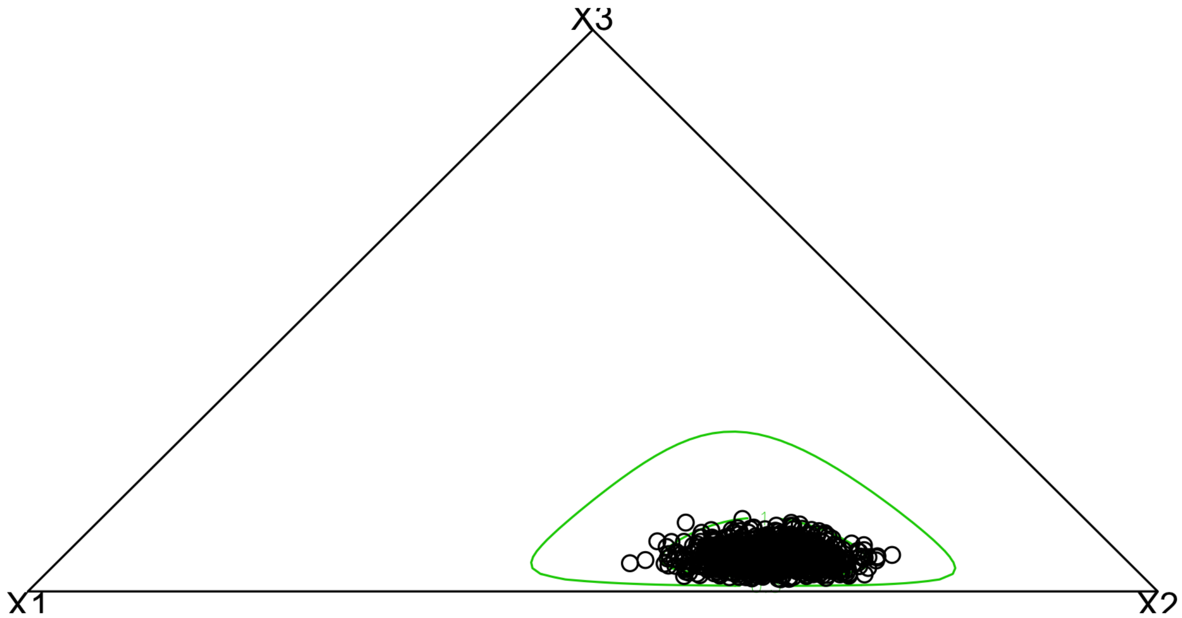
Dirichlet examples

$\text{Dirichlet}(1, 10, 1)$



Dirichlet examples

$\text{Dirichlet}(50, 100, 10)$



Likelihood

- Let $Y_i, \ldots, Y_n | \boldsymbol{\theta} \sim \text{Categorical}(\boldsymbol{\theta})$.
- Recall

$$\Pr[y_i = d | \boldsymbol{\theta}] = \prod_{d=1}^{D} \theta_d^{1[y_i = d]}.$$

- Then,

$$p[Y|\boldsymbol{\theta}] = p[y_1, \ldots, y_n|\boldsymbol{\theta}] = \prod_{i=1}^{n} \prod_{d=1}^{D} \theta_d^{1[y_i=d]} = \prod_{d=1}^{D} \theta_d^{\sum_{i=1}^{n} 1[y_i=d]} = \prod_{d=1}^{D} \theta_d^{n_d}$$

where n_d is just the number of individuals in category d.

- Maximum likelihood estimate of θ_d is

$$\hat{\theta}_d = \frac{n_d}{n}, \;\; d = 1, \ldots, D$$

Posterior

- Set $\pi(\boldsymbol{\theta}) = \text{Dirichlet}(\alpha_1, \ldots, \alpha_D)$.

$$\begin{aligned}
\pi(\boldsymbol{\theta}|Y) &\propto p[Y|\boldsymbol{\theta}] \cdot \pi[\boldsymbol{\theta}] \\
&\propto \prod_{d=1}^{D} \theta_d^{n_d} \prod_{d=1}^{D} \theta_d^{\alpha_d - 1} \\
&\propto \prod_{d=1}^{D} \theta_d^{\alpha_d + n_d - 1} \\
&= \text{Dirichlet}(\alpha_1 + n_1, \ldots, \alpha_D + n_D)
\end{aligned}$$

- Posterior expectation:

$$\mathbb{E}[\theta_d|Y] = \frac{\alpha_d + n_d}{\sum_{d^\star=1}^{D}(\alpha_{d^\star} + n_{d^\star})}.$$

Combining information

- For the prior, we have

$$\mathbb{E}[\theta_d] = \frac{\alpha_d}{\sum_{d^\star=1}^{D} \alpha_{d^\star}}$$

- We can think of
  - $\theta_{0d} = \mathbb{E}[\theta_d]$ as being our **"prior guess"** about θ_d, and
  - $n_0 = \sum_{d^\star=1}^{D} \alpha_{d^\star}$ as being our **"prior sample size"**.
- We can then rewrite the prior as $\pi(\boldsymbol{\theta}) = \text{Dirichlet}(n_0\theta_{01}, \ldots, n_0\theta_{0D})$.

Combining information

- We can write the posterior expectation as:

$$\begin{aligned}\mathbb{E}[\theta_d|Y] &= \frac{\alpha_d + n_d}{\sum_{d^\star=1}^{D}(\alpha_{d^\star} + n_{d^\star})} \\ &= \frac{\alpha_d}{\sum_{d^\star=1}^{D}\alpha_{d^\star} + \sum_{d^\star=1}^{D} n_{d^\star}} + \frac{n_d}{\sum_{d^\star=1}^{D}\alpha_{d^\star} + \sum_{d^\star=1}^{D} n_{d^\star}} \\ &= \frac{n_0\theta_{0d}}{n_0 + n} + \frac{n\hat{\theta}_d}{n_0 + n} \\ &= \frac{n_0}{n_0 + n}\theta_{0d} + \frac{n}{n_0 + n}\hat{\theta}_d.\end{aligned}$$

since MLE is

$$\hat{\theta}_d = \frac{n_d}{n}$$

- Once again, we can express our posterior expectation as a weighted average of the prior expectation and MLE.

- We can also extend the Dirichlet-multinomial model to more variables (contingency tables).

Example: pre-election polling

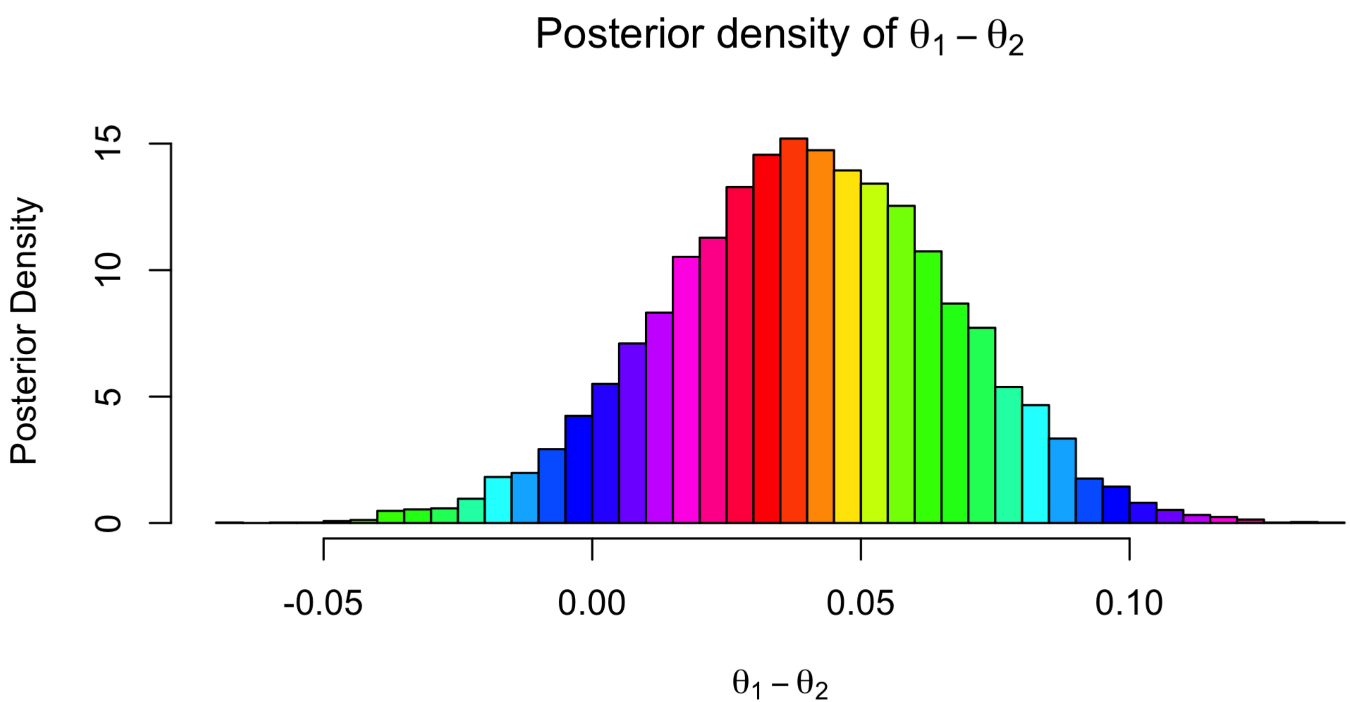
- Fox News Nov 3-6 pre-election survey of 1295 likely voters for the 2016 election.
- For those interested, FiveThirtyEight is an interesting source for pre-election polls.
- Out of 1295 respondents, 622 indicated support for Clinton, 570 for Trump, and the remaining 103 for other candidates or no opinion.
- Drawing inference from pre-election polls is way more complicated and nuanced that this. We only use the data here for this simple illustration.
- Assuming no other information on the respondents, we can assume simple random sampling and use a multinomial distribution with parameter $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$, the proportion, in the survey population, of Clinton supporters, Trump supporters and other candidates or no opinion.

Example: pre-election polling

- With a noninformative uniform prior, we have $\pi(\boldsymbol{\theta}) = \text{Dirichlet}(1, 1, 1)$.
- The resulting posterior is $\text{Dirichlet}(1 + n_1, 1 + n_2, 1 + n_3) = \text{Dirichlet}(623, 571, 104)$.
- Suppose we wish to compare the proportion of people who would vote for Trump versus Clinton, we could examine the posterior distribution of $\theta_1 - \theta_2$.
- We can even compute the probability $\Pr(\theta_1 > \theta_2 | Y)$.

Example: pre-election polling

```
#library(gtools)
PostSamples <- rdirichlet(10000, alpha=c(623,571,104))
#dim(PostSamples)
hist((PostSamples[,1] - PostSamples[,2]),col=rainbow(20),xlab=expression(theta[1]-theta[2])
     ylab="Posterior Density",freq=F,breaks=50,
     main=expression(paste("Posterior density of ",theta[1]-theta[2])))
```



Example: pre-election polling

- Posterior probability that Clinton had more support than Trump in the survey population, that is, $\Pr(\theta_1 > \theta_2|Y)$, is

```
#library(gtools)
mean(PostSamples[,1] > PostSamples[,2])
```

```
## [1] 0.9311
```

- Once again, this is just a simple illustration with a very small subset of the 2016 pre-election polling data.

- Inference for pre-election polls is way more complex and nuanced that this.

What's next?

Move on to the readings for the next module!