Probability review and one parameter models

Dr. Olanrewaju Michael Akande

Jan 15, 2020

Announcements

- No make-up for Monday's lab.
- Final exam will be either online or take home. Not in class.
- Homework one soon...but here are some readings to keep you busy:

1. Efron, B., 1986. Why isn't everyone a Bayesian?. The American Statistician, 40(1), pp. 1-5.
2. Gelman, A., 2008. Objections to Bayesian statistics. Bayesian Analysis, 3(3), pp. 445-449.
3. Diaconis, P., 1977. Finite forms of de Finetti's theorem on exchangeability. Synthese, 36(2), pp. 271-281.
4. Gelman, A., Meng, X. L. and Stern, H., 1996. Posterior predictive assessment of model fitness via realized discrepancies. Statistica sinica, pp. 733-760.
5. Dunson, D. B., 2018. Statistics in the big data era: Failures of the machine. Statistics & Probability Letters, 136, pp. 4-9.

Outline

- Probability review
  - Random variables
  - Joint distributions
  - Independence and exchangeability
- Introduction to Bayesian Inference (Cont'd)
  - Conjugacy
  - Kernels
  - Bernoulli and binomial data
  - Selecting priors
  - Truncated priors

Probability review

Discrete random variables

- A random variable is discrete if the set of all possible outcomes is countable.
- The probability mass function (pmf) of a discrete random variable Y, $p(y)$ describes the probability associated with each possible value of Y.
- $p(y)$ has the following properties:
  1. $0 \leq p(y) \leq 1$ for all values $y \in \mathrm{Y}$.
  2. $\sum_{y \in \mathrm{Y}} p(y) = 1$.

Bernoulli distribution

- The Bernoulli distribution can be used to describe an experiment with two outcomes, such as
  - Flipping a coin (heads or tails);
  - Vote turnout (vote or not); and
  - The outcome of a basketball game (win or loss).
- In all cases, we can represent this as a binary random variable where the probability of "success" is θ and the probability of "failure" is $1 - \theta$.
- We usually write this as: $Y \sim \text{Bernoulli}(\theta)$, where $\theta \in [0, 1]$.
- It follows that

$$\Pr(Y = y \,|\, \theta) = \theta^y (1-\theta)^{1-y}; \quad y = 0, 1.$$

- What is the mean of this distribution? What is the variance?

Binomial distribution

- The binomial distribution describes the number of successes from n independent Bernoulli trials.
- That is, $Y =$ number of "successes" in n independent trials and θ is the probability of success per trial.
- We usually write this as: $Y \sim \text{Bin}(n, \theta)$, where $\theta \in [0, 1]$.
- The pmf is

$$\Pr(Y = y \mid \theta, n) = \binom{n}{y} \theta^y (1-\theta)^{n-y}; \quad y = 0, 1, \ldots, n.$$

- **Example**: $Y =$ number of individuals with type I diabetes out of a sample of n surveyed.
- Binomial likelihoods are commonly used in collecting data on proportions.
- What is the mean of this distribution? What is the variance?

Poisson distribution

- $Y \sim \mathrm{Po}(\theta)$ denotes that Y is a Poisson random variable.
- The Poisson distribution is commonly used to model count data consisting of the number of events in a given time interval.
- The Poisson distribution is parameterized by θ and the pmf is given by

$$\Pr[Y = y \mid \theta] = \frac{\theta^y e^{-\theta}}{y!}; \quad y = 0, 1, 2, \ldots; \quad \theta > 0.$$

- Similar to binomial but with no limit on the total number of counts.
- What is the mean of this distribution? What is the variance?

General discrete distributions

- Useful to consider general discrete distributions having an arbitrary form.
- Suppose $Y \in \{y_1^\star, \ldots, y_k^\star\}$. Then define $\Pr(Y = y_h^\star) = \pi_h$ for each $h = 1, \ldots, k$. That is,

$$\Pr[Y = y \mid \boldsymbol{\pi}] = \prod_h \pi_h^{1[Y = y_h^\star]}; \;\; y \in y_1^\star, \ldots, y_k^\star$$

  where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)$.

- $(y_1^\star, \ldots, y_k^\star)$ are "atoms" representing possible values for Y.
- For example, these may be words in a dictionary or values for education as a categorical variable. Useful for text data, categorical observations, etc.
- Can also write as $Y \sim \sum_{h=1}^k \pi_h \delta_{y_h^\star}$, where $\delta_{y_h^\star}$ denotes a unit mass at $y_h^\star$.
- Often called the categorical distribution or generalized Bernoulli distribution. Also, see the multinomial distribution.

Continuous random variables

- The probability density function (pdf), $p(y)$ or $f(y)$, of a continuous random variable Y has slightly different properties:

    1. $0 \leq f(y)$ for all $y \in \mathrm{Y}$.

    2. $\int_{y \in \mathrm{R}} p(y) \mathrm{d}y = 1$.

- The pdf for a continuous random variable is not necessarily less than 1.

- Also, $p(y)$ is NOT the probability of value y.

- However, if $p(y_1) > p(y_2)$, we say informally that y_1 has a "higher probability" than y_2.

Uniform density

- The simplest example of a continuous density is the uniform density.
- $Y \sim \text{Unif}(a, b)$ denotes density is uniform in interval (a, b).
- The pdf is simply

$$f(y) = \frac{1}{b-a}; \quad y \in (a, b).$$

- The cdf is

$$F(y) = \Pr(Y \leq y) = \int_a^y \frac{1}{b-a} \mathrm{d}z = \frac{y-a}{b-a}$$

- The mean (expectation) is

$$\frac{a+b}{2}$$

- What is the variance? Also, can you prove the formula for the mean?

Beta density

- The uniform density can be used as a prior for a probability if $(a, b) \subset (0, 1)$.
- However, it is very inflexible clearly.

Why?

- An alternative for $y \in \mathcal{Y}$ is the beta density, written as $Y \sim \text{Beta}(a, b)$, with

$$f(y) = \frac{1}{B(a,b)} y^{a-1}(1-y)^{b-1}; \quad y \in (0,1),\ a > 0,\ b > 0.$$

where $B(a, b) = \dfrac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$. $\Gamma(n) = (n-1)!$ for any positive integer n.

- As we have already seen, the beta density is quite flexible in characterizing a broad variety of densities on $(0, 1)$.
- Beta(1,1) is the same as Unif(0,1). Workout the pdfs to convince yourself!

Gamma density

- The gamma density will be useful as a prior for parameters that are strictly positive.
- For random variables $Y \sim \mathrm{Ga}(a, b)$, we have the pdf

$$f(y) = \frac{b^a}{\Gamma(a)} y^{a-1} e^{-by}; \quad y \in (0, \infty), \; a > 0, \; b > 0.$$

- Properties:

$$\mathrm{E}[Y] = \frac{a}{b}; \quad \mathrm{V}[Y] = \frac{a}{b^2}.$$

- **Note**: parameterizations of the gamma distribution vary!
- Under this parameterization, if $Y \sim \mathrm{Ga}(1, \theta)$, then $Y \sim \mathrm{Exp}(\theta)$, that is, the exponential distribution.

Continuous joint distributions

- Suppose we have two random variables $\theta = (\theta_1, \theta_2)$.

- Their **joint** distribution function is

$$\Pr\left(\theta_1 \leq a, \theta_2 \leq b\right) = \int_{-\infty}^{a}\int_{-\infty}^{b} p(\theta_1, \theta_2) \mathrm{d}\theta_1 \mathrm{d}\theta_2,$$

  where $p(\theta_1, \theta_2)$ is the joint probability density function (pdf).

- The **marginal** density of θ_1 can be obtained by

$$p(\theta_1) = \int_{-\infty}^{\infty} p(\theta_1, \theta_2) \mathrm{d}\theta_2,$$

  which is referred to as marginalizing out θ_2.

- We will be doing a lot of "marginalizations" so take note!

FACTORIZING JOINT DENSITIES AND INDEPENDENCE

- The joint density $p(\theta_1, \theta_2)$ can be factorized as

$$p(\theta_1, \theta_2) = p(\theta_1 \mid \theta_2)p(\theta_2), \quad \text{or} \quad p(\theta_1, \theta_2) = p(\theta_2 \mid \theta_1)p(\theta_1).$$

- For independent random variables, the joint density equals the product of the marginals:

$$p(\theta_1, \theta_2) = p(\theta_1)p(\theta_2).$$

- This implies that $p(\theta_2 \mid \theta_1) = p(\theta_2)$ and $p(\theta_1 \mid \theta_2) = p(\theta_1)$ under independence.

- These relationships extend automatically to $\theta = (\theta_1, \ldots, \theta_p)$. That is,

$$p(\theta_1, \ldots, \theta_p) = \prod_{j=1}^{p} p(\theta_j),$$

under mutual independence of the elements of the θ vector.

Conditional independence

- Suppose $y_i \overset{iid}{\sim} f(\theta)$ for $i = 1, \ldots, n$.
- Data $\{y_i\}$ are independent & identically distributed draws from distribution $f(\theta)$.
- The data are said to be conditionally independent given θ.

$$L(y; \theta) = \prod_{i=1}^{n} f(y_i; \theta),$$

where $L(y; \theta) =$ likelihood of the data conditionally on θ.

- The marginal likelihood of the data is

$$L(y) = \int L(y; \theta) p(\theta) \mathrm{d}\theta.$$

- $L(y)$ can no longer be written as a product of densities as in $\prod_{i=1}^{n} h(y_i)$; we lose independence when we marginalize out θ.

Exchangeability

- In marginalizing out θ, the observations $\{y_i\}$ are no longer independent.

- $\{y_i\}$ are exchangeable if $p(y_1, \ldots, y_n) = p(y_{\pi_1}, \ldots, y_{\pi_n})$, for all permutations π of $\{1, \ldots, n\}$.

- de Finetti's Theorem: Suppose $\{y_i\}$ are exchangeable under above definition for any n. Then

$$p(y_1, \ldots, y_n) = \int \left[\prod_{i=1}^{n} f(y_i; \theta) \right] p(\theta) \mathrm{d}\theta.$$

  for some θ, prior distribution $p(\theta)$ and sampling model $f(y_i; \theta)$.

- Simply put, de Finetti's Theorem states that exchangeable observations are conditionally independent relative to some parameter.

- de Finetti's Theorem is critical in providing a motivation for using parameters and for putting priors on parameters.

Introduction to Bayesian Inference (Cont'd)

Frequentist inference

- Given **data** $\{y_i\}$ and an **unknown parameter** θ, estimate said θ.
- How to estimate θ under the frequentist paradigm?
    - Maximum likelihood estimate (MLE)
    - Method of moments
    - and so on...
- Frequentist ML estimation finds the one value of θ that maximizes the likelihood.
- Typically uses large sample (asymptotic) theory to obtain confidence intervals and do hypothesis testing.

Bayesian inference

- Once again, given **data** $\{y_i\}$ and an **unknown parameter** θ, estimate said θ.
- Bayesians update their prior information for θ with information in the data $\{y_i\}$, to obtain the posterior density $p(\theta | y)$.
- Personally, I prefer being able to obtain posterior densities that describe my parameter, instead of estimated summaries (usually measures of central tendency) along with confidence intervals.
- Bayes' theorem - reminder:

$$p(\theta | y) = \frac{p(\theta)L(y; \theta)}{\int_\Theta p(\tilde{\theta})L(y; \tilde{\theta})\mathrm{d}\tilde{\theta}} = \frac{p(\theta)L(y; \theta)}{L(y)}$$

Comments on the posterior density

- The posterior density is more concentrated than the prior & quantifies learning about θ.
- In fact, this is the optimal way to learn from data - see discussion in Hoff chapter 1.
- As more & more data become available, posterior density will converge to a normal (Gaussian) density centered on the MLE (Bayes central limit theorem).
- In finite samples for limited data, the posterior can be highly skewed & noticeably non-Gaussian.

Conjugacy

- Starting with an arbitrary prior density $p(\theta)$ & likelihood $L(y;\theta)$ we may encounter problems in calculating the posterior density $p(\theta \mid y)$.
- In particular, you may notice the denominator in the Bayes' rule:

$$L(y) = \int_{\Theta} p(\tilde{\theta}) L(y; \tilde{\theta}) \mathrm{d}\tilde{\theta}.$$

  This integral may not be analytically tractable!

- When the prior is conjugate however, the marginal likelihood can be calculated analytically.
- Conjugacy $\Rightarrow$ the posterior has the same form as the prior.
- Often useful to think of hyperparameters of a conjugate prior distribution as corresponding to having observed a certain number of (historical) pseudo-observations with properties specified by the parameters.
- Conjugate priors make calculations easy but may not represent our prior information well.

Kernels

- In Bayesian statistics, the kernel of a pdf omits any multipliers that do not depend on the random variable or parameter we care about.
- For many distributions, the kernel is in a simple form but the normalizing constant complicates calculations.
- If one recognizes the kernel as that matching a known distribution, then the normalizing factor can be reinstated. This is a very MAJOR TRICK we will use to calculate posterior distributions.
- For example, the normal density is given by

$$p(y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

but the kernel is just

$$p(y \mid \mu, \sigma^2) \propto e^{-\frac{(y-\mu)^2}{2\sigma^2}}.$$

Bernoulli data

- Back to our example: suppose $\theta \in (0, 1)$ is the population proportion of individuals with diabetes in the US.
- Suppose we take a sample of n individuals and record whether or not they have diabetes (as binary: 0,1).
- Then we can use the Bernoulli distribution as the sampling distribution. also, we already established that we can use a beta prior for θ.

Bernoulli data

- Generally, it turns out that if
  - $f(y_i; \theta) : y_i \overset{iid}{\sim} \text{Bernoulli}(\theta)$ for $i = 1, \ldots, n$, and
  - $p(\theta) : \theta \sim \text{Beta}(a, b)$,

  then the posterior distribution is also a beta distribution.

- Can we derive the posterior distribution and its parameters? Let's do some work on the board!

- Updating a beta prior with a Bernoulli likelihood leads to a beta posterior - we have conjugacy!

- Specifically, we have.

$$p(\theta | \{y_i\}) : \theta | \{y_i\} \sim \text{Beta}(a + \sum y_i, b + n - \sum y_i).$$

- This is the beta-Bernoulli model. More generally, this is just the beta-binomial model.

Beta-binomial in more detail

- Suppose the likelihood of the data is

$$L(y;\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}.$$

- Suppose also that we have a $\mathrm{Beta}(a, b)$ prior on the probability θ.
- Then the posterior density then has the beta form

$$\pi(\theta \mid y) = \mathrm{Beta}(a+y, b+n-y).$$

- The posterior has expectation

$$\mathrm{E}(\theta \mid y) = \frac{a+y}{a+b+n} = \frac{a+b}{a+b+n} \times \text{prior mean} + \frac{n}{a+b+n} \times \text{sample mean}.$$

- For this specification, sometimes a and b are interpreted as "prior data" with a interpreted as the prior number of 1's, b as the prior number of 0's, and $a+b$ as the prior sample size.
- As we get more and more data, the majority of our information about θ comes from the data as opposed to the prior.

Binomial data

- For example, suppose you want to find the Bayesian estimate of the probability θ that a coin comes up heads.
- Before you see the data, you express your uncertainty about θ through the prior $p(\theta) = \text{Beta}(2, 2)$
- Now suppose you observe 10 tosses, of which only 1 was heads.
- Then, the posterior density $p(\theta \mid y, n)$ is $\text{Beta}(3, 11)$.

Binomial data

- Recall that the mean of $\text{Beta}(a, b)$ is $a/(a+b)$.
- That means, before you saw the data, you thought the mean for θ was 2/(2+2) = 0.5.
- However, after seeing the data, you believe it is 3/(3+11) = 0.214.
- The variance of $\text{Beta}(a, b)$ is $ab/[(a+b)^2(a+b+1)]$.
- So before you saw data, your uncertainty about θ (i.e., your standard deviation) was $\sqrt{4/[4^2 \times 5]} = 0.22$.
- However, after seeing 1 Heads in 10 tosses, your uncertainty is 0.106.
- Clearly, as the number of tosses goes to infinity, your uncertainty goes to zero.

Operationalizing data analysis

We will explore another example soon but first, how should we approach data analysis in general?

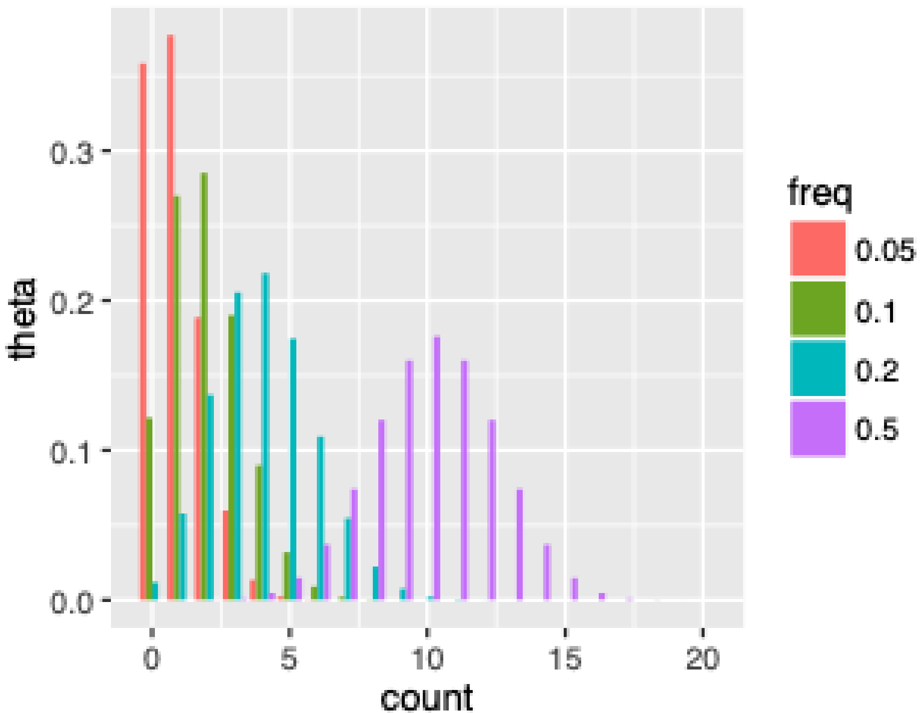
- Step 1. State the question.
- Step 2. Collect the data.
- Step 3. Explore the data.
- Step 4. Formulate and state a modeling framework.
- Step 5. Check your models.
- Step 6. Answer the question.

Example: rare events

- Step 1. State the question:
  - What is the prevalence of an infectious disease in a small city?
  - Why? High prevalence means more public health precautions are recommended.
- Step 2. Collect the data:
  - Suppose you collect a small random sample of 20 individuals.
- Step 3. Explore the data:
  - Let Y denote the unknown number of infected individuals in the sample.

Example: rare events

- Step 4. Formulate and state a modeling framework:
  - Parameter of interest: θ is the fraction of infected individuals in the city.
  - Sampling model: a reasonable model for Y can be $\mathrm{Bin}(20, \theta)$

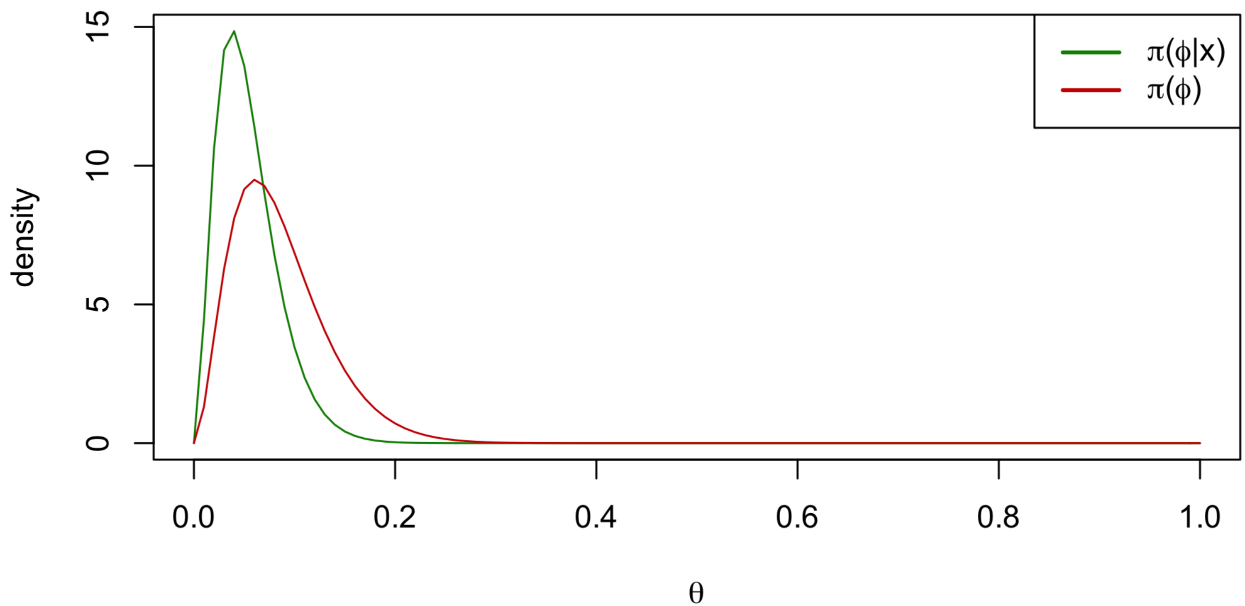


Example: rare events

- Step 4. Formulate and state a modeling framework:
  - Prior specification: information from previous studies – infection rate in "comparable cities" ranges from 0.05 to 0.20 with an average of 0.10. So maybe a standard deviation of roughly 0.05?
  - What is a good prior? The **expected value** of θ close to 0.10 and the **variance** close to 0.05.
  - Possible option: $\mathrm{Beta}(3.5, 31.5)$ or maybe even $\mathrm{Beta}(3, 32)$?

Example: rare events

- Step 4. Formulate and state a modeling framework:
  - Under $\text{Beta}(3, 32)$, $\Pr(\theta < 0.1) \approx 0.67$.
  - Posterior distribution for the model: $(\theta \mid Y = y) = \text{Beta}(a + y, b + n - y)$
  - Suppose $Y = 0$. Then, $(\theta \mid Y = y) = \text{Beta}(3, 32 + 20)$



Example: rare events

- Step 5. Check your models:
  - Compare performance of posterior mean and posterior probability that $\theta < 0.1$.
  - Under $\text{Beta}(3, 52)$,
    - $\Pr(\theta < 0.1 \mid Y = y) \approx 0.92$. More confidence in low values of θ.
    - For $E(\theta \mid Y = y)$, we have

$$E(\theta \mid y) = \frac{a + y}{a + b + n} = \frac{3}{52} = 0.058.$$

    - Recall that the prior mean is $a/(a + b) = 0.09$. Thus, we can see how that contributes to the prior mean.

$$\begin{aligned} E(\theta \mid y) &= \frac{a + b}{a + b + n} \times \text{prior mean} + \frac{n}{a + b + n} \times \text{sample mean} \\ &= \frac{a + b}{a + b + n} \times \frac{a}{a + b} + \frac{n}{a + b + n} \times \frac{y}{n} \\ &= \frac{35}{52} \times \frac{3}{35} + \frac{20}{52} \times \frac{0}{n} = \frac{3}{52} = 0.058. \end{aligned}$$

Example: rare events

- Step 6. Answer the question:
  - People with low prior expectations are generally at least 90% certain that the infection rate is below 0.10.
  - $\pi(\theta \mid Y)$ is to the left of $\pi(\theta)$ because the observation $Y = 0$ provides evidence of a low value of θ.
  - $\pi(\theta \mid Y)$ is more peaked than $\pi(\theta)$ because it combines information and so contains more information than $\pi(\theta)$ alone.
  - The posterior expectation is 0.058.
  - The posterior mode is 0.04.
    - Note, for $\text{Beta}(a, b)$, the mode is $(a-1)/(a+b-2)$.
  - The posterior probability that $\theta < 0.1$ is 0.92.