36-617: Applied Linear Models

Bayes, Shrinkage, and Multi-Level Models

Brian Junker
132E Baker Hall
brian@stat.cmu.edu

Announcements

- No new HW, no quiz this week
  - Work on your IDMRAD rough drafts (Due Weds / Grace Fri)
  - IDMRAD papers submitted to Canvas, not Gradescope
    - Details will be on Canvas
- See handout on rough draft IDMRAD papers
- Reading: Please at least skim
  - Lynch Ch 3: Basics of Bayesian Statistics
    - Worth reading a little more carefully than a skim
  - Lynch Ch 4: Modern Model Estimation Part 1: Gibbs Sampling
    - Read 4.1, 4.2 more carefully; skim rest of chapter

(these are in the week12 folder on canvas!)

Outline

- Today:
  - Shrinkage
  - Review of MLE
  - Crash course in Bayes
  - Normal-Normal Model & Shrinkage
  - MLM's and Shrinkage
- Project Discussion
- After Thanksgiving:
  - A little practical Bayes / MCMC for multi-level models

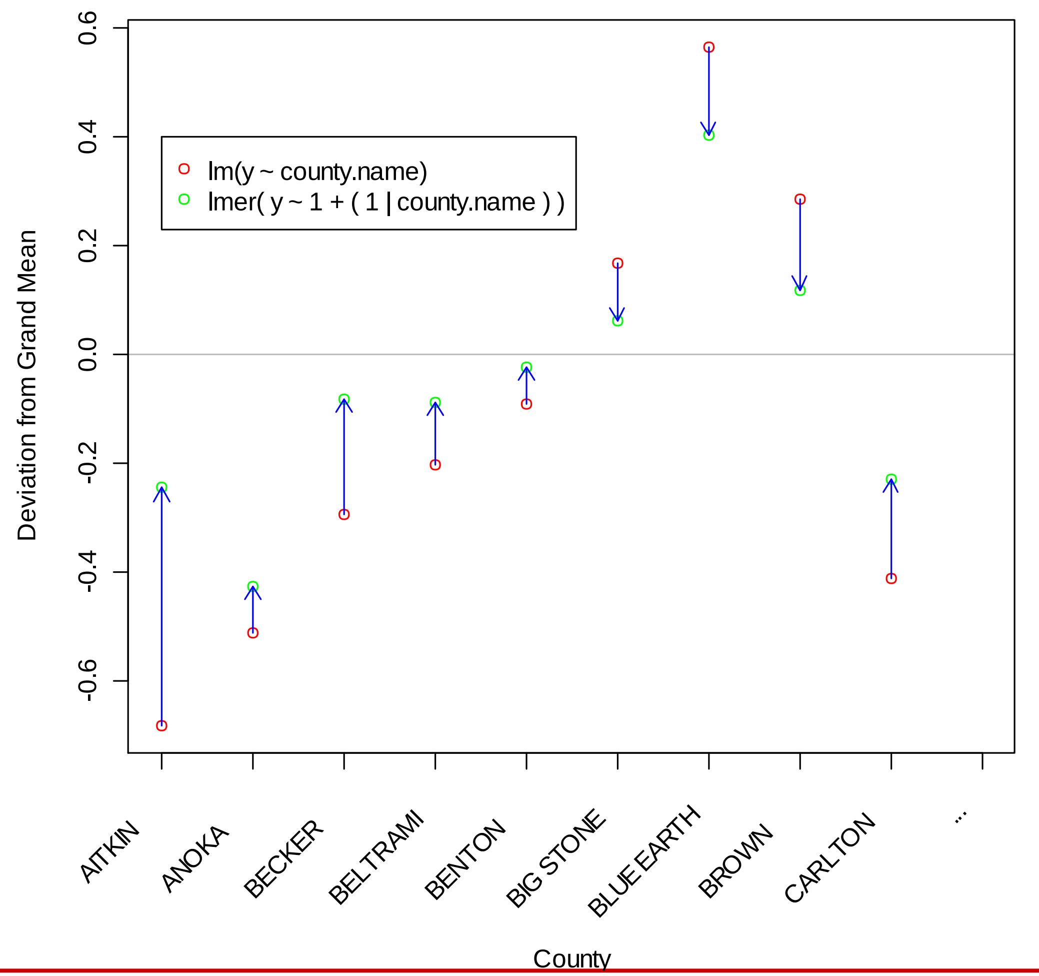
An MLM phenomenon: Shrinkage

The fitted multilevel model underpredicts high obs's and overpredicts low ones.

The distribution assumptions underlying lmer() "smooth out" extreme observations!

Multi-level models provide more smoothing/shrinkage to groups with smaller sample sizes (since there is less evidence that their values should be different from "grand mean".)

We'll talk about ***why*** today…



Methods of Estimation – How can we systematically construct "good" estimators?

- Several methods have proven useful:
  - *Method of Moments (MoM)*: The k^th moment of X is $E[X^k]$. MoM estimators combine unbiased estimates of moments of X.
  - *Least Squares (LS)*: Obtained by minimizing squared error $\sum_{i=1}^{n}(Y_i - E[Y_i])^2$. Ordinary linear regression!
  - *Maximum likelihood (ML)*: The likelihood is the probability of the data we observed. ML estimators (MLE's) choose parameter values that maximize the likelihood.
  - *Bayesian Estimation (Bayes)*: Treat the parameters as random variables, and use Bayes' rule to pick the parameter value most likely, given the data (the "reverse" of ML!)

Maximum Likelihood Estimators (MLE's)

- Let $X_1, \ldots, X_n$ be an iid sample from $f_X(x;\theta)$, $x_1, \ldots, x_n$ are the observed values
- The *likelihood* of the sample is the joint density

$$\begin{aligned} L(\theta) &= f(x_1, \ldots, x_n;\ \theta) = f(x_1;\ \theta) f(x_2;\ \theta) \cdots f(x_n;\ \theta) \\ &= \prod_{i=1}^{n} f(x_i;\ \theta) \end{aligned}$$

- The *maximum likelihood estimate* $\hat{\theta}_{MLE}$ maximizes $L(\theta)$:

$$L(\hat{\theta}_{MLE}) \geq L(\theta) \quad \forall\ \theta$$

- Strategy: It's usually (but not always) easier to work with the *log likelihood*

$$LL(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(x_i;\ \theta)\ .$$

Pennsylvania, Pre-Midterm Poll

- John Fetterman (D) running for election to the US Senate against Mehmet Oz ®
- In a Suffolk University Poll (October 27-30):
  - 457 of 500 voters expressed a preference for Fetterman or Oz.
  - Of those 457: 233 prefer Fetterman.
- In most polling, weights are attached to each response, to adjust the "representativeness" of the response for things like
  - who is likely to be home when survey worker calls
  - who refuses to answer
  - etc
- We will ignore weights etc and treat the 457 as a simple random sample.

Possible models for the data

- 457 individual Bernoulli coin flips, x_i = 1 for Fetterman, x_i = 0 for Oz

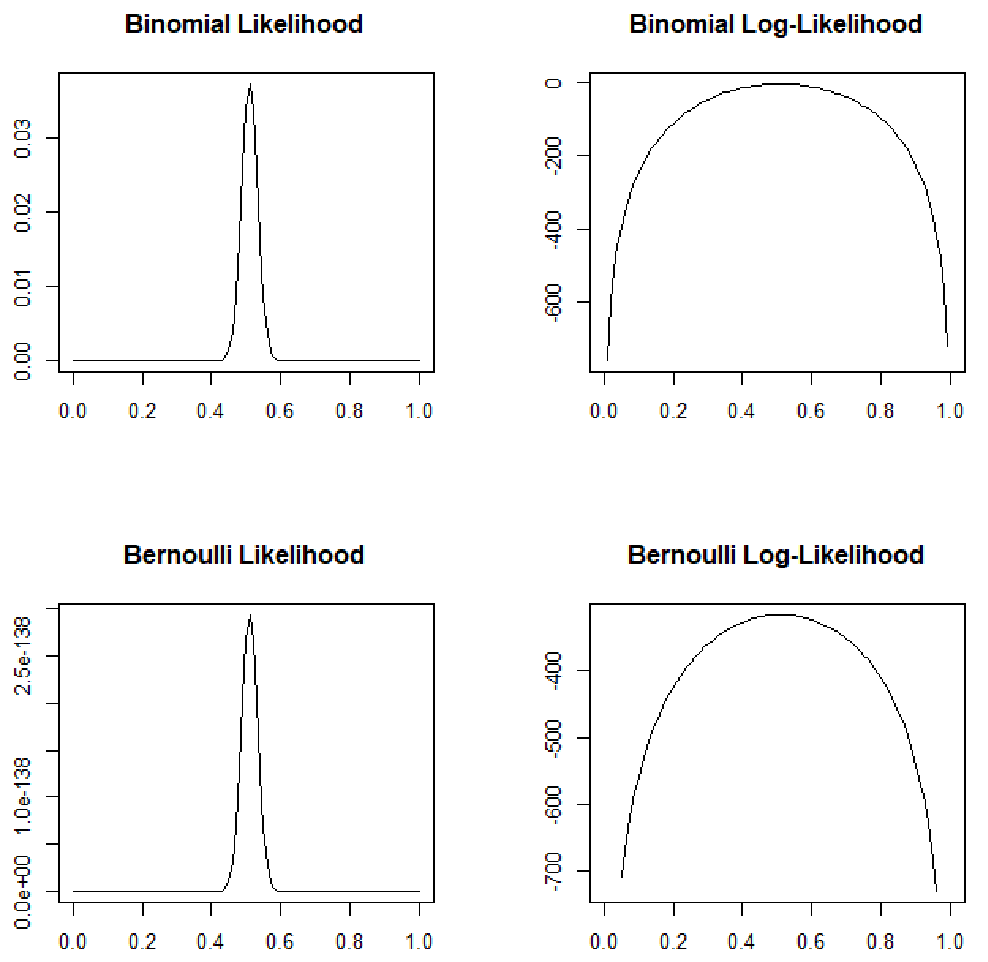
$$L_{ber}(p) = \prod_{i=1}^{457} p^{x_i}(1-p)^{1-x_i} = p^{233}(1-p)^{224}$$

- 457 trials, 233 “successes” (Fetterman voters)

$$L_{bin}(p) = \binom{457}{233} p^{233}(1-p)^{224}$$

- What matters for MLE and SE is ***shape***, not ***size***!

Binomial and Bernoulli Likelihoods



Finding the MLE…

- If we use the Bernoulli likelihood,

$$\begin{aligned} LL_{ber}(p) &= \log L_{ber}(p) \\ = \log p^k(1-p)^{n-k} &= k\log p + (n-k)\log(1-p) \end{aligned}$$

- If we use the Binomial likelihood

$$\begin{aligned} LL_{bin}(p) &= \log L_{bin}(p) \\ = \log \binom{n}{k} p^k(1-p)^{n-k} &\propto k\log p + (n-k)\log(1-p) \end{aligned}$$

- Either way we want to maximize

$$k\log p + (n-k)\log(1-p)$$

with k = 233, n=457

MLE: Point Estimate

- Differentiating and setting to zero…

$$\begin{aligned} 0 &= LL'(p) = \frac{d}{dp}\left[k\log p + (n-k)\log(1-p)\right] \\ &= \frac{k}{p} - \frac{n-k}{1-p} = \frac{k-pn}{p(1-p)} \end{aligned}$$

- so, clearly,

$$\hat{p} = \frac{k}{n} = \frac{233}{457} = 0.510$$

Bayes' Rule (a.k.a. Bayes' Theorem)

- A very simple idea with very powerful consequences
- We often start with information like P[A|B] and what we really want is P[B|A]. Bayes' Theorem lets us "turn the conditioning around":

$$\mathbf{P}[\mathbf{B}|\mathbf{A}] = \frac{P[A\&B]}{P[A]} = \frac{\mathbf{P}[\mathbf{A}|\mathbf{B}]P[B]}{P[A]}$$

- See https://arbital.com/p/bayes_rule/ for lots of examples and proselytizing.

Conditional probability & conditional density

- P[A|B] = P[A&B]/P[B]
- P[B] = P[B|A]P[A] + P[B|A^c]P[A^c]
- P[A & B] = P[B|A]P[A]

- Bayes' Theorem:

$$
\begin{aligned}
P[B|A] &= \frac{P[A\&B]}{P[A]} = \frac{P[A|B]P[B]}{P[A]} \\
&= \frac{P[A|B]P[B]}{P[A|B]P[B] + P[A|B^c]P[B^c]}
\end{aligned}
$$

- f(x|y) = f(x,y)/f(y)
- $f(y) = \int f(y|x) f(x) dx$
- f(x,y) = f(y|x) f(x)

- Bayes' Theorem:

$$
\begin{aligned}
f(y|x) &= \frac{f(x,y)}{f(x)} = \frac{f(x|y)f(y)}{f(x)} \\
&= \frac{f(x|y)f(y)}{\int f(x|y^*)f(y^*)dy^*}
\end{aligned}
$$

Bayes' Theorem for Data

- Bayes' Theorem

$$f(y|x) = \frac{f(x,y)}{f(x)} = \frac{f(x|y)f(y)}{f(x)}$$

$$= \frac{f(x|y)f(y)}{\int f(x|y^*)f(y^*)dy^*}$$

Dummy variable of integration

- Let x = data, y = θ (parameter!); then

$$f(\theta|\text{data}) = \frac{f(\text{data},\theta)}{f(\text{data})} = \frac{f(\text{data}|\theta)f(\theta)}{f(\text{data})}$$

$$= \frac{f(\text{data}|\theta)f(\theta)}{\int f(\text{data}|\theta^*)f(\theta^*)d\theta^*}$$

Bayes' Theorem for Data

- We call
  - f(θ) the *prior distribution*
  - f(data|θ) = L(θ) the *likelihood*
  - f(θ|data) the *posterior distribution*
- So Bayes' Theorem says

$$f(\theta|\text{data}) = \frac{f(\text{data}|\theta)f(\theta)}{f(\text{data})} \propto f(\text{data}|\theta)f(\theta)$$

- Slogan: (posterior) $\propto$ (likelihood)$\times$(prior)

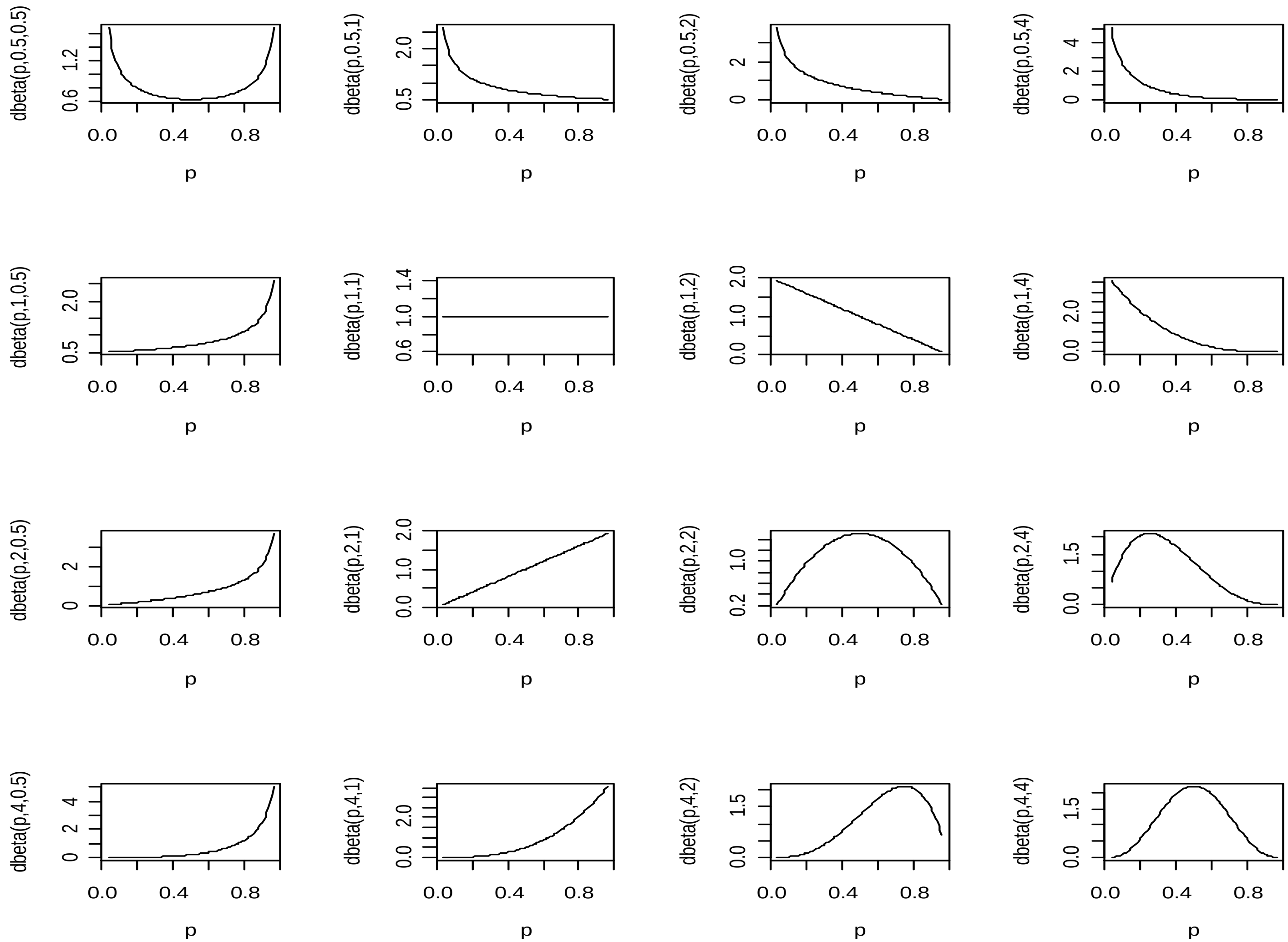
Back to 2022 PA pre-midterm poll

- The *likelihood* is the same as before:

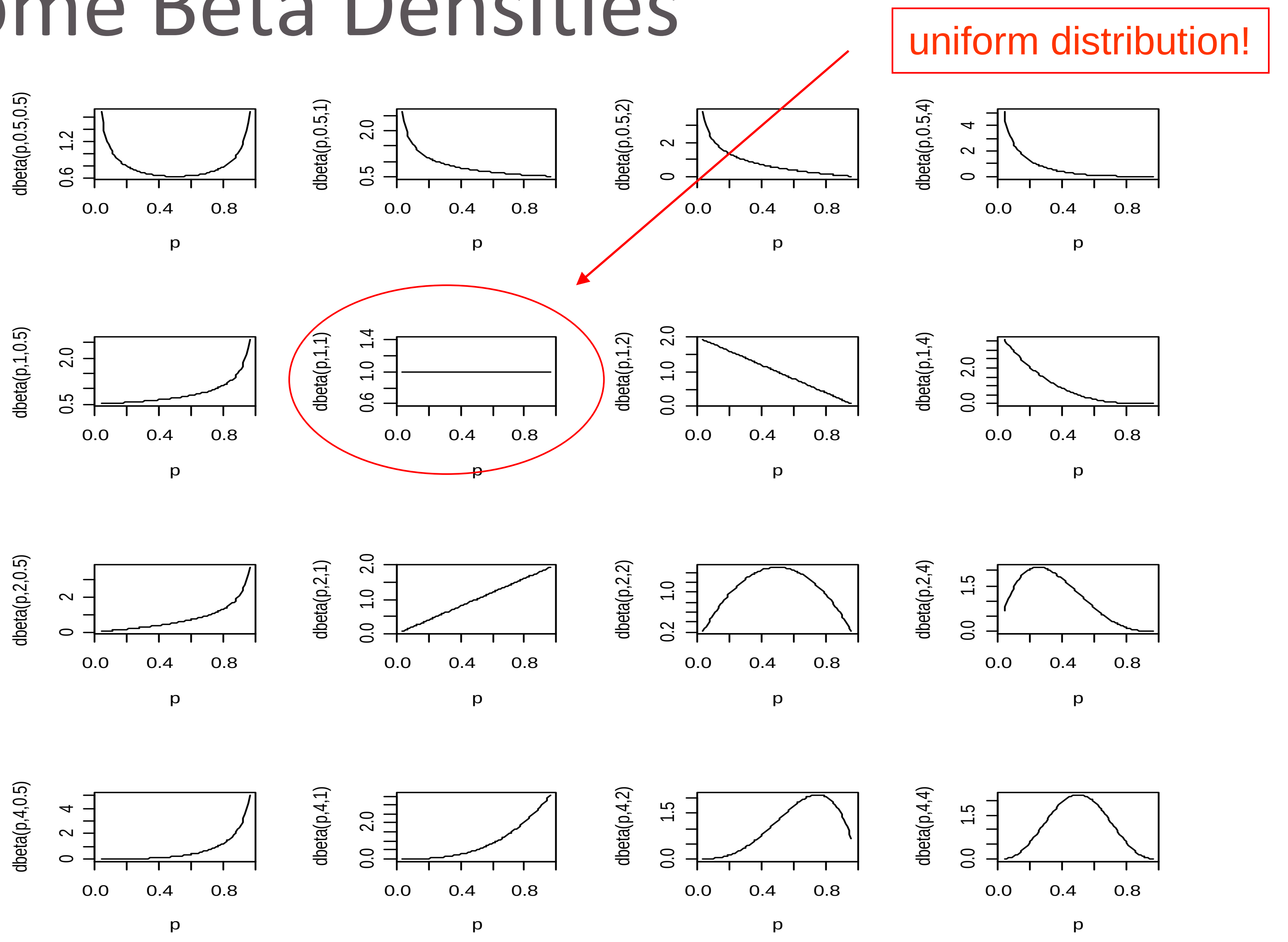
$$L(p) \propto p^k (1-p)^{n-k}$$

- We need a *prior distribution*. One good choice is a *beta distribution*, with
  - Density $f(p|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$
  - Mean $E[p] = \frac{\alpha}{\alpha+\beta}$
  - Variance $\text{Var}(p) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
- Some graphs of beta densities appear on the next slide

Some Beta Densities



Some Beta Densities



Choosing prior parameters…

- The *likelihood* is the same as before:

$$L(p) \propto p^k (1-p)^{n-k} = p^{226}(1-p)^{225}$$

- The *prior distribution* is a beta distribution

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

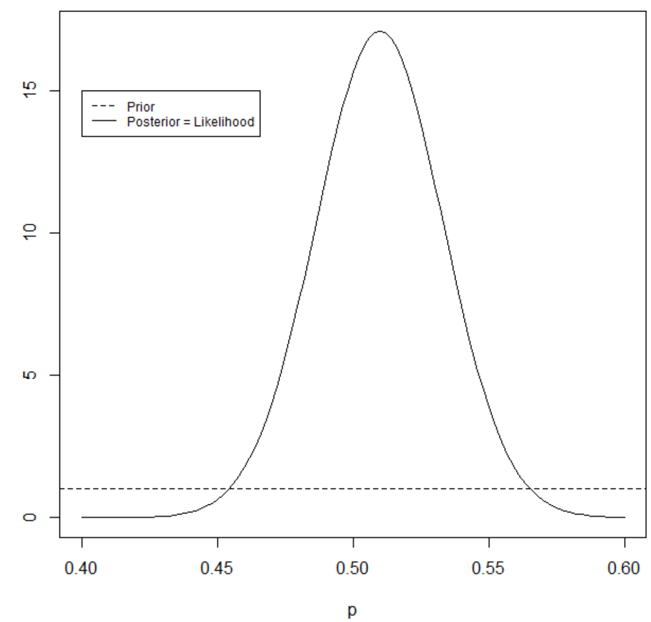
  - α = 1, β = 1 gives a uniform distribution – no preference for one p over another!
  - Suppose that in a previous poll, 942 prefer Fetterman and 1008 prefer Oz. Could set α=942, β=1008

If α=1 and β=1...

- (posterior) $\propto$ (likelihood)$\times$(prior):

$$f(p|\text{data}) \quad \propto \quad L(p) \times 1 \;=\; p^{233}(1-p)^{224}$$

- Since f(p|data)=L(p), posterior mode = MLE = 233/457 = 0.5098468
- Since f(p|data) is a beta with α=234, β=225, E[p|data] = 234/459 = 0.5098039



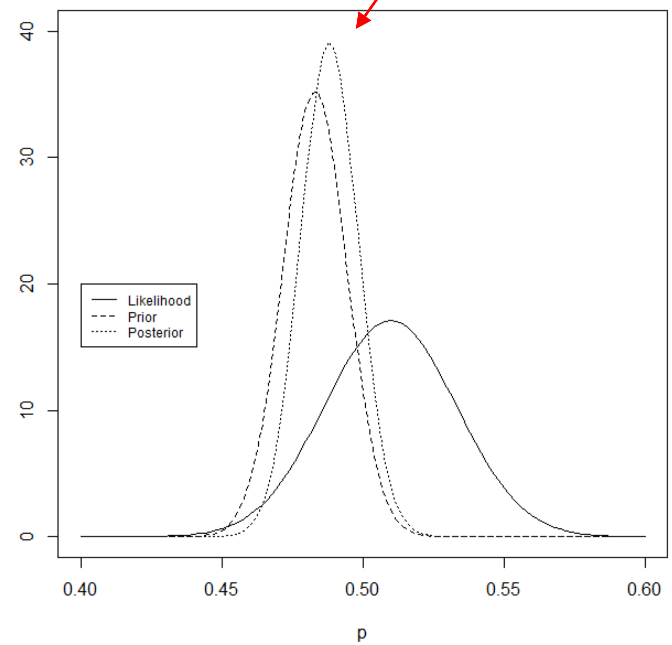
If α=942, β=1008…

"shrinkage": posterior between prior & likelihood

- (posterior) $\propto$ (likelihood)$\times$(prior):

$$f(p|\text{data}) \;\propto\; L(p) \times p^{941}(1-p)^{1007}$$
$$= \; p^{1174}(1-p)^{1231}$$

- Since f(p|data) = beta(p,1175,1232), E[p|data] = 1175/2406 = 0.488 vs MLE=0.510



Normal Model: Estimate μ, with σ^2 Known, One Observation y $\sim$ N(μ,σ^2)

- For our prior distribution, we'll assume $\mu \sim N(\mu_0,\tau_0^2)$:

$$
\begin{aligned}
f(y|\mu) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\mu)^2} \\
f(\mu) &= \frac{1}{\sqrt{2\pi}\tau_0} e^{-\frac{1}{2\tau_0^2}(\mu-\mu_0)^2} \\
f(\mu|y) &\propto f(y|\mu)f(\mu) \propto \exp\left\{-\frac{1}{2}\left[\frac{(y-\mu)^2}{\sigma^2} + \frac{(\mu-\mu_0)^2}{\tau_0^2}\right]\right\}
\end{aligned}
$$

- Posterior must be normal for μ (quadratic in μ!); to identify it, complete the square…

- The exponent of f(μ|y) looks like -1/2 times

$$\begin{aligned}
\frac{(y-\mu)^2}{\sigma^2} + \frac{(\mu-\mu_0)^2}{\tau_0^2} &= \frac{\tau_0^2+\sigma^2}{\tau_0^2\sigma^2}\left[\mu^2 - \frac{2y\mu\tau_0^2 + 2\mu\mu_0\sigma^2}{\tau_0^2+\sigma^2} + \frac{y^2\tau_0^2 + \mu_0^2\sigma^2}{\tau_0^2+\sigma^2}\right] \\
&= \frac{\tau_0^2+\sigma^2}{\tau_0^2\sigma^2}\left[\left(\mu - \frac{y\tau_0^2+\mu_0\sigma^2}{\tau_0^2+\sigma^2}\right)^2 + \text{junk}(y,\sigma^2,\mu_0,\tau_0^2)\right] \\
&= \frac{1}{\tau_1^2}(\mu-\mu_1)^2 + (\text{known junk})
\end{aligned}$$

so that μ|y $\sim$ N(μ_1, $\tau_1{}^2$), where

$$\begin{aligned}
\tau_1^2 &= \frac{\tau_0^2\sigma^2}{\tau_0^2+\sigma^2} = \frac{1}{1/\sigma^2 + 1/\tau_0^2} \\
\mu_1 &= \frac{y\tau_0^2+\mu_0\sigma^2}{\tau_0^2+\sigma^2} = \left(\frac{\tau_0^2}{\tau_0^2+\sigma^2}\right)y + \left(\frac{\sigma^2}{\tau_0^2+\sigma^2}\right)\mu_0
\end{aligned}$$

- The exponent of f(μ|y) looks like -1/2 times

$$\frac{(y-\mu)^2}{\sigma^2} + \frac{(\mu-\mu_0)^2}{\tau_0^2} = \frac{\tau_0^2+\sigma^2}{\tau_0^2\sigma^2}\left[\mu^2 - \frac{2y\mu\tau_0^2 + 2\mu\mu_0\sigma^2}{\tau_0^2+\sigma^2} + \frac{y^2\tau_0^2+\mu_0^2\sigma^2}{\tau_0^2+\sigma^2}\right]$$

$$= \frac{\tau_0^2+\sigma^2}{\tau_0^2\sigma^2}\left[\left(\mu - \frac{y\tau_0^2+\mu_0\sigma^2}{\tau_0^2+\sigma^2}\right)^2 + \text{junk}(y,\sigma^2,\mu_0,\tau_0^2)\right]$$

$$= \frac{1}{\tau_1^2}(\mu-\mu_1)^2 + \text{(known junk)}$$

so that $\mu|y \sim N(\mu_1, \tau_1^2)$, where

$$\tau_1^2 = \frac{\tau_0^2\sigma^2}{\tau_0^2+\sigma^2} = \frac{1}{1/\sigma^2 + 1/\tau_0^2}$$

$$\mu_1 = \frac{y\tau_0^2+\mu_0\sigma^2}{\tau_0^2+\sigma^2} = \left(\frac{\tau_0^2}{\tau_0^2+\sigma^2}\right)y + \left(\frac{\sigma^2}{\tau_0^2+\sigma^2}\right)\mu_0$$

n Observations $y_i \sim N(\mu,\sigma^2)$

- Since

$$p(y_1,\ldots,y_n|\mu) = N(y_1,\ldots,y_n|\mu,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2}$$

$$\propto N(\overline{y}|\mu,\sigma^2/n) \equiv \frac{1}{\sqrt{2\pi\sigma^2/n}} e^{-\frac{1}{2\sigma^2/n}(\overline{y}-\mu)^2}$$

we can apply the results for one observation

- $p(y_1, \ldots, y_n|\mu) \propto N(\overline{y}|\mu, \sigma_n^2)$, $\sigma_n^2 = \sigma^2/n$
- $p(\mu) = N(\mu|\mu_0, \tau_0^2)$
- $p(\mu|\text{data}) = N(\mu|\mu_n, \tau_n^2)$ where

$$\tau_n^2 = \frac{1}{1/\sigma_n^2 + 1/\tau_0^2} = \frac{1}{n/\sigma^2 + 1/\tau_0^2}$$

$$\mu_n = \frac{\overline{y}/\sigma_n^2 + \mu_0/\tau_0^2}{1/\sigma_n^2 + 1/\tau_0^2} = \left(\frac{\tau_0^2}{\tau_0^2 + \sigma^2/n}\right)\overline{y} + \left(\frac{\sigma^2/n}{\tau_0^2 + \sigma^2/n}\right)\mu_0$$

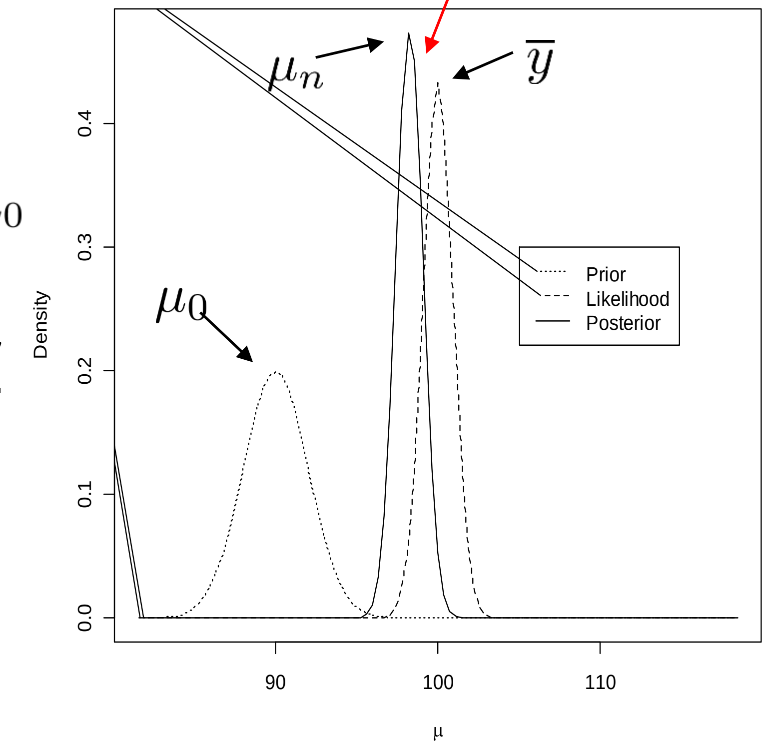
Normal Mean, Example

“shrinkage”: posterior between prior & likelihood

- Suppose we know σ=12, we look at n=169 IQ scores, and we find $\bar{y}$ = 100.
- We use as prior N(μ_0, τ_0^2) with μ_0=90, τ_0^2 = 4
- Shrinkage determined by

$$\mu_n = \left(\frac{\tau_0^2}{\tau_0^2 + \sigma^2/n}\right)\bar{y} + \left(\frac{\sigma^2/n}{\tau_0^2 + \sigma^2/n}\right)\mu_0$$

- $\dfrac{\tau_0^2}{\tau_0^2 + \sigma^2/n}$ is the ***reliability***
- n larger $\Rightarrow$ reliability larger $\Rightarrow$ less shrinkage



Minnesota Radon Example

- Emphasize Distribution Structure

$$\text{Level 2:} \quad \mu_j \overset{iid}{\sim} N(\mu_0, \tau_0^2)$$

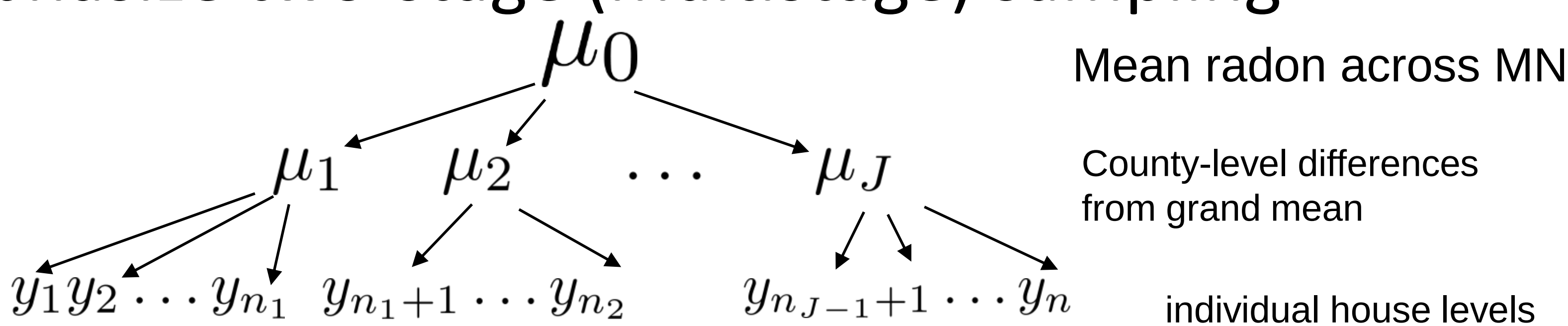
$$\text{Level 1:} \quad y_i \overset{indep}{\sim} N(\mu_{j[i]}, \sigma^2)$$

- Emphasize Bayesian point of view

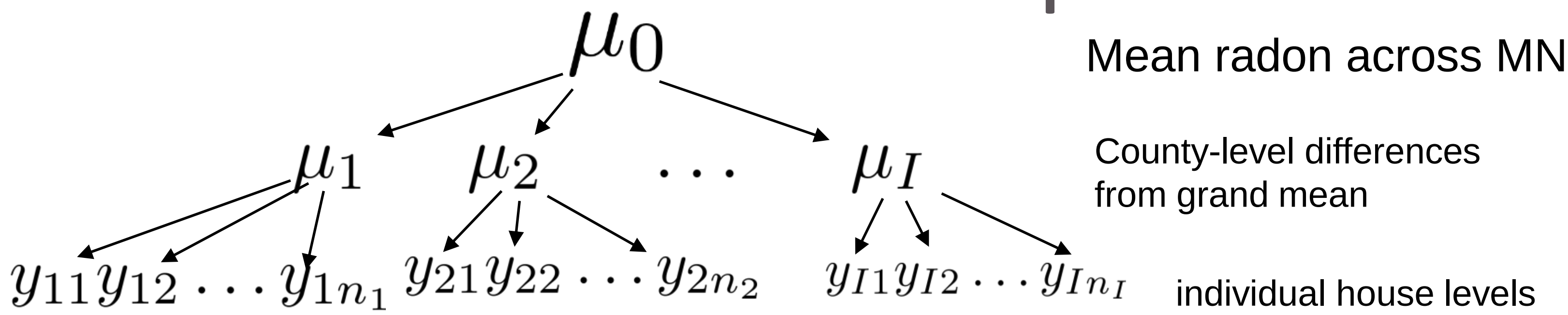
$$\text{Prior:} \quad \mu_j \overset{iid}{\sim} N(\mu_0, \tau_0^2)$$

$$\text{Likelihood:} \quad y_i \overset{indep}{\sim} N(\mu_{j[i]}, \sigma^2)$$

- Emphasize two-stage (multistage) sampling



Minnesota Radon Example



- In each county i with n_i houses, the posterior mean radon level $E[\mu_i | y_{i1}, \dots, y_{in_i}]$ will be

$$\mu_i^{post} = \left(\frac{\tau_0^2}{\tau_0^2 + \sigma^2/n_i}\right)\bar{y}_i + \left(\frac{\sigma^2/n_i}{\tau_0^2 + \sigma^2/n_i}\right)\mu_0$$

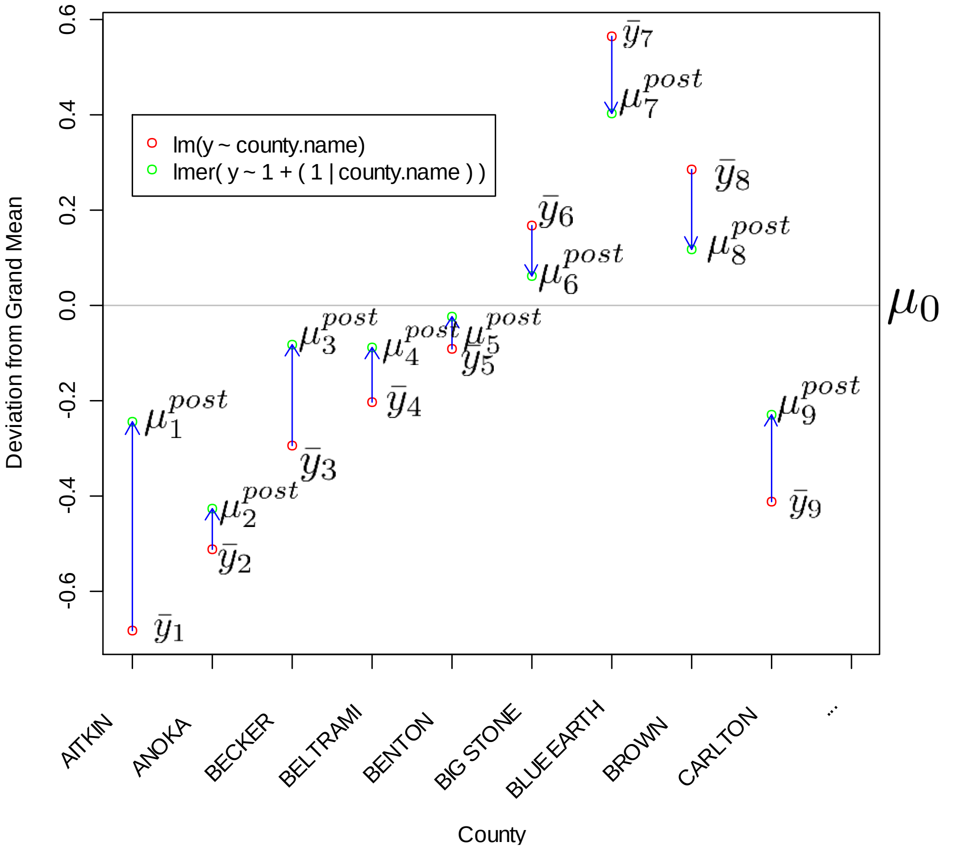
  - When n_i large, $\mu_i^{post} \approx \bar{y}_i$
  - When n_i small, $\mu_i^{post} \approx \mu_o$

Minnesota Radon Example

- In the figure, the grand mean is μ_0
- In each county i with n_i houses, posterior mean is

$$\mu_i^{post} = \left(\frac{\tau_0^2}{\tau_0^2 + \sigma^2/n_i}\right)\bar{y}_i + \left(\frac{\sigma^2/n_i}{\tau_0^2 + \sigma^2/n_i}\right)\mu_0$$

  - When n_i large, $\mu_i^{post} \approx \bar{y}_i$
  - When n_i small, $\mu_i^{post} \approx \mu_0$



MLM's and Shrinkage

- The random effect "estimates" that `lmer` produces with `ranef()` are a form of posterior means $E[\eta|\text{data}]$ for each η.
- The posterior means $E[\eta|\text{data}]$ are always shrunk toward the prior mean 0, so that the random effects α are always shrunk toward the corresponding fixed effects β.
- The Bayesian pov not only provides insights, but also
  - Novel ways to expand the multi-level model framework
  - Simulation-based methods of estimation (MCMC with `jags`, Hamiltonian MC with `stan`, etc.)

(we will take a look at some of this next week!)

Summary

- Today:
  - Shrinkage
  - Review of MLE
  - Crash course in Bayes
  - Normal-Normal Model & Shrinkage
  - MLM's and Shrinkage
- Project discussion
- After Thanksgiving:
  - A little practical Bayes / MCMC for multi-level models