Introduction to Bayesian

Lin Zhang

Department of Biotatistics
School of Public Health
University of Minnesota

Course Outline

- What is Bayesian? Why Bayesian?
- Single and Multi-Parameter Models
- Hierarchical Model and Bayesian Computation
- Model Checking and Comparison
- Bayesian Linear and Generalized Linear Models
- Bayesian Methods in Various Applications
    - hierarchical variable selection
    - spatial modeling
    - clinical trial design

Frequentist versus Bayesian Statistics

- The goal of statistical inference is, broadly, to answer scientific questions using data, which is often translated into estimating a parameter(s) of interest.
  - Does reducing the nicotine content of cigarettes reduce cigarette comsumption in smokers?
  - What environmental factors are associated with obesity?

Frequentist versus Bayesian Statistics

- The goal of statistical inference is, broadly, to answer scientific questions using data, which is often translated into estimating a parameter(s) of interest.
  - Does reducing the nicotine content of cigarettes reduce cigarette comsumption in smokers?
  - What environmental factors are associated with obesity?
- To a frequentist, these unknown model parameters are fixed, and only estimable by replications of data from some experiment.
- A Bayesian thinks of unknown parameters as random, and thus having distributions (just like the data). We make inference based on distribution of the parameters conditional on the observed data.

Frequentist versus Bayesian Statistics

- The goal of statistical inference is, broadly, to answer scientific questions using data, which is often translated into estimating a parameter(s) of interest.
  - Does reducing the nicotine content of cigarettes reduce cigarette comsumption in smokers?
  - What environmental factors are associated with obesity?
- To a frequentist, these unknown model parameters are fixed, and only estimable by replications of data from some experiment.
- A Bayesian thinks of unknown parameters as random, and thus having distributions (just like the data). We make inference based on distribution of the parameters conditional on the observed data.

◊ We can have a "guess" of the unknown parameters even without data available!

Frequentist versus Bayesian Statistics

- Practice Bayesian thinking:

$$\text{Consider } \theta = \text{proportion of Minnesota children (between the ages of 6 months and 17 years) who will receive influenza vaccinations in year 2020}$$

Bayesian Inference Process

- Three steps to estimate a parameter θ:
  1. Writes down a prior guess, $\pi(\theta)$

     $$\Big\Downarrow + \text{data}, X$$

  2. Obtain the posterior distribution, $p(\theta|X)$
  3. Perform statistical inferences (point and interval estimates, hypothesis tests) by appropriately summarizing the posterior.

Bayesian Inference Process

- Three steps to estimate a parameter θ:
  1. Writes down a prior guess, $\pi(\theta)$

     $$\Big\Downarrow + \text{data}, X$$

  2. Obtain the posterior distribution, $p(\theta|X)$
  3. Perform statistical inferences (point and interval estimates, hypothesis tests) by appropriately summarizing the posterior.
- Note that

  $$\text{posterior information} \geq \text{prior information} \geq 0\,,$$

  with the second "$\geq$" replaced by "$=$" only if the prior is noninformative (which is often uniform, or "flat").

Another example of Bayesian thinking

- From *Business Week*, online edition, July 31, 2001:
  "Economists might note, to take a simple example, that American turkey consumption tends to increase in November. A Bayesian would clarify this by observing that Thanksgiving occurs in this month."

Another example of Bayesian thinking

- From *Business Week*, online edition, July 31, 2001:
  "Economists might note, to take a simple example, that American turkey consumption tends to increase in November. A Bayesian would clarify this by observing that Thanksgiving occurs in this month."
- Data: Plot of turkey consumption by month

Another example of Bayesian thinking

- From *Business Week*, online edition, July 31, 2001:
  "Economists might note, to take a simple example, that American turkey consumption tends to increase in November. A Bayesian would clarify this by observing that Thanksgiving occurs in this month."
- Data: Plot of turkey consumption by month
- Prior:
  - location of Thanksgiving in the calendar
  - knowledge of Americans' Thanksgiving eating habits

Another example of Bayesian thinking

- From *Business Week*, online edition, July 31, 2001:
  "Economists might note, to take a simple example, that American turkey consumption tends to increase in November. A Bayesian would clarify this by observing that Thanksgiving occurs in this month."
- Data: Plot of turkey consumption by month
- Prior:
  - location of Thanksgiving in the calendar
  - knowledge of Americans' Thanksgiving eating habits
- Posterior: Understanding of the pattern in the data!

Yet Another Example

- Suppose you are about to make your first submission to a particular academic journal
- You assess your chances of your paper being accepted (you have an opinion, but the "true" probability is unknown)
- You submit your article and it is accepted!
- Question: What is your revised opinion regarding the acceptance probability for papers like yours?

Yet Another Example

- Suppose you are about to make your first submission to a particular academic journal
- You assess your chances of your paper being accepted (you have an opinion, but the "true" probability is unknown) – your prior guess
- You submit your article and it is accepted! – observed data
- Question: What is your revised opinion regarding the acceptance probability for papers like yours?

Yet Another Example

- Suppose you are about to make your first submission to a particular academic journal
- You assess your chances of your paper being accepted (you have an opinion, but the "true" probability is unknown) – your prior guess
- You submit your article and it is accepted! – observed data
- Question: What is your revised opinion regarding the acceptance probability for papers like yours?
- If you said anything other than "1", you are a Bayesian!

Yet Another Example

- Suppose you are about to make your first submission to a particular academic journal
- You assess your chances of your paper being accepted (you have an opinion, but the “true” probability is unknown) – your prior guess
- You submit your article and it is accepted! – observed data
- Question: What is your revised opinion regarding the acceptance probability for papers like yours?
- If you said anything other than “1”, you are a Bayesian!

Bayes mean revision of estimates

Why Bayesian?

- Incorporate prior information
- The Bayesian approach expands the class of models we can fit to our data, enabling us to handle:
  - repeated measures
  - unbalanced or missing data
  - complex correlations (longitudinal, spatial, or cluster sample) / multivariate data
  - and many other settings that are awkward or infeasible from a classical point of view.
- Ease the interpretation of statistical inference result, e.g.

95% *confidence interval* vs 95% *credible interval*

Bayes can account for structure

County-level breast cancer rates per 10,000 women:

79	87	83	80	78
90	89	92	99	95
96	100	$\star$	110	115
101	109	105	108	112
96	104	92	101	96

- With no direct data for $\star$, what estimate would you use?

Bayes can account for structure

County-level breast cancer rates per 10,000 women:

79	87	83	80	78
90	89	92	99	95
96	100	$\star$	110	115
101	109	105	108	112
96	104	92	101	96

- With no direct data for $\star$, what estimate would you use?
- Is 200 reasonable?

Bayes can account for structure

County-level breast cancer rates per 10,000 women:

79	87	83	80	78
90	89	92	99	95
96	100	$\star$	110	115
101	109	105	108	112
96	104	92	101	96

- With no direct data for $\star$, what estimate would you use?
- Is 200 reasonable?
- Probably not: all the other rates are around 100

Bayes can account for structure

County-level breast cancer rates per 10,000 women:

79	87	83	80	78
90	89	92	99	95
96	100	$\star$	110	115
101	109	105	108	112
96	104	92	101	96

- With no direct data for $\star$, what estimate would you use?
- Is 200 reasonable?
- Probably not: all the other rates are around 100
- Perhaps use the average of the "neighboring" values (again, near 100)

Bayes can account for structure

County-level breast cancer rates per 10,000 women:

79	87	83	80	78
90	89	92	99	95
96	100	$\star$	110	115
101	109	105	108	112
96	104	92	101	96

- With no direct data for $\star$, what estimate would you use?
- Is 200 reasonable?
- Probably not: all the other rates are around 100
- Perhaps use the average of the "neighboring" values (again, near 100)
- You incorporate the structure in your prior guess.

Accounting for structure (cont'd)

- Now assume that data become available for county $\star$: 100 women at risk, 2 cancer cases. Thus

$$rate = \frac{2}{100} \times 10,000 = 200$$

Would you use this value as the estimate?

Accounting for structure (cont'd)

- Now assume that data become available for county $\star$: 100 women at risk, 2 cancer cases. Thus

$$\mathit{rate} = \frac{2}{100} \times 10,000 = 200$$

  Would you use this value as the estimate?

- Probably not: The sample size is very small, so this estimate will be unreliable. How about a compromise between 200 and the rates in the neighboring counties?

Accounting for structure (cont'd)

- Now assume that data become available for county $\star$: 100 women at risk, 2 cancer cases. Thus

$$rate = \frac{2}{100} \times 10,000 = 200$$

  Would you use this value as the estimate?

- Probably not: The sample size is very small, so this estimate will be unreliable. How about a compromise between 200 and the rates in the neighboring counties?
- Now repeat this thought experiment if the county '$\star$' data were 20/1000, 200/10000, ...
- What happened?

Accounting for structure (cont'd)

- Now assume that data become available for county $\star$: 100 women at risk, 2 cancer cases. Thus

$$rate = \frac{2}{100} \times 10,000 = 200$$

  Would you use this value as the estimate?

- Probably not: The sample size is very small, so this estimate will be unreliable. How about a compromise between 200 and the rates in the neighboring counties?
- Now repeat this thought experiment if the county '$\star$' data were 20/1000, 200/10000, ...
- What happened?
  – *Bayesian methods obtain posterior estimates by weighting the data and prior information appropriately, and allow the data to dominate as the sample size becomes large.*

Bayes allow for early stopping

Example from Berry(2006, Nature Reviews 5:27-36):

- A Phase II clinical trial to study the effectiveness of concurrent administration of trastuzumab with standard chemotherapy on HER2/neu breast cancer patients.
- Patients were equally randomized to two arms: (1) standard chemotherapy and (2) standard chemotherapy + transtuzumab. A patient's response to treatment was assessed and called a "pathological complete response" (pCR) if the pathologist found no tumor in the excised tissue after treatment.
- Target accrual was 164 with equal randomization by a frequentist design, and one interim analysis was planned after 82 patients evaluated.

Bayesian Early Stopping (cont'd)

Results from the first 34 patients:

	pCR	Total	pCR%
stand	12	18	67%
stand+trast	4	16	25%

- Accural was slower than expected, $<$ 2 patients per month.

Bayesian Early Stopping (cont'd)

Results from the first 34 patients:

	pCR	Total	pCR%
stand	12	18	67%
stand+trast	4	16	25%

- Accural was slower than expected, $<$ 2 patients per month.
- Problem: Takes 2 or 3 more years for a formal interim analysis!

Bayesian Early Stopping (cont'd)

Results from the first 34 patients:

	pCR	Total	pCR%
stand	12	18	67%
stand+trast	4	16	25%

- Accural was slower than expected, $<$ 2 patients per month.
- Problem: Takes 2 or 3 more years for a formal interim analysis!
- Bayesian predictive probability of (standard frequentist) statistical significance was calculated assuming the trial were to continue to 164 patients, which was 95%.
- Considering the compelling evidence and extremely slow accrual rate, the trial was stopped.

Bayesian Early Stopping (cont'd)

Results from the first 34 patients:

	pCR	Total	pCR%
stand	12	18	67%
stand+trast	4	16	25%

- Accural was slower than expected, $<$ 2 patients per month.
- Problem: Takes 2 or 3 more years for a formal interim analysis!
- Bayesian predictive probability of (standard frequentist) statistical significance was calculated assuming the trial were to continue to 164 patients, which was 95%.
- Considering the compelling evidence and extremely slow accrual rate, the trial was stopped.
- This is an example of Bayesian adaptive design.

Confidence Interval vs Credible Interval

Consider $X_i \sim N(\theta, \sigma^2)$, $i = 1, \ldots n$.

Confidence Interval vs Credible Interval

Consider $X_i \sim N(\theta, \sigma^2)$, $i = 1, \ldots n$.

- The frequentist 95% confidence interval of θ is

$$\delta(x) = \bar{x} \pm 1.96s/\sqrt{n}$$

  - **Interpretation:** If we generate 100 samples of size n and compute $\delta(x)$ for each sample, about 95 of all the obtained intervals will capture the true value of θ.

Confidence Interval vs Credible Interval

Consider $X_i \sim N(\theta, \sigma^2)$, $i = 1, \ldots n$.

- The frequentist 95% confidence interval of θ is

$$\delta(x) = \bar{x} \pm 1.96s/\sqrt{n}$$

  - **Interpretation:** If we generate 100 samples of size n and compute $\delta(x)$ for each sample, about 95 of all the obtained intervals will capture the true value of θ.
  - *Under the frequentist paradigm, the confidence interval $\delta(x)$ is random and depends on (θ, σ^2). Uncertainty is quantified by investigating how $\delta(x)$ would vary in repeated data sampling from the same population.*

Confidence Interval vs Credible Interval

Consider $X_i \sim N(\theta, \sigma^2)$, $i = 1, \ldots n$.

- The frequentist 95% confidence interval of θ is

$$\delta(x) = \bar{x} \pm 1.96 s/\sqrt{n}$$

  - **Interpretation:** If we generate 100 samples of size n and compute $\delta(x)$ for each sample, about 95 of all the obtained intervals will capture the true value of θ.
  - *Under the frequentist paradigm, the confidence interval $\delta(x)$ is random and depends on (θ, σ^2). Uncertainty is quantified by investigating how $\delta(x)$ would vary in repeated data sampling from the same population.*

- For a Bayesian 95% credible interval $\delta(x)$
  - **Interpretation:** Conditional on the observed data, the probability of $\delta(x)$ covering the true value of θ is 0.95.

Confidence Interval vs Credible Interval

Consider $X_i \sim N(\theta, \sigma^2)$, $i = 1, \ldots n$.

- The frequentist 95% confidence interval of θ is

$$\delta(x) = \bar{x} \pm 1.96 s/\sqrt{n}$$

  – **Interpretation:** If we generate 100 samples of size n and compute $\delta(x)$ for each sample, about 95 of all the obtained intervals will capture the true value of θ.

  – *Under the frequentist paradigm, the confidence interval $\delta(x)$ is random and depends on (θ, σ^2). Uncertainty is quantified by investigating how $\delta(x)$ would vary in repeated data sampling from the same population.*

- For a Bayesian 95% credible interval $\delta(x)$

  – **Interpretation:** Conditional on the observed data, the probability of $\delta(x)$ covering the true value of θ is 0.95.

  – *To a Bayesian, data sets which might have been, but were not, observed are irrelevant to making inferences about the unknown parameters. The only data set of any relevance is the one that was actually observed.*

Bayes/frequentist controversy in hypothesis testing

- Frequentist hypothesis testing utilizes a pre-specified test-statistic, which, again, is assumed as random and depends on θ.
- Hypothesis testing is completed by comparing the observed test statistic to the sampling distribution of the test statistic under the null $\rightarrow$ p-value.

Bayes/frequentist controversy in hypothesis testing

- Frequentist hypothesis testing utilizes a pre-specified test-statistic, which, again, is assumed as random and depends on θ.
- Hypothesis testing is completed by comparing the observed test statistic to the sampling distribution of the test statistic under the null $\rightarrow$ p-value.
- However, this could violates the Likelihood Principle!
- **Likelihood Principle:** In making inferences or decisions about θ after $\mathbf{x}$ is observed, all relevant experimental information is contained in the likelihood function for the observed $\mathbf{x}$. Furthermore, two likelihood functions contain the same information about θ if they are proportional to each other as functions of θ

Binomial vs. Negative Binomial

Example due to Pratt (comment on Birnbaum, 1962 *JASA*):
Suppose 12 independent coin tosses: 9H, 3T. Test:

$$H_0 : \theta = 0.5 \text{ vs. } H_A : \theta > 0.5$$

- Two possibilities for $f(x|\theta)$:

Binomial vs. Negative Binomial

Example due to Pratt (comment on Birnbaum, 1962 *JASA*):
Suppose 12 independent coin tosses: 9H, 3T. Test:

$$H_0 : \theta = 0.5 \text{ vs. } H_A : \theta > 0.5$$

- Two possibilities for $f(x|\theta)$:
  - Binomial: $n = 12$ tosses (fixed beforehand)

$$\Rightarrow \qquad X = \#H \sim Bin(12, \theta)$$

$$\Rightarrow \qquad L_1(\theta) = p_1(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} = \binom{12}{9}\theta^9(1-\theta)^3.$$

Binomial vs. Negative Binomial

Example due to Pratt (comment on Birnbaum, 1962 *JASA*):
Suppose 12 independent coin tosses: 9H, 3T. Test:

$$H_0 : \theta = 0.5 \text{ vs. } H_A : \theta > 0.5$$

- Two possibilities for $f(x|\theta)$:
  - Binomial: $n = 12$ tosses (fixed beforehand)

$$\Rightarrow \qquad X = \#H \sim Bin(12, \theta)$$

$$\Rightarrow \qquad L_1(\theta) = p_1(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} = \binom{12}{9}\theta^9(1-\theta)^3.$$

  - Negative Binomial: Flip until we get $r = 3$ tails

$$\Rightarrow \qquad X \sim NB(3, \theta)$$

$$\Rightarrow \qquad L_2(\theta) = p_2(x|\theta) = \binom{r+x-1}{x}\theta^x(1-\theta)^r = \binom{11}{9}\theta^9(1-\theta)^3.$$

Binomial vs. Negative Binomial

Example due to Pratt (comment on Birnbaum, 1962 *JASA*):
Suppose 12 independent coin tosses: 9H, 3T. Test:

$$H_0 : \theta = 0.5 \text{ vs. } H_A : \theta > 0.5$$

- Two possibilities for $f(x|\theta)$:
  - Binomial: $n = 12$ tosses (fixed beforehand)

$$\Rightarrow \qquad X = \#H \sim Bin(12, \theta)$$

$$\Rightarrow \qquad L_1(\theta) = p_1(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} = \binom{12}{9}\theta^9(1-\theta)^3.$$

  - Negative Binomial: Flip until we get $r = 3$ tails

$$\Rightarrow \qquad X \sim NB(3, \theta)$$

$$\Rightarrow \qquad L_2(\theta) = p_2(x|\theta) = \binom{r+x-1}{x}\theta^x(1-\theta)^r = \binom{11}{9}\theta^9(1-\theta)^3.$$

- Adopt the rejection region, "Reject H_0 if $X \geq c$."

Binomial vs. Negative Binomial

- *p*-values:
  - $\alpha_1 = P_{\theta=\frac{1}{2}}(X \geq 9) = \Sigma_{j=9}^{12}\binom{12}{j}\theta^j(1-\theta)^{12-j} = \mathbf{.075}$
  - $\alpha_2 = P_{\theta=\frac{1}{2}}(X \geq 9) = \Sigma_{j=9}^{\infty}\binom{2+j}{j}\theta^j(1-\theta)^{3} = \mathbf{.0325}$

Binomial vs. Negative Binomial

- *p*-values:
  - $\alpha_1 = P_{\theta=\frac{1}{2}}(X \geq 9) = \Sigma_{j=9}^{12} \binom{12}{j} \theta^j (1-\theta)^{12-j} = \mathbf{.075}$
  - $\alpha_2 = P_{\theta=\frac{1}{2}}(X \geq 9) = \Sigma_{j=9}^{\infty} \binom{2+j}{j} \theta^j (1-\theta)^3 = \mathbf{.0325}$
- So at $\alpha = .05$, two different decisions! Violates the Likelihood Principle, since $L_1(\theta) \propto L_2(\theta)$!!

Binomial vs. Negative Binomial

- <u>p-values:</u>
  - $\alpha_1 = P_{\theta=\frac{1}{2}}(X \geq 9) = \Sigma_{j=9}^{12}\binom{12}{j}\theta^j(1-\theta)^{12-j} = \mathbf{.075}$
  - $\alpha_2 = P_{\theta=\frac{1}{2}}(X \geq 9) = \Sigma_{j=9}^{\infty}\binom{2+j}{j}\theta^j(1-\theta)^3 = \mathbf{.0325}$
- So at $\alpha = .05$, two different decisions! Violates the Likelihood Principle, since $L_1(\theta) \propto L_2(\theta)$!!
- What happened?

Binomial vs. Negative Binomial

- p-values:
  - $\alpha_1 = P_{\theta=\frac{1}{2}}(X \geq 9) = \Sigma_{j=9}^{12} \binom{12}{j} \theta^j (1-\theta)^{12-j} = \mathbf{.075}$
  - $\alpha_2 = P_{\theta=\frac{1}{2}}(X \geq 9) = \Sigma_{j=9}^{\infty} \binom{2+j}{j} \theta^j (1-\theta)^3 = \mathbf{.0325}$
- So at $\alpha = .05$, two different decisions! Violates the Likelihood Principle, since $L_1(\theta) \propto L_2(\theta)$!!
- What happened? The probability of the unpredicted and non-occurring set $X \geq 10$ has been used as evidence against H_0 !
- Jeffreys (1961): *"...a hypothesis which may be true may be rejected because it has not predicted observable results which have not occurred."*

Conditional (Bayesian) Perspective

- Always condition on data which has actually occurred; the long-run performance of a procedure is of (at most) secondary interest. Fix a prior distribution $\pi(\boldsymbol{\theta})$, and use Bayes' Theorem (1763):

$$
\begin{array}{rcl}
p(\boldsymbol{\theta}|\mathbf{x}) & \propto & f(\mathbf{x}|\theta)\pi(\boldsymbol{\theta}) \\
(\text{“posterior} & \propto & \text{likelihood} \times \text{prior”})
\end{array}
$$

Conditional (Bayesian) Perspective

- Always condition on data which has actually occurred; the long-run performance of a procedure is of (at most) secondary interest. Fix a prior distribution $\pi(\boldsymbol{\theta})$, and use Bayes' Theorem (1763):

$$\begin{array}{rcl} p(\boldsymbol{\theta}|\mathbf{x}) & \propto & f(\mathbf{x}|\theta)\pi(\boldsymbol{\theta}) \\ (\text{“posterior} & \propto & \text{likelihood} \times \text{prior”}) \end{array}$$

- Thus, the Bayesian formalism strictly follows the **Likelihood Principle**.
  - Given a pre-specified prior, the coin-tossing data give the same posterior distributions for both *Binomial* and *Negative Binomial* model assumptions, and thus the same inferential conclusions!

Advantages to Bayesian Inference

- Provides an intuitive approach to specifying complex models
- Ability to formally incorporate prior information
- The reason for stopping experimentation does not affect the inference
- Intuitive interpretation of results (e.g. confidence intervals)
- Inferences are conditional on the actual data
- Does not rely on asymptotics – all calculations are exact!

Disadvantages to Bayesian Inference

- Can be dependent on the prior distribution! Two experimenters could get different answers with the same data!
  - Questions: How to pick the prior $\pi(\theta)$? How to control influence of the prior? How to get objective results (say, for a court case, scientific report,....)?

- No direct connection with Type-I error rate (which regulators care about).
  - Although, we will see that many Bayesian procedures have good frequentist properties.

- Can require extensive and time-consuming computational algorithms.
  - But computing keeps improving: MCMC methods, `BUGS`, `JAGS`, `Stan` software....