36-463/663: Multilevel & Hierarchical Models

From Maximum Likelihood to Bayes

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

Outline

- 2016 Pre-election poll in Ohio
- Binomial and Bernoulli MLE
- Bayes' Rule
- Bayes for densities
- Bayesian inference
- 2016 Pre-election poll in Ohio

Ohio, 2016 Pre-Election Poll

- Donald Trump (R) running for election to the presidency against Hillary Clinton (D)
- In a Suffolk University Poll (Sept 12-14, 2016):
  - 401 of 500 voters expressed a preference for Trump or Clinton.
  - Of those 401: 208 prefer Donald Trump.
- 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 401 as a simple random sample.

Possible models for the data

- 401 individual Bernoulli coin flips, $x_i = 1$ for Trump, $x_i = 0$ for Clinton

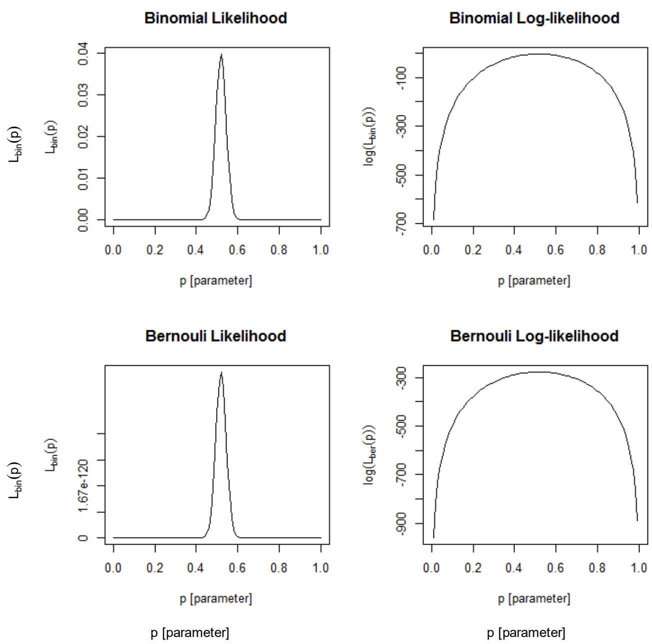
$$L_{ber}(p) = \prod_{i=1}^{401} p^{x_i}(1-p)^{1-x_i} = p^{208}(1-p)^{193}$$

- 401 trials, 208 "successes" (Trump voters)

$$L_{bin}(p) = \binom{401}{208} p^{208}(1-p)^{193}$$

- What matters for MLE and SE is <u>*shape*</u>, not <u>*size*</u>!

Binomial and Bernoulli Likelihoods



Proportionality and log-proportionality…

- f(θ) $\propto$ g(θ) ["f(θ) *<u>is proportional to</u>* g(θ)"] if

  f(θ) = cg(θ)
- Clearly L_{bin}(p) $\propto$ L_{ber}(p), with c = $\binom{401}{208}$
- For log-likelihoods we also write "$\propto$":

  LL_{bin}(p) $\propto$ LL_{ber}(p)

  because LL_{bin}(p) = LL_{ber}(p) + log$\binom{401}{208}$

  (weird, huh?)

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 = 208, n=401

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,

$$\hat{p} = \frac{k}{n} = \frac{208}{401} = 0.52$$

MLE: Standard Error & CI

- First we calculate the expected Information

$$
\begin{aligned}
I(p) &= E\left[-\frac{d^2}{dp^2}LL(p)\right] = E\left[-\frac{d}{dp}LL'(p)\right] \\
&= E\left[-\frac{d}{dp}\left(\frac{k}{p}-\frac{n-k}{1-p}\right)\right] = E\left[\frac{k}{p^2}+\frac{n-k}{(1-p)^2}\right] \\
&= \frac{np}{p^2}+\frac{n(1-p)}{(1-p)^2} = \frac{n}{p(1-p)}
\end{aligned}
$$

- and then

$$
\begin{aligned}
SE(p) &= 1/\sqrt{I(\hat{p})} = \sqrt{\hat{p}(1-\hat{p})/n} \\
&= \sqrt{0.52(1-0.52)/401} = 0.025
\end{aligned}
$$

- A CI for p is then (0.47,0.57), uncertain who wins!

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":

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

- See http://yudkowsky.net/rational/bayes for a ton of examples and geeky proselytizing.

Finding Terrorists

- According to http://wiki.answers.com/Q/How_many_people_fly_in_a_year , US airlines carry 561.9 million passengers per year
- According to http://www.rand.org/pubs/occasional_papers/2010/RAND_OP292.pdf , 42 people were indicted in the US for jihadists activities in 2009. About 2000 people are under surveillance in the UK (http://www.videojug.com/interview/the-structure-of-al-qaeda) so let's generously assume that about 10,000 are under surveillance in the US.
- Let's assume (again generously) that all 10,000 will try to fly once in the US in a year, carrying a detectable weapon.
- Now suppose TSA methods are 99.99% accurate:
  - P[red light | terrorist] = 0.9999 = P[green light | not terrorist]
- *What is P[terrorist | red light]? P[not terrorist | green]?*
- *How many travellers will be red-lighted?*

Terrorists and Bayes

- B = terrorist;
  B^c = not terrorist
  - $P(B) = 10{,}000/(561.9*10^6)$
    $= 1.78*10^{-15}$
- A = red light; A^c = green light
  - $P[A|B]=0.9999$
  - $P[A^c|B^c] = 0.9999$
- $P[A] = P[A\&B] + P[A\&B^c]$
  $= P[A|B]P[B] + P[A|B^c]P[B^c]$
  $= (0.9999)(1.78*10^{-15})$
  $+ (1-0.9999)(1-1.78*10^{-15})$
  $= 0.00012$

- $P[B|A] = P[A|B]P[B]/P[A]$
  $= (0.9999)(1.78*10^{-15}) /$
  $0.00012 = 1.5*10^{-11}$
- $P[B^c|A^c] =$
  $P[A^c|B^c]P[B^c]/P[A^c]$
  $= (0.9999)(1-1.78*10^{-15})) /$
  $(1-0.00012) \approx 1$
- $E[\#A] = P[A] * (561*10^6)$
  $= (0.00012)(561*10^6) =$
  66,188
- There better be other ways!

Conditional probability & conditional density

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

- Bayes' Theorem:

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

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

- Bayes' Theorem:

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

Bayes' Theorem for Data

- Bayes' Theorem

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

- 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)

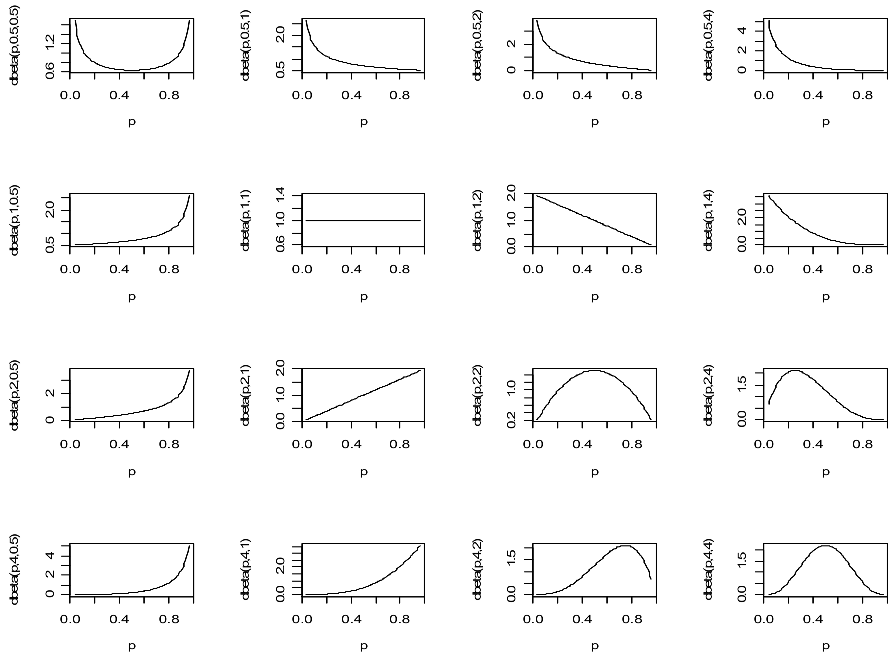
Back to 2016 Ohio pre-election poll

- The ***likelihood*** is the same as before:

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

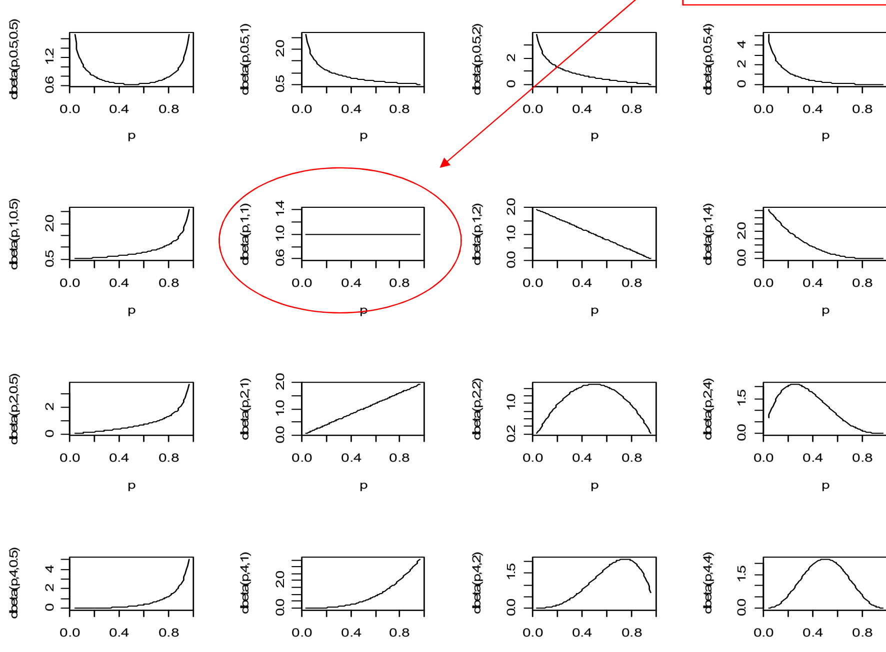
- We need a ***prior distribution***. One good choice is a *beta distribution*, with
  - Density $f(p|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$
  - Mean $E[p] = \frac{\alpha}{\alpha+\beta}$
  - Variance $\text{Var}(p) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$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

Some Beta Densities



Some Beta Densities

uniform distribution!



Choosing prior parameters…

- The *likelihood* is the same as before:

$$L(p) \propto p^k (1-p)^{n-k} = p^{208}(1-p)^{193}$$

- The *prior distribution* is a beta distribution

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

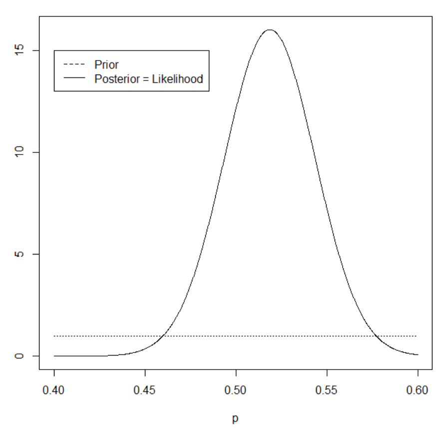
  - α = 1, β = 1 gives a uniform distribution – no preference for one p over another!
  - Suppose that in a previous poll, 942 prefer Trump and 1008 prefer Clinton. Could set α=942, β=1008

If α=1 and β=1…

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

$$f(p|\text{data}) \quad \propto \quad L(p) \times 1 \quad = \quad p^{208}(1-p)^{193}$$

- Since f(p|data)=L(p), posterior mode = MLE = 208/401 = 0.52
- Since f(p|data) is a beta with α=209, β=194, E[p|data] = 209/403



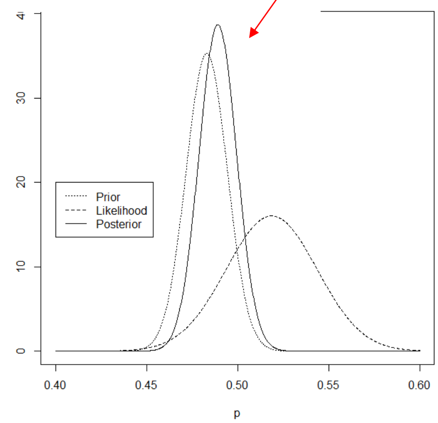
If α=942, β=1008…

“shrinkage”: posterior between prior & likelihood

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

$$f(p|\text{data}) \propto L(p) \times p^{941}(1-p)^{1007}$$
$$= p^{1149}(1-p)^{1201}$$

- Since f(p|data) = beta(p,1150,1202), E[p|data] = 1150/2352 = 0.489 vs MLE=0.519



Standard Errors (α=942, β=1008)

- Since

$$Var(p) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$
$$= \frac{(1150)(1202)}{(2352)^2(2353)} = 1.06 \times 10^{-4}$$

then $SE(\hat{p}) = \sqrt{1.06 \times 10^{-4}} = 0.0103$

(compare to SE=0.018 from MLE…)

- Approx 95% interval from $\hat{p} \pm 2SE$:

  (0.47, 0.51) … still can’t decide…

Alternative interval for p…

- Since we know the posterior distribution of p, we can calculate the 2.5%-ile and 97.5%-ile and get another 95% interval:

```
> nsim <- 10000
> p <- rbeta(nsim,1150,1202)
> quantile(p,c(0.025,.5,.975))
     2.5%       50%     97.5% 
0.4688515 0.4889114 0.5086120
```

- Gives almost the same 95% interval:

  (0.47, 0.51) ... *still* can't decide...

Summary

- For MLE
  - Need a function proportional to L(θ)
  - Calculate MLE by setting 0 = LL'(θ)
  - Calculate $SE = 1/\sqrt{I(\theta)}$ where I(θ) = E[-LL''(θ)]
- For Bayes
  - Need a function proportional to L(θ)
  - Need a prior distribution
  - Calculate (posterior) $\propto$ (likelihood)$\times$(prior)
  - Calculate posterior mean, SE
    - Use formula if you have one
    - Use simulation if you don't!

What we did…

- 2016 Pre-election poll in Ohio
- Binomial and Bernoulli MLE
- Bayes' Rule
- Bayes for densities
- Bayesian inference
- 2016 Pre-election poll in Ohio
- Summary!