Statistics 225
Bayesian Statistical Analysis (Part 4)

Hal Stern

Department of Statistics
University of California, Irvine
sternh@uci.edu

March 28, 2019

Model checking
Introduction

- So far:
  - build probability models
  - compute/simulate posterior distn
- Now:
  - model checking (does the model fit the data)
  - sensitivity analysis (are conclusions sensitive to assumptions)
  - model selection (which is the best model)
  - robust analysis (are conclusions sensitive to data)

Model checking
General ideas

- Don't ask if the model is true
- Does the model fit and provide useful inferences
- Remember the model includes
  - sampling distribution
  - prior distribution
  - hierarchical structure
  - explanatory variables
- More than one model can fit (sensitivity analysis)

Model checking: types of checks

- Classical ideas
  - Check whether parameter estimates make sense
  - Check whether predictions make sense
  - Does the model generate data like "my data" (simulation approach, residual analysis)
  - Embed in a larger model
- Bayesian ideas
  - Compare posterior distribution of parameters to substantive knowledge
  - Compare posterior predictive distribution of future data to substantive knowledge
  - Compare posterior predictive distribution of future data to observed data
  - Evaluate sensitivity of inferences to other model specifications (e.g., alternate priors or sampling distributions, embed in larger model)

Posterior predictive model checking

- y^{rep} = replicate data that might have occurred
- Replicated under same model as original data (e.g., same covariate values) with same values for unknown parameters θ
- Posterior predictive distribution of y^{rep}

$$\begin{aligned} p(y^{rep}|y) &= \int p(y^{rep}, \theta|y)\; d\theta \\ &= \int p(y^{rep}|\theta, y)p(\theta|y)d\theta \\ &=? \int p(y^{rep}|\theta)p(\theta|y)d\theta \end{aligned}$$

- Last equality is generally (but not always) true
- Easy to obtain simulations of y^{rep} given posterior simulations of θ
- Other possible definitions of replications (more on this later)

Posterior predictive model checking

- $T(y, \theta)$ is a test quantity or discrepancy measure
- Compare posterior predictive distribution of $T(y^{rep}, \theta)$ to posterior distribution of $T(y, \theta)$
- One possible summary (but not the only one) is the posterior predictive P-value

$$
\begin{aligned}
P_b &= \Pr(T(y^{rep}, \theta) > T(y, \theta) | y) \\
&= \int \int I_{[T(y^{rep}, \theta) > T(y, \theta)]} p(y^{rep} | \theta) p(\theta | y) dy^{rep} d\theta
\end{aligned}
$$

- Special case $T(y, \theta) = T(y)$ is a test statistic
  - compare posterior predictive distribution of $T(y^{rep})$ to observed $T(y)$
- Diagnostics such as plots of residuals are special cases of posterior predictive checks

Posterior predictive model checking

Relation to traditional tests

- Example:
  - suppose $y_1, \ldots, y_n$ are iid $N(\mu, \sigma^2)$
  - believe $\mu = 0$, so fit $N(0, \sigma^2)$ model
  - want to check fit of $N(0, \sigma^2)$ model
  - weak example because obvious model checking approach is to fit the "bigger" $N(\mu, \sigma^2)$ model and check whether $\mu = 0$ is plausible
- Frequentist approach
  - test statistic: $T(y) = \bar{y}$
  - begin by assuming σ^2 is fixed

$$
\begin{aligned}
\text{p-value} &= P(\overbrace{T(y^{rep})}^{r.v.} \geq \overbrace{T(y)}^{\text{obs.value}} \; |\sigma^2) \\
&= P(\bar{y}^{rep} \geq \bar{y}|\sigma^2) \\
&= P\left(\frac{\sqrt{n}\bar{y}^{rep}}{S} \geq \frac{\sqrt{n}\bar{y}}{S}|\sigma^2\right) = P\left(t_{n-1} \geq \frac{\sqrt{n}\bar{Y}}{S}\right)
\end{aligned}
$$

  - last equality because distn no longer depends on σ^2
  - it is not always possible to get rid of nuisance parameters in this way

Posterior predictive model checking

Relation to traditional tests (cont'd)

- Posterior predictive approach

$$
\begin{aligned}
\text{p-value} = &\ P(T(y^{rep}) \geq T(y)|y) \\
= &\ \int\int I_{[T(Y^{rep}) \geq T(y)]} p(Y^{rep}|\sigma^2) p(\sigma^2|y) dy^{rep} d\sigma^2 \\
= &\ \int \underbrace{P(T(y^{rep}) \geq T(y)|\sigma^2)}_{\text{classical p-value}} p(\sigma^2|y) d\sigma^2
\end{aligned}
$$

  - if the classical *p*-value is independent of σ^2, as for $T(y) = \bar{y}$ in the example, then the posterior predictive *p*-value is equal to classical *p*-value
  - if not, then formula above shows how the Bayesian approach handles nuisance parameters

Posterior predictive model checking
Defining replications

- Defining replications y^{rep}
  - usually keep features of original data fixed (e.g., sample size)
  - different definitions are possible in hierarchical models
    - replications of the same units

$$p(\phi|y) \rightarrow p(\theta|\phi, y) \rightarrow p(y^{rep}|\theta)$$

    - replicate data for new units

$$p(\phi|y) \rightarrow p(\theta|\phi) \rightarrow p(y^{rep}|\theta)$$

Posterior predictive model checking

Defining test measures

- Defining test statistics or discrepancies
  - measure features of data not directly included in the model (bad to use $T(y) = \bar{y}$ if the model includes a location parameter)
  - may define a number of test measures
  - difficult to speak in general terms because good test measures depend on context
  - examples
    - to check for autocorrelation in a sequence of Bernoulli trials, use a count of the number of runs
    - to check for new predictor in regression model, use $\text{corr}(y - X\beta, x_{new})$
    - to check for asymmetry in a normal model, use $|y_{.9} - \theta| - |y_{.1} - \theta|$
    - to check overall fix in a complex model, use $T(y;\theta) = \sum \left[(y_i - E(y_i|\theta))^2 / \text{Var}(y_i|\theta) \right]$ (Note: asympt χ^2 for known θ but here no reliance on asymptotic distn)

Related ideas

- Parametric bootstrap (e.g., Efron, 1979)
  - plug in point estimate $\hat{\theta}$
  - simulated replicate data sets from $p(y|\hat{\theta})$
- Marginal distribution (Box, 1980)
  - reference distribution is $p(y) = \int p(y|\theta)p(\theta)d\theta$
  - note this is prior predictive distribution
  - requires proper prior distribution

Criticisms of pp model checks

- Unobserved data ($y^{(rep)}$) is not relevant for Bayesian inference
- Posterior predictive checks are too conservative ("double-counting(?)" the data)
- Main concern is that posterior predictive p-values are not uniformly distributed under the null hypothesis
- Critics complain that it is difficult to interpret because of above ... what is an unusually high or low value in practice
- Alternatives have been proposed, e.g., conditional predictive distn or partial posterior predictive distn (Bayarri and Berger in JASA 2000)
  - avoid some of the criticisms by conditioning on "some" of the data but not all
  - can be hard to compute
- Counterpoint: Post. pred. p-values are posterior probabilities of relevant quantities and can be interpreted as probabilities

On the conservatism of pp model checks

- Suppose that $Y \sim N(\mu, 1)$ and $\mu \sim N(0, 9)$
- Observe $Y_{obs} = 10$. Is this unusual?
- Prior predictive approach
  - marginal distn of Y is $N(0, 10)$
  - p-value $= 1 - \Phi(10/\sqrt{10}) = .008$
  - don't believe model
  - the observed value 10 is not consistent with this prior distn and data model
- Posterior predictive approach
  - posterior distn of μ is $N(0.9Y_{obs}, 0.9) = N(9, .9)$
  - posterior predictive distn of Y is $N(9, 1.9)$
  - p-value $= .23$
  - model cares about posterior fit (this minimizes the effect of the prior)
  - would this approach ever reject the model (yes, $Y_{obs} = 23$)

Posterior predictive model checking

- Easy to execute
- Analogous to usual model checking ideas
- Can be somewhat conservative in practice .. but can argue appropriately so because it does not reject a model that generates data like my data

Sensitivity analysis

- Generally true that many models can be fit to the same data
- Question is how sensitive the inferences we draw are to the different models
- Different types of inferences may have different sensitivity
  - posterior mean or median for parameter of interest is typically not sensitive
  - extreme percentiles are more sensitive
- Approaches
  - fit different models
  - expand model/embed model in larger family (more on this later)
    - examp: consider normal distn as part of $t_\nu(\mu, \sigma^2)$ family (normal distn corresponds to $\nu = \infty$)

Model comparison / Model Selection / Model Averaging

- Model checking assesses the fit of a single model
- Sensitivity analysis considers multiple models with a focus on whether the inference changes
- We next consider approaches the choose between (or average over) a set of models
- We address three topics
    - Comparing models
    - Model selection (via the Bayes factor)
    - Model averaging

Model comparison

- Given one or more models it is natural to assess performance in terms of predictive accuracy
- This also provides a mechanism for comparing models
- Goal is to predict new data $\tilde{y}$ from the same data generating process
- How do we measure accuracy?
- Need a scoring rule that assesses the quality of the predictive density
- The log predictive density $\log p(\tilde{y}|\theta)$ is a common choice
  - matches squared error in normal models
  - related to Kullback-Leibler information

Model comparison

- Ideal measure would be out-of-sample predictive performance (i.e., assess on new data from the same process)
- Let f be true data generating model, y be observed data and $\tilde{y}$ future data
- Out-of-sample predictive fit for a single new data point using logarithmic score is

$$\log p_{post}(\tilde{y}_i) = \log E_{post}(p(\tilde{y}_i|\theta)) = \log \int p(\tilde{y}_i|\theta)p_{post}(\theta)d\theta$$

  where p_{post} is shorthand notation for the posterior distribution $p(\theta|y)$ (this notation keeps formulas a bit neater)
- Of course we don't have future data so ideally would average this over the distribution f
  elpd = expected log predictive density for a new data point
  elpd $= E_f(\log p_{post}(\tilde{y}_i) = \int (\log p_{post}(\tilde{y}_i))f(\tilde{y}_i)d\tilde{y}_i$

Model comparison

- Recall
  elpd = expected log predictive density for a new data point
  elpd = $E_f(\log p_{post}(\tilde{y}_i) = \int(\log p_{post}(\tilde{y}_i))f(\tilde{y}_i)d\tilde{y}_i$
- There is a choice to be made between evaluating predictive performance for the joint distn of $\tilde{y}$ or evaluating predictive performance by considering the sum over individual points $\tilde{y}_i$
  - These are the same if model for y is independent given parameters
  - Many common precedures use pointwise (so we will do that)

  elppd = exp. log pointwise predicitve density for new data set
  elppd = $\sum_{i=1}^{n} E_f(\log p_{post}(\tilde{y}_i)$
- As in model checking there is some ambiguity in defining what predictive performance means in hierarchical models (e.g., predictions at the same schools or at new schools from the population distn)
- Both are plausible and it will depend on context; we don't worry about this issue further
- .... So how do we estimate elppd or other relevant quantity?

Model comparison

- Given that f is unknown and we only have our data set y, the most natural idea is to summarize the predictive accuracy of the fitted model by
  lppd = log pointwise predictive density
  $\text{lppd} = \log \prod_{i=1}^{n} p_{post}(y_i) = \sum_{i=1}^{n} \log \int p(y_i|\theta) p_{post}(\theta) d\theta$
  $\text{lppd} \approx \sum_{i=1}^{n} \log(\frac{1}{S} \sum_{s=1}^{S} p(y_i|\theta_s))$
  where $\theta_s, s = 1, \ldots, S$ are posterior simulations
- The lppd is an overestimate (biased high) of the target elppd
  - same data is used to fit the model and assess the model
  - bias likely to depend on number of parameters in the model
- We consider approaches to correcting for this bias

Model Comparison Measures

- For historical reasons measures of predictive accuracy are
  - described as information criteria
  - based on the deviance (log predictive density multiplied by -2)
- Akaike information criteria (AIC)
  - uses plug-in estimate (MLE) for θ rather than posterior distribution
  - applies penalty to predictive accuracy based on asymptotic normal posterior distribution
  - $\widehat{elpd}_{AIC} = \log p(y|\hat{\theta}_{mle}) - k$
  - $AIC = -2 \log p(y|\hat{\theta}_{mle}) + 2k$
  - if iid data, then $\log p(y|\hat{\theta}_{mle}) = \sum_i \log p(y_i|\hat{\theta}_{mle})$
  - number of parameters is not always well defined (e.g., strong prior distns, hierarchical models)

Model Comparison Measures

- Deviance information criteria (DIC)
  - Makes two changes to AIC:
    replaces MLE with posterior mean $\hat{\theta}_{Bayes} = E(\theta|y)$
    replaces k with data-based bias correction
  - $\widehat{elpd}_{DIC} = \log p(y|\hat{\theta}_{Bayes}) - p_{DIC}$
  - p_{DIC} is effective number of parameters
    $p_{DIC} = 2(\log p(y|\hat{\theta}_{Bayes}) - E_{post}(\log p(y|\theta)))$
    estimated as $p_{DIC} = 2(\log p(y|\hat{\theta}_{Bayes}) - \frac{1}{S}\sum_s \log p(y|\theta^s))$
  - $DIC = -2\log p(y|\hat{\theta}_{Bayes}) + 2p_{DIC}$

Model Comparison Measures

- Watanabe-Akaike information criteria (WAIC)
  - More fully Bayesian approach
    - uses $lppd = \sum_i \log(\frac{1}{S}\sum_s p(y_i|\theta^s))$ as starting point rather than plugging in an estimate
    - alternative definition(s) of estimated number of parameters
    - derived as approximation to cross-validation (discussed below)
  - p_{WAIC} is effective number of parameters
    $p_{WAIC} = 2\sum_{i=1}^{n}(\log(E_{post}\, p(y_i|\theta)) - E_{post}(\log p(y_i|\theta)))$
    estimated as
    $p_{WAIC} = 2\sum_{i=1}^{n}(\log(\frac{1}{S}\sum_s p(y_i|\theta^s)) - \frac{1}{S}\sum_s \log p(y|\theta^s))$
  - $\widehat{elppd}_{WAIC} = lppd - p_{WAIC}$
  - $WAIC = -2\ lppd + 2\ p_{WAIC}$
- another (often better) expression for p_{WAIC} is in the text
- WAIC relies on pointwise calculations
  (others don't because they use point estimate for θ)

Model Comparison Measures

- Bayesian information criteria (BIC)
  - You may have heard of BIC (or SBC)
  - Often provided with AIC
  - $BIC = -2 \log p(y|\hat{\theta}) + k \log n$ (for some estimate, often MLE)
  - Motivation is different
  - BIC derived as an approximation to marginal probability density under the model, $p(y)$, not predictive accuracy
  - Relevant to Bayes factors (discussed below) but not here

Model Comparison Measures

- Leave-one-out cross-validation (LOOCV)
  - Cross validation
    - idea is to partition data into training y_{train} and holdout $y_{holdout}$ data sets
    - model is fit to training data yielding posterior distribution $p_{train}(\theta) = p(\theta|y_{train})$
    - fit evaluated by examining $\log p_{train}(y_{holdout}) = \log \int p(y_{holdout}|\theta) p_{train}(\theta) d\theta$
    - typically estimated by simulations from $p_{train}(\theta)$
  - LOOCV is the special case with n repetitions, each having holdout set equal to a single point
  - More details on the next slide

Model Comparison Measures

- Leave-one-out cross-validation (LOOCV)
  - Define $p_{post(-i)} = p(\theta | y_{(-i)})$
  - Assume we have posterior simulations from each $p_{post(-i)}$, denoted $\theta^{is}, s = 1, \ldots, S$
  - $lppd_{loo-cv} = \sum_{i=1}^{n} \log p_{post(-i)}(y_i)$, calculated as $\sum_{i=1}^{n} \log(\frac{1}{S} \sum_{s=1}^{S} p(y_i | \theta^{is}))$
  - Slight bias because this estimates predictive accuracy of model based on $n - 1$ observations (rather than n)
  - Bias correction addressed in text but not usually applied
  - Can define effective number of parameters (in analogy with other approaches), $p_{loo-cv} = lppd - lppd_{loo-cv}$
  - Then (trivially), $lppd_{loo-cv} = lppd - p_{loocv}$

Model selection / Bayes factors

- Model selection is a limiting case of model comparison in which the goal is to formally decide between the models
- Suppose there are two competings models M_1 and M_2 for a data set
  - different prior distns $p_1(\theta_1)$ and $p_2(\theta_2)$
  - different data models $p_1(y|\theta_1)$ and $p_2(y|\theta_2)$
  - note θ_1 and θ_2 may be of different dimension
- Consider a full Bayesian analysis
  - begin with prior probability $p(M_1) = 1 - p(M_2)$
  - then posterior odds of M_1 relative to M_2 are

$$\frac{p(M_1|y)}{p(M_2|y)} = \frac{p(y|M_1)}{p(y|M_2)} \frac{p(M_1)}{p(M_2)}$$

  - posterior odds are the product of prior odds and a form of likelihood ratio $p(y|M_1)/p(y|M_2)$
  - the ratio $p(y|M_1)/p(y|M_2)$ is known as the Bayes factor
  - it is a measure of how much the data changes the odds in favor of M_1 vs M_2

Bayes Factors

- Bayes factor of model 1 relative to model 2

$$BF_{12} = \frac{p(y|M_1)}{p(y|M_2)} = \frac{\int p(y|\theta_1, M_1)p(\theta_1|M_1)\, d\theta_1}{\int p(y|\theta_2, M_2)p(\theta_2|M_2)\, d\theta_2}$$

  - notation: M_1 and M_2 are not events they merely identify models
  - Bayes factor is only defined when the marginal density of y under each model is proper
    (requires a proper prior distn)

Bayes Factor
Computation

- To compute Bayes factors we need to be able to compute marginal likelihoods

$$p(y) = \int p(y|\theta)p(\theta)\, d\theta$$

- There are a number of approaches
- Simple Monte Carlo approach
  - simplest concept but doesn't work very well
  - draw G values of θ from $p(\theta)$, call them $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(G)}$
  - $\hat{p}(y) = \frac{1}{G}\sum_{g=1}^{G} p(y|\theta^{(g)})$
  - problem: prior distn may not have probability where $p(y|\theta)$ is substantial $\rightarrow$ poor estimate

Bayes Factor
Computation (cont'd)

- Alternative Monte Carlo approach
  - consider the following identity (true for any pdf $h(\theta)$)

$$p(y)^{-1} = \int \frac{h(\theta)}{p(y|\theta)p(\theta)} p(\theta|y) d\theta$$

  - draw G values of θ from $p(\theta|y)$
  - $\hat{p}(y) = \left[\frac{1}{G}\sum_{g=1}^{G} \frac{h(\theta^{(g)})}{p(y|\theta^{(g)})p(\theta^{(g)})}\right]^{-1}$
  - $h(\theta)$ could be prior distribution or normal approx to the posterior distn
  - problem: not a stable calculation because of the possibility of small numbers in the denom

Bayes Factor
Computation (cont'd)

- Chib's marginal likelihood method
  - note that $p(y) = p(y|\theta)p(\theta)/p(\theta|y)$
  - idea: evaluate above at one value of θ, say the posterior mean or the posterior mode
  - numerator terms are easy
  - need to estimate denominator at chosen θ
  - can use a density estimate derived from a posterior sample
  - Chib proposes an alternative approach using Gibbs sampling
    - suppose target is $p(\theta^*|y)$ with $\theta = (\theta_1, \theta_2)$
    - then $p(\theta_1^*, \theta_2^*|y) = p(\theta_1^*|y)p(\theta_2^*|\theta_1^*, y)$
    - we know the last term (since we have the density available for Gibbs samling)
    - we can estimate the first from available posterior draws as $\frac{1}{N}\sum_{i=1}^{N} p(\theta_1^*|\theta_2^{(i)}, y)$
    - Chib shows how to generalize this to more components of θ

Bayes Factor

Improper prior distributions

- Consider $y|\theta \sim N(\theta, 1)$ with $p(\theta) \propto 1$

$$p(y) \propto \int \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-\theta)^2} d\theta = 1$$

- Looks OK **but** $p(y) = 1$ for $y \in (-\infty, \infty)$ is not a valid marginal distn
- Ideas:
  - approx improper prior with proper prior (Unif$(-c, c)$) but BF is very sensitive to choice of c
  - partial Bayes factor: use part of the data to build a proper prior distn and then compute BF on the rest of the data, e.g., use y_1 and flat prior to define "new" prior

$$p(\theta) = N(\theta|y_1, 1)$$

    and then can define a Bayes Factor for $y_2, \ldots, y_n$
  - fractional Bayes factor

Bayes Factor
Asymptotic approximation

- If sample size n is large, then

$$\log(BF) \approx log(p(y|\hat{\theta}_2, M_2)) - log(p(y|\hat{\theta}_1, M_1)) - \frac{1}{2}(d_1 - d_2)log(n)$$

where

  - $\hat{\theta}_i$ = posterior mode under M_i ($i = 1, 2$)
  - d_i = dimension of the parameter space of M_i

- Equivalent to ranking models based on the BIC (Bayes information criterion)

$$\text{BIC} = -log(p(y|\hat{\theta}, M) + \frac{1}{2}d\ log(n)$$

Bayes Factors

Bayes factors and model averaging - I

- Given m models $(M_1, \ldots, M_m)$ with parameter vectors $\theta_1, \ldots, \theta_M)$ and prior probabilities $P(M_1), \ldots, P(M_m)$
- Suppose that each model is used to estimate a quantity of interest Δ (exists in all models)
- This could be a relevant summary for the scientific problem being studied or perhaps a prediction for a future observable quantity
- Then $P(\Delta|y) = \sum_{j=1}^{m} p(\Delta|M_j, y)p(M_j|y)$
- Posterior probability for model j is

$$p(M_j|y) = \frac{p(y|M_j)p(M_j)}{\sum_k p(y|M_k)p(M_k)}$$

- Notes:
  - numerator is just marginal likelihood for model j
  - $p(M_j|y)/p(M_i|y) = BF_{ji}\frac{p(M_j)}{p(M_i)}$
  - can write $p(M_j|y) = p(M_j)/\left(\sum_k BF_{kj}p(M_k)\right)$

Bayes Factors

Bayes factors and model averaging - II

- The previous slide envisions fitting each model separately and is completely general
- If the models are related, e.g., regression models with different predictors from a fixed list, then one can average in a different way
- Build a single "super" model that includes (M_j, θ_j) as parameters and average over this model
- Computation - a single MCMC incorporating all models (reversible jump MCMC)

Classical ideas and Bayesian Inference

- Model selection is closely related to traditional hypothesis testing
- Makes this a good time to check in on some classical ideas and their Bayesian counterparts
- Some general comments on classical/Bayesian
  - Bayesian = classical for some problems (large samples, small number of parameters with noninformative prior distns)
  - Standard methods often correspond to a Bayesian model for some prior (e.g., in hierarchical models we saw that complete pooling and no pooling correspond to specific (extreme) choices of the prior distribution on the random effects)
  - Big differences on some issues (e.g., p-values)
    - p-values are based on probability distribution over possible values of y
    - Bayesian ideas all condition on the single fixed observed y

Classical ideas and Bayesian Inference

- Asymptotics
  - $\hat{\theta}_{MLE}$ is asymptotic efficient and consistent
  - $\hat{\theta}_{post.mode}$ is asymptotic efficient and consistent
- Point estimation
  - optimal Bayes point estimates depend on the specification of a loss function
  - classical inference relies on MLE (or occasionally other estimation strategies)
  - Bayes estimators are not generally unbiased ....
    but then again neither are MLEs
    (recall defn of unbiasedness: $E(\hat{\theta}(y)|\theta) = \theta$)

Classical ideas and Bayesian Inference

- Confidence intervals
  - interpretation of Bayes and frequentist intervals are very different
  - most people want the Bayesian interpretation
- Hypothesis testing
  - Frequentist setup:

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_a : \theta > \theta_0$$

$$\text{p-value} \quad = \quad P(\bar{Y} \text{ is unusually large} | H_0 \text{ is true})$$

    - only assessing H_0 vs data
    - p-value depends on unobserved values
    - likelihood ratio tests work for nested models only

Classical ideas and Bayesian Inference

- Hypothesis testing (cont'd)
  - Bayesian view:
    - need a prior distn $p(\theta)$ under both hypotheses
    - Bayes factor $BF = p(y|H_0)/p(y|H_a)$ where $p(y|H) = \int p(y|\theta, H)p(\theta|H)d\theta$
    - alternative for simple situation (like previous slide), just compute $\Pr(\theta > \theta_o|y)$

Classical ideas and Bayesian Inference

Hypothesis testing - an interesting example

- Discussion due to Morris (JASA 1987)
- Consider binomial sampling: $y|\theta \sim \text{Bin}(n, \theta)$

$$H_0 : \theta \leq 0.5 \qquad H_a)\theta > 0.5$$

n	y	$\hat{\theta}$	t	p-value
20	15	0.750	2.03	0.02
200	115	0.575	2.05	0.02
2000	1064	0.523	2.03	0.02

- Simple Bayesian analysis
  - model: $\hat{\theta} \sim N(\theta, 0.25/n)$ (normal approximation to binomial)
  - prior: $\theta \sim N(0.5, (0.05)^2)$

$$p(\theta > 0.5|y) = \begin{cases} 0.796 & (n = 20) \\ 0.953 & (n = 200) \\ 0.976 & (n = 2000) \end{cases}$$

Classical ideas and Bayesian Inference

- Multiple comparisons
  - e.g., effect of performing many hypothesis tests
  - tempting to say that Bayesian's don't care about multiple comparisons but there is a price to modeling many parameters
- Stopping rules/data collections
  - recall binomial/neg.binomial example
  - more on this later
- Nonparametrics
  - many nonparametric tests/procedures have been developed
  - Bayesian non-parametrics is more and more popular (not covered here)