# DATA.STAT.720 Introduction to Bayesian Analysis I

HYON-JUNG KIM

Tampereen yliopisto

# Contents

# 1 Introduction

**Bayesian statistics**

'Bayesian': named after Thomas Bayes (1702-1761).

The basic tool of inference is the Bayes' theorem, which was published posthumously in 1763 in 'An essay towards solving a problem in the doctrine of chance'. It was rediscovered and systematically exploited later by Laplace.

Based on an idea of subjective probability, it provides a natural, intuitively plausible way to think about statistical problems by revising previous information based on data observations.

- Having some knowledge or beliefs about matter in hand
⇒ Express these as a (prior) probability distribution.

- Collect some data (likelihood).

- Use Bayes' theorem to combine prior knowledge and data information to find a new (posterior) probability distribution about the unknowns (parameters).

# 2 Probability

- Randomness : most statistical modeling is based on an assumption of random observations from some probability distribution.

- Definition: Probability $P(A)$ is a measure of the chance that an event $A$ will happen.

- Axioms of Probability

  1. $P(A) \geq 0$ for any event $A$

  2. $P(S) = 1$ where $S$ is an universal set.

  3. $P(\sim A) = 1 - P(A)$

  4. If A and B have no outcomes in common then $P(A \cup B) = P(A) + P(B)$.

- Interpretations of probability

  1. Classical: Probability is a ratio of favorable cases to total (equipossible) cases

  The fundamental assumption is that the game is fair (based on the game theory) and all outcomes are equally likely.

  2. Frequentist: Probability is the limiting value of the frequency of some event as the number of trials becomes infinite.

  It can be legitimately applied only to repeatable problems and is believed as an objective property in the real world.

  3. Subjectivist: Probabilities may represent some numerical values assigned as to some degrees of personal belief. Most events in life are not repeatable. Probabilities are essentially conditional and there is no one correct probability.

- Frequency probability inference:

  - Data are drawn from a distribution of known form but with an unknown parameter and often this distribution arises from explicit randomization.

  - Inferences regard the data as random and the parameter as fixed (even though the data are known and the parameter is unknown)

- Subject probability inference:

  - Probability distributions are assumed for the unknown parameters and for the observations (i.e. both parameters and observations are random quantities).

  - Inferences are based on the prior distribution and the observed data.

- Comparison/Generality

  - Frequentists are disturbed by the dependence of the posterior results on the subjective prior distribution

  - Bayesians say that the prior distribution is not the only subjective element in an analysis. The assumptions about the sampling distributions are also subjective.

  - Whose probability distribution should be used? When there are enough data, a good Bayesian analysis and a good frequentist analysis will tend to agree. If the results are sensitive to prior information, a Bayesian analyst is obligated to report this sensitivity and to present different results obtained from a wide range of prior information.

  - Bayesians can often handle problems the frequentist approach cannot. Bayesians often apply frequentist techniques but with a Bayesian interpretation. Most untrained people interpret results in the Bayesian way more easily. (Often the Bayesian answer is what the decision maker really wants to hear.)

- Conditional Probability: the conditional probability of $B$ given $A$ is

$$P(B|A) = P(A \cap B)/P(A),$$

  where $P(A \cap B) = P(AB)$ is the joint probability that both $A$ and $B$ occur.

  **Independence** of events:

  $A$ and $B$ are independent if $P(A|B) = P(A)$ or $P(B|A) = P(B)$.

  **Multiplication Rule**:

$$P(AB) = P(A)P(B|A).$$

  Then, by definition:

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{P(B|A)P(A)}{P(B)} \quad : \text{This is the } \textbf{Bayes' theorem}$$

  Applying the **Law of Total Probability**:

$$P(B) = P(B, A) + P(B, \sim A) = P(B|A)P(A) + P(B| \sim A)P(\sim A)$$

$$\Rightarrow P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B| \sim A)P(\sim A)}$$

  This result is referred to as the expanded form of the Bayes' theorem.

cf. $P(A, B|C) = P(A|B, C)P(B|C)$

Example: Diagnostic test for a rare disease

A young person was diagnosed as having a certain disease with a medical test (i.e. his test result was positive). Suppose that the probability of having such disease in general is about 0.0001 and suppose further that the probability of positive test result when the person really has such disease is 0.9. Also, the probability of positive test result when the person does not have such disease is 0.01.

What is the probability that the person has the disease given a positive test result?

Example: Inspector Smith is waiting for Holmes and Watson who are both late for an appointment. Smith is worried that if the roads are icy, one or both of them may have crashed his car.

P( roads are icy)= .7  P( roads are not icy)=.3

P(One crashes | icy) =.8  P( One crashes | not icy)= .1

i) What is the probability that Holmes crashes his car?

ii) Suddenly Smith learns that Watson has crashed his car. What is the probability that Holmes crashes his car given the information that Watson has crashed?

Example: A patient goes to a doctor due to some discomfort. The doctor believes that he may have a disease 'A'. Doctor's experience: P(having a disease A) = 0.7.

The patient undertakes an examination: P(positive result given that he does not have the disease) = 0.4  P(positive result given that he has the disease) = 0.95

i) What is the probability that he has the disease A given the positive result?

A more reliable test 'W': P(W gives positive result given that he does not have the disease)= 0.04  P(W gives positive result given that he has the disease)= 0.99

ii) What is the probability of the patient being ill given that the first test was positive and that the second test was negative?

# 3 Bayesian Inference

- Random variables: represent outcomes of an uncertain phenomenon

  - Formally, it is a function from the sample space to the set of outcomes.

  - In Bayesian thinking, the sample space can be thought of as a set of possible worlds in which the random variable may have different values.

  To a frequentist, the height of a person is not a random variable, but it is to a Bayesian if you do not know his/her height.

- Probability distribution

  A probability distribution for a random variable quantifies how likely the different values are and summarizes all the information about the random variable.

  - Summaries of the distribution: Measures of central tendency, variability, and the shape of distribution.

  - Parameterized families of distributions: much of statistics is based on constructing models of phenomena using parameterized families of distributions.

  Observations $X = (X_1, \ldots, X_n)$ are sampled randomly from the distribution $f(X|\theta)$ where $\theta$ denotes the parameter of a model for the phenomenon generating the observations

  e.g. Normal, Poisson, Gamma distributions, etc.

  - Conditional/marginal distributions

- Bayesian Inference for parameters of probability models

  - Often we model a set of observations $X = (X_1, X_2, ..., X_n)$ as independent trials from a probability distribution with unknown parameters. The joint distribution of $X$ given a parameter $\theta$ viewed as a function of $\theta$ is called the likelihood function. To draw inferences about $\theta$, a Bayesian statistician specifies a prior distribution $g(\theta)$ for $\theta$. The $X_i$'s are usually independent given $\theta$, but are not independent marginally.

- **Bayes' theorem** in parametric distribution

  It is essentially a formula for learning from new data. It tells us how to convert our prior beliefs for a proposition into posterior beliefs after learning the information (in addition to the background prior information). It is fundamental to the Bayesian approach to statistics.

  Suppose we have an initial or prior belief about a hypothesis $H$ and suppose that we observe some data $D$. Then we can calculate our revised or posterior belief about the truth of $H$ in the light of the data evidence $D$ using Bayes' theorem.

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

  In parallel,

$$\begin{aligned} [\theta|X] &= \frac{[X|\theta][\theta]}{[X]} \\ g(\theta|x) &= \frac{f(x|\theta)g(\theta)}{f(x)} = \frac{f(x|\theta)g(\theta)}{\int f(x|\theta)g(\theta)d\theta} \quad \text{for continuous } \theta \end{aligned}$$

  i.e. **Posterior distribution** $\propto$ **likelihood** $\times$ **prior** ($\propto$ means "proportional to")

  - The marginal likelihood $f(x)$ is also called the predictive distribution for observations. (Predictive distribution incorporates both uncertainty about $X$ given $\theta$ and uncertainty about $\theta$. We use $f(x)$ to predict the observations we expect before observing them.)

  - A fundamental identity of Bayesian inference: $f(x|\theta)g(\theta) \propto f(x)g(\theta|x)$

  - Frequentists condition on parameters and base inferences on data distribution. Bayesians condition on knowns and treat unknowns probabilistically.

- Learning

  - Free (prior) information can never hurt. Decide what information to collect. Determine cost and information gain from sample.

  - The general principle is that in elaboration we try to use probability theory to construct elaborative measurements of probability that are difficult to measure and therefore liable to large measurement errors out of other probabilities that are easier to measure and hence hopefully more accurate.

# 4 Likelihood

The problem of statistical inference is to use observed data to learn about unknown features of the process that generated those data.

In order to make inference, it is essential to describe the link between $X$ and $\theta$, and this is done through a statistical model. The purpose of the statistical model is to describe this relationship by deriving the probability $P(\boldsymbol{x}|\theta)$ with which we would have obtained the observed data $\boldsymbol{x} = (x_1, ..., x_n)$ for any value $\theta$ (that is true).

**Definition**: a Likelihood function $L(\theta : \boldsymbol{x})$ is defined as any function of $\theta$ such that $L(\theta) \equiv L(\theta : \boldsymbol{x}) = k \cdot f(\boldsymbol{x}|\theta)$ for some constant $k$.

**Definition:** For a random variable $X$ with density(mass) function $f(\boldsymbol{x}|\theta)$ if $f(\boldsymbol{x}|\theta)$ can be expressed in the form $cq(\boldsymbol{x}|\theta)$ where $c$ is a constant, not depending upon $\boldsymbol{x}$, then any such $q(\boldsymbol{x}|\theta)$ is a kernel of the density $f(\boldsymbol{x}|\theta)$. The constant $c$ is called a normalizing constant with the fact

$$\int_S f(\boldsymbol{x}|\theta)d\boldsymbol{x} = \int_S cq(\boldsymbol{x}|\theta)d\boldsymbol{x} = 1 \quad \Rightarrow 1/c = \int_S q(\boldsymbol{x}|\theta)d\boldsymbol{x}.$$

In practice, when the likelihood is not given by the usual known probability distribution, it is not always straightforward to compute this integral and no closed form for it may exist.

In Bayesian statistics spotting kernels of distributions can be very useful in computing posterior distributions.

- Likelihood may not be enough for inference. The Bayesian approach is based also on some prior information.
- Prior distribution ( $g(\theta)$ or $P(\theta)$ ): formulates your prior beliefs about the parameters.

  Note that frequency probability is not able to represent such beliefs since parameters are referred as unknown but not random. The prior distribution is the major source of disagreement between two approaches - Bayesian and frequentist's
- Posterior distribution ( $g(\theta|\boldsymbol{x})$ or $P(\theta|\boldsymbol{x})$ ):

presents the probability distribution of the unknown parameter after we take the prior information and learn from the data

Note again that the posterior distribution has no meaning in the frequentist theory.

- **Bayesian Methods for Inference**

  i) Model a set of observations with a probability distribution with unknown parameters.

  ii) Specify prior distributions for the unknown parameters.

  iii) Use the Bayes' theorem to combine these two parts into the posterior distribution.

  iv) Use the posterior distribution to draw inferences about the unknown parameters of interest.

Example: Prior to Posterior

Suppose that there are three states of nature $A_1, A_2, A_3$ and two possible data $D_1, D_2$:

| | $P(D\|A)$ | | |
|---|---|---|---|
| | $D_1$ | $D_2$ | Prior |
| $A_1$ | 0.0 | 1.0 | 0.3 |
| $A_2$ | 0.7 | 0.3 | 0.5 |
| $A_3$ | 0.2 | 0.8 | 0.2 |

What happens to our belief about $A_1, A_2, A_3$ if we observe $D_2$? (if we observe $D_1$?)

Example:

A black male mouse is mated with a female black mouse whose mother had a brown coat.

$B$ and $b$ are alleles of the gene for coat color. The gene for black fur is given the letter $B$ and the gene for brown fur is given the letter $b$ where $B$ is the dominant allele to $b$.

The male and female have a litter with 5 pups that are all black. We want to determine the male's genotype. The prior information suggests $P(BB) = 1/3$ and $P(Bb) = 2/3$. What is the posterior probability that the male's genotype is $BB$?

Examples of Probability Distributions:

- $Y_i|\lambda \sim$ Poisson $(\lambda)$ distribution

$$f(y_i|\lambda) = \lambda^{y_i} e^{-\lambda}/y_i! \;\; y_i = 0, 1, 2, ...$$

$E[Y_i|\lambda] = \lambda, \quad Var[Y_i|\lambda] = \lambda.$

- $Y_i|\beta \sim$ exponential $(\beta)$ distribution

$$f(y_i|\beta) = \frac{1}{\beta} e^{-y_i/\beta}, \quad y_i \geq 0$$

$E[Y_i|\beta] = \beta, \quad Var[Y_i|\beta] = \beta^2.$

- $Y_i|(\alpha, \beta) \sim$ Gamma$(\alpha, \beta)$ distribution

$$f(y_i|\alpha, \beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} y_i^{\alpha-1} \exp(-y_i/\beta) \quad y_i > 0, \quad \alpha > 0, \beta > 0$$

$E[Y_i|(\alpha, \beta)] = \alpha\beta, \quad Var[Y_i|(\alpha, \beta)] = \alpha\beta^2.$

- $Y_i|(\alpha, \beta) \sim$ Inverse Gamma$(\alpha, \beta)$ distribution

$$f(y_i|\alpha, \beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} y_i^{-(\alpha+1)} \exp(1/(-y_i\beta)) \quad y_i > 0, \quad \alpha > 0, \beta > 0$$

$E[Y_i|(\alpha, \beta)] = \frac{1}{(\alpha-1)\beta}, \quad Var[Y_i|(\alpha, \beta)] = \frac{1}{(\alpha-1)^2(\alpha-2)\beta^2}.$

- $Y_i|(\mu, \sigma^2) \sim$ Normal $(\mu, \sigma^2)$ distribution

$$f(y_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \mu)^2}{2\sigma^2}\right), \quad -\infty < y_i < \infty \;\; \mu > 0, \sigma^2 > 0.$$

$E[Y_i|(\mu, \sigma^2)] = \mu, \quad Var[Y_i|(\mu, \sigma^2)] = \sigma^2.$

- $Y_i|p \sim$ Binomial $(n, p)$ distribution

$$f(y_i|n, p) = \binom{n}{y_i} p^{y_i}(1-p)^{n-y_i}, \quad y_i = 0, 1, 2, ...$$

$E[Y_i|p] = np, \quad Var[Y_i|p] = np(1-p).$

- $Y_i|(\alpha, \beta) \sim$ Beta $(\alpha,\beta)$ distribution

$$f(y_i|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y_i^{\alpha-1}(1-y_i)^{\beta-1}, \quad 0 < y_i < 1 \quad \alpha > 0, \beta > 0.$$

$E[Y_i|(\alpha, \beta)] = \frac{\alpha}{\alpha+\beta}, \quad Var[Y_i|(\alpha, \beta)] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$