L1: Bayesian Theory–Introduction and Motivation.

September 28, 2020

1 Scientific Method and Bayesian Statistics

The scientific method can be viewed as a framework for learning about the world: nature, biological and physical processes, human beings and activities. Cowles (2013) (taken from Berry (1996)) describes the scientific method as the following five step process:

1. Define the question or the problem to be addressed.

   For example: water temperatures in freshwater streams with bare banks (no vegetation) can sometimes get too hot for aquatic life. If vegetation is planted along the banks (habitat restoration), how much might temperatures be lowered, and how much might aquatic life survival increase?

2. Assess the relevant available information. Decide whether it is sufficient for the purpose at hand:

   (a) If yes, draw appropriate conclusions, make appropriate decisions, and take appropriate actions.

   (b) If no, go to the next step.

   Example: literature review in another location indicates that planting vegetation x at density y may lower water temperatures during hottest summer month by 2^o C, and increase survival of fish species z by 5%. But that vegetation, stream environment, and fish species are very different from our situation: need more information.

3. Determine what additional information is needed and design a sample, a study, or an experiment to get it.

   Example: may use an "artificial" setting with aquariums and controlled water temperatures, and different fish species, or conduct a field experiment with vegetation planting along randomly selected stream banks.

4. Collect the additional information: select the sample or carry out the study or experiment from step 3.

5. Use the data obtained in step 4 to update what was previously known. Return to step 2.

Statistics in general are central to steps 2, 3, and 5. For example, with step 2 one might have data from previous samples or studies or experiments that could be analysed. With step 3, statistics are used to draw probability samples from one or more existing populations, or use methods for setting up a designed experiment for comparing two or more procedures, for example. Step 5 involves the analysis of sample or experimental data using a variety of methods, e.g., estimating parameters.

Bayesian statistics is particularly well suited for steps 2 and 5 as it providing a quantitative framework for assessing current knowledge (step 2) and then updating or revising that knowledge by incorporating the new information gained (step 5).

A general schematic for Bayesian statistics:
Current Knowledge Prior to Study
$\Rightarrow$ Carry out Study
$\Rightarrow$ Combine Prior Knowledge with Study Results to Yield New Knowledge, Updated Knowledge.

2 Prerequisites

2.1 Probability Basics

You are expected to know the basics of probability including:

The general mathematical structure of probability: a Probability Space or Probability Triple, (Ω, $\mathcal{F}$, Pr).

1. A sample space, Ω, which is the set of all possible outcomes.
2. A set of events (a σ-algebra) $\mathcal{F}$, where each event is a subset from Ω that contains zero or more outcomes.
3. A function Pr that maps events in $\mathcal{F}$ to numbers between 0 and 1 (inclusive), i.e., probabilities.

For any event E in $\mathcal{F}$, $0 \leq \Pr(E) \leq 1$, and $\Pr(\Omega)=1$.

Given an event A in $\mathcal{F}$, the complement of A, denoted A^c is that part of Ω that does not include A: thus $A \cap A^c=\emptyset$ and $A \cup A^c=\Omega$.

Conditional probability. Given two events A and B in $\mathcal{F}$, where $\Pr(B) > 0$, the conditional probability of A given B is written $\Pr(A|B)$ and is defined as $\Pr(A \cap B)/\Pr(B)$. An alternative definition: $\Pr(A \cap B) = \Pr(B)\Pr(A|B)$.)

General probability results and definitions. Given events A and B in $\mathcal{F}$:

- Probability *at least one occurs*

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$$

  If events are *Mutually Exclusive*, then $\Pr(A \cap B) = 0$. And the *Addition Rule* applies: $\Pr(A \cup B) = \Pr(A) + \Pr(B)$.

- Probability *that both occur*:

$$\Pr(A \cap B) = \Pr(A|B)\Pr(B) = \Pr(B|A)\Pr(A)$$

  If the events are *Independent*, then $\Pr(A|B) = \Pr(A)$. And the *Multiplication Rule* applies: $\Pr(A \cap B) = \Pr(A)\Pr(B)$.

  Note: the term *joint probability* is a synonym for the probability that two, or more, events both occur.

  Note: We will typically write $\Pr(AB)$ for $\Pr(A \cap B)$; $\Pr(A, B)$ is another expression.

- Law of Total Probability: if events A_1, A_2, ..., A_K are mutually exclusive and exhaustive[1]

$$\Pr(B) = \sum_{i=1}^{K} \Pr(A_i \cap B) = \sum_{i=1}^{K} \Pr(B|A_i)\Pr(A_i)$$

[1] Mutually exclusive means $A_i \cap A_j=\emptyset$ and exhaustive means $\cup_{i=1}^{K} A_i=\Omega$.

2.2 Random variables (rv's)

Random variables, e.g., X, are real-valued functions defined on probability spaces, i.e., mappings from the outcomes of a random process, the events, to the real number line:

$$X : \mathcal{F} \to \mathcal{R}$$

More loosely stated, they are numbers associated with random events.

- The mapping induces a probability distribution for the random variables; e.g., $\Pr(X = x) \equiv \sum_{i \in K} \Pr(\omega_i)$, where ω_i are events in $\mathcal{F}$ (or Ω) such that $X(\omega_i)$=x.
- Random vectors of length p, $\mathbf{X} = (X_1, \ldots, X_p)^T$, are mappings from $\mathcal{F}$ to $\mathcal{R}^p$.

When the random variables are discrete values, e.g., counts, the induced probability distribution is a *probability mass function*, pmf. And when the random variables are continuous values, e.g., lengths, the induced probability distribution is a *probability density function*, pdf. Random variables that can be discrete or continuous are possible, e.g., values below some detection limit are assigned the value 0, and values at or above the detection limit have the detected value.

We assume familiarity with the following discrete random variables, their pmfs, expected values, and variances.

- $X \sim$ Bernoulli(θ): $E[X]$=θ, $V[X]$=$\theta(1-\theta)$
- $X \sim$ Binomial(n, θ): $E[X]$=$n\theta$, $V[X]$=$n\theta(1-\theta)$
- $X \sim$ Poisson(μ): $E[X]$=μ, $V[X]$=μ

And the discrete random vector:

- $\mathbf{X} = X_1, X_2, \ldots, X_k \sim$ Multinomial(n,θ_1,θ_2,...,θ_k): $E[X_i]$=$n\theta_i$, $V[X_i]$=$n\theta_i(1-\theta_i)$, $Cov(X_i, X_j) = -n\theta_i\theta_j$, for $i \neq j$.

And the following continuous random variables and their pdfs.

- $X \sim$ Uniform(α, β): $E[X]$=$\frac{\alpha+\beta}{2}$, $V[X] = \frac{(\beta-\alpha)^2}{12}$
- $X \sim$ Beta(α, β): $E[X]$=$\frac{\alpha}{\alpha+\beta}$, $V[X] = \frac{\alpha}{\alpha+\beta}\frac{\beta}{\alpha+\beta}\frac{1}{\alpha+\beta+1} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
- $X \sim$ Normal(μ, σ^2): $E[X]$=μ, $V[X]$=σ^2
- $X \sim$ Gamma(α, β): $E[X] = \frac{\alpha}{\beta}$, $V[X]$=$\frac{\alpha}{\beta^2}$; *(the shape-rate parameterisation).*
- $X \sim$ Exponential(λ): $E[X]$=$\frac{1}{\lambda}$, $V[X]$=$\frac{1}{\lambda^2}$. Recall: Exponential(λ) $\equiv$ Gamma(1,λ).

And the continuous random vector:

- $\mathbf{X} \sim$ Multivariate Normal (μ, Σ), where μ is p by 1 column vector and Σ is a p by p symmetric matrix.

Notes.

- See Appendix A of the notes by King and Ross for other probability distributions that will be used in the course.
- Given a *joint* probability distribution for two or more random variables, say $p(X, Y)$, we will refer to the distribution for one variable alone as a *marginal* distribution, e.g., $p(X)$ is the marginal distribution for X.
- We will sometimes refer to the *Coefficient of Variation* or CV of a random variable (when that makes sense), where $CV(X) = \sqrt{V[X]}/E[X]$ and is often expressed in percent; e.g., $CV(X)$=15%.

2.3 Some Notation

- Population parameters will usually be denoted by Greek letters, with θ a generic representation. Scalar, single parameters will often be represented by lower case letter, e.g.,

  $$\mu, \sigma^2, \alpha, \beta, \rho, \gamma$$

  And upper case letter sometimes denoting vectors of parameters, e.g., $\Theta = (\theta_1, \ldots, \theta_p)^T$.
- Random variables (that are not parameters) will usually be denoted by Roman letters with upper case letters often denoted unrealized, yet-to-be-observed, random variables and lower case letters denoted realised, observed values. For example, $X \sim$ Binomial(10,0.3) and x=6.

A notational distinction will not always be made between a probability density function (pdf) and a probability mass function (pmf) for a random variable; e.g., $p(Y)$ is the probability distribution function for a random variable Y.

A generic representation of probability distributions (from Gelfand and Smith, 1990) is based on squared brackets, as in $[X]$ is the probability distribution for X, $[X|Y]$ is the conditional probability distribution for X given Y, and $[X, Y]$ is the joint distribution for X and Y.

To indicate that a rv follows a specific distribution, the notation $\sim$ followed by a distribution name or shorthand will be used:

$$y \sim \text{Normal}\left(\mu, \sigma^2\right) \equiv N(\mu, \sigma^2)$$

Reminders.

- A pdf evaluated at a single value, $p(y)$, does not yield the probability of the event y as the probability that a specific value occurs is 0. However the probability that Y falls in some interval $[a, b]$, $\Pr[a \leq Y \leq b]$ is found by integrating the pdf over the interval, $\int_a^b f(y)dy$.
- Another perspective on pdfs: the limit of $\Pr(y - \epsilon \leq Y \leq y + \epsilon)$ as ϵ goes to 0, is the pdf $p(y)$.
- When conditioning is done on a continuous random variable, the result is a conditional density not a probability.

3 Real World Meaning of Probability

While the mathematical notion of probability seems clear, there are a variety of real world interpretations or definitions of probability. These interpretations have relevance to Bayesian statistics. Here we just consider two: long-run frequency and subjective.

Long-run frequency interpretation. The long-run frequency interpretation of probability is that the probability of an event A is the fraction of times the event A would occur if the same random process is repeatedly carried out many many times ("infinitely" often). For example, if one rolls a die many, many times, Pr(2) is the fraction of times that the side with 2 pips lands face up. Another example is that if a simple random sample of two people is drawn without replacement from a list of six people, the probability that a given person will be in the sample is 2/6—where such draws could be done repeatedly.

This definition runs into difficulties when a process, for which there is uncertainty about its outcome, cannot be repeated many times. For example, will there be an earthquake above 3.0 Richter here tomorrow? There either will or there won't, but we are uncertain about the outcome, and while we can't repeat "tomorrow", we would like to have quantitative assessment or measure of our belief about whether an earthquake will occur or not.

Subjective probability. The subjective probability of an event is defined to be a number between 0 and 1 that quantifies an individual's belief that the event will occur (or did occur but the outcome is unknown to that individual). Thus subjective probabilities are *personal* and will differ between people. My subjective probability for an earthquake above 3.0 occurring tomorrow is 1/1,000,000. A geoscientist would have a different subjective probability.

A key component of Bayesian statistical inference is a statement of the probability, or probabilities, about some state of nature, what is called the prior probability. While not the only way to define prior probabilities, the subjective definition or interpretation is generally more applicable than a long-run frequency definition. How one defines prior probabilities is a *hugely important* issue in Bayesian statistics.

4 Bayes Theorem and Examples

Bayes Theorem can be viewed as a particular formulation of a conditional pmf, for discrete random outcomes, or a conditional probability density, for continuous random outcomes.

4.1 Discrete case

Let X and Y be two discrete valued random variables with possible values denoted $x_1,\ldots,x_n$ and $y_1,\ldots,y_m$, respectively. The conditional probability that $X = x_i$ given $Y = y_j$ is defined by:

$$\Pr(X = x_i | Y = y_j) = \frac{\Pr(X = x_i, Y = y_j)}{\Pr(Y = y_j)}$$

which can be rewritten as

$$\Pr(X = x_i | Y = y_j) = \frac{\Pr(Y = y_j | X = x_i) \Pr(X = x_i)}{\Pr(Y = y_j)}$$

The law of total probability can be used to rewrite the denominator $\Pr(Y = y_j)$:

$$\Pr(X = x_i|Y = y_j) = \frac{\Pr(Y = y_j|X = x_i)\Pr(X = x_i))}{\sum_{k=1}^{n}\Pr(X = x_k, Y = y_j)} = \frac{\Pr(Y = y_j|X = x_i)\Pr(X = x_i))}{\sum_{k=1}^{n}\Pr(Y = y_j|X = x_k)\Pr(X = x_k)} \quad (1)$$

Equation 1 is often the way that Bayes Theorem is defined, for discrete valued random variables. Sometimes it is written without denoting the exact value of X and Y, namely:

$$\Pr(X|Y) = \frac{\Pr(Y|X)\Pr(X))}{\sum_{x}\Pr(X = x, Y)} \quad (2)$$

Bayes Theorem also holds for outcomes that are random events but not random variables, in other words, numerical values are not defined on the events. For example, X could refer to the presence of a disease and X^c is the absence of the disease, and similarly Y could be the result of a positive test result for the disease and Y^c is a negative test result.

Example 1. Karen has two housemates, Jill and Mary, who do all the cooking. Mary cooks about 2/3s of the time while Jill cooks 1/3 of the time. Mary tends to burn the dinner less often than Jill, about 1/6th of the time, while Jill burns the dinner 1/4th of the time. Karen comes home, steps through the door, smells burned food, and yells: "Jill, you must be cooking!". What is the probability that Karen guessed correctly? We use Bayes Theorem to find out the probability that it was Jill. Letting M and J denote Mary and Jill and B=burnt dinner

$$\Pr(J|B) = \frac{\Pr(J \cap B)}{\Pr(B)} = \frac{\Pr(B|J)\Pr(J)}{\Pr(B|J)\Pr(J) + \Pr(B|M)\Pr(M)} = \frac{\frac{1}{4}\frac{1}{3}}{\frac{1}{4}\frac{1}{3} + \frac{1}{6}\frac{2}{3}} = 0.429$$

Thus, she has more likely guessed wrong.

Note that *prior* to acquiring additional information the probability that Jill is cooking is 1/3. But given additional information, namely the dinner has been burnt, changes our probability for Jill doing the cooking, it is now 0.429.

Example 2. (From Efron 2005, Jr of the Am. Stat. Assoc.) A woman was pregnant and a sonogram showed that she was going to have two twin boys. Her doctor told her that in the general population of all twins, fraternal and identical, the proportion of identical twins is 1/3. Given that she knew she had two boys, she wanted to know what the probability was that *her* boys would be identical twins. Assume that if twins are fraternal the probability of either sex for each twin is 1/2, the probability of two boys is 1/4, two girls is 1/4, and a boy and a girl is 1/2.

$$\begin{aligned}
\Pr(\text{Identical}|\text{Twin Boys}) &= \frac{\Pr(\text{Twin Boys} \cap \text{Identical})}{\Pr(\text{Twin Boys})} \\
&= \frac{\Pr(\text{Twin Boys}|\text{Identical})\Pr(\text{Identical})}{\Pr(\text{Twin Boys}|\text{Identical})\Pr(\text{Identical}) + \Pr(\text{Twin Boys}|\text{Fraternal})\Pr(\text{Fraternal})} \\
&= \frac{\frac{1}{2}\frac{1}{3}}{\frac{1}{2}\frac{1}{3} + \frac{1}{4}\frac{2}{3}} \\
&= \frac{\frac{1}{6}}{\frac{1}{6} + \frac{1}{6}} = \frac{1}{2}
\end{aligned}$$

Thus a 50% probability that they would be identical.

Note that *prior* to acquiring additional information the probability that the woman will have identical twins is 1/3. But given additional information, namely that she has twin *boys*, changes our probability for identical twins, it is now 1/2.

Comment. Note that in these two examples, the applications of Bayes Theorem were simply exercises in probability, not in statistics, as there are no unknown parameters to estimate nor hypotheses to test. The focus in this course will, of course, be on statistical inference, particularly parameter estimation and hypothesis testing.

4.2 Continuous case

Here we consider random variables, thus numerical outcomes, not categorical. Let X and Y be two continuous random variables with corresponding pdf's $f(x)$ and $g(y)$. The continuous version of Bayes Theorem is then expressed as a conditional probability density function:

$$f(X|Y) = \frac{f(X,Y)}{g(Y)} = \frac{g(Y|X)f(X)}{\int g(Y|X)f(X)dX} \tag{3}$$

Bayes Theorem can be applied to hybrid cases, where X is discrete and Y is continuous, and one of the examples below is such a case.

5 Bayesian Statistical Inference

The most basic setup for Bayesian statistical inference has two components:

1. A *likelihood* for the parameter, θ, $L(\theta|y)$, which is the sampling model for the data, $f(y|\theta)$, viewed as a function of θ, thus the data are known constants.
2. A *prior probability distribution* for θ, $\pi(\theta)$.

Bayesian inference for θ produces a third component, the conditional probability of θ given the data, what is called the *posterior probability distribution* for θ:

$$p(\theta|y) = \frac{p(\theta,y)}{m(y)} \tag{4}$$

where $m(y)$, often written $p(y)$, is sometimes called the *marginal distribution* for y or the *normalizing constant*.

Eq'n 4 can be rewritten as

$$p(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{\int p(\theta,y)d\theta} = \frac{f(y|\theta)\pi(\theta)}{\int f(y|\theta)\pi(\theta)d\theta} \propto f(y|\theta)\pi(\theta) \tag{5}$$

where formulation 5 makes clear the basis for the "name" Bayesian in Bayesian inference.

Comments.

1. Bayesian inference can be viewed as taking the knowledge that one has about θ before seeing sample data and *updating* or *modifying* that knowledge after seeing data.
2. Given that we are considering probability distributions for θ, prior and posterior, implies that θ is being viewed as a random variable. This does not mean that θ is necessarily a random quantity in nature. It is a random variable in the sense that one is uncertain about the value of the parameter.

3. Bayesian inference may look deceptively simple. Just calculate the conditional probability distribution. But it's not—which is why there are many books and papers written on it. There are two complications that we will focus on that make up a large part of this course.
   (a) Complication 1: Calculation of $m(y)$, an integration over θ, can be extremely difficult, in particular when θ is a multi-dimensional parameter.
   (b) Complication 2: Specification of the prior distribution can be both difficult, particularly for high dimensional θ, and controversial.
4. Sequential updating of one's knowledge about θ given new information, e.g., a new sample, is carried out by applying Bayes formula again, using the posterior from the last iteration as the prior for the current iteration. Letting y_{first} denote the first sample and y_{second} denote the second sample:
$$p(\theta|y_{first}, y_{second}) = \frac{f(y_{second}|\theta, y_{first})p(\theta|y_{first})}{\int p(\theta, y_{first}, y_{second})d\theta} = \frac{f(y_{second}|\theta, y_{first})p(\theta|y_{first})}{\int f(y_{second}|\theta, y_{first})p(\theta|y_{first})d\theta} \quad (6)$$
5. In a few simple settings the posterior distribution can be deduced by just looking at the numerator of Eq'n 5 and examining terms that include the parameter θ, the so-called "kernel". The kernel for a well-known pdf or pmf might be evident. Example 3 will demonstrate this.

Example 3. A mobile phone company is releasing a new phone. Based on experience with the previous model, only 85% of the phones still work after two years. The company advertises that 95% of the *new* models will still be working after two years. A consumer advocacy group is skeptical that the new model will be much better than the old model. It plans to randomly sample n=200 purchasers of the new model and after 2 years and find out how many still have working phones. To give the phone company some allowance, the group chooses a prior for the probability of working, θ, that has a mean value of 0.90 and a CV of 20%. In particular the prior for $\theta \sim$ Beta(1.6, 0.178). After waiting two years, the sample is taken and 172 phones, or 86%, are still working. How reasonable is the phone company's claim?

Assuming independence between the phones and an in-common "survival" probability, the number of phones, y, still working after two years has a Binomial(n=200,θ) distribution. This is the sampling distribution for the data or the likelihood for θ.

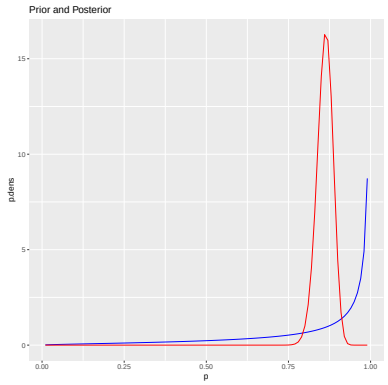
The posterior distribution for θ is then:
$$\begin{aligned} p(\theta|y) &\propto \pi(\theta)f(y|n,\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\binom{n}{y}\theta^y(1-\theta)^{n-y} \\ &\propto \theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^y(1-\theta)^{n-y} = \theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1} \end{aligned}$$
which is recognized as the kernel of a Beta($\alpha+y$, $\beta+n-y$), in this case Beta(1.6+172, 0.178+200-172) = Beta(173.6,28.178).

Notes:

- Figure 1 shows the prior (in blue) and posterior (in red) densities for the probability a mobile phone is still working after 2 years. Note that the inclusion of the data has shifted the prior distribution noticeably to the left. This is an example of the informal expression that data "dominated" the prior.
- Given the posterior probability distribution it is easy to calculate various summaries such as the mean of the posterior distribution, in this case 173.6/(173.6+28.178)= 0.86.
- Further, one can examine the claim made by the phone company and calculate that the probability that the survival probability θ is 90% or higher is 0.042. Referring to the phone company claim "more accurately", the probability that $\theta \geq 0.95$ is 8.450503e-07.

Note: the parameters in the prior distribution for θ are called *hyperparameters*.

Figure 1: Prior (in blue) and posterior (in red) distribution for the probability that mobile phone is still working after 2 years.



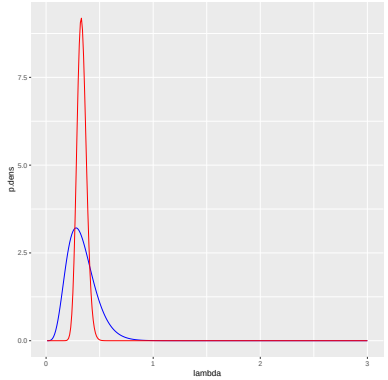
Example 4. After entering a queue at a bank on Fridays between 11:30AM and 1:00PM, the waiting time for seeing a teller follows a exponential distribution with rate parameter λ. Thus the average waiting time is $1/\lambda$. You are considering switching to a new bank and will visit that bank next Friday. In your experience at your current bank you currently wait from 2 to 6 minutes. You select a prior distribution for λ where the expected value for λ is 1/3 minutes with a CV of 40%, in particular the prior for λ is Gamma(6.25, 18.75). You then visited the bank and observed 52 customers enter the queue during the 11:30-1:00PM period. The total amount of waiting time was 156 minutes. The posterior distribution for λ:

$$p(\lambda|y) \propto \pi(\lambda) f(y_1, \ldots, y_n|\lambda) = \lambda^{\alpha-1} \exp(-\beta\lambda) \prod_{i=1}^{n} \lambda \exp(-\lambda y_i)$$

$$= \lambda^{\alpha+n-1} \exp(-\lambda(\beta + \sum_{i=1}^{n} y_i))$$

which is the kernel for a Gamma($\alpha + n$, $\beta + \sum_{i=1}^{n} y_i$) distribution. Thus the posterior distribution for λ is Gamma(6.25+52, 18.75+156) = Gamma(58.25, 174.75). See Figure 2.

Figure 2: Prior (in blue) and posterior (in red) distribution for the exponential waiting time parameter λ.



6 Prior and Posterior Predictive Distributions

The *predictive distribution* is another name for the marginal distribution for the data, i.e., the distribution for the data after the parameter has been integrated out.

There are two predictive distributions: the prior predictive and the posterior predictive. Letting y^{new} denote a new, not-yet-observed data point, the *Prior Predictive* distribution is

$$p(y^{new}) = \int_\theta f(y^{new}|\theta)\pi(\theta)d\theta \tag{7}$$

And, letting $\mathbf{y}^{old}$ denote already observed data, e.g., an available sample, and the *Posterior Predictive* distribution is

$$p(y^{new}|y^{old}) = \int_\theta f(y^{new}|\theta, y^{old})p(\theta|y^{old})d\theta \tag{8}$$

where $p(\theta|y^{old})$ is the *posterior* probability distribution for θ.

Demonstration. Consider the binomial sampling model with a Beta prior on θ and the prior predictive distribution. Let n be the binomial sample size.

$$\begin{aligned}
p(y^{new}) &= \int_\theta f(y^{new}|\theta)\pi(\theta)d\theta = \int \binom{n}{y^{new}}\theta^{y^{new}}(1-\theta)^{n-y^{new}}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta \\
&= \binom{n}{y^{new}}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int \theta^{\alpha+y^{new}-1}(1-\theta)^{\beta+n-y^{new}-1}d\theta \\
&= \binom{n}{y^{new}}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+y^{new})\Gamma(\beta+n-y^{new})}{\Gamma(\alpha+\beta+n)}\int \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+y^{new})\Gamma(\beta+n-y^{new})}\theta^{\alpha+y^{new}-1}(1-\theta)^{\beta+n-y^{new}-1}d\theta \\
&= \binom{n}{y^{new}}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+y^{new})\Gamma(\beta+n-y^{new})}{\Gamma(\alpha+\beta+n)}
\end{aligned}$$

This is a Beta-Binomial distribution, which we denote Beta-Binomial(n, α, β).

For the posterior predictive distribution, given that the posterior distribution for θ is still Beta, one can substitute $\alpha^{post} = \alpha + y^{old}$ and $\beta^{post} = \beta + n - y^{old}$ in the above results. Letting n^{new} be the new sample size, the posterior predictive distribution is then:

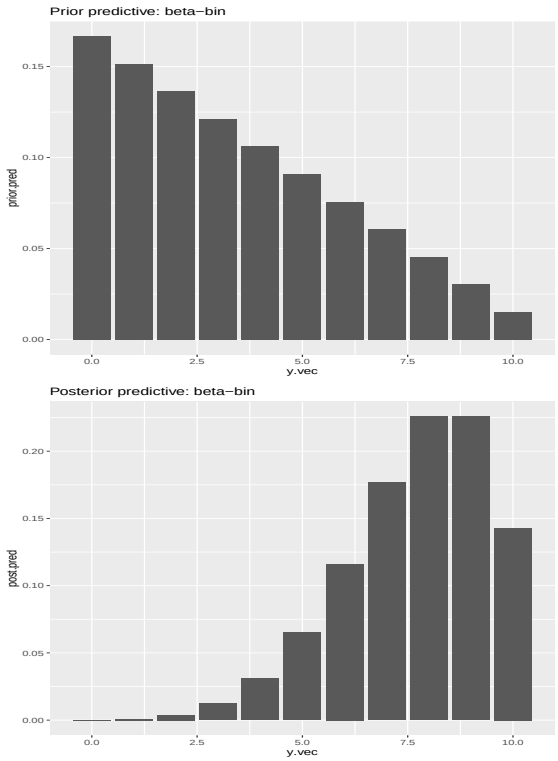
$$p(y^{new}|y^{old}) = \binom{n^{new}}{y^{new}}\frac{\Gamma(\alpha^{post}+\beta^{post})}{\Gamma(\alpha^{post})\Gamma(\beta^{post})}\frac{\Gamma(\alpha^{post}+y^{new})\Gamma(\beta^{post}+n^{new}-y^{new})}{\Gamma(\alpha^{post}+\beta^{post}+n^{new})}$$

namely, a Beta-Binomial$(n^{new}, \alpha^{post}, \beta^{post})$.

Example 4. Assume a binomial sampling model, Binomial(n, θ), with n=10 and the prior for θ is Beta(1,2). Suppose that y=9 is observed, then the posterior for θ is Beta(10,3). Letting n^{new}=10, Figure 3 shows the prior and the posterior predictive distributions for y^{new}. Before observing the data, the predicted probability of y^{new}=9 was 0.0303. After observing the data, the predicted probability of y^{new}=9 is now 0.2256.

Comparing prior and posterior predictive distributions is one way of evaluating the influence of the prior.

Figure 3: Prior and posterior predictive distributions for y^{new} where $y|\theta \sim$ Binomial(n=10,θ). Thus the prior and posterior predictive distributions are both Beta-Binomial distributions. The prior for θ was Beta(1,2) and y=9 was observed (the posterior for θ was then Beta(10,3)).



7 Likelihood principle and Bayesian vs Frequentist inference

Of relevance to comparisons between Bayesian and Frequentist inference is the Likelihood Principle. The Likelihood Principle says that given a sample of data, $\mathbf{y}$, any two sampling models for $\mathbf{y}$, say $f_1(\mathbf{y}|\theta)$ and $f_2(\mathbf{y}|\theta)$, that have the same likelihood function yield the same inference for θ. Thus f_1 and f_2 might have different normalizing constants but the same likelihood. The main point is that inference for θ depends on the observed $\mathbf{y}$ alone not on unobserved values of $\mathbf{y}$.

One aspect of Frequentist inference that does not obey the Likelihood Principle is the use of p-values in hypothesis tests. Suppose that the sampling model for the data is Poisson(θ) and that there are two hypotheses about θ: a null hypothesis, $H_0 : \theta = 1$, and an alternative hypothesis, $H_1 : \theta = 2$. Suppose that a single observation is made yielding the value y=2. The standard frequentist approach is to calculate the p-value: the probability of the observed value and any values in a direction away from H_0 in the direction of H_1. In this case the p-value is $\Pr(Y \geq 2|H_0)$ = 1- $\Pr(Y = 0|H_0)$ - $\Pr(Y = 1|H_0)$ = 1-exp(−1) - exp(−1) ∗ 1 = 0.264. Thus, one would not reject H_0.

This procedure is violating the Likelihood Principle, however, in that inference is being based on more than the likelihood of the data. The probability of events that *did not occur* are included in the inference.

A more straightforward approach to testing these two hypotheses would be to compare the likelihoods of θ under the two hypotheses. Under H_0, θ=1 and $\Pr(Y = 2|\theta = 1) = L(\theta = 1|Y = 2) = \exp(-1) * 1/2! = 0.184$. Under H_1, θ=2 and $\Pr(Y = 2|\theta = 2) = L(\theta = 2|Y = 2) = \exp(-2) * 2^2/2! = 0.271$. Thus, given a value of y=2, H_1 seems more consistent with the data than does H_0, as can be seen by the likelihood ratio

$L(\theta = 1|Y = 2)/L(\theta = 2|Y = 2) = 0.6796$.

Part of the problem with p-values in hypothesis testing is the "inherent" special status of one hypothesis, the Null Hypothesis, over another hypothesis, the Alternative Hypothesis. The p-value is calculated *conditional* on H_0 *being true*.

We will discuss hypothesis testing in a Bayesian framework in detail later, but here we will discuss a Bayesian approach for this particular setting. To do so we will bring the notion of Odds. The odds for an event is the ratio of the probability of the event to the probability of its complement. For example, if a fair die is thrown, the odds of the sides 2, 3, 4, 5, or 6 landing face up is $\Pr(2 \cup 3 \cup 4 \cup 5 \cup 6)/\Pr(1) = (5/6)/(1/6) = 5$.

A Bayesian approach to testing the above hypotheses is to specify prior probabilities for both hypotheses. In this case, assuming no expert knowledge, prior probabilities of 0.5 for H_0 and H_1 seem reasonable and the prior *odds* is the ratio $0.5/0.5 = 1$. The posterior probability for H_0 is proportional to the product of the prior and the likelihood, namely 0.5*0.184 = 0.092, and the posterior probability for H_1 is proportional to 0.5*0.271 = 0.1355. Thus the posterior odds of H_0 to H_1 is 0.092/0.1335 = 0.6796, in this case simply the likelihood ratio because the prior probabilities were equal. Thus the Bayesian conclusion would be that there is more evidence for H_1 than H_0.

Bayesian inference is consistent with the Likelihood Principle in that the posterior distribution is proportional to the product of the prior distribution and the likelihood alone, i.e., the effect of the data on the posterior is only through the likelihood.

8 Some History

Quoting from Wikipedia: "Bayes' theorem is named after Reverend Thomas Bayes (1701—1761), who first provided an equation that allows new evidence to update beliefs in his *An Essay towards solving a Problem in the Doctrine of Chances* (1763). It was further developed by Pierre-Simon Laplace, who first published the modern formulation in his 1812 "Theorie analytique des probabilites". Sir Harold Jeffreys put Bayes' algorithm and Laplace's formulation on an axiomatic basis. Jeffreys wrote that Bayes' theorem 'is to the theory of probability what the Pythagorean theorem is to geometry'."

Thus Bayes Theorem has been around for over 250 years. However, widespread use of Bayesian statistics is a relatively recent development–over the last 30 years.

While statistics were certainly gathered and used in the 1800s, statistics started becoming a more established and mature discipline in the early 1900s due to contributions by Karl Pearson, R.A. Fisher, Jerzy Neyman, Egan Pearson and others. Contributions from Fisher, which included the development of maximum likelihood theory, and the hypothesis testing framework of Neyman and E. Pearson became the roots of Frequentist Statistical methods.

Bayesian methods, however, did not thrive in the 1900s. One reason was concern over the influence of priors on inference. There were heated arguments over the merits of frequentist and Bayesian inference in the 60's, 70's, and 80's—but in a practical sense these arguments had little effect because people generally could not carry out Bayesian inference except for a relatively restricted range of problems. The hang-up was calculating the posterior distribution. Specifically, the integration involved in calculating the normalizing constant or the marginal distribution for the data, $m(\mathbf{y}) = \int_\theta f(\mathbf{y}|\theta)\pi(\theta)d\theta$, was simply too difficult.

All this began to change in the early 1990s with the application of, and further development of Monte Carlo simulation methods, which provided ways of generating samples from the posterior distribution. Note this is in contrast to being able to writing down a closed-form expression for the posterior density, or being able to exactly evaluate the posterior pdf for a given value of θ (or a vector of θ's). Given a large enough sample from the posterior distribution, however, gives one the ability to make all sorts of approximate inferences

about the posterior, e.g., the mean from a large enough sample from the posterior distribution can be a "good enough" approximation of the true posterior mean.

Since then many problems that have proven quite hard to solve in a Frequentist framework have become more readily solved in a Bayesian framework.

Controversy over the influence of priors remains and considerable thought and research into formulating and selecting priors continues as well, but the ability to solve problems previously unsolveable has led to a revolution of usage of Bayesian methods.

Among the things that we will discuss in this class are

1. Key "classical" Bayesian statistical principles and methods that have been around for decades;
2. Various perspectives on prior distributions;
3. Some of the Monte Carlo procedures used for making Bayesian inference possible.

Optional, but interesting reading

- Lavine, M. "What is Bayesian statistics and why everything else is wrong."
  `http://people.math.umass.edu/~lavine/whatisbayes.pdf`
- Efron, B. 2005. "Bayesians, frequentists, and scientists."
  `https://courses.physics.ucsd.edu/2015/Fall/physics210b/REFERENCES/Efron_Bayesians_Frequentists.pdf`

R code to produce figures

Example 3: Mobile phone failure.

```
#-- Mobile phone failure example
library(ggplot2)
xseq <- seq(0.01,0.99,by=0.01)
prior.a <- 1.6; prior.b <- 0.178
n <- 200; y <- 172
post.a  <- prior.a+y; post.b <- prior.b + n - y
prior.pdf <- dbeta(xseq,prior.a,prior.b)
posterior.pdf <- dbeta(xseq,post.a,post.b)
df <- data.frame(x=xseq,prior=prior.pdf,post=posterior.pdf)
g1 <- ggplot(df,mapping=aes(x,prior)) + geom_line(col="blue") +
  geom_line(mapping=aes(y=post),col="red") + xlab(expression(theta))
g1

cat("Pr(survival)>0.9=",1-pbeta(0.9,post.a,post.b),"\n")
# Pr(survival)>0.9= 0.04176059
```

Example 4: waiting time at bank.

```
 #-- Example 4: Waiting time at bank
```

```
xseq <- seq(0.01,3,by=0.01)
prior.a <- 6.25; prior.b <- 18.75
n <- 52; ttl.wait <- 156
post.a <- prior.a+n; post.b <- prior.b + ttl.wait
prior.pdf <- dgamma(xseq,shape=prior.a,rate=prior.b)
posterior.pdf  <- dgamma(xseq,shape=post.a,rate=post.b)
df <- data.frame(x=xseq,prior=prior.pdf,post=posterior.pdf)
g1 <- ggplot(df,mapping=aes(x,prior)) + geom_line(col="blue") +
  geom_line(mapping=aes(y=post),col="red") + xlab(expression(lambda)) +
  ggtitle("Prior and Posterior: Exponential dist parameter")
g1
```

Example: prior and posterior predictive with Beta-Binomial distributions.

```
# Code to evaluate Beta-Binomial pmf
dbetabin <- function(y,size,alpha,beta) {
  bin.coef <- choose(n=size,k=y)
  num <- beta(alpha+y,beta+n-y)
  den <- beta(alpha,beta)
  out <- bin.coef*num/den
  return(out)
}

library(ggpubr)
n <- 10; y.obs <- 9

#prior and posterior
a.prior <- 1; b.prior  <- 2
a.post <- a.prior+y.obs; b.post <- b.prior+n-y.obs

y.vec <- 0:n
#prior predictive
prior.pred <- dbetabin(y=y.vec, size=n, alpha=a.prior, beta=b.prior)
#posterior predictive
post.pred <- dbetabin(y=y.vec,size=n,alpha=a.post,beta=b.post)

df <- data.frame(y.vec=y.vec,prior.pred=prior.pred,post.pred=post.pred)

g.prior <- ggplot(df,aes(x=y.vec,y=prior.pred)) + geom_col() +
             labs(title="Prior predictive: beta-bin")
g.post <- ggplot(df,aes(x=y.vec,y=post.pred)) + geom_col() +
             labs(title="Posterior predictive: beta-bin")
ggarrange(g.prior,g.post)
```

September 28, 2020