Bayesian Statistics
Note 1
Bayesian Paradigm, Prior and Posterior distributions

Keunbaik Lee

Sungkyunkwan University

Some preliminary Q & A I

- What is the philosophical difference between *classical ("frequentist")* and *Bayesian* statistics?
  - To a frequentist, unknown model parameters are *fixed* and unknown, and only estimable by replications of data from some experiments.
  - A Bayesian thinks of parameters as *random*, and thus having distributions (just like the data). We can thus think about unknowns for which no reliable frequentist experiment exists, e.g.

$$\theta = \text{ proportion of US men with untreated atrial fibrillation}$$

Some preliminary Q & A II

- How does it work?
  - A Bayesian writes down a *prior* guess for θ, $p(\theta)$, then combines this with the information that the data X provide to obtain the *posterior* distribution of θ, $p(\theta|X)$. All statistical inferences (point and interval estimates, hypothesis tests) then follow as appropriate summaries of 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").

Some preliminary Q & A III

- Is the classical approach "*wrong*"?
  - While a "hardcore" Bayesian might say so, it is probably more accurate to think of classical methods as merely "limited in scope"!
  - The Bayesian approach expands the class of models we can fit to our data, enabling us to handle
    - repeated measures
    - unbalanced or missing data
    - nonhomogenous variances
    - multivariate data

      - and many other settings that are awkward or infeasible from a classical point of view
  - The approach also eases the interpretation of and learning from those models once fit.

Bayes means revision of estimates I

Humans tend to be Bayesian in the sense that most revise their opinions about uncertain quantities as more data accumulate. For 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?
- If you said anything other than "1", you are a Bayesian!

Bayes can account for structure I

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

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

- With no direct data for *, 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 II

- Now assume that data become available for county *: 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 * data were 20/1000, 200/10000, ...

Bayes can account for structure III

- Bayes and empirical Bayes methods can incorporate the structure in the data, weight the data and prior information appropriately, and allow the data to dominate as the sample size becomes large.

Motivating Example I

- From Berger and Berry (1988, Amer. Scientist): Consider a clinical trial to study the effectiveness of Vitamin C in treating the common cold.
- Observations are matched pairs of subjects (twins?), half randomized (in "double bind" fashion) to vitamin C, half to placebo. We count how many pairs had C giving superior relief after 48 hours.

Motivating Example II

- **Two Designs**
  - Design # 1: Sample $n = 17$ pairs, and test

$$H_0 : P(C \text{ better}) = \frac{1}{2} \quad vs. \quad H_A : P(C \text{ better}) \neq \frac{1}{2}$$

    Suppose we observe $x = 13$ preferences for C. Then

$$p - value = P(X \geq 13 \text{ or } X \leq 4) = .049$$

    So if $\alpha = .05$, stop and reject H_0.
  - Design # 2: Sample $n_1 = 17$ pairs. Then:
    - if $x_1 \geq 13$ or $x_1 \leq 4$, stop.
    - otherwise, sample an additional $n_2 = 27$ pairs. Reject H_0 if $X_1 + X_2 \geq 29$ or $X_1 + X_2 \leq 15$.

Motivating Example III

- We choose this second stage since under H_0,

$$P(X_1 + X_2 \geq 29 \text{ or } X_1 + X_2 \leq 15) = .049$$

  - the same as Stage 1!
- Suppose we again observe $X_1 = 13$. Now:

$$\begin{aligned} p - value &= P(X_1 \geq 13 \text{ or } X_1 \leq 4) \\ &\quad + P(X_1 + X_2 \geq 29 \text{ and } 4 < X_1 < 13) \\ &\quad + P(X_1 + X_2 \leq 15 \text{ and } 4 < X_1 < 13) \\ &= .085 \quad \leftarrow \text{ no longer significant at } \alpha = .05! \end{aligned}$$

- Yet the observed data was exactly the same; all we did was contemplate a second stage (no effect on data), and it changed our answer!

Additional Q & A I

- Q: What if we kept adding stages?
  A: p-value $\to 1$, even though x_1 still 13!
- : Q: So are p-values really "objective evidence"?
  A: No, since extra info (like design) critical!
- What about unforeseen events?
  Example: First 5 patients develop an allergic reaction to the treatment - trial is stopped by clinicians.
  Can a frequentist analyze thses data?
  A: No: This aspect of design wasn't anticipated, so p-values not computable!
- Q: Can a Bayesian?
  A: (obviously) Yes - as we shall see....

Binomial vs. Negative Binomial I

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

$$H_0 : \theta = \frac{1}{2} \quad vs. \quad H_A : \theta > \frac{1}{2}.$$

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

$$=> \quad X = \#H \sim Bin(12, \theta)$$

$$L_1(\theta) = p_1(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} = \binom{12}{9}\theta^9(1-\theta)^2$$

Binomial vs. Negative Binomial II

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

$$=> \quad X \sim NB(3, \theta)$$

$$L_2(\theta) = p_2(x|\theta) = \binom{4+x-1}{x}\theta^x(1-\theta)^r = \binom{11}{9}\theta^9(1-\theta)^3$$

- Adopt the rejection region: "Reject H_0 if $X \geq c$,"
- $p-$values:
  - $\alpha_1 = P_{\theta=\frac{1}{2}}(X \geq 9) = \sum_{j=9}^{12}\binom{12}{j}\theta^j(1-\theta)^{12-j} = 0.075$
  - $\alpha_2 = P_{\theta=\frac{1}{2}}(X \geq 9) = \sum_{j=9}^{\infty}\binom{2+j}{j}\theta^j(1-\theta)^{12-j} = 0.0325$
- So at $\alpha = 0.05$, two different decisions! Violates the Likelihood Principle, since $L_1(\theta) \propto L_2(\theta)$!!
- What happened? Besides the observed $x = 9$, we also took into account the "more extreme" $X \geq 10$.

Binomial vs. Negative Binomial III

- Jeffreys (1961): "... a hypothesis which may be true may be rejected because it has not predicted observed results which have not occurred."
- In our example, the probability of the unpredicted and non-occurring set $X \geq 10$ has been used as evidence against H_0.

Bayesians have a problem with $p-$values I

- $p = P$(results as surprising as you got *or more so*).
  The "or more so" part gets us in trouble with:
  - *The Likelihood Principle:* When making decisions, only the observed data can play a role.

  This can lead to bad decisions (especially, false positives).
- Are $p-$values at least more objective, because they are not influenced by any prior distribution?
  - No, because they are influenced crucially by the design of the experiment, which determines the reference space of events for the calculation.

Bayesians have a problem with $p-$values II

- Purely practical problems also plague $p-$values:
  - Ex: Unforseen events: First 5 patients develop a rash, and the trial is stopped by clinicians.
    $=>$ this aspect of design was not anticipated, so strictly speaking, the $p-$value is not computable.
- 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 $p(\theta)$, and use Bayes Theorem (1763):

$$\begin{aligned} p(\theta|x) &\propto f(x|\theta)p(\theta) \\ (\text{Posterior} &\propto \text{likelihood} \times \text{prior}) \end{aligned}$$

Bayesians have a problem with $p-$values III

- Indeed, it often turns out that using the Bayesian formalism with relatively vague priors produces procedures which perform well using traditional *frequentist* criteria (e.g., low mean squared error over repeated sampling)!

Formal Probability Basics I

- We like to think of "Probability formally as a *function* that assigns a real number to an *event*"
- We denote by H the basic experimental context in which events will arise. Very often H will be a *hypothesis*. Its complement, is denoted by H^c or $\bar{H}$.
- Let E and F be any events that might occur under H. Then a probability function $P(E|H)$ (spoken as E given H) is defined as:

P1 $0 \leq P(E|H) \leq 1$ for all E, H.
P2 $P(H|H) = 1$ and $P(H^c|H) = 0$.

Formal Probability Basics II

P3 $P(E \cup F|H) = P(E|H) + P(F|H)$ whenever $E \cap F \cap H = \{\emptyset\}$ - whenever impossible for any two of the events E, F and H to occur. Usually consider: $E \cap F = \{\emptyset\}$ and say they are *mutually exclusive*.

- If E is an event, then we denote its complement by E^c. Since $E \cap E^c = \emptyset$, we have from P3:

$$P(E^c) = 1 - P(E).$$

Formal Probability Basics III

- *Conditional Probability* of E given F:

$$P(E|F \cap H) = \frac{P(E \cap F|H)}{P(F|H)}$$

  We will often write EF for $E \cap F$.

- Compound probability rule: write the above as

$$P(E|FH)P(F|H) = P(EF|H)$$

Formal Probability Basics IV

- Independent Events: E and F are said to be *independent* (we will write $E \perp F$) if the occurrence of one does not imply the occurrence of the other. Then, $P(E|FH) = P(E|H)$ and we have the following *multiplication rule*:

$$P(EF|H) = P(E|H)P(F|H).$$

- Marginalization: We can express $P(E|H)$ by "marginalizing" over the event F:

$$\begin{aligned} P(E|H) &= P(EF|H) + P(EF^c|H) \\ &= P(F|H)P(E|FH) + P(F^c|H)P(E|F^cH). \end{aligned}$$

Statistical Inference I

- Object: To draw conclusions from data about quantities that are not observed.
  ex) Clinical trial of a new cancer drug for estimating the five year survival probability in a population as compared to a standard drug administered to a similar but different population. It is neither feasible nor ethically accepted to experiment.
- Two kinds of unobserved quantities:
  1. potentially observed quantities. ex) future observation of a process.
  2. quantities that are not directly observed; parameters that govern the hypothetical process leading to the observed data. ex) regression coefficients etc.

Statistical Inference II

- Classical Inference: Use a model and then estimate or test hypothesis for unknown parameters of the model.
- Bayesian Inference:
  1. Use probability models for both the observed and unobserved quantities.
  2. The key point in a Bayesian analysis is explicit use of models in making inferences.

Statistical Inference III

- Three key steps in a Baysian analysis
  1. Set up a full proability model for both observables and nonobservables.
  2. Conditioning on the observed data: Find the conditional distribution of the unobserved quantities from the unobserved quantities.
  3. Model diagnostics: Evaluate the fit of the model. Is the model sensible? If not, go for some other model.
- The importance of Bayesian thinking
  - Common-sense interpretation of statistical conclusion
  - The pragmatic advantages of a Bayesian framework.

Statistical Inference IV

- In the classical frequentist framework, the notion of a confidence interval is based on the idea of repeatability of an experiment. For example, a 95% confidence interval for a parameter of interest means that if the experiment is performed under similar conditions a large number of times, then the interval will contain the true parameter approximately 95% of times.
- In contrast, a Bayesian probability interval for a parameter of interest has direct interpretation in the sense that it is based on the conditional distribution of the parameter of interest given the observed data.

Statistical Inference V

- Exchangeability and Explanatory Variables
  - Often in statistics we assume the observations to be iid. A more general assumption (which includes iid as a special case) is that the observations are **exchangeable**. i.e., the joint pdf $p(y_1, \cdots, y_n)$ of the observations is invariant under any one of the $n!$ permutation of the arguments.
  - A more realistic model is the regression model where the response variables y_i's can be explained in terms of certain covariates, say x_i. For example, when the response is income of an individual, a useful covariate is education of the person.

Bayes Theorem I

$$
\begin{aligned}
P(A|B) &= \frac{P(AB)}{P(B)} = \frac{P(AB)}{P(AB) + P(A^c B)} \\
&= \frac{P(A)P(B|A)}{P(A)P(B|A) + P(A^c)P(B|A^c)}
\end{aligned}
$$

Examples I

		Diagnosis	
		Psychosis (D)	Not Psychosis(D^c)
Test	Positive (T)	374	10
	Negative (T^c)	14	680
	Total	388	690

Sensitivity: $P(T|D) = 374/388 = .964$;
Specificity: $P(T^c|D^c) = 680/690 = .986$
Positive Predictive Value (PPV): $P(D|T)$
Negative Predictive Value (NPV): $P(D^c|T^c)$

$$P(D|T) = \frac{P(D)P(T|D)}{P(D)P(T|D) + P(D^c)P(T|D^c)} = \frac{.964P(D)}{.964P(D) + .014P(D^c)}$$

where $P(D) = 388/(388 + 690)$ is correct if the cases are samples from a bigger population.

On the other hand, if these totals are predetermined, estimates may not be very correct.

Examples II

- In store A, the ratio of sales of products x and y is 6:4, and in store B, the ratio of sales of products x and y is 3:7. The sales ratio of stores A and B is said to be 4:6. What is the probability that any product x and product y are from store A?
  Sol.) Let H_1 and H_2 be respectively sales of stores A and B. We also let D be the sales of product x. Then $P(H_1) = 4/10$, $P(H_2) = 6/10$, $P(D|H_1) = 6/10$, and $P(D|H_2) = 3/10$. Now we have

$$\begin{aligned} P(H_1|D) &= \frac{P(H_1)P(D|H_1)}{P(H_1)P(D|H_1) + P(H_2)P(D|H_2)} \\ &= \frac{4/10 \times 6/10}{4/10 \times 6/10 + 6/10 \times 3/10} = 24/42. \end{aligned}$$

  Similarly,

$$\begin{aligned} P(H_1|D^c) &= \frac{P(H_1)P(D^c|H_1)}{P(H_1)P(D^c|H_1) + P(H_2)P(D^c|H_2)} \\ &= \frac{4/10 \times 4/10}{4/10 \times 4/10 + 6/10 \times 7/10} = 16/58. \end{aligned}$$

Examples III

- The chance of getting any particular cancer is said to be 0.1%. In cancer screening, people who have the cancer have a 95% chance of getting a positive test. And there is a 2% chance that a healthy person will be misdiagnosed for positive. What is the probability of getting this cancer when diagnosed as benign in these cancer tests?
  Sol.) Let H_1 and H_2 be respectively events that a subject has a cancer and does not have a cancer. We also let D be positive diagnosis of the cancer. Then $P(H_1) = 0.001$, $P(H_2) = 0.999$, $P(D|H_1) = 0.95$, and $P(D|H_2) = 0.02$. We also have

$$\begin{aligned} P(H_1|D) &= \frac{P(H_1)P(D|H_1)}{P(H_1)P(D|H_1) + P(H_2)P(D|H_2)} \\ &= \frac{0.001 \times 0.95}{0.001 \times 0.95 + 0.999 \times 0.02} = 0.045. \end{aligned}$$

Examples IV

- A factory produces certain products. If the machine is in good condition, the defect rate is 7%; otherwise, the defect rate is 30%. On average, the rate of the machine being good is 90%. If the machine is not in good condition, stop production and repair the machine. One day, the results of production are as follows.

$$G, G, B, G, B, G, G, G, G: \quad G - Good, \ B - Bad$$

Obtain the posterior probability of the machine's condition.
Sol.) Let H_1 and H_2 be respectively events the machine is good and not good. We also let G be a good product. Then $P(H_1) = 0.90 = 1 - P(H_2)$, $P(G|H_1) = 1 - 0.07 = 0.93$, and $P(G|H_2) = 1 - 0.30 = 0.70$. We first calculate the posterior probability that the machine is in good given the first product is good (G):

$$\begin{aligned} P(H_1|G) &= \frac{P(H_1)P(G|H_1)}{P(H_1)P(G|H_1) + P(H_2)P(G|H_2)} \\ &= \frac{0.90 \times 0.93}{0.90 \times 0.93 + 0.10 \times 0.70} = 0.923. \end{aligned}$$

Now we sequentially calculate the posterior probabilities that the machine is in good given $G, G, B, G, B, G, G, G, G$:

0.923, 0.941, 0.788, 0.831, 0.535, 0.604, 0.670, 0.729, 0.782 (See R code)

Examples V

- Monty Hall Problem
  Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice? Let A_1, A_2 and A_3 be the events that the car is behind the door 1, 2, and 3, respectively. We also let the event O that the host open the door 3.

$$P(A_1) = P(A_2) = P(A_3) = 1/3,$$

Examples VI

and

$$P(O|A_1) = 1/2, \quad P(O|A_2) = 1, \quad P(O|A_3) = 0.$$

Then by Bayes Theorem,

$$P(A_1|O) = \frac{P(O|A_1)P(A_1)}{\sum_{i=1}^{3} P(O|A_i)P(A_i)} = 1/3,$$
$$P(A_2|O) = \frac{P(O|A_2)P(A_2)}{\sum_{i=1}^{3} P(O|A_i)P(A_i)} = 2/3.$$

Notations I

- θ: unobservable real or vector quantities
- y: real or vector-valued observation data
- $\tilde{y}$: real or vector-valued unknown but potentially observable quantities
- $P(y|\theta)$: likelihood=conditional pdf of y given θ
  The function $P(y|\theta)$ is a function of θ for fixed y
- $P(\theta)$: prior=marginal pdf of θ
- $P(\theta|y)$: posterior of θ given y (conditional pdf of θ given y)

Notations II

- Bayesain inference is based on

$$P(\theta|y) = \frac{P(\theta, y)}{P(y)} = \frac{P(y|\theta)P(\theta)}{P(y)}$$

or

$$P(\theta|y) \propto P(\theta)P(y|\theta)$$

where $P(y) = \begin{cases} \sum_\theta P(y|\theta)P(\theta), & \text{if } \theta \text{ is discrete;} \\ \int P(y|\theta)P(\theta)d\theta, & \text{if } \theta \text{ is continuous.} \end{cases}$ Thus the Bayesian inference depends only on the likelihood function and the prior.

Example I

Males: 1 X-chromosome and 1 Y-chromosome; Females: 2 X-chromosomes
Each child inherits one chromosome from each parent. Hemophilia is a disease that exhibits X-chromosome linked recessive inheritance. A male who inherits the gene which causes the disease on the X-chromosome is affected, while a female who inherits the gene in one of her X-chromosomes is not affected.

Q1: Suppose there is a female who has an affected brother, but a non-affected father, what is the probability that the female is a carrier of the gene?

Sol. Let $\theta = \begin{cases} 1, & \text{if the female is a carrier of the the gene;} \\ 0, & \text{if the female is not a carrier of the the gene.} \end{cases}$ Without any further information, $P(\theta = 1) = P(\theta = 0) = 1/2$. In the Bayesian terminology, this is the law of equal ignorance. Also, this is the **prior**.

Example II

Q2: Suppose we provide the additional information that the woman has a son who is not affected. What is the conditional probability that the woman is a carrier of the gene?
Sol.: Let

$$y_1 = \begin{cases} 1, & \text{if the son is affected;} \\ 0, & \text{if the son is not affected.} \end{cases}$$

Find

$$\begin{aligned} & P(\theta = 1 | y_1 = 0) \\ = & \frac{P(y_1 = 0|\theta = 1)P(\theta = 1)}{P(y_1 = 0|\theta = 1)P(\theta = 1) + P(y_1 = 0|\theta = 0)P(\theta = 0)} \\ = & \frac{\frac{1}{2} \cdot \frac{1}{2}}{\frac{1}{2} \cdot \frac{1}{2} + 1 \cdot \frac{1}{2}} = 1/3. \end{aligned}$$

Example III

Q3: Suppose further that the woman has a second son who is not affected. What is the conditional probability that the woman is a carrier of the gene?
Sol.) Then we calculate

$$
\begin{aligned}
& P(\theta = 1 | y_1 = 0, y_2 = 0) \\
= & \frac{P(y_1 = 0, y_2 = 0 | \theta = 1) P(\theta = 1)}{P(y_1 = 0, y_2 = 0 | \theta = 1) P(\theta = 1) + P(y_1 = 0, y_2 = 0 | \theta = 0) P(\theta = 0)} \\
= & 1/5
\end{aligned}
$$

- The idea is to keep updating the posterior, and use yesterday's posterior as today's prior.

Example IV

- Note also that y_1 and y_2 are conditionally independent given θ (Assumption).
- Are they marginally independent? Answer: No

$$P(y_1 = 0, y_2 = 0) = 5/8; \quad P(y_1 = 0) = 3/4 = P(y_2 = 0)$$

- Although y_1 and y_2 are assumed to be conditionally independent given θ, they are not marginally independent.

Posterior predictive distribution I

- Notations:
  y =observed data; θ =unknown parameter
  $\tilde{y}$ =future observation
  $P(y|\theta)$ =sampling distribution of y given θ or likelihood of θ
  $P(\theta)$ =marginal pdf of θ or prior for θ
  $P(\theta|y)$ =conditional pdf of θ given y or posterior pdf of θ
  $P(\tilde{y}|y)$ =posterior pdf of $\tilde{y}$ given y
  $P(y)$ = marginal pdf of y $= \int P(y|\theta)P(\theta)d\theta$

Posterior predictive distribution II

- Find $P(\tilde{y}|y)$. Then

$$\begin{aligned} P(\tilde{y}|y) &= \frac{P(\tilde{y}, y)}{P(y)} \\ &= \int P(\tilde{y}|y, \theta) P(\theta|y) d\theta. \end{aligned}$$

If we further assume that $\tilde{y}$ and y are conditionally independent given θ, then the above further simplies to

$$P(\tilde{y}|y) = \int P(\tilde{y}|\theta) P(\theta|y) d\theta.$$

Posterior predictive distribution III

- Remark: The function $P(y|\theta)$ when regarded as a function of θ for fixed y is called the likelihood function. Thus, the Bayesian inference depends only on the likelihood function and the prior.