STATS 225: Bayesian Analysis
Introduction

Babak Shahbaba

Department of Statistics, UCI

Why Bayesian Analysis?

The role of statistics in science

- Statistical methods are mainly inspired by applied scientific problems.
- The overall goal of statistical analysis is to provide a robust framework for designing scientific studies, collecting empirical evidence, and analyzing the data, in order to understand unknown phenomena, answer scientific questions, and make decisions.
- To this end, we rely on the observed data as well as our *domain knowledge*.

Domain knowledge

- Our domain knowledge, which we refer to as *prior* information, is mainly based on previous empirical evidence.
- For example, if we are interested in the average normal body temperature, we would of course measure body temperature of a sample of subjects from the population, but we also know, based on previous data, that this average is a number close to $98.6\,^{\circ}\mathrm{F}$.
- In this case, our prior knowledge asserts that values around 98 are more plausible compared to values around 90 or 110.

Objective vs. subjective

- We could of course attempt to minimize our reliance on prior information.
- Most frequentist methods follow this principle and use the domain knowledge to decide which characteristics of the population are relevant to our scientific problem (e.g., we might not include height as a risk factor for cancer), but avoid using priors when making inference.
- Note that this should not give us the illusion that our frequentist methods are entirely objective.

Easy to understand, hard to implement

- Bayesian methods on the other hand provide a mathematical framework to incorporate prior knowledge in the process of making inference.
- This is based on the philosophy that if the prior is in fact informative, this should lead to more accurate inference and better decisions. Also, the way we incorporate our prior knowledge in the analysis should be explicit.
- The counterargument is that this makes our analysis more prone to mistakes.
- This is of course true! While the underlying concept for Bayesian statistics is quite simple, implementing Bayesian methods might be more difficult compared to their frequentist counterparts.
- Therefore, while Bayesian methods provide a coherent and robust framework for statistical inference, they require a careful specification of models and priors. Additionally, you need strong computational skills to implement them.

Just Likelihood

Likelihood-based inference

- As mentioned above, we also define the underlying mechanism that generates data, y, using a probability model, $P(y|\theta)$, which depends on the unknown parameter of interest, θ.
- Frequentist methods typically use this probability for inference.
- To estimate model parameters, we can find their values such that the probability of the observed data is maximum.
- For this, we first need to construct the corresponding *likelihood function*, and then maximize the likelihood function with respect to model parameters θ.
- The likelihood function is defined by plugging-in the observed data in the probability distribution and expressing it as a function of model parameters, i.e., $f(\theta, y)$.

Likelihood-based inference

- Under weak regularity conditions, the MLE demonstrates attractive properties as the sample size n increases (i.e., $n \to \infty$). These include its asymptotic normality, consistency, and efficiency.

- We can also use the likelihood function to device standard tests (Wald test, score test, and likelihood ratio test) to perform hypothesis testing.

- Strong Likelihood principle: Denote the observed sample from a random variable, X, with $p_1(x|\theta)$ as $\boldsymbol{x}$, and the observed sample from another random variable, Y, with $p_2(y|\theta)$ as $\boldsymbol{y}$. If the corresponding likelihood functions are proportional, $f_1(\theta, \boldsymbol{x}) \propto f_2(\theta, \boldsymbol{y})$, then inference for θ should be the same whether we observe $\boldsymbol{x}$ or $\boldsymbol{y}$.

Violation of the strong likelihood principal

- The following example is discussed in David MacKay's book.
- A scientist has just received a grant to examine whether a specific coin is fair (i.e., $P(H) = P(T) = 0.5$) or not.
- He sets up a lab and starts tossing the coin. Of course, because of his limited budget, he can only toss the coin a finite number of times.
- He tosses the coin 12 times, of which only 3 are heads.
- He hires a frequentist statistician and ask him to estimate the p-value hoping that the result could be published in one of the journals that only publish if the p-value is less than 0.05!
- The statistician says: "you tossed the coin 12 times and you got 3 heads. The one-sided p-value is 0.07".

Violation of the strong likelihood principal

- The scientist says: "Well, it wasn't exactly like that... I actually repeated the coin tossing experiment until I got 3 heads and then I stopped".
- The statistician say: "In that case, your p-value is 0.03".
- Note that in the first scenario, we use a binomial model, and in the second scenario, we use a negative-binomial model with the following likelihood functions respectively,

$$\begin{aligned} f_1(\theta, x) &= \binom{n}{x} \theta^x (1-\theta)^{n-x} \\ f_2(\theta, x) &= \binom{n-1}{x-1} \theta^x (1-\theta)^{n-x} \end{aligned}$$

  which are proportional.
- We will see the answer by a Bayesian statistician later.

It's all about making decisions

- Gelman and Nolan (2002) proposed the following experiment.
- I am going to keep tossing a coin and you are going to guess the outcome, you would win \$1 every time you guess "head" and the outcome is "head". What would be your strategy?
- Would you change your strategy if I tell you the coin is not a fair coin and the probability of head is only 0.1?

It's all about making decisions

- It is clear that to make decisions, we need more than just probability: we need a measure of loss or gain for each possible outcome.
- For this, we usually use a *loss function* that assigns to each possible outcome a number that represents the cost and the amount of regret (e.g., loss of profit) we endure when that outcome occurs.
- Decision theory plays a key role in Bayesian statistical inference.
- To make decisions, we follow the *expected loss* principle and choose the option whose expected loss is minimum.

A curious case of inadmissibility!

- Consider the following model:

$$\begin{aligned} \boldsymbol{x}|\boldsymbol{\mu} &\sim N_p(\boldsymbol{\mu}, I), \qquad \boldsymbol{x} = (x_1, x_2, \ldots, x_p) \\ \boldsymbol{\mu} &\sim N(0, \tau^2 I) \end{aligned}$$

- A reasonable estimate of $\boldsymbol{\mu}$ is $\hat{\boldsymbol{\mu}}^{(MLE)} = \boldsymbol{x}$, i.e., the maximum likelihood estimator.

- Note that with respect to the squared error loss function, $L(\boldsymbol{\mu}, \delta) = \|\delta - \boldsymbol{\mu}\|^2$, the Bayes risk for this estimator is

$$E[E[L(\boldsymbol{\mu}, \delta)]] = p$$

where the first expectation is with respect to $\boldsymbol{x}$, and the second expectation is with respect to $\boldsymbol{\mu}$.

A curious case of inadmissibility!

- While the above estimator is commonly used (e.g, ANOVA, regression), the statistics community was shocked when Stein and James showed that the following estimator, known as James-Stein estimator, dominates MLE for $p > 2$:

$$\hat{\boldsymbol{\mu}}^{(JS)} = (1 - \frac{p-2}{\|\boldsymbol{x}\|^2})\boldsymbol{x}$$

by proving that

$$E[\|\hat{\boldsymbol{\mu}}^{(JS)} - \boldsymbol{\mu}\|^2] < E[\|\hat{\boldsymbol{\mu}}^{(MLE)} - \boldsymbol{\mu}\|^2]$$

The expectation is with respect to $\boldsymbol{x}$.

- We can show that (Efron, 2010)

$$E[\|\hat{\boldsymbol{\mu}}^{(JS)} - \boldsymbol{\mu}\|^2] = p - E[\frac{(p-2)^2}{\|\boldsymbol{x}\|^2}]$$

for every choice of μ (i.e., regardless of the prior).

A curious case of inadmissibility!

- We will show later that the Bayes estimator has even lower risk.
- Such shrinkage estimators are the main inspiration behind the field of empirical Bayes, which was created to help frequentist methods to achieve full Bayesian efficiency in large scale studies (i.e., having a Bayesian omelet without breaking Bayesian eggs!).
- For more details, see Efron's book on Large-Scale Inference.
- In this course, we follow a more formal (and principled) Bayesian framework, which provides similar benefits through the shrinkage of parameter estimates.

Bayesian Analysis in a Nutshell

Bayesian inference

- Bayesian inference is making statements about unknown quantities in terms of probabilities given the observed data and our prior knowledge.
- Our *prior* knowledge represents the extent of our belief and uncertainty regarding the value of unobservables. We express our prior using probability models.
- We also use probability models to define the underlying mechanism that has generated the data.

Bayesian inference

- Bayesian inference therefore starts by defining the joint probability for our prior opinion and the mechanism based on which the data are generated.
- To make inference, we update our prior opinion about unobservables given the observed date. We refer to this updated opinion as our *posterior* opinion, which itself is expressed in terms of probabilities.
- As we can see, probability has a central role in Bayesian statistics.
- For deriving Bayesian methods and making statistical inference, probability provides a coherent and axiomatic framework.

Probability: It's personal!

- In the Bayesian paradigm, probability is a measure of uncertainty.
- "Coins don't have probabilities, people have probabilities", Persi Diaconis.
- "The only relevant thing is uncertainty–the extent of our own knowledge and ignorance," Bruno deFinetti.
- In this view, all that matters is uncertainty, and all uncertainties are expressed in terms of probability.
- Therefore, we use probability models for random variables that change and those that might not change (e.g., the population mean) but we are uncertain about their value .

Probability: It's personal!

- Consider the well-known coin tossing example. What is the probability of head in one toss?
- There are only two possibilities for the outcome: head and tail. Assuming symmetry (i.e., a fair coin), head and tail equal probability $1/2$.
- In the frequentist view, probability is assigned to an event by regarding it as a class of individual events (i.e., trials) all equally probable and stochastically independent.
- For the coin tossing example, we assume a sequence of *iid* tosses, and the probability of head is $1/2$ since the number of times we observe head divided by the number of trials reaches $1/2$ as the number of trials grows.

Probability: It's personal!

- Note that while Bayesians and frequentists provide the same answer, there is a fundamental and philosophical difference in how they view probability.
- Bayesians feel comfortable to assign probabilities to events that are not repeatable.
- For example, I can show you a picture of a car and ask "what is the probability that the price of this car is less than $5,000?".

Prior probability

- As mentioned above, within the Bayesian framework, we use probability not only for the data, y, where our certainty is due to data variation, but also for model parameters, θ, where our uncertainty is due to the fact that θ is the population parameter, and it is almost always unknown.
- Therefore, before doing statistical inference, we need to specify the extent of our belief and our uncertainty about the possible values of the parameter using prior probability.
- We denote this probability as $P(\theta)$.
- We usually use our (or other's) domain knowledge, which is accumulated based on previous scientific studies.
- We almost always have such information, although it could be vague.

Prior probability

- For example, consider the study conducted by Mackowiak, et al. to find whether the average normal body temperature is the widely accepted value of $98.6\,^{\circ}\mathrm{F}$.
- Their hypothesis was that the average normal body temperature is, in fact, less than $98.6\,^{\circ}\mathrm{F}$.
- For the average normal body temperature, for example, we know that it should be close to $98.6\,^{\circ}\mathrm{F}$.
- Let's denote the average normal body temperature for the population as θ. We know that θ should be close to $98.6\,^{\circ}\mathrm{F}$; that is, values close to $98.6\,^{\circ}\mathrm{F}$ are more plausible than values close to $90\,^{\circ}\mathrm{F}$ for example.
- We assume that as we move away from the $98.6\,^{\circ}\mathrm{F}$ the values become less likely in a symmetric way (i.e., it does not matter if we go higher or lower).

Prior probability

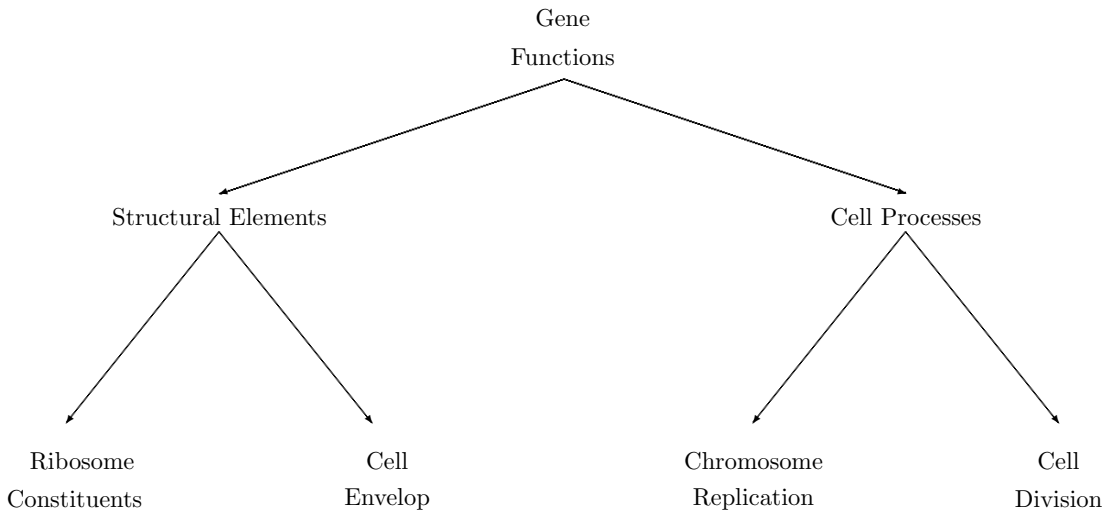
- Based on the above assumptions (and ignoring the fact that body temperature cannot be negative), we can set $\theta \sim N(98.6, \tau^2)$.
- In the above prior, τ^2 determines how certain we are about the average normal body temperature being around $98.6\,^{\circ}\mathrm{F}$.
- If we believe that it is almost impossible that the average normal body temperature is above 113.6 and below 83.6, we can set $\tau = 5$ so the approximate 99.7% interval includes all the plausible values from 83.6 to 113.6.
- A general advise is that we should keep an open mind, consider all possibilities, and avoid using very restrictive priors.

More on priors

- Sometimes, our prior opinion is based on what we know about the underlying structure of the data.
- For example, in many classification problems, we have prior knowledge about how classes can be arranged in a hierarchy.
- Hierarchical classification problems of this sort are abundant in statistics and machine learning.
- One such example is prediction of genes biological functions.

More on priors

- As shown in the following figure, gene functions usually are presented in a hierarchical form, starting with very general classes (e.g., cell processes) and becoming more specific in lower levels of the hierarchy (e.g., cell division).



A part of a gene annotation hierarchy proposed by Riley (1993) for *E. coli* genome.

- In the Bayesian framework, we can incorporate such information in our model.

Course outline

- So far, we tried to establish why we use Bayesian analysis.
- Throughout this course, we will discuss different Bayesian models and their applications for analyzing scientific problems.
- We first start with simple models with one unknown parameter.
- We then move to more complex models such as hierarchical Bayesian models and generalized linear models.
- Inference for these models tends to be difficult. We will discuss some advanced computational methods for Bayesian inference with complex models.
- Finally, we discuss Bayesian nonparametric. More specifically, we will discuss Gaussian process and Dirichlet process models.