The Zipper Example[1]

STA442/2101 Fall 2017

[1] See last slide for copyright information.

Overview

Preparation: Indicator functions

Conditional expectation and the Law of Total Probability

$I_A(x)$ is the *indicator function* for the set A. It is defined by

$$I_A(x) = \begin{cases} 1 & \text{for } x \in A \\ 0 & \text{for } x \notin A \end{cases}$$

Also sometimes written $I(x \in A)$

$$\begin{aligned} E(I_A(X)) &= \sum_x I_A(x)p(x), \text{ or} \\ & \int_{-\infty}^{\infty} I_A(x)f(x)\,dx \\ &= P\{X \in A\} \end{aligned}$$

So the expected value of an indicator is a probability.

Applies to conditional probabilities too

$$
\begin{aligned}
E(I_A(X)|Y) &= \sum_x I_A(x)p(x|Y), \text{ or} \\
& \int_{-\infty}^{\infty} I_A(x)f(x|Y)\,dx \\
&= Pr\{X \in A|Y\}
\end{aligned}
$$

So the conditional expected value of an indicator is a *conditional* probability.

Double expectation: $E\left(g(X)\right) = E\left(E[g(X)|Y]\right)$

$$E\left(E[I_A(X)|Y]\right) = E[I_A(X)] = Pr\{X \in A\}, \text{ so}$$

$$\begin{aligned} Pr\{X \in A\} &= E\left(E[I_A(X)|Y]\right) \\ &= E\left(Pr\{X \in A|Y\}\right) \\ &= \int_{-\infty}^{\infty} Pr\{X \in A|Y=y\} f_Y(y)\, dy, \text{ or} \\ & \quad \sum_y Pr\{X \in A|Y=y\} p_Y(y) \end{aligned}$$

This is known as the *Law of Total Probability*

The Zipper Example

Members of a Senior Kindergarten class (which we shall treat as a sample) try to zip their coats within one minute. We count how many succeed.

How about a model?

$Y_1, \ldots, Y_n \stackrel{i.i.d.}{\sim} B(1, \theta)$, where θ is the probability of success.

A better model than $Y_1, \ldots, Y_n \stackrel{i.i.d.}{\sim} B(1, \theta)$

- Obviously, the probability of success is not the same for each child.
- Some are almost certain to succeed, and others have almost no chance.

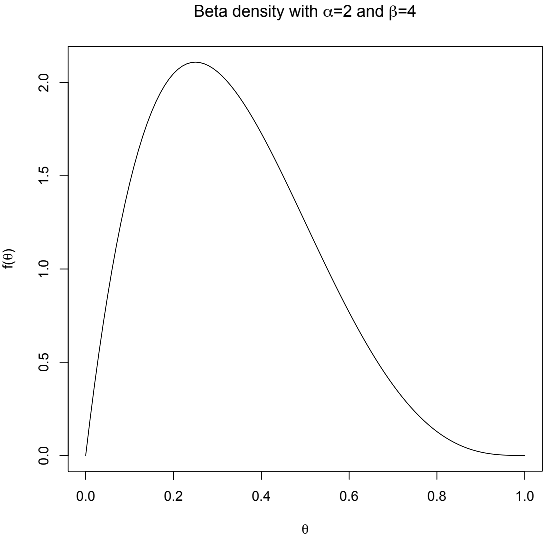
Alternative Model: $Y_1, \ldots, Y_n$ are independent random variables, with $Y_i \sim B(1, \theta_i)$.

$Y_1, \ldots, Y_n$ independent $B(1, \theta_i)$

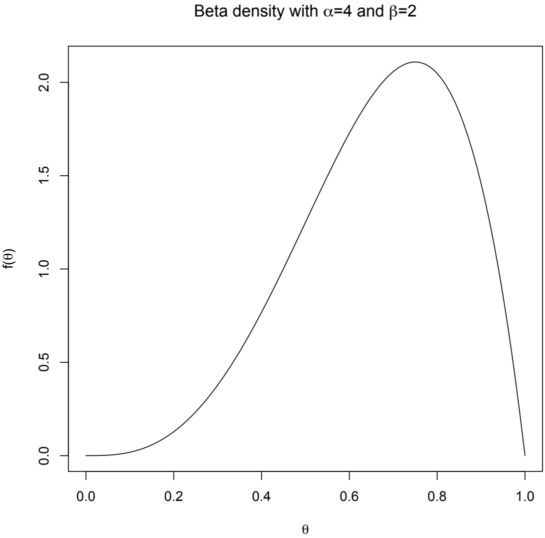
- This is a two-stage sampling model.
- First, sample from a population in which each child has a personal probability of success.
- Then for child i, use θ_i to generate success or failure.
- Note that $\theta_1, \ldots, \theta_n$ are random variables with some probability distribution.
- This distribution is supported on $[0, 1]$
- How about a beta?

$$f(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

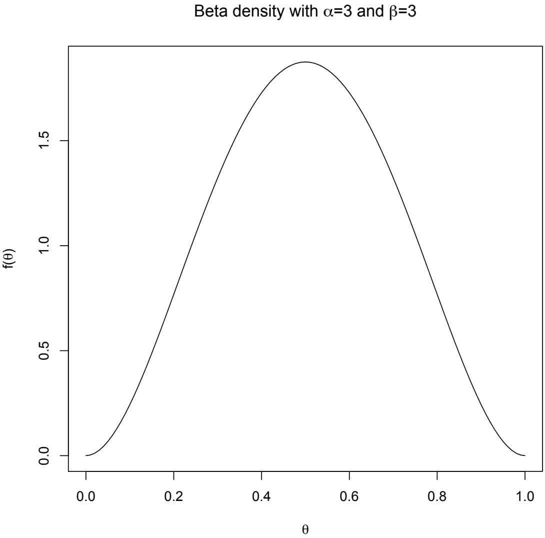
Beta density is flexible



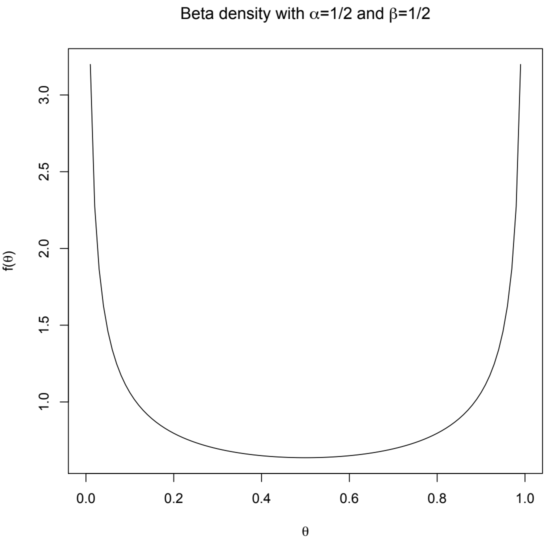
Beta density is flexible



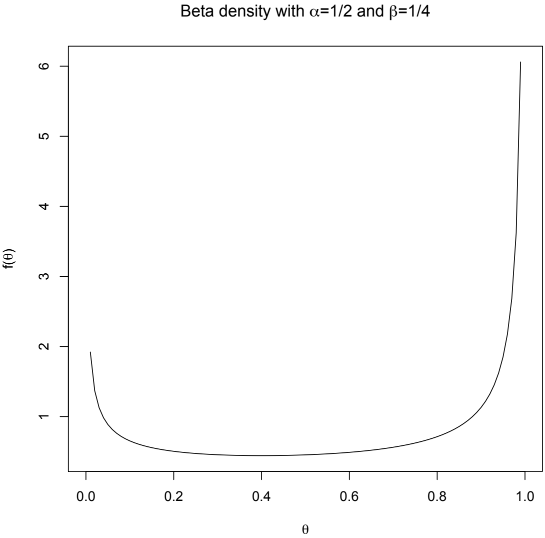
Beta density is flexible



Beta density is flexible



Beta density is flexible



Law of total probability

Double expectation

$$\begin{aligned}
P(Y_i = 1) &= \int_0^1 P(Y_i = 1|\theta_i)\, f(\theta_i)\, d\theta_i \\
&= \int_0^1 \theta_i\, f(\theta_i)\, d\theta_i \\
&= \int_0^1 \theta_i\, \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_i^{\alpha-1}(1-\theta_i)^{\beta-1}\, d\theta_i \\
&= \frac{\alpha}{\alpha+\beta}
\end{aligned}$$

Distribution of the observable data

$$P(\mathbf{Y}=\mathbf{y}|\alpha,\beta)=\prod_{i=1}^{n}\left(\frac{\alpha}{\alpha+\beta}\right)^{y_i}\left(1-\frac{\alpha}{\alpha+\beta}\right)^{1-y_i}$$

- Distribution of the observable data depends on the parameters α and β only through $\frac{\alpha}{\alpha+\beta}$.
- Infinitely many (α,β) pairs yield the same distribution of the data.
- How could you use the data to decide which one is right?

Parameter Identifiability

The general idea

- The parameters of the Zipper Model are not *identifiable*.
- The model parameters cannot be recovered from the distribution of the sample data.
- And all you can ever learn from sample data is the distribution from which it comes.
- So there will be problems using the sample data for estimation and inference about the parameters.
- This is true *even if the model is completely correct.*

Definitions

Connected to parameter identifiability

- A *Statistical Model* is a set of assertions that partly specify the probability distribution of the observable data.
- Suppose a statistical model implies $\mathbf{D} \sim P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta$. If no two points in Θ yield the same probability distribution, then the parameter $\boldsymbol{\theta}$ is said to be *identifiable.*
- That is, identifiability means that $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$ implies $P_{\boldsymbol{\theta}_1} \neq P_{\boldsymbol{\theta}_2}$.
- On the other hand, if there exist distinct $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ in Θ with $P_{\boldsymbol{\theta}_1} = P_{\boldsymbol{\theta}_2}$, the parameter $\boldsymbol{\theta}$ is *not identifiable.*

An equivalent definition

Equivalent to $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2 \Rightarrow P_{\boldsymbol{\theta}_1} \neq P_{\boldsymbol{\theta}_2}$

- The probability distribution is always a function of the parameter vector.
- If that function is one-to-one, the parameter vector is identifiable, because then $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$ yielding the same distribution could not happen.

- That is, if the parameter vector can somehow be recovered from the distribution of the data, it is identifiable.

Theorem

If the parameter vector is not identifiable, consistent estimation for all points in the parameter space is impossible.

- Suppose $\theta_1 \neq \theta_2$ but $P_{\theta_1} = P_{\theta_2}$
- $T_n = T_n(D_1, \ldots, D_n) \stackrel{p}{\rightarrow} \theta$ for all $\theta \in \Theta$.
- Distribution of T_n is identical for θ_1 and θ_2.

Why don't we hear more about identifiability?

- Consistent estimation indirectly proves identifiability.
- Because without identifiability, consistent estimation would be impossible.
- Any *function* of the parameter vector that can be estimated consistently is identifiable.

Maximum likelihood fails for the Zipper Example

It has to fail.

$$
\begin{aligned}
L(\alpha,\beta) &= \left(\frac{\alpha}{\alpha+\beta}\right)^{\sum_{i=1}^n y_i} \left(1-\frac{\alpha}{\alpha+\beta}\right)^{n-\sum_{i=1}^n y_i} \\
\ell(\alpha,\beta) &= \log\left(\left(\frac{\alpha}{\alpha+\beta}\right)^{\sum_{i=1}^n y_i} \left(1-\frac{\alpha}{\alpha+\beta}\right)^{n-\sum_{i=1}^n y_i}\right)
\end{aligned}
$$

Partially differentiate with respect to α and β, set to zero, and solve.

Two equations in two unknowns

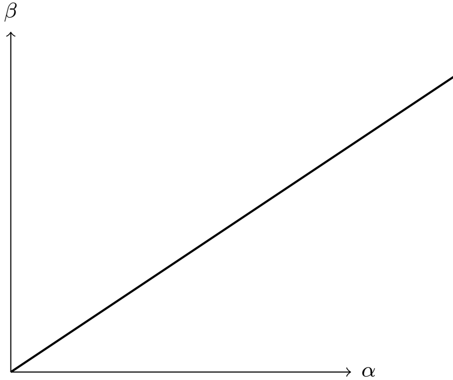
$$\frac{\partial \ell}{\partial \alpha} \stackrel{set}{=} 0 \quad \Rightarrow \quad \frac{\alpha}{\alpha+\beta} = \overline{y}$$
$$\frac{\partial \ell}{\partial \beta} \stackrel{set}{=} 0 \quad \Rightarrow \quad \frac{\alpha}{\alpha+\beta} = \overline{y}$$

Any pair (α, β) with $\frac{\alpha}{\alpha+\beta} = \overline{y}$ will maximize the likelihood.

The MLE is not unique.

What is happening geometrically?

$$\frac{\alpha}{\alpha+\beta} = \overline{y} \Leftrightarrow \beta = \left(\frac{1-\overline{y}}{\overline{y}}\right)\alpha$$



Fisher Information: $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = \left[E\{ -\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log f(Y|\boldsymbol{\theta})\} \right]$

The Hessian of the minus log likelihood approximates n times the Fisher Information.

$$\begin{aligned} \log f(Y|\alpha,\beta) &= \log\left(\left(\frac{\alpha}{\alpha+\beta}\right)^Y \left(1-\frac{\alpha}{\alpha+\beta}\right)^{1-Y} \right) \\ &= Y\log\alpha + (1-Y)\log\beta - \log(\alpha+\beta) \end{aligned}$$

$\mathcal{I}(\alpha,\beta) = \left[E\{-\frac{\partial^2}{\partial\alpha\partial\beta}\log f(Y|\alpha,\beta)\}\right]$

Where $\log f(Y|\alpha,\beta) = Y\log\alpha + (1-Y)\log\beta - \log(\alpha+\beta)$

$$\begin{aligned}
\mathcal{I}(\alpha,\beta) &= E\left(\begin{array}{cc} -\frac{\partial^2 \log f}{\partial\alpha^2} & -\frac{\partial^2 \log f}{\partial\alpha\partial\beta} \\ -\frac{\partial^2 \log f}{\partial\beta\partial\alpha} & -\frac{\partial^2 \log f}{\partial\beta^2} \end{array}\right) \\
&= \ldots \\
&= \frac{1}{(\alpha+\beta)^2}\left(\begin{array}{cc} \frac{\beta}{\alpha} & 1 \\ 1 & \frac{\alpha}{\beta} \end{array}\right)
\end{aligned}$$

- Determinant equals zero.
- The inverse does not exist.
- Large sample theory fails.
- Second derivative test fails.
- The likelihood is flat (in a particular direction).

Look what has happened to us.

- We made an honest attempt to come up with a better model.
- And it *was* a better model.
- But the result was disaster.

There is some good news.

Remember from earlier that by the Law of Total Probability,

$$P(Y_i = 1) = \int_0^1 \theta_i \, f(\theta_i) \, d\theta_i = E(\Theta_i)$$

- Even when the probability distribution of the (random) probability of success is completely unknown,
- We can estimate its expected value (call it μ) consistently with $\overline{Y}_n$.
- So that *function* of the unknown probability distribution is identifiable.
- And often that's all we care about anyway, say for comparing group means.
- So the usual procedures, based on a model nobody can believe (Bernoulli), are actually informative about a much more realistic model whose parameter is not fully identifiable.
- We don't often get this lucky.

One more question about the parametric version

What would it take to estimate α and β successfully?

- Get the children to try zipping their coats twice, say on two consecutive days.
- Assume their ability does not change, and conditionally on their ability, the two tries are independent.
- That will do it.

- This kind of thing often happens. When the parameters of a reasonable model are not identifiable, maybe you can design a different way of collecting data so that the parameters *can* be identified.

Moral of the story

- If you think up a better model for standard kinds of data, the parameters of the model may not be identifiable. You need to check.
- The problem is not with the model. It's with the data.
- The solution is better *research design*.

Copyright Information