Bayesian data analysis

Central component is modeling

Generative model is a story: observed data are realization from a probability dist.
God-like figure draws \( \theta \) from an urn, within \( \theta \) is the essence of how to draw \( y \).
"God created the world in 7 days and we haven't seen much of him since."
Vector of hyperparameters \( \phi \) is specified or itself modeled.

Inference, model checking, model improvement.

"People don't go around inviting you to their ex-wives." (why model improvement doesn't go into papers)

Bayes inference represented by matrix of posterior simulations

Post-processing

Inference for \( y \) and \( \theta \) (not necessary)

Model checking

Decision analysis

"The confidence interval can exclude \( \theta \), in which case I can submit it to a really good journal, or it can include \( \theta \), in which case I can look really hard and throw out some bad data points." (example of decision analysis)

PERC metabolism model

Goal: how much PERC metabolized at low doses

Want population distribution (since some people are more susceptible than others)

Experimental data: expose 6 volunteers to PERC for 4 hrs, then measure concentrations in blood and air for 2 hrs

4 compartment model of metabolism: well-perfused, poorly perfused, liver, fat

15 parameter per person.

Plot looks like mixture of Expo decays. Looks like Expo in log scale.

PERC released more slowly out of fat than out of well-perfused tissues.

Modeling

Sometimes model comes first, based on substantive considerations

Sometimes model chosen based on data collection (e.g., traditional statistics of survey and experiments - "design driven analysis")

Other times data come first (e.g., protein binding, data-fitting)

Usually a mix.
For ideas that did not work:

Fitting 4-compartment model directly to data. Nonlinear least squares.

\[
\sum_{j=1}^{2} \sum_{i=1}^{3} \left( \delta_{ij} - \mathcal{E}(y_{ij} | x_i, e_j) \right)^2 \frac{1}{\sigma_{ij}^2}
\]

Individuals repeat?

Does not allow for time variation of parameters. (Discussion of how we always know model is wrong)

Fitting separately to each person is unstable by 20 data pts. for 15 parameters.

Picky data and estimating parameters for the "standard man" defeats goal of learning about population distribution.

"We always have prior information, of course. The only debate is how to use the prior information."

Assisted model fit.

Set some parameters to fixed values from pharma literature, estimate the others.

couldn't fit the data well

using stick to adjust parameters, with resistance indicating strength of prior.

"It would be cool even for people who aren't blind. I would use it. Just because blind people can do it... I'm not too proud."

difficult to get fixed (prior) values for PEC to specific parameters we not much prior info.

"I'll get rid of the well-formed tissues, get rid of poorly perfused tissues. All you are is just liver and fat. Probably true for some of you."

what makes a good statistical analysis? other people can repeat what you did for other applications, and it becomes default. can do it with the click of a button.

"Statistics is said to be the science of defaults. One of our challenges is to defamiliarize things."

1-2 compartment model

1 compartment \[ y = A e^{-kt} \]

2 compartments \[ y = A e^{-kt} + B e^{-bt} \]

doesn't fit data well. Most PEC leaves right away. But some stays in body after a week or more.

w/ 2 compartments. can do slow vs. fast, but then fit is bad in the middle.

not realistic for low-dose extrapolation, which is our goal.

Problem with poor-fitting model is inability to extrapolate.
Expected more about data collection.

"We don't think about x. We're all like θ | y | x. What about x? It's because it's not modeled. y is the data, not x. [Someone leaves] Don't take it personally!"

Problem: Least squares gives bad fit for low doses because we have more measurements from the beginning, where there are still high doses. Change measurements are weighted too much.

However, we can't just downweight the undesirable points - it's ad hoc from a Bayesian perspective.

"How could that happen? It's least squares!" [said in the room, pleading voice]

"Forget genetic - baking. Someone needs to design the new financial scenarios tomorrow."

"That's not in the likelihood. That's not in the prior. Where is it? It's nowhere!"

(Why we can't assign weights as we please)

"We want to downweight that. But we can't, [jiggling up and down] case it's cheating!" How to solve? Good question.

> Survey sampling and causal inference are the same thing.

Missing data: respondent vs. non-respondent, treatment vs. control.

"In my head, I have these simulations of Don Rubin and Jennifer Hill running in a loop.

"What would Jennifer do? What would Don say?"

Here, can segment x axis, analogous to stratification. Subclassification in causal inference.

Simulation from prior distribution.

Get prior info on parameters from previous lit, trying to fit data within prior constraints, but not make prior info for same person.

Bayesian models with hierarchical prior distributions.

Example: prior dist. for a rate parameter in the metabolism, θj for person j.

$$\log \theta_j \sim N(\mu, \tau^2)$$ population dist.

$$\mu \sim N(\log 16, (\log 10)^2)$$

$$\tau^2 \sim \log 2$$

Large uncertainty, small variance.

Can learn about θ using data from several people. Use people to find out where θ is. Pasting.

16 1.6 16 160 1600

Only works with hierarchical prior because for any given person, they could be anywhere.
example: parameterizing so independence seems reasonable.

\[ A, B, C, D \]

\[
\Theta_1 = \frac{A}{A+B+C+D}, \quad \Theta_2 = \frac{B}{A+B+C+D}, \quad \ldots \quad \Theta_1 \Theta_2 + \Theta_3 + \Theta_4 = 1.
\]

**trick:**

\[
\Theta_1 = \frac{e^{y_1}}{e^{y_1} + e^{y_2} + e^{y_3} + e^{y_4}}, \quad \Theta_2 = \frac{e^{y_2}}{e^{y_1} + e^{y_2} + e^{y_3} + e^{y_4}}
\]

ten put indep. Normal priors on \( y_1, \ldots, y_4 \). Ends up only 7 parameters (8 is overpecified). induces a correlated prior on \( \Theta \)'s.

called soft max.

course website:  http://stat.columbia.edu/~gelman/arm/data.cour01/

"This guy comes to me and says, ‘I have prior information and data, and I’d like to combine them, and I heard Bayes is a good way to do that.’ Well, that’s as good as it gets! Normally you want to do Bayes but they won’t let you because they’re like [in rapid voice] ‘Ugh, it’s subjective, I’m not allowed to, it’s subjective.’ But here this guy is saying, ‘I have prior information and data and I want to combine them!’ I’m like, ‘I can do that! I was trained to do that!’"

\[ 9/7/12 \]

Want something that is consistent with model and data... but what if it’s like finding an apt in NYC?

\[
\text{SOL'}
\]

in our case, prior and data are probabilistic, so the boundaries are fuzzier.

"In statistics it’s enough for our results to be cool. In psychology they’re supposed to be correct. In economics they’re supposed to be correct and consistent with your ideology."

Back to example from last time: fit the model w/ Gibbs and Metropolis, re-run model under hypothetical lower cost exposures.

scatterplot of inferences for 6 individuals: each point is drawn from posterior dist.

uncertainty: within-person variability.

variance: between-person variability.

scale of uncertainty and variance are the same.

if variance were much higher than uncertainty (people far apart from each other), we would want to know why and would want to put it in the model. try to explain unexplained variance.
used model to predict future data

Sample from posterior distr, run model, add noise, rinse and repeat to get theoretical
quantities, plotted against actual data.

Prediction is poor for 0-15 min due to assumption of instantaneous mixing between
compartments, but otherwise data falls within 10-90th quantiles.

Putting it all together

(a) physiological model - a model where the parameters have names.

This is the best way to get prior info, combine info across diff. people.

\[ y = Ae^{-ax} + Be^{-bx} \]

is a "phenomenological model".

A and B don't have meaning. Difficult to choose prior for them.

(b) hierarchical model - allows people to vary

without common parameters, not enough data to estimate params separately
for each individual

but don't want complete pooling b/c tree is a lot of variance even among
healthy young male Dutch volunteers.

(c) prior info

(d) data

Need prior info on physiological params, data to learn about PDES in particular.

(e) Bayesian inference

Find params consistent with both prior and data, if possible

automatically includes uncertainty and variability.

(f) computation

(g) model checking - check reasonableness, consistency with prior, fit to current data,

fit to new data.

"we need all of these things, any f of them would not be enough."

With sriver parameterization, won't be much variance across people (example: use % of 

body mass in compartments instead of raw mass).

Bayesian inference automatically separates uncertainty and variancem

"sometimes classical statistics gives up, [says there are probabilities we can't estimate]."

Bayes never gives up. [..] so we're under more responsibility to check our models.

"you have to worry to falsify the model, if you love somebody, set them free."
Bayesian data analysis

"Inference" too narrow - doesn't include model-checking

"Statistics" too broad - includes design and data collection

Wanted to call it "statistics using conditional probability," but thought "wouldn't put the butts in the seats."

Bayesian: condition on data and prior information

What is Bayes?

- data + regularization (not our bad stuff)
- data + prior info (add good stuff)
- logical probabilistic reasoning ("not compelling, more an aspiration")

1. Probability and inference

Different approaches to statistics:

- Traditional likelihood

- Pure nonparametric (machine learning - no model)

- Robust (econometrics - how well does least squares work under general smoothness conditions, no generative model for data)

- Bayes (very model-dependent, must be willing to throw out model)

"A chicken is an egg's way of making another egg."

Inference

Goal is to use model to figure out what's wrong with the model and get a better model. Inferences from

1. Overview

Three steps to Bayesian data analysis:

- Set up probability model
- Inference
- Model checking

"Inference is the glamorous bug."

Then go back and improve the model.

2. General rationale for statistical inference

x (unmodeled data), like sample data, and y (modeled data)

Rubin's philosophy: all statistics is inference about missing data

- parameters are missing data - some observable, some inherently unobservable

In a world of prediction, what is the role of parameters?

- Some say: parameters don't exist, only past data and future data
- Parameters allow conditional independence, making the model simpler

Rubin's two questions:

1. What would you do if you had all the data?
2. What were you doing before you had any data? (i.e., what's your prior?)
1.3. Bayesian inference

\[ N(\mu | \mu_0, \sigma^2) \] is the Normal PDF.

"Xiao-Li thinks our notation is better."

1.4. Example: about spelling suggestions for "Koffee"

Prior probabilities of coffee, Kofi, koffee depend on reference set you choose, how much information you get in, more information = better prior.

Likelihood involves research too! To model the likelihood, may want to conduct experiments about who makes what typo, etc.

Point: both prior and likelihood involve modeling choices.

1.5. Probability as a measure of uncertainty

Frequency reference sets = Bayesian probability.

1.6 and 1.7. Examples to support his argument that probability is empirical, like height or weight.

"No, it's inside the exp, you can't touch that."

Modeling using conditional probabilities:

Early 20th century: have data, find what dist. the data look like, and then learn about reality.

Example: heights look like mixture of two Normals, 95% of boy births looks like \( \text{Bin}(n, p) \).

Late 20th century: regression modeling, conditional distributions.

21st century: hierarchical non-parametric modeling.

2a

Ch 2.


The basics:
- Data model (distinct from likelihood b/c different data models can have same likelihood)
- Prior density
- Posterior density

Why does likelihood come before prior?

Suggested answer: usually you're trying to explain some data. Gelman: "usually you're trying to explain some data, you statistician."

Likelihood tells you where you need to care about your prior.

But prior beliefs do determine design and data collection.

2.1. Estimating a probability from Binomial data

Example of bovine spongiform - Uniform prior seems inappropriate, if we observe 0 out of 75, is the posterior mean \( \frac{1}{77} \)?

Really, we'd want to give an interval estimate \( [0, \frac{3}{75}] \).

Dependence of \( \theta \) and \( n \): might need large \( n \) to estimate small \( \theta \).

2.2. Posterior as compromise between data and prior information

When is posterior variance higher than prior variance?

- bad luck. 70.1, 70.2, 69.9, 73.5, but observation brings the variance up.
- bad model. If prior is bad, you'll learn that from the data.

But even with bad model, posterior variance can still get smaller.
Research problem: when do models have "warning lights"?

"There are two types of models. Good models, if they don't fit, you get a large standard error. Bad models, if the model doesn't fit, it goes... 'no problem.' [laughter] 'I have a great compromise for you.' The model isn't able to tell you it's bad.' Need a name for this property.

Ex. 1: It can be viewed as mixture of Normals with unequal variances. Seeing an extreme observation is evidence of a high-variance Normal component, so it prior will inflate posterior variance when we have an extreme observation, but how much do we want? Need to decide if of it.

"As you know from teaching introductory statistics, 30 is infinity."

Ex. 2: Acceptance/confidence region obtained by inverting $\chi^2$ goodness-of-fit test. Consists of all goodness-of-fit not rejected by $\chi^2$ test.

---

This leads to a scary situation where right before the model rejects everything, it gives you a very narrow confidence band.

---

"If rejects everything, okay? The $\chi^2$ acceptance region is the goddamn empty set."

---

So that's a bad thing.

Problem here is using the $\chi^2$ test to get uncertainty statements and also to check fit of model.

"Your computations are your inference." - Rubin.

If model is good but computations are wrong, the model you actually fit is the one you computed. Conversely, "computations could save your butt." If model has bad mode but computer doesn't find it, it's as if it isn't there.

2.3 Summarizing posterior inference

Central vs. shortest intervals
- Same for symmetric distribution. For one-sided dist., central interval is silly because it excludes $0$, the most likely value. In general, shortest interval is drawn toward the region with highest posterior density.
- Central interval invariant to transformations.
- Shortest interval harder to compute. Empirical shortest interval is a little bit biased and variable.

"because it's the winner in a noisy competition." Noisier than central interval.
Goals of posterior intervals:
- Sometimes want nested intervals: 50% inside 95% inside 95%. Won’t always happen with multimodal distributions, so has to be refinement of procedure.

Two purposes of interval estimation:
- Accept what’s inside.
- Reject what’s outside.

Accept what’s inside because it is a positive summary of where params are likely to be, but don’t reject 0 just bc it’s outside the interval.

Finding shortest posterior intervals: combination of bootstrapping and smoothing works best, but uncontrolled tradeoffs - yet more simulations if you can!

2.4. Informative prior distributions

Interpretations:
- Population - “the urn full of trials”
- State of knowledge - represents our assumptions
- Software defaults (statistician in a box)

“Conceptually there is a true prior.” True prior is not true subjective belief, but rather the distribution of all possible θs for which you would apply a given statistical procedure; it’s “behavioral rather than cognitive.”

“We can imagine there’s some sort of a base.”

Bin model: Is Beta(α, β) prior equivalent to α+β claim points, or α+β-2?

Probability of a girl birth:
P(girl birth) = 0.485 in general population. Compensates for the fact that at every age, boys die more than girls. 55% boys at embryo stage, then equal sex ratio at age 20. "Which I’m told is convenient," “and then eventually you’ve got grandma by herself...”

Variation in sex ratio completely explainable by binomial variance.

Constructing a prior distribution:
Believe p should be between 0.4 and 0.5.

"But that’s a soft upper bound. We prefer soft power. We’re like Bill Clinton; we’re not like George W. Bush."

Beta(α, β) prior with mean 0.485 and SD 0.01 turns out to correspond to α+β = 2500 data points, whereas n = 980, so prior is very strong."
P(girl) for beautiful vs. ugly parents

Data: difference in P(girl) estimated from 3000 respondents

$0.03 \pm 0.02$ (selected comparison)

$0.047 \pm 0.043$ (linear regression)

Prior: $N(0, 0.003^2)$.

Two reasons for prior zero mean:
1. Don't want to bias results.
2. No prior reason to think beautiful parents have more girls.

Prior variance loosely based on previous research ("we need a probabilistic model of the scientific process").

Equivalent sample size:

Consider survey with $n$ parents, compare sex ratio of prettest $\delta_1$ to ugliest $\delta_2$.

$$\delta = \sum \frac{(x_i - \bar{x})^2}{x_i - \bar{x}}$$

Further support comparisons get more weight.

Shouldn't compare top half and bottom half because points in middle give noise; no leverage.

More statistically efficient to do upper 3rd/lower 3rd; get rid of noise.

$$SE = \sqrt{\frac{3}{5} \cdot \frac{\mu^2}{n} + \frac{3}{5} \cdot \frac{\sigma^2}{\bar{x}}} = \frac{0.5 \cdot \sigma}{\bar{x}} \implies n = 166,000.$$  

So survey with 166,000 people would be equally weighted with prior.

2.6. Normal mean, known variance

$$\Theta \sim N(\mu_0, \sigma^2), \quad \bar{y}_n \sim N(\Theta, \sigma^2/n)$$

$$\Theta \sim N \left( \frac{\bar{x}_n \cdot \sigma^2_n + \sigma^2_n \cdot \bar{y}_n}{\sigma^2_n + \sigma^2_n}, \quad \frac{1}{\sigma^2_n + \sigma^2_n} \right)$$

2.7. Other standard single-parameter models

Poison vs. exponential - almost never do $y \sim \text{Poisson} (\Theta)$, instead do $y \sim \text{Poisson} (x; \Theta)$

2.8. Estimating kidney cancer rates

Low-population counties shank more to prior mean.

Pairable of the two exams:

Option 1: 100 questions

Option 2: 1 question, score is 0 or 100.

When hiring, go with option 1: that's like the Bayes estimate.

Low-population counties "don't get a chance to show their stuff." We don't 'have' (believe) them.

Estimating prior dist. from data:

Prior is population distribution - represents true kidney cancer death rates. Data more variable because noisy.

For any given county, prior is supposed to represent my knowledge of the rate in that county.

Match mean of rates to mean of black curve.

Variance of rates to variance of black curve + average Poisson sampling variance.
Connecting cancer rate example to Bayesian themes

- Model \( y_j \sim \text{Pois}(100y_j \theta_j) \)

  Assumes independence between individuals
  Independence between counties, conditioned on \( \theta_j \)
  "you can't die twice of something it's double jeopardy."

  Likelihood based on Poisson approximation to Binomial

  Poisson actually has fewer assumptions because \( \theta_j \)'s for Binomial have to be the same for people.

  For Poisson, can have different probabilities and integrate \( \theta_j \) as average.

- Informative Gamma prior: equivalent sample size (in each county) of 20, prior mean of \( \frac{20}{430000} \) people has posterior mean falling between prior and data

- One prior, many different datasets (each county as its own dataset)

- Data-modified prior distribution

Comparing Bayesian inferences to other estimates

- Do parameter estimates look reasonable?
- Do predictions look reasonable?

- Artifacts:
  Raw rates lead to artifacts - highest rates all in low-pop countries.
  Posterior means have artifacts too - highest rates still in high-pop. counties.

Cross-validation

- Take 1st half of 80's. Use 3 estimators (raw rates, Bayes estimate, prior mean) to predict 2nd half of 80's.

  Bayes does best.

- What if data not broken down by year?
  \( n = 200,000, y = 30 \), artificially break up into \( y_j, y_j \) with \( y_j \sim \text{Bin}(30, \frac{1}{5}) \)

  If Poisson model is true, this is exactly what it assumes! So can prove theorem that Bayes does best under mean squared error, assuming model is true and using this cross-validation.

  Bayesian inference is implicit cross-validation.

  Raw rate works best if \( y_j, y_j \) always half and half, but that's not what Poisson model assumes.

2.9. Noninformative prior distributions

"Another way of saying a prior is proper is that it's a generative model; you can use it to generate data."

An improper prior is not a generative model, either because it

1. doesn't integrate to 1, or because it

2. depends on the data, need data to set the prior, so can't generate data with it.

In Bayes, you can set prior based on the functional form of the data, but not the realization of the data themselves.

"Noninformative" depends on scale:

\( p(\theta) = 1 \)

1. Uniform

\( p(\log \theta) = 1 \)

1. Uniform (improper)

\( \text{Logistic} \)

\( \text{Beta}(0, 0) \)
2.10. Weakly informative priors

WIP: prior that contains some info (is proper) but less than you really have.

Recall the idea of the unknown true prior $p(\theta) = N(\theta | \mu_0, \sigma_0^2)$.

Assumed prior (subjective prior) is $p(\theta) = N(\theta | \mu_1, \sigma_1^2)$.

WIP inflates variance by factor of $k^2$, $k > 1$.

Need to formalize the idea of using less info than is available.

Ch 3.

3. Introduction to multivariate models

3.1. Averaging over nuisance parameters

"Suppose there’s someone you want to get to know better, but you have to talk to all her friends too.

They’re like the nuisance parameters."

$p(\theta_1 | y) p(\theta_2 | \theta_1, y)$ “First I have to learn $\theta_1$. Then I can go in for the kill.”

Def. of nuisance depends on context. Example: producer/user of a scale cares about mean, manufacturer cares about variance.

3.2. Normal data with noninformative prior

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

Prior $p(\mu, \sigma^2) \propto \sigma^{-2}$

Equivalent to $p(\mu, \sigma) \propto \sigma^{-1}$, $p(\mu, \log \sigma) \propto 1$

Integrate out $\mu$ from joint posterior density: $\sigma^2 | y \sim Inv-\chi^2(n+1, s^2)$

$\mu | (\sigma^2, y) \sim N(\bar{y}, \frac{s^2}{n})$.

3.3. Normal data with conjugate prior

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

Conjugate family

$\sigma^2 \sim Inv-\chi^2$

$\mu | \sigma^2 \sim$ Normal with variance proportional to $\sigma^2$

Use conjugate priors for understanding (prior as extra data) and computation reasons.
3.4. Multinomial model

\[ y_i - x_i \sim \text{Mult} (n, \theta_1, \ldots, \theta_k) \]

unknown probabilities \( \theta_1, \ldots, \theta_k \) constrained to unit simplex, so can't be independent.

Noninformative priors:

\( \theta_1, \ldots, \theta_k \sim \text{Dir} (1, \ldots, 1) \) uniform on \( \Theta \)'s

\( \sim \text{Dir} (0, \ldots, 0) \) uniform on log \( \Theta \)'s

\( \sim \text{Dir} (\frac{1}{k}, \ldots, \frac{1}{k}) \)

Example of joint prior on regression coefficients

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon \quad \text{each group has 4 coefficients} \]

"why is it Normal? Because that's the only continuous multinomial distribution we have."

"oh, we have the multinomial. Uh, as if that's a different distribution."

3.6. Multivariate Normal with unknown mean and variance.

good prior for \( \Sigma \) matrix is hard ble of dimensionality + positive-definiteness constraint.

"This is a paper we have. That's making the rounds of getting rejected."

3.7 Example: analysis of biasing experiment

\[ y_{1i} \sim \text{Bin}(n_i, \theta_i), \ i = 1, \ldots, 4 \]

assumptions: independence within and between groups, conditional on \( \theta_i \)

\( \text{Same } \theta_i \) for each rat within a group.

\( \text{If rats are fighting for food, underdispersion. If contagion, overdispersion.} \)

"In cage 1, they all die, and then in cage 2 they all hear about it, and they're like, 'Don't eat that shit, man.'"

\( \theta_i = \text{logit}^{-1}(\alpha + \beta x_i), i = 1, \ldots, 4 \)

\( \text{p}(\alpha, \beta) = 1 \)

assumptions: monotonicity of dose-response.

\( \text{in particular, follows logistic curve} \)

\( \theta_i \text{ deterministic function of } \alpha, \beta \)

\( \theta_i \text{ between 0 and 1} \)

\( \text{no group-level error} \)

\( \alpha \text{ and } \beta \text{ could be anything} \)

information in logistic regression is \( (\text{number of successes}) \times (\text{probability of success}) \). At the extremes, the probability is too small, and right in the middle, the news is nothing. So the most information is in the intermediate points.


the others are "ties" or "obvious".

posterior dist. for \( \beta \) is asymmetric: \( \beta \) could be really large, but we're sure of the sign (positive).

GLM vs. Bayes: GLM overfits but gets a really good fit in this case "by luck".
\( \alpha \) and \( \beta \) are positively correlated a posteriori because the data happen to be such that the curve points around a negative value of \( x \). Since \( \alpha \) is the intercept (\( x = 0 \)) and higher \( \beta \) means higher intercept, we get positive corr. but the pivot pt. is close to 0, so corr. is weak.

Should parameterize so intercept is meaningful: "If you height is 0, you're not going to make any money."

L) against arbitrary histogram smoothing:

"I disagree with those people. Those straw people."

"The best histograms have the seeds of their own destruction."

Difficulties with ratios as 00's when denominator can be positive or negative.

\[ LD50 = \frac{-x}{\beta} \] moving from \( \beta > 0 \) to \( \beta < 0 \) doesn't actually correspond to a logical sequence of models.

This is not a continued

logical progression.

"they're two completely different models that happen to be connected by a common parameterization."

Incremental cost-effectiveness ratio

old treatment: cost \( C_1 \), efficacy \( E_1 \)

new treatment: cost \( C_2 \), efficacy \( E_2 \)

\[ \frac{C_2 - C_1}{E_2 - E_1} \] incremental cost-effectiveness ratio

Some cost-effectiveness ratios are gradients I and III, with completely different interpretations, meaningless even.

Instrumental variables

\( IV \) estimate: \( \frac{\text{cost of regression of } y \text{ on } I}{\text{cost of regression of } z \text{ on } I} \)

need to restrict demand to be positive and admit you're proceeding under that assumption.

Feiler-Crespy problem

\( x \sim x \sim N(0, \sigma^2) \), \( y \sim y \sim N(\theta, \sigma^2) \), hard to get internal estimates for \( \theta_0 \) with any variance.

but what does \( \frac{\partial \theta_0}{\partial x} \) mean if \( \partial x \) can be positive or negative?

folk theorem of computational stats: when you have computational problems, often there's a problem with your model.

Pinocho principle: a model created solely for computational reasons can take on a life of its own.
work with log density instead of actual density

"raise your hand if you haven't heard this principle before. Hey, you heard a new principle!"

computing posterior density on a grid

compute unnormalized log density on grid

log-post

rescale and exponentiate

$a \exp (\log \text{post} - \text{max} (\log \text{post}))$

normalize to sum to 1

$a / \text{sum}(a)$

Chicken brain data

"I called them out, they told me to fuck off, basically."

Sham treatment can be thrown out after it's determined that sham has no effect

Est. 1 $y_{\text{exposed}} - y_{0_{\text{exposed}}} = (y_{0_{\text{ sham}}} - y_{0_{\text{ sham}}})$

Est. 2 $y_{\text{exposed}} = y_{0_{\text{exposed}}}$

Est. 3 $y_{i_{\text{c}}} - y_{0_{\text{c}}} = \lambda (y_{0_{\text{c}}} - y_{0_{\text{ Sham}}})$

Ch. 4.

4.1 Large-sample inference and frequency properties

Normal approximations to posterior dist.

Large-sample theory

Counterexamples to the theorems - "it's not a theorem until you have counterexamples"

Frequency evaluation of Bayesian inferences

4.1 Normal approximation

$\log p(\theta | y) = \log p(\theta | y) - \frac{1}{2} (e^\theta - \bar{e}^\theta)^T \left[ \frac{d^2}{d\theta^2} \log p(\theta | y) \right] (e^\theta - \bar{e}^\theta) + \cdots$

Normal centered at $\hat{\theta}$ with inverse variance $\uparrow$

4.2 Large-sample theory

claim $y_i \sim y_0 \sim f(y)$, modeled as $p(y | \theta)$, which may not contain $f$

as $n \to \infty$:

* if $\theta$ discrete and finite, $p(\theta | y) \to$ Dirac mass at $\theta$ or model closest to $f$ in K-L distance
* if $\theta$ continuous and compact set, $p(\theta | y) \to$ point mass at $\theta$ or closest model
* under some conditions, $p(\theta | y)$ approaches Normal dist.

4.3 Counterexamples to the theorems

unidentified parameters $(y = \theta, r\theta_2)$

model changing with sample size

unbounded likelihood, e.g. mixture models

solution is to bound the variance ratio or the actual variance

there's a sense in which the mixture model includes a class of models we're not interested in,

as with the cost-effectiveness ratio

improper posteriors

constrained priors

boundary estimates

tails - the further out in the tails, the more data you need to fit it well.
4.4 Frequentist evaluations of Bayesian inferences

Example where unbiasedness does not make sense:

\( \theta = \text{height of a woman (inches)} \),
\( y = \text{height of her adult daughter} \)

\[ \hat{\theta}(y) = y \text{ is unbiased. bias} = -0.5(\theta - 64) \]

\[ \hat{\theta}(y) = 64 + 0.5(y - 64) \text{ is really biased. bias} = -0.75(\theta - 64) \]

"This is big-ass bias."

Unbiased estimate is \( \hat{\theta}(y) = 64 + 2(y - 64) \): need to anti-shrink!

Classical statistician says \( \theta \) is not a parameter; it's a "predictive quantity" but it has a distribution.

In frequentist inference, don't condition on \( Z \); bias of prediction is \( E(2\theta) - E(2\hat{\theta}) \).

4b

Ch 5

5. Hierarchical models

No harm example: want more precise estimate of \( P(\text{tumor}) \) at dose of 6.

5.1 Constructing a parameterized prior dist.

\( y_j \sim \text{Bin}(n_j, \theta_j), j = 1, \ldots, 71, \quad \theta_j \sim \text{Beta}(\alpha, \beta) \)

5.2 Exchangeability

\( \theta_1, \ldots, \theta_j \) exchangeable if \( p(\theta_1, \ldots, \theta_j) \) symmetric.

Non-exchangeable models:

- Time effects

\[ \theta_j = \logit^{-1}(\alpha + \beta x_j + e_j), \quad x_j = \text{date of study} j \]

No point to write \( \alpha + \beta x_j + e_j \); index shouldn't carry information.

"Having the index carry information is like having the lamp hanging from the wire ... you don't want to

Use in electrical connection as a mechanical connection."

- Having some experiments are in different settings

- Markov chain

"Exchangeability is a fiction not just of reality, but of the information you have."

Say we know the 70 previous experiments are different from the new one.

Exchangeability violated: use above model, with \( x_j \) the indicator of the new experiment.

Old: \( \alpha + \varepsilon_1, \alpha + \varepsilon_2, \ldots, \alpha + \varepsilon_{71} \)

New: \( \alpha + \beta + \varepsilon_{71} \)

Not much info about \( \beta \) in the likelihood.

If prior on \( \beta \) is noninformative, then not using the others to help with \( \varepsilon_{71} \). If noninformative, then will pool.

Group-level predictors - many non-exchangeable models fall into this category.

9/26/12
5.3 Fully Bayesian analysis of conjugate hierarchical models

\[
p(\psi | y) \propto p(\psi) p(\theta | \psi) p(y | \theta, \psi)
\]

marginal posterior of hyperparameters

\[
p(\psi | y) = \int p(\psi | \theta, y) \, d\theta
\]

\[
\propto p(\psi) \int p(\theta | \psi) p(y | \theta, \psi) \, d\theta
\]

If we can do this integral, then just need to compute \(p(\psi | y)\) on grid of \(\psi\),

single \(\theta^5\) from grid and then single \(\theta^5\) from \(p(\theta^5 | \theta, y)\).

Rat tumor model: 
Model: \(y_j \sim \text{Bin}(n_j, \theta_j)\)

\[
\theta_j \sim \text{Beta}(\alpha, \beta)
\]

Assumptions: independence conditional on \(\Theta\), exchangeability

\(\Theta\) could have come from a Beta! (improper from assumptions - not bimodal)

is exchangeability problematic? it's just inference conditional on no information distinguishing the \(\theta_j\), if you have info, exchangeability just ignores the info.

prior on \((\alpha, \beta)\).

Uniform prior on \((\log(\frac{\alpha}{\beta}), \log(\alpha + \beta))\) doesn't work (improper posterior)

Unit over large box doesn't work either - mass pulled toward edges of box

Instead use \(p\left(\log\left(\frac{\alpha}{\beta}\right), \log(\alpha + \beta)\right) \propto \alpha \beta (\alpha \beta)^{-1/2}\)

go from joint posterior \(p(\Theta, \alpha, \beta | y)\) to marginal posterior \(p(\alpha, \beta | y)\) so can compute on 2D grid

need to work hard for good parametrization so that we can come up with a good prior

\(\theta\) back to time effects? should measure \(\bar{y}_j\) in years/decades relative to average time of experiment

- otherwise integrals "will be the log probability of death in the year that class was born"

- posterior mean of \(\bar{y}\) as function of observed rate

- no pooling is 45° line, complete pooling is horizontal line, our model is a compromise.

5.4 Exchangeable parameters from Normal model

Model: \(y_j \sim \mathcal{N}(\theta_j, \sigma^2_j)\)

\[
\theta_j \sim \mathcal{N}(\mu, \tau^2)
\]

Second statement is more restrictive because assumes \(\theta_j\) can be modeled as Normal -

no reason that this should be true.

get conditional posterior \(\Theta | \mu, \tau, y\), average over marginal posterior of \(\mu, \tau\).

5.5 Example: parallel experiments in 8 schools

8 schools, 60 students in each, randomized to earn coaching or no coaching

Separate estimates unsatisfactory, but if experiments had been replicated, would not expect school A to get

28 again.

Pooled estimate also feels unsatisfactory - we would want to send kid to school A over school C in

case for parallel pooling.
and parameters, half from N(1,1) and half from N(-1,1). If it were a mixture of ind. components, the T in each mixture should be a r.v., not fixed.

![Graphs showing distributions and expected values](image)

bootstrap doesn't make sense because T in each mixture component is pretty much fixed.

Instead, get standard errors by running things twice to side a little bit.

"The Ames said you guys had a lot of bad models. They blamed themselves. And I blame them too."

"It's not like the door is open. It's like where is the door? I can't even see the wall. Maybe this describes most of my research."

Sweet spot for statistics:

- lot of data - don't need fancy methods.
- no data - can't do anything.
- in between - sweet spot.

Theoretical models continued:

5 schools example:

data consistent with $T \leq 0$

Robin's approach: draw from posterior of $T$ and then average over the different draws:

- what if model were applied to 5 unrelated objects?
- partial pooling wouldn't happen because $T$ would be large.
- if exchangeability is inappropriate, don't do it.
- can be mixture model ($7$ schools + $1$ grand-diagnostic observation, etc.)
- dependent prior on $\mu$, $T$ would shrink $\mu$ toward $0$ for low $T$

5.6 Theoretical modeling applied to a meta-analysis

transformation to log odds, then apply 5 schools model

3 steps of inference:

- null effect
- effect is a single study (existing or new)
- precision for a new person (in existing or new study)
Why are doses more effective in experiments than real world?

1. Researchers chose medication that is most likely to benefit.
   less expensive to multiply θ by 2 ran a by 4.
2. Signal dose is higher in experiment because of better compliance, controlled conditions, etc.
3. Statistical significance filter

5.7 WIP for hierarchical variance parameter

log τ ~ Uniform(-\infty, \infty) leads to improper posterior

"you see this sucker - I should sing 'this sucker' - you're gonna die."

τ ~ Uniform(0, \infty)

τ ~ ICam(10^{-3}, 10^{-3}) \left\{ \begin{array}{l}
\text{problems}
\end{array} \right.

τ ~ Cauchy(0, 10)

ICam(1,1) prior cuts off (dead zone near 0), and ICam(10^{-3}, 10^{-3}) is much too strong.

Unit prior with 3 schools doesn't cut off any tail. ("we want to cut off the tail, we're not dogs here.")

must shrink a lot; and moreover, posterior predictive has huge variance

implications:
1. to make inference from 3 schools, you only have to use prior info.
2. prior info, even a little bit, can help.

5b

Ch. 6

6. Model checking

Why wouldn't we put all our prior info into the model?

1. Sometimes want to express prior info in probabilistic model
2. Could add too many parameters, make model more complicated
3. Concerns about human psychology - don't want to fool ourselves.
   example: priors for ESP experiments.
4. Statistics as science of defaults. Might want something that everyone can use.

5. not wanting to cheat. inference is part of a larger process.

Here is a whole class of problems where you don't want to put down what you believe the most.

- cards data - need to preserve anonymity
- sports tournament - can't award victory based on prior belief
- assigning grades - presumably we want to assign grades based on actual understanding of material, and final exam is a noisy measure, so should we add the prior score in to the model?

What these examples have in common is that the inference is embedded within a larger societal or scientific process, which necessitates different rules. We care about fairness, replicability of results, etc.
"A Bayesian wants everyone else to be non-Bayesian – if every wisnew used Bayesian inference to assign grades, my lifetime funding as someone who wants to hire is tremendously complicated.

if you have to save the world or humanity will be extinguished, use your prior. but normally it's embedded in a process, entangled with social goals, so there may be reasons to omit prior info.

"if I'm only an experimenter to save the world, I better use my prior."

model checking is:

- comparing estimates and predictions to substantive knowledge
- comparing predictions to observed data (your original data can test your model)
- graphical and numerical tests

6.1 Model checking in applied Bayesian statistics

With great power comes great responsibility

example: with only 3 schools, should you just give up? can't estimate group-level variance

Sensitivity analysis

not just sensitivity to the prior, but sensitivity to the likelihood (which enters n times!)

Gelman on sensitivity analysis: "I just never get around to it."

- paper on the boxer, wrestler, and coin flip.

  two r.v.s: indicator of heads in a coin toss and indicator of who wins in a boxer vs wrestler fight to the death.

- different priors do change your inference.

if you were fully committed to sensitivity analysis, you would consider models

\[ p_1(\theta), p_2(\theta), p_3(\theta), \ldots \] which could be expressed as \( p(\theta | Y) \),

but then you put a prior on \( p(Y) \), and you're back at a Bayesian model again!

in 3 schools, for example, robins put prior on \( \theta \) and averaged over it.

- there's no way around this because it doesn't make sense to compare across values of \( Y \) without a probability dist. on \( Y \) - otherwise how would you know if the sensitivity is awesome or not?

All models are false: "in the grand scheme of things, the kitten is already dead."

Bayes vs. Super-Bayes

real Bayes: build the models, fit the models, check the models, expand the model, refine, refine, and repeat.

Super-Bayes: all the models are already there! start with mixture of all possible models.

"but you don't have to be George Cartman to know there's always some model that's not in your super-model."

how far can we take super-Bayes?

6.2 Do the inferences from the model make sense?

8-12 9-10 6-7 5-6 when you find a problem with the model, don't just tweak prior, change the whole model (incorporate intentions, tones, etc.)

Choosing a discrepancy statistic: choose what you care about
6.3 Posterior predictive checking

Compare observed data to replications simulated from model

\[ \begin{align*}
\theta &\rightarrow y \\
&\rightarrow y_{\text{rep}} \end{align*} \] Compare

Replicating and p-values

Classical p-value: \( P(T(y_{\text{rep}}) \geq T(y) | y, \theta) \) if test statistic is positive, doesn't depend on \( \theta \)

Bayesian posterior p-value

\( P(T(y_{\text{rep}}) \geq T(y) | y) = \int \text{p-value}(y_{\text{rep}} | y) \, p(\theta | y) \, d\theta \)

Compute by simulation

Idea: generate 19 replicated datasets and put real dataset in random place. Can you identify the real one? What about 99 false datasets? How's a way to get a p-value based on the entire posterior distribution?

Picture of 20 replications under normal model - many of the datasets don't look normal at all.

6.4 Graphical posterior predictive checks

Today: picture forces us to "notice what we never expected to see."

Exploratory and confirmatory data analysis are the same thing - they're both about checking model for problems.

EDA and p-values being in the same chapter: They're not inference.

"Inference is not the inverse of a hypothesis test."

Definition of inference vs. checking: you have a model, you believe it, you have a willing suspension of disbelief; you do inference; then you do model-checking / hypothesis tests.

EDA is a type of model check - same as a hypothesis test; difference is that hypothesis test uses p-value as summary, and with EDA it's graphical, checked against a vague sense of expectation (Tukey quote): "All graphics are model checks."

Residual plot is posterior predictive check where you don't have to draw 20 reps because you already know what reps are supposed to look like. Res plot is a test setup with known symmetry properties.

Loa

Bayesian additive regression trees (BART)

Nonparametric fitting \( E(y | x) = \frac{1}{k} \sum_{j=1}^{k} j_{\lambda}(x) \)

Studies the space into regions, can get step functions (very flexible) which are smoothed via uncertainty/averages

Jennifer likes BART: We can model all the x's and y's in their high-dimensional glory instead of collapsing into a p-value.

"I hope she's getting my work in her classes."

On priors: "They don't have to be weakly informative, they can just be shitty."
Why to relax a fully informative prior?

1. Could be wrong, even after adjusting for the fact you could be wrong.
2. Don't want to cheat.
3. Better generalization to other contexts.

Can start noninformative and then tighten up, or start upright and then relax.

Introducing p for non-cumulative, "noninformative" prior is actually too informative b/c pulls things to the right.

**Dog models**

\[ y_{ij}^* = \begin{cases} 
1 & \text{if } j \text{ gets sick at time } t \\
0 & \text{if } j \text{ avoids sick at time } t 
\end{cases} \quad t = 0, 1, 2, \ldots \]

**Model 1:**

\[ P(y_{ij}^* = 1) = \beta^t, \quad \beta \in [0, 1]. \]

**Model 2:**

\[ P(y_{ij}^* = 1) = \logit^{-1}(\alpha + \beta t), \quad \beta \text{ should be negative, or positive.} \]

**Model 3:**

\[ P(y_{ij}^* = 1) = A \ (# prev. sheets) B \ (# prev. avoidance) \]

\[ A \text{ and } B \text{ between } 0 \text{ and } 1. \text{ Need to constrain when fitting, e.g., get estimates } > 1. \]

\[ A \approx .9, B \approx .8. \text{ Days tend to learn more from avoidance than sheets, plus there's the psychological impact of learning the dog is smart if it avoids right away.} \]

Dogs-specific effects missing from all three - would prefer hierarchical model.

**Model checking continued.**

6.6 Connections to classical testing

**p-values and u-values**

\[ p-value = P(T(y) > T(y_0, y|y_0)) \text{ is a random variable, function of } y \text{ (which has prior predictive dist.)} \]

\[ u-value \text{ same function } u(y) \text{ that has uniform distribution, averaging over prior predictive dist.)} \]

**Poisson predictive p-value:** Typically concentrated around .5, so is conservative, won't reject as often.

**Example:**

\[ pPP \text{ is much near } 0.5 \text{ and this makes sense.} \]

**Data:** \[ y \sim N(\theta, 1) \]

**Prior:** \[ \theta \sim N(0, \sigma^2) \], \[ \sigma = 100 \]

**Test stat:** \[ T(y) = y \text{ (sample mean)} \]

**Poisson predictive p-value:** \[ pPP = \Pr \left( \frac{y}{y^*} > 1 \right) \]

\[ \text{is essentially never reject, if prior were strong, it could reject.} \]
“You can’t stand on the bench of the sea of uncertainty with the waves lapping at your ankles. You have to jump into the sea and swim and check your head underwater and blow some bubbles.”

6.7 Model checking for the & schools.

Two kinds of replications:
- New data from some schools: \( \Theta \rightarrow \gamma_{(1)} \)
- New data from new schools: \( (u, v) \rightarrow \Theta^{(r)} \rightarrow \gamma_{(r)}^{(r)} \)

\[
(\Theta, r) \rightarrow \Theta \rightarrow \gamma_{(1)} \rightarrow \gamma_{(r)}^{(r)}
\]

as you move up the ladder, priors become more concentrated around 0.5. When you get more similar to the actual data, the priors get less extreme.

Can imagine \( \gamma \) coming out of the page - everything stacked together like macaroni.

Digression to talk on graphical models.
- Some people think estimation is too boring, so they try to "learn" connections.
- "Learn" is a scary word.

\[
(\Theta, r) \rightarrow \Theta \rightarrow \gamma_{(1)} \rightarrow \gamma_{(r)}^{(r)}
\]

is not a posterior predictive check. \( \gamma_{(1)}^{(r)} \) is pointing the wrong way.

For inference within a model, it is enough to have a joint distribution on everything, but for checking the model, need to take into account the data-generating process, so the axiom matters.

"Inference is normal science. Model checking is revolutionizing science."

6b: Correlation between pre and post is higher among control than among treated.
- \( y = \alpha + \beta x + \Theta x \) is wrong - constant treatment effect never happens.
- Blood pressure drug.

\[
\begin{align*}
\text{before} & : \quad y_i = w_i + \epsilon_i \\
\text{after} & : \quad \begin{cases} 
\epsilon_i + \epsilon_i^2 & \text{if control} \\
\lambda w_i + \epsilon_i^2 & \text{if treated} \\
\epsilon_i & \text{less than 1 to capture variance reduction}
\end{cases}
\end{align*}
\]
Stat 220 6b

- "additive treatment effect"
- teaching German
  Control: no one learns German
  treatment: some kids learn a little, some learn a lot. creates more variance.
  Still, correlation higher among control.

Ch. 7.

7. Evaluating, comparing, and expanding models

7.1. Evaluating the predictive accuracy of a model

log predictive density as a measure of fit
- harder to interpret than mean squared error
- more to compare models than evaluate single model

log data density rather than log posterior density?
- "the answer is that this is a fruitful ambiguity"
- schools example - ambiguous what is prior and what is likelihood
- non-pooling model does no prediction for new schools

Observed fit, cross-validation, and external validation

difference between observed fit and LOOCV is measure of overfitting
- it's the absolute difference that matters, can't say "10% decrease"
- we could add anything to the log density
- can also do 5-fold cross-validation - divide dataset into five pieces many times and take average.

7.2. Information criteria and effective number of parameters

estimates of out-of-sample predictive accuracy
- within-sample predictive accuracy (not good because of overfitting)
- adding an adjustment
- cross-validation
- simulation

Akaike information criterion (AIC)

\[ \hat{\text{elpd}}_{\text{AIC}} = \log p(y|\theta_{\text{mle}}) - \chi \]

\[ \text{elpd} = \text{expected log predictive density} \]

\[ \text{elpd} = \mathbb{E}(\log p(y|\theta_{\text{mle}})) \]

effective # of parameters

fitting a function with 20 parameters gives 30 data points

\[ y_i \sim \text{Poiss}(\lambda_i), \quad \lambda = 35, \quad y_i \]

\[ \text{prob of death at given age} \]

uniform prior: \( p(\theta) = 1 \)

constraint of increasing convexity reduces effective # of parameters

problem: posterior doesn't look right. turns out this prior is informative. uniform on 0's is uniform on second differences. So, prior is concentrated on quadratic curves.
Denance information criterion (DIC)
\[
\hat{\text{DIC}} = \log p ( \tilde{\theta} | \text{data}) - \text{post} = E \left( \log p ( \tilde{\theta} | \text{data}) \right) - \text{post} = E \left( \log p ( \theta | \text{data}) \right)
\]
\[= \begin{cases} 2 \left[ \log \text{posterior density given } \hat{\text{theta}} \right] - \log \left[ \log \text{posterior density given } \theta \right] \end{cases}
\]
\[
\text{PDIC, } \text{NRE} = \text{NRE} \left( \log p ( \theta | \text{data}) \right)
\]

Winsorize-Aitken information criterion (WAIC)
\[
\hat{\text{WAIC}} = \log p ( \text{post}) - \text{WAIC} = E \text{post} \text{ posterior predictive density}
\]
\[
\text{WAIC} = \sum_{i=1}^{S} \text{Var} \left( \log p ( \theta | \text{data}) \right)
\]

Need to predict data - not easy for network data, time series, etc., where data points are dependent.

"Bayesian" information criterion (BIC)
\[
\text{BIC} : \quad \hat{\text{BIC}} = \log p ( \theta | \text{data}) - k
\]
\[
\text{BIC} : \quad \text{BIC} = \text{BIC} - \frac{k}{2} \text{log } \text{N}
\]

Cross-validation

for many portions of the data into training and holdout:

for each to training set, get posterior sims

compare log posterior predictive density of holdout

average over simulations to get \( \hat{\text{elpd}} \text{model} \).

imperfect because still depends on observed data only, so it's still a random variable - the number that gets earned out is for this dataset only.

7a

8.


5.1 Bayesian inference requires a model for data collection.

Full model is \( p( \text{data} | \text{parameters}) \).

Need a model for the data collection process.

<table>
<thead>
<tr>
<th>Observed data</th>
<th>Complete data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scoping</td>
<td>( n \text{ units in sample} )</td>
</tr>
<tr>
<td>Experiment</td>
<td>outcomes under observed treatment</td>
</tr>
<tr>
<td>Rating</td>
<td>rounded observations</td>
</tr>
<tr>
<td>Unintentional missingness</td>
<td>observed values</td>
</tr>
</tbody>
</table>
8.2 Formulated models for data collection
data \( y \), missing indicators \( X \): we can allow for partial information

\[
\text{expected data} \quad y = 10, 10, 12, 11, 9.
\]

\[
\text{likelihood} \quad \prod_{i=1}^{n} \left[ \mathbb{E} \left( \frac{y_i - \mu_i}{\sigma} \right) - \mathbb{E} \left( \frac{y_i - \mu_i}{\sigma} \right) \right]
\]

\[
\text{latent variable formulation} \quad p(\epsilon, \sigma, z | y) = p(\epsilon | \sigma) \cdot p(z | \theta, \sigma) \cdot p(y | z, \theta, \sigma)
\]

\[
\overset{\text{data model}}{\overset{\text{measurement model}}{\overset{\text{distribution model}}{\overset{\text{missing data}}{\overset{\text{incomplete}}{\text{data collection}}}}}}
\]

8.3 Ignorability:

"Why is this chapter different from all other chapters?"

in chapter 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, ...

we write \( p(\epsilon | y) = p(\epsilon) p(y | \epsilon) \)

in chapter 8

\[
p(\theta, \phi | y, x) = p(\theta, \phi) p(y | \theta, \phi) p(x | \theta, \phi)
\]

\[
\text{missing at random:} \quad p(y | x) = p(y | x, \theta, \phi)
\]

\[
\text{distinct parameters:} \quad p(\theta, \phi) = p(\theta) p(\phi)
\]

\[
\text{violated if we purposely get bigger sample sizes for more things}
\]

\[
\text{this is how missingness here. "selection model" - select what gets observed.}
\]

\[
\text{I depends on \( y \) because, for example, old people round ages more.}
\]

\[
\text{amount of rounding can itself be a r.v.} \]

\[
\text{amount of rounding can itself be a r.v.}
\]

if these are satisfied, then the RMM factors as \( p(\theta) p(\phi | \theta) p(y | \theta) p(x | y) \)

\[
\text{So we can just analyze the data, ignoring the data collection process.}
\]

\[
\text{this is called ignorability: distinct parameters + MAR.}
\]

\[
\text{when we just look at } p(\theta) p(y | \theta), \text{ we're assuming ignorability.}
\]

now add covariates:

\[
p(\theta, \phi, y, x) = p(\theta, \phi | y) p(y | x) p(x | y, \phi, x)
\]

\[
\text{distinct means:} \quad p(\theta, \phi | y) = p(\phi | x) p(\phi | x).
\]

\[
\text{learning } x \text{ could be informative about } \theta \text{ and } \phi, \text{ but conditional on } x, \text{ get no additional info about } \theta \text{ from } \phi.
\]

\[
\text{MAR: } p(x | y, \phi, x) = p(x | y, \phi, x)
\]

\[
\text{missingness can depend on info you have.}
\]

\[
\text{true as many } x \text{ es possible into the model, so ignorability is reasonable assumption.}
\]

\[
\text{double-blindness is a form of conditional independence - puts a wall between information and decision.}
\]

\[
\text{this is a thread connecting classical stats to what we do!}
\]
8.4 Sample surveys

Simple random sampling

Finite-population inference

$$ \bar{y} = \frac{N - 1}{N} \bar{y}_{obs} + \frac{N - 1}{N} \bar{y}_{ms} $$

to do inference, assume

$$ \bar{y}_{obs} \sim \text{the same c.s.} \quad \text{(same variance)} $$

then can just use $\bar{y}_{obs}$ to infer $\Theta$. get inference for $\bar{y}_{ms}$ based on posterior predictive.

covariates $\bar{y}_{j} \sim N_{n-1} (\bar{y}_{j,0}, (1 - \frac{r}{N}) s^2_{ms})$.

Simplified sampling

$$ \bar{y} = \frac{\bar{y}}{N_j} $$

$$ \frac{\bar{y}}{N_j} \sim \text{stoch. sizes} $$

8.5 Designed experiments

"An experiment means doing something you wouldn't otherwise do."

we want $p(\Xi | Y, x) = p(Y | x)$, so include enough $x$'s to make this true

for randomized block design, $x$ needs to include block membership

need hierarchical model of $y$ given $x$.

Let's designs that check.

lots of ignorable designs! consider the following sequential designs.

a. $n = 20$

b. $2$ hours

c. $n \sim 10 \times 20$

d. $n_1 = 10$

e. $|n_1| = 4$

all are ignorable. all have the same likelihood function!

for (a) and (c), if $\Theta$'s close together, takes forever to reach stopping rule, get precise estimate near 0.

if $\Theta$'s far apart, get more right away.

if interval is $[0.0001, 0.0003]$, $\Theta$ cannot possibly include 0

"maybe we think there's nothing going on because it's conventional research."

"it's so great when you guys laugh - it's like blood to a vampire!"
Example: matched pairs design.

| item | gp | treat | y|v| y|v|
|------|----|-------|---|---|---|
| 1    | 1  | 1     | ✓ | ✓ | ✓ |
| 2    | 1  | 2     | ✓ | ✓ | ✓ |
| 3    | 2  | 1     | ✓ | ✓ | ✓ |
| 4    | 2  | 2     | ✓ | ✓ | ✓ |
| 5    | 3  | 1     | ✓ | ✓ | ✓ |
| 6    | 3  | 2     | ✓ | ✓ | ✓ |
| 7    |    |       |   |   |   |

The estimator is \( \tilde{Z} = \tilde{y}_1 - \tilde{y}_2 \) where \( \tilde{y}_j = \bar{y}_j - y_j \), \( j \) indexing the groups.

\[
\begin{align*}
\text{right:} & \quad \frac{S_x^2}{n} \\
\text{wrong:} & \quad \frac{S_x^2}{n} + \frac{S_y^2}{n}
\end{align*}
\]

\( \tilde{y} \) an error \( y = \alpha + \beta T + \varepsilon \).

Also right: reg \( y \) on \( T \) and pg y indicator gives same estimate.

Bayesian (also right): put prior on \( \alpha \) group coefficients \( \theta_1, \ldots, \theta_n \sim N(\mu, \tau^2) \), then do this regression.

\[ \text{"this is a textbook way too. it's just my textbook."} \]

will only make a difference when \( T \) is small (not much group variation).

by partially pooling, get more precise estimate, because you're partially pooling toward the "wrong" analysis that gives more degrees of freedom.

in practice, the real benefits are for more complicated designs, since this generalizes to 3 people per group, etc.

Example: experiment on 50 cows:

cows: amount of milk fat produced.
treatments: 4 levels of feed additive.
background variables: location, age, initial cow weight.
repeated measurement within balance.
how to analyze?

conditional on location, age, and cow weight, it's ignorable. those are the only covariates used in the assignment. "he wasn't faced to face with the cows, as it were."

so just throw those covariates into the model, regress treatment on the covariates. then it's irrelevant how many times he randomized.

Q. if sampling and the role of randomization

why randomized?

with no item-level background variables, randomization is the only ignorable design.

with background variables:

design ABABACBABA uses background variable of location.

advantages of randomization are for model checking (regression future data) and robustness.
8.7 Observational studies

Balance (on observed and unobserved variables)

Lack of complete overlap in high dimensions - hard to detect and deal with.

Matching and poststratification.

Matching and regression together: matching deals with lack of overlap, and regression helps further with balance.

Propensity score matching.

Poststratification (example: mother's education)

8.8 Censoring and truncation

N observations from \( f(y;10) \), only observed when \( y \leq 200 \), observe \( n \) cases where \( y < 200 \).

Scenario 1: \( N \) unknown, truncated-data likelihood

\[
p(\theta | y) \propto p(\theta) \prod_{i=1}^{n} f(y_i;10) \frac{1}{F(200;\theta)^n}
\]

Scenario 2: \( N \) known, censored-data likelihood

\[
p(\theta | y, N) \propto p(\theta) (1 - F(200;\theta))^{N-n} \prod_{i=1}^{n} f(y_i;10)
\]

For a Bayesian, if \( N \) is unknown, we should be able to use the censored-data model + a prior on \( N \), average over \( N \) to get \( p(\theta | y) \):

\[
p(\theta | y) \propto \sum_{N=n+1}^{\infty} p(N) p(\theta) (1 - F(200;\theta))^{N-n} \prod_{i=1}^{n} f(y_i;10)
\]

but it turns out this reduces to the truncated-data model if and only if \( p(N) \propto \frac{1}{N} \).

Summary of Chapter 8

The method of data collection dictates the minimal level of modeling required for a valid Bayesian analysis.

Condition on all information used in the design

- In a survey: stratum and cluster indicators, any variable that determines probability of sampling or possibility of non-response.
- In an experiment: pairs, blocks, any variable used in treatment assignment.

Mr. P: regression analysis, multi-level b/c many unknown parameters, and poststratification.

"Survey weights are like McDonald's chicken nuggets - you don't know what goes into them."

Open problem - poststratification on many variables.
Latin square homework problem

"before you do anything, you gotta take the... you gotta take the..."

\[ \begin{array}{ccc}
    0 & 1 & 2 \\
    1 & 0 & 2 \\
    2 & 2 & 0 \\
\end{array} \]

because the effects are probably multiplicative, and in my case this is more interpretable.

"you take the log so first that you don't even see the actual data. plus you take the log because you can, because they're all positive."

\[ y_i = \alpha + \alpha_{row[i]} + \beta_{column[i]} + \gamma_{treatment[i]} + \epsilon_i \]

proportionality models:

\[ y_i \sim N(\mu + \alpha_{row[i]} + \beta_{column[i]} + \gamma_{treatment[i]}), \sigma_y^2) \quad i = 1, 2, ..., 25 \]

\[ \alpha_j \sim N(0, \sigma_{\alpha}^2) \]

\[ \beta_k \sim N(0, \sigma_{\beta}^2) \quad j, k, l = 1, ..., 5 \]

\[ \gamma_i \sim N(0, \sigma_{\gamma}^2) \]

\[ \rho(\mu, \sigma_{\alpha}, \sigma_{\beta}, \sigma_{\gamma}, \sigma_y) \propto 1 \]

Petri dish homework problem

6 cultures per dish, 5 dishes: individual-level analysis or dish-level analysis?

\[ \bar{y}_A = \frac{\sum A}{5}, \quad \bar{y}_B = \frac{\sum B}{5} \quad \text{vs.} \quad \bar{y}_A = \frac{\sum A}{5}, \quad \bar{y}_B = \frac{\sum B}{5} \]

right thing is to do dish-level analysis because if she's right and there are no dish effects, then \( \sigma_y^2 \) will be 6 times less variable anyway. but if she's wrong, she should do the dish analysis!

while the sham data from the chicken brains, there the less from the dish analysis is small.

even better would be to do partial pooling.

Ch 9

9. Decision analysis.

data, model > inference > decision analysis > decision.

9.1 Bayesian decision theory in different contexts

decision trees

- \( d_1 \)
- \( d_2 \)
- \( d_3 \)

\[ p(\theta | d_1), p(\theta | d_2), p(\theta | d_3) \]

\[ U(d, \theta) \]

JITT question about microkilo: 2.6 cents per microkilo is likely -- people will accept some amount of money to increase risk, but it doesn't make sense to talk about people accepting 2600 to die.
DeNovo analysis vs. "statistical decision theory" - evil twin, says things like "use posterior mode for 0-1 loss" but isn't connected to real-world decision problem.

Dave Krantz: goal-based framework.

ex: effects of incentives for telephone surveys.

6 factors: incentive or not, voice, form, timing, route, burden.

Fit model, see if inferences make sense, pipe them into decision analysis.

ex: radar measurements - remediate, measure, or do nothing.

10/26/12

Ch. 10.

10. Overview of computation

10.1 Crude estimation by ignoring some information

Simpler model

set hyperparameters to fixed values

Note: don't always want to use na""ivy or complete ignoring as simple model.

computational unstable since you're estimating a ton of parameters.

quick insertion of missing data

network of models.

fit multiple models to:

1. compare
2. check
3. understand what we're doing - "scaffolding our understanding"

"exploratory model analysis"

open question: how to do this systematically?

drawing a "network of models" - models connected to other models in a network, and edge exists if they differ by only one thing. Define "operations" you can perform on a model: adding a predictor, adding a hyperparameter, adding a more general distribution (t instead of Normal), etc.

15. General nature of scientific revolution

The process: understanding new model

fig: improvement in understanding over time

idea is that you throw in some data and the computer program builds the model by starting with simple things and adding things as you need a model-checking module.
10.2 Direct simulation

...approximation $M(y(0))$ that dominates $p(0|y)$ can be computed and drawn from.

For $s=1$ to $S$:
- Draw $\theta$ from density proportional to $y$.
- Accept with probability $\frac{f(\theta|y)}{M(y(0))}$.

Usual doesn't work because acceptance rates absurdly low.

10.3 Numerical integration

Goal: posterior expectation $E(h(0)|y) = \int h(0) \cdot p(0|y) \, d\theta$.

Estimate by $\frac{1}{S} \sum_{i=1}^{S} h(0)_{i}$ if can draw directly.

Laplace's method:
- Approximate $h(0) \cdot p(0|y)$ by quadratic in $\theta$.
- Fit to mode and curvature of the mode.

Using unnormalized density:

$$E(h(0)|y) = \frac{\int h(0) \cdot q(0|y) \, d\theta}{\int q(0|y) \, d\theta}$$

apply Laplace to numerator and denominator.

10.4 Importance sampling

Goal: $E(h(0)|y) = \frac{\int h(0) \cdot q(0|y) \, d\theta}{\int q(0|y) \, d\theta}$

But can't draw from $q$, only from $y$.

Compute importance weights $w(\theta_{i}) = \frac{q(\theta_{i}|y)}{g(\theta_{i})}$

Estimate of $E(h(0)|y)$:

$$\frac{1}{S} \sum_{i=1}^{S} h(0)_{i} \cdot w(\theta_{i})$$

Want $g$ to have heavier tails than $q$.

Scenario 1: $g$ may be heavy tailed.

Scenario 2: $g$ may not be heavy tailed.

0.5 Computing normalizing factors

$p(y|0) = \frac{1}{Z(0)} q(y|0)$. $Z(0)$ is the normalizing factor.

Typically want to calculate $Z(0)$ offline - get $z(\theta)$ as a function of $\theta$.
models in BDA: “Normal, Poisson, Binomial, Gamma, that’s it.”

\[ p(y|\theta) \propto e^{-\theta \sum c_{ij}(y_i-y_j)^2} \]
\[ y_i \text{ is banded} \]
\[ \text{if } i \text{ and } j \text{ are neighbors, } 0 \text{ otherwise} \]
\[ \theta > 0, \text{ says neighbors have to be similar} \]

spatial equivalent of AR model in time series
finding normalizing factor is hard.

10.6 Use of posterior simulations in Bayesian data analysis

estimating more event probabilities by combining simulations and analytic probabilities

10.8 Practical issues

Take-data debugging

\[ M \rightarrow \theta \rightarrow y \]

\[ \theta \text{ prior} \rightarrow y \text{ prior} \]

\[ \theta \text{ infer} \rightarrow y \text{ infer} \]

debugging: bridge between

simple models that can be fit successfully \[ \rightarrow \]
complex models that don’t fit

- work with smaller datasets
- strip down the model
- fixed parameter values, then strong priors, then weak priors

Week 9 - Hurricane Sandy.

10a.

read Ch 20 of Gelman and Hill.

JITT

\[ p = .6 \]

\[ \text{SE } \frac{.5}{\sqrt{n}} \]

\[ \frac{.5}{\sqrt{n}} \]

"the SE is .75 times (1-.75) over... oh, give me a break,
this works just fine."

set cliff \[ \geq 2.8 \times \text{SE} \]

\[ .15 \geq 2.8 \times \frac{.5 \times \sqrt{2}}{\sqrt{n}} \]

solve for \[ n \].

Ch. 13.

13. Approximations to the posterior distribution

rule of approximations has changed - used to be the easy thing to do, but now MCMC is sometimes easier

than the approximation. Approximation is more scalable

def. of approximation: doesn’t converge to the right stationary distribution

multinomial posteriors often arise in discrete data (generics - one gene vs. the other)
13.2 Boundary-avoiding priors for model summaries

A prior that's best for the posterior node might be fine for full Bayes.

Conversely, if we knew we're going to use PM, we might want to choose a different prior.

E.g., Gamma(3, 0.1) prior for \( \theta \) in 8-schools example

13.3 Normal and related mixture approximations

Fit to make and approximate the mode

Mixture of Normals or autoregressive

Importance sampling (with replacement)

"Rubin calls this sampling importance resampling, which is silly because once you resample, you must have sampled already. It's like you say e1, you can't say e2 to e4 because of course it was at e2, it's not like it was at e3!

I guess e4 is old-fashioned nowadays. I guess I should do something like g3."

13.7 Variational Bayes

Goal: approximate \( p(\theta | y) \) by \( q(\theta) \)

Typically assume \( q(\theta) = \prod_j q_j(\theta_j) \)

Approximating each marginal distribution:

\[ p(\theta | y) = \int_{\theta_1} \cdots \int_{\theta_j} p(\theta_1, \ldots, \theta_j | y) d\theta_1 \cdots d\theta_j \]

Converges to a distributional estimate (many times is just a lower bound)

VB for 8 schools

\[ \log p(\theta, \mu, \tau | y) = \text{const} - \frac{1}{2} \sum_j \left( \frac{y_j - \theta_j}{\sigma_j^2} \right)^2 - \frac{1}{2} \frac{(\theta - \mu)^2}{\tau^2} - \frac{1}{2} \frac{(\sigma_j^2)^2}{\tau^2} - 8 \log \tau \]

Look at \( \theta_i \):

\[ \text{const} - \frac{1}{2} \left( \frac{y_i - \theta_i}{\sigma_i^2} \right)^2 - \frac{1}{2} \left( \frac{\theta_i - \mu}{\tau^2} \right)^2 \]

\( \theta_i \) is the star! Only care about terms involving \( \theta_i \)

\( \mu, \tau \): average over, using \( g(\mu), g(\tau) \)

VB underestimates variance.

Research project: VB + particle filtering.

10b

11/7/12

WITT: if we want to find the slope of a dose-response curve and we know it's linear, just measure at endpoints!

Three sources of error: measurement error, model error, natural variability.

Xinked on Bayesian vs. frequentists: "I don't like this."

Frequentism is a conservative viewpoint.

Bayesianism is conservative in that it respects prior beliefs.

Frequentism is conservative in the sense of distilling any overall/overriding philosophy. Seeing a situation like this, they would say the prior is not appropriate.

Bayesians have this idea that their methods should work everywhere. Frequentists are more willing to say their method doesn't work in this situation.

But isn't that an unprincipled way to incorporate prior information? Sure, but then the fair question is: how much do you lose from a discrete approach to prior information as opposed to a prior distribution?
praise is not the only aspect of a statistical test.
what do frequentists give up on? frequentists might say "i can't give you a probability." bayesians will give you something, but it could be bad, so it would be easy to make a claim where the bayesian looks silly:

freq: "i can't give you a probability."
bayesian: "i use a flat prior and the probability is 1/36."

"but if someone's going to be unfairly attacked, i'd rather it not be me."

causality and statistical learning
forward vs. reverse causal questions
rubin hates reverse causal questions, but it's easy to think in terms of reverse causality.
now, some things are easy to think about yet be wrong, like anthropomorphizing objects.
or folk physics: people think a baseball trajectory looks like.
but in this case it seems there's something to these questions - it's a fruitful inquiry.
what way to think about reverse causal questions is as a way to generate a list of forward causal questions, which are easier to handle from a statistical perspective.
does it make sense to formalize reverse causal inference? is it possible?
"the way i think about causal inference now is in terms of model checking."
if attachment is related to earnings and that's a surprise, then it's a degenerate from our usual framework.

different perspectives on causal inference
humans: wired to ask reverse questions
macroeconomists: state-space models
applied micro: forward causal inference - think in terms of intercausal.
skeptics: fitting models
computer scientists: modeling everyday reasoning
example: traveling salesman, really hard optimization problem, and yet people figure out how to get places.
similarity, causality is hard, but everyone does it every day, so it's possible for computers to learn it too, an optimistic view!
statement like "rain causes mud, mud does not cause rain" sound silly, but it's about thinking of the relationships between variables in terms of how people think about them.

example: deterrent effect of the death penalty
low school prof with no stats background trusts "sophisticated econometricians"
indicators regression models with additional assumptions
economists don't want assumptions, don't want bias, so end up with low power and high variance, so they throw a ton of data at it, including inappropriate pooling across time, which creates bias.

example: u-shaped of happiness
graph from Suspense shows middle-aged people are sad, but GSS happiness is exactly the other way around.
Difficulties with the research program of learning causal structure

"I don't like the idea of static variables causing each other."

"learning causal structure" is all about discovering zeros, but there are no true zeros.

a lot of social science proceeds by trying to discover stylized facts and reason from them,

but the statement that something has no effect isn't generally reasonable.

11a

Ch. 16

16. Generalized linear models

16.1 Standard GLM likelihoods

16.2 Setting up and interpreting GLMs

effects - like log population size in Poisson regression - coefficient should be 1 when you do the regression.

idea in regression - if have 3 ethnic groups and 5 grade levels, set one of each to 0.

expanded example: ETS, predicting Calc II grades from AP scores.

\[ y_i = a + b_j[i] + \text{error} \]

\( y_j \) = group of obs. i

\[ y_i = a + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_5 x_{i5} + \text{error} \]

indicator variables

\[ b_j \sim N(\beta (\mu_j - \bar{\mu}), \sigma_b^2) \]

not exchangeable

\[ \sigma_b = \{1, 2, 3, 5, 6, 6.5\} \]

if \( \sigma_b = 0 \), then the \( b_j \)'s are all exactly on the line.

SJEs are just from ordinary linear regression.

errors within a group

\[ \text{avg}_j = a + \beta_j + \epsilon_j \]

errors confounded with \( a \), so SSe too wide.

instead of \( a, \beta, \text{error} \), have \( a + \text{error} \), \( \beta + \text{slope}(\text{error}) \), residual. These are more stable.

parameters don't mean what you think.

with few groups, you have basically no info on group-level variance \( \sigma_b^2 \)

by question, is appropriate post-processing.
takeaway: multilevel modeling lets you estimate individual coefficients and group-level effects all at once.

"the computation is going to be hard for CMs no matter what, so you're already dead. okay, not dead, but that bus is already gone."

No. 3 Computation

Separation - will always happen with enough predictors. Only happens for discrete, not continuous.

11b

JITT

ordered logit

\[
\begin{array}{cccc}
14 & 15 & 15 & 0 \\
S_1 & S_2 & S_3 & S_4 \\
\end{array}
\]

using mode for \((\beta, \delta)\), \(\hat{S}_3 = \hat{S}_4\).

so predicted probability is 0.

pilot study is to figure out your design, not to measure effect.

sampling words at random:

words \(x_i\);

actual probability \(p(x_i)\);

target probability \(p(x_i) \approx 1\);

use importance resampling: \(w_i = \frac{p_i}{g_i}\), yet 2000 and resample 1000.

Ch 20

20. Nonparametric regression

is more important than density estimation.

20.1 Splines and other basis functions

"because cubically is closed under addition - doesn't that sound good?"

Cubics allow you to fit a sequence of points piecewise and have continuous 1st derivative.

Divide x-axis into pieces, fit a cubic within each piece.

Can express spline as sum of basis functions, then model is \(y_i = \sum \frac{1}{j} \beta_j \phi_j(x_i) + \text{error}\).

Gaussian kernels:

by adding scaled versions of these, can get increasing functions or other shapes.

20.2 Prior distributions for basis function models

Instead of estimating the knots, think of there being a large but fixed number of knots, put a prior on how many you need, and only keep how many you need.

Similar to variable selection in regression.

improper symmetric prior means scale mixture of Normals.
20.3 Multivariate regression surfaces

additive model (not good for hi prem because that would assume constant difference between men and women)

multivariate kernels for multivariate response

"tensor products, which are some sort of product... of tensors, I believe."

20.4 Gaussian processes

\[ y_i = g(x_i) + \text{error} \]

\[ g(x) \]

\[ \text{defined on discrete points } y_1, y_2, \ldots, y_T. \]

\[ \begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_T \end{pmatrix} \sim N \left( \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_T \end{pmatrix}, \begin{pmatrix} \sigma_1 & \sigma_1 & \cdots & \sigma_1 \\ \sigma_1 & \sigma_2 & \cdots & \sigma_2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_1 & \sigma_2 & \cdots & \sigma_T \end{pmatrix} \right) \]

mean process: \[ \uparrow \]

can't regression: \[ \uparrow \text{ covariance.} \]

can have hyperparameters. high correlations for nearby values, low correlations for far away values.

want local smoothness but globally uncorrelated. Unlike AR process.

Covariance function:

\[ c(x, x') = \phi_2 \exp(-\phi_2 \|x-x'\|^2) \]
\[ c(x, x') = \phi_1 \exp(-\frac{d}{\sum_{j=1}^d} \phi_j (x_j - x_j')^2) \]

Gaussian process prior for coefficients of basis expansion

12a

Bayes helps if you're allowed to use prior information or if you have a lot of parameters.

"one of the most influential Bayesian books ever written by ld savage, and it's full of good ideas.

there isn't a single good idea in the whole book."

"saying seems like a more serious lecturer, which is good, you get the soft jokes now and then next semester

it's time to learn some shit."

Ch. 21

21. Finite mixture models

Example: identifying a three-component mixture

"you're not going to go to heaven because you only needed two." (mixture components)

motivation: want to learn about

problem bins = are democrats getting more seats than they deserve?

if they got 50% of the vote, how many seats would they get?

electoral responsiveness = how responsive is congress to the voters?

if there were a 1% shift in the vote, how would that affect the # of seats assigned to each party?"
hypothesised election shift things and add noise

\[ y_i = x_i + \epsilon_i, \quad \epsilon_i \sim N(0, 0.06^2) \]

\[ y_i, y_i^p = x_i, \quad \epsilon_i, \epsilon_i^p \]

0.06 is unexplained variation in vote share, obtained as

\[ T_S = \sqrt{\frac{1}{n-1} \sum \left( \frac{y_i - \bar{y}_i}{n - \bar{y}_i} - \frac{(y_i^p - \bar{y}_i^p)}{n - \bar{y}_i^p} \right)^2} \]

21.4 More general formulation

in schizophrenia reach time example, mixture components actually mean something, as distinguished from
just fitting clusters because you can.

meaning of a cluster depends on context,

might make sense to arbitrarily break into clusters to explain variation in outcome

example: clustering the 50 states according to welfare policy.

21.5 Label switching and posterior computation

for identifiability, impose constraints like \( \mu_1, \sigma_1, \mu_2, \sigma_2 \)

21.6 Unspecified \( K \) of mixture components

average over unknown \( K \)

choose a prior

marginal likelihood of \( K \) or trans-dimensional MCMC (propose merging clusters)

or choose a large (but not huge) \( K \) and maybe some of the clusters will be empty

prior on membership probabilities

\( (\pi_1, \ldots, \pi_K) \sim \text{Dir} \left( \frac{1}{K}, \ldots, \frac{1}{K} \right) \).

"I’ve never done it. I just read about it in my own book."

12b

question on hand to compare models - writes in point letters "Out of Sample Prediction Error."

("So there’s that.")

"the likelihood function never tells you the whole story, because you don’t know where the data come from."

What can you do with 21 data points?

1. See whether the data are consistent with your prior

2. Do minimise analyses. in this case, found that none of the 16 predictors was highly correlated

"better to have analyzed and lost them never to have analyzed the data at all."

"a Bayesian version will usually make things better."
"Hierarchical models: just like what you did before, but the standard errors are a little bit better."

"Political scientists have no pride: they'll use anything that works, or doesn't work - it's not like they would know."

Ch. 22

22.1 Bayesian histograms, also better histograms

Histogram without specifying # of bins or kernel locations

more bags where there's more variability - solves the hist (renchy 1000) problem.

Is hierarchical Bayes nonparametric?

Bayesian models are always parametric at the top level, but not at intermediate stages in the sense that the curve may not be restricted to any substance of the space of possible curves

line between parametric and nonparametric is blurry

research principle: "you take what someone is doing and pretend they're being Bayesian."

Bayesian data analysis

The three steps!

Building confidence in classes of models.

"I've climbed the 8 schools. As we develop, we climb more and more things. We'll be able to choose DP's and CRPs."

What is the role of theory?

"The full name of theoretical statistics is the theory of applied statistics."

"Statistics is applied statistics."

"The gambler's ruin problem, that's a theory about what happens when you're a gambler."

proper with demur or multiple comparisons

Tukey: the model is irrelevant. you construct the model, it leads to a method, the method had statistical properties, and it's the properties that matter, not the model.

"like he advocated plotting a histogram instead of a histogram. you take the square root of the counts, because if the counts follow a Poisson and you... [goops, come back!] oops, I said Poisson!"

Open questions

Systematic model choice

Grammar of models

Computation - reprogramming

Where will statistics be in 20 years?

understanding

you don't perceive the change, but it's there.