Intro to Hierarchical Models

ISTA 410 / INFO 510: Bayesian Modeling and Inference

U. of Arizona School of Information

September 21, 2020

Public Service Announcement

Necessary PSA on voting:

- If you're eligible to vote in AZ, you must be registered as of Monday 10/5
- Check your registration at: `https://my.arizona.vote/WhereToVote.aspx?s=individual`

Voting is important, you should do it if you can.

Outline

Last week

- DAGs as probabilistic models
- Causal inference and paradoxes

Now:

- Hierarchical (multilevel) models
- More simulation in Python

Example

Exercise 3.8

We'll use a problem from the previous HW: the bike lane problem.

Problem:

- Analyze the proportion of vehicles on residential streets that are bicycles
- Compare streets with bike lanes vs. no bike lanes

Exercise 3.8

In the HW problem, we:

- set up a model with parameters θ_y (proportion of bicycles on bike-lane streets) and θ_z (proportion of bicycles on no-bike-lane streets)
- compute a posterior distribution for θ_y, θ_z
- draw from the posterior to estimate the expected difference between groups

Exercise 3.8

The model:

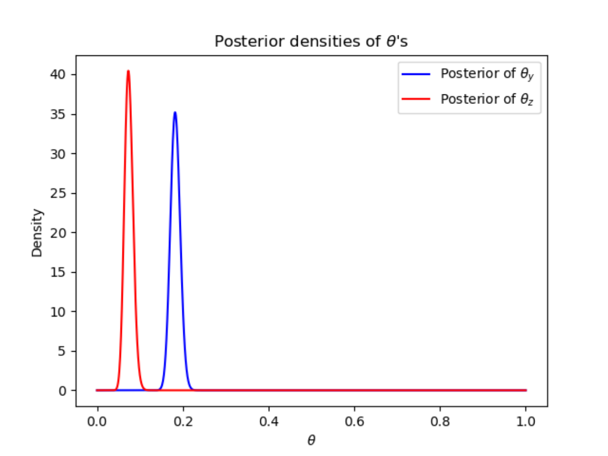
$$y_j \sim \mathrm{Binomial}(\theta, n_j)$$
$$\theta \sim \mathrm{Beta}(\alpha_0, \beta_0)$$

for fixed α_0, β_0.

- Choosing $\alpha_0 = 1, \beta_0 = 1$ gives a completely noninformative (flat) prior
- Weakly informative prior also reasonable, e.g. $\alpha_0 = 1, \beta_0 = 3$ for prior mean of 25% bicycle traffic

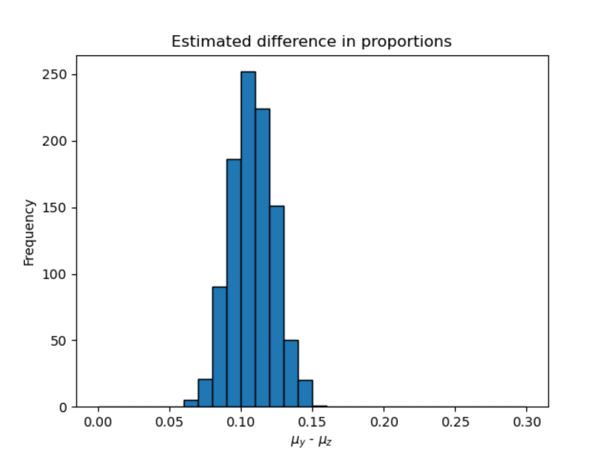
Exercise 3.8

Posteriors for the streets with and without bike lanes:



Exercise 3.8

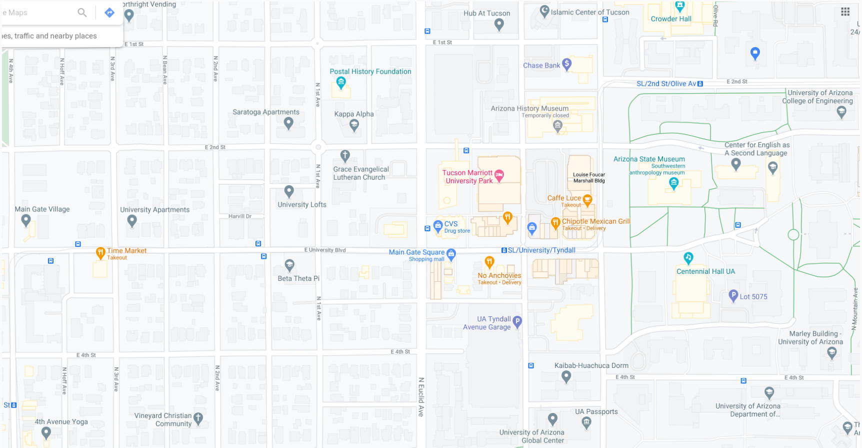
Then, by sampling from the posterior, we can estimate the distribution of quantities of interest, such as $\theta_y - \theta_z$:



A hierarchical model

Why a hierarchical model?

The model we wrote in the previous section treats all streets as the same; each street's observation is an observation of the same underlying proportion.



Why a hierarchical model?

As an alternative, we could treat each street as an independent entity:

Why a hierarchical model?

As an alternative, we could treat each street as an independent entity:

$$y_j \sim \mathrm{Binomial}(\theta_j, n_j)$$
$$\theta_j \sim \mathrm{Beta}(\alpha_0, \beta_0)$$

for fixed α_0, β_0.

- Exactly like the previous model, except we now have 10 independent θ_js for the 10 streets
- Same considerations for choice of prior

Call this the separate-effects model.

Why a hierarchical model?

Choosing between the two models: classically, do an analysis of variance

- Compare variance within groups (streets) to variance between streets
- Test against the null hypothesis that all streets are the same
- If we reject the null, take the separate-effects model
- If we don't take the pooled model

Problem: false dichotomy!

Why a hierarchical model?

In reality, it is most plausible that both of the following are true:

- The streets are not identical; some of the streets are more popular with cyclists
- Observations of one street can inform our knowledge of the others

So: neither side of this dichotomy is preferable.

Analogy: cafes

Imagine you're walking into a cafe; how long will it take you to get your coffee?

- Varies among cafes / franchises
- Not completely independent

The Bayesian solution

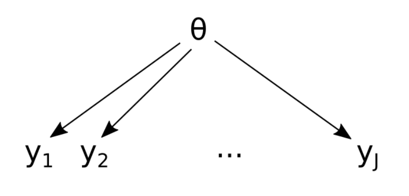
With a Bayesian approach, we can find a compromise.

- We have a θ for each street
- However, instead of being fully independent, each θ is drawn from a common probability distribution
- This probability distribution, a *hyperprior*, depends on *hyperparameters* which we estimate from the data

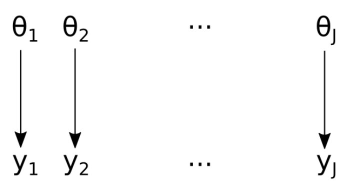
(note: slightly different sense of the term *hyperparameter* from its common use in ML)

Examining this graphically

Pooled model:

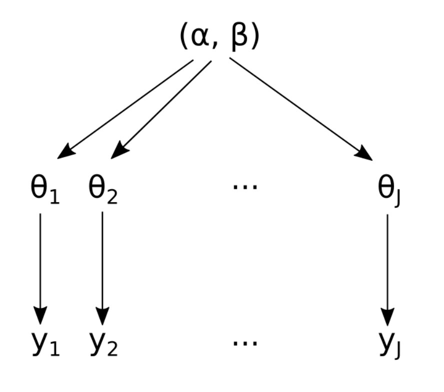


Separate model:



Examining this graphically

Hierarchical model combines the features of these two:



Note the usual DAG properties still apply: θ_js are no longer fully independent, but they are *conditionally* independent given α, β.

Setting up the model

This is conceptually only a slight difference from our previous model:

$$\begin{aligned} y_j &\sim \mathrm{Binomial}(\theta_j, n_j) \\ \theta_j &\sim \mathrm{Beta}(\alpha, \beta) \\ p(\alpha, \beta) &\propto ??? \end{aligned}$$

Choosing a hyperprior

We need a prior distribution for α, β; this can be a tricky part of this sort of modeling, because the interpretation of these parameters is not so simple compared to θ_j.

- Flat prior on α, β? We have a lot of data...

Choosing a hyperprior

We need a prior distribution for α, β; this can be a tricky part of this sort of modeling, because the interpretation of these parameters is not so simple compared to θ_j.

- Flat prior on α, β? We have a lot of data...
- ...but the posterior isn't integrable

So we need to put a little thought into a prior.

Choosing a hyperprior

BDA suggests the following as a prior for a similar example:

$$p(\alpha, \beta) \propto (\alpha + \beta)^{-5/2}$$

What's the intuition behind this?

Choosing a hyperprior

BDA suggests the following as a prior for a similar example:

$$p(\alpha, \beta) \propto (\alpha + \beta)^{-5/2}$$

What's the intuition behind this?

In a beta distribution, interpretation of parameters as "pseudocounts":

Choosing a hyperprior

BDA suggests the following as a prior for a similar example:

$$p(\alpha, \beta) \propto (\alpha + \beta)^{-5/2}$$

What's the intuition behind this?

In a beta distribution, interpretation of parameters as "pseudocounts":

- If we start with $\mathrm{Beta}(\alpha, \beta)$ and make binomial observations, we update to the posterior $\mathrm{Beta}(\alpha + n_s, \beta + n_f)$, with n_s successes and n_f failures
- So, we can think of α and β as "counts" of imaginary observations

Choosing a hyperprior

Goal: prior is noninformative on the mean value of θ_j and the spread, or scale, of that mean

- Mean is $\frac{\alpha}{\alpha+\beta}$
- Scale parameters (standard errors) for means are distributed like $n^{-1/2}$ where n is the sample size
- Our "sample size"

So: set up a prior distribution that is uniform on $\left(\frac{\alpha}{\alpha+\beta}, (\alpha+\beta)^{-1/2}\right)$

Choosing a hyperprior

Define:

$$
\begin{aligned}
w &= \frac{\alpha}{\alpha + \beta} \\
z &= (\alpha + \beta)^{-1/2} \\
p(w, z) &\propto 1
\end{aligned}
$$

Do some calculus...

Choosing a hyperprior

Define:

$$
\begin{aligned}
w &= \frac{\alpha}{\alpha + \beta} \\
z &= (\alpha + \beta)^{-1/2} \\
p(w, z) &\propto 1
\end{aligned}
$$

Do some calculus...

...and we arrive at

$$p(\alpha, \beta) \propto (\alpha + \beta)^{-5/2}$$

which is an integrable (proper) prior.

The full model

So, now we have a fully-specified probability model:

$$\begin{aligned} y_j &\sim \text{Binomial}(\theta_j, n_j) \\ \theta_j &\sim \text{Beta}(\alpha, \beta) \\ p(\alpha, \beta) &\propto (\alpha + \beta)^{-5/2} \end{aligned}$$

Inference the hard way

As usual, we can make inferences by sampling from the posterior distribution. This can be done the hard way (directly), or the easy way (MCMC).

Hard way:

1. Calculate the posterior density $p(\alpha, \beta|y)$ on a grid of α and β values.
2. Sum over the β values to get an estimate of the marginal posterior $p(\alpha|y)$; use this to draw samples of α.
3. For each sampled value of α, use the conditional posterior $p(\beta|\alpha, y)$ (which is a slice of)
4. For each sampled pair (α_i, β_i), draw values of θ_j from the beta distribution $\mathrm{Beta}(\alpha_i + y_j, \beta_i + y_j - n_j)$

Inference the easy way

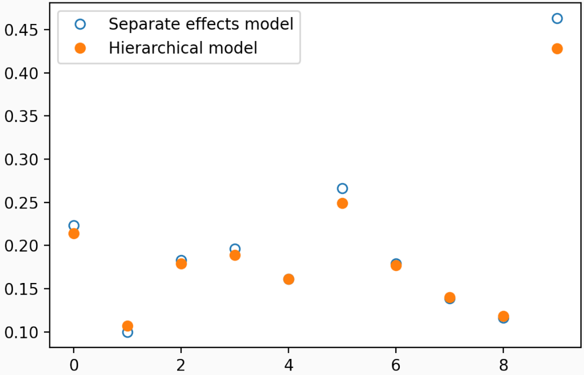
The easier approach: use MCMC to sample from the posterior.

Let's see this in action...

Comparison

What is the difference in the results?

Let's compare point estimates:



Shrinkage and regularization

The shrinkage effect we see is a form of regularization:

- Most extreme observations “shrunk” toward an overall average
- Amount of shrinkage tuned to relative sample size

Difference: we learned the strength of regularization from the data

Underfitting and overfitting

Another way to think about this, in terms of underfitting and overfitting:

- The pooled model: maximum underfitting
- The separate-effects model: maximum overfitting
- Hierarchical model: adaptive regularization

With enough observations the seperate effects model will estimate each street similarly to the hierarchical model.

Summary

Hierarchical models:

- Have several “levels” of parameters stacked
- Perform adaptive regularization – learn priors from the data

Next time: more models

Appendix: hyperprior calculation

Reminder

As a reminder, our prior distribution was uniform on

$$\left(\frac{\alpha}{\alpha+\beta}, (\alpha+\beta)^{-1/2}\right)$$

Define $w = \frac{\alpha}{\alpha+\beta}, z = (\alpha+\beta)^{-1/2}$, and set $p(w, z) \propto 1$.

Changing variables for probability densities comes from changing variables for integrals, because the PDF is defined by the property that

$$\mathrm{Pr}(x_1, \ldots, x_n \in A) = \int_A p(x_1, \ldots, x_n) dx_1 \ldots dx_n$$

To perform the change of variables, we need to multiply by the absolute determinant of the Jacobian matrix

$$J = \begin{pmatrix} \frac{\partial w}{\partial \alpha} & \frac{\partial w}{\partial \beta} \\ \frac{\partial z}{\partial \alpha} & \frac{\partial z}{\partial \alpha} \end{pmatrix}$$

To perform the change of variables, we need to multiply by the absolute determinant of the Jacobian matrix

$$J = \begin{pmatrix} \frac{\partial w}{\partial \alpha} & \frac{\partial w}{\partial \beta} \\ \frac{\partial z}{\partial \alpha} & \frac{\partial z}{\partial \alpha} \end{pmatrix}$$

$$J = \begin{pmatrix} \frac{\beta}{(\alpha+\beta)^2} & \frac{-\alpha}{(\alpha+\beta)^2} \\ -\frac{1}{2}(\alpha+\beta)^{-3/2} & -\frac{1}{2}(\alpha+\beta)^{-3/2} \end{pmatrix}$$

so $|\det J| = \frac{1}{2}(\alpha+\beta)^{-5/2}$ (and we can drop the 1/2 because it's a constant)