For typed text, my vectors and matrices are denoted with bold font. On your homework, please explicitly denote vectors using the arrow notation!

Also, you're welcome to drop the subscripts on distributions! I provide them for clarity, but at this point you're all hopefully able to tell the difference between $f(\mathbf{y})$ and $f(\theta)$.

1. Let's now consider count data. Let Y_i represent counts from some process, $i = 1, \ldots, n$. For example, Y_i might represent number of fish caught at Lake Dunmore in an hour on the i-th fishing trip! A common distribution for count data, and the one we will use here, is the Poisson:

$$Y_i|\theta \overset{\text{iid}}{\sim} \text{Poisson}(\theta), \qquad i = 1 \ldots, n$$

Take a second now to look at the PMF of the Poisson and remind yourself of the support!

   (a) The *rate* parameter θ of the distribution must be positive. A distribution with support on the positive real line is the Gamma distribution. Take a second now to look at the PDF of the Gamma!

   We will assume the following prior distribution for θ:

$$\theta \sim \text{Gamma}(a, b)$$

   Under the proposed sampling model, obtain the posterior distribution for θ by explicitly deriving/evaluating the marginal likelihood along the way. You should be able to name the posterior exactly!

   (b) Clearly relate the posterior mean of θ to the sample mean and the prior mean. Can you provide some way to "interpret" the a and b in the context of this statistical model?

2. We will consider the general case of the Bernoulli distribution, which is the **Categorical distribution**. We have a discrete random variable Y that can take one of the values $1, 2, \ldots, J$, each with a specified probability. The PMF of this distribution is

$$f_{y|\boldsymbol{\theta}}(y|\boldsymbol{\theta}) = \Pr(Y = j|\theta_1, \ldots, \theta_J) = \begin{cases} \theta_j & j = 1, 2, \ldots J \\ 0 & o.w. \end{cases},$$

.

   There is a constraint on the θ_j: they must live on the $(J-1)$-*dimensional simplex* Δ^{J-1}. This means the probabilities θ_j are such that:

   - $\theta_j \geq 0$ for all $j = 1, \ldots, J$, and
   - $\sum_{j=1}^{J} \theta_j = 1$.

In the context of the Categorical distribution, this ensure that the probability of obtaining category j is non-negative, and the experiment will always result in one of the J categories.

We denote this distribution as $Y \sim \text{Categorical}_J(\boldsymbol{\theta})$ or $Y \sim \text{Categorical}(\theta_1, \ldots, \theta_J)$. This is like a Bernoulli, but with J possible outcomes, rather than just 2. *Note: the Categorical distribution is equivalent to the Multinomial distribution with 1 trial.*

The θ_j are unknown, and we would like to obtain a posterior distribution for $\boldsymbol{\theta}$ based on data from a Categorical sampling model. Because $\boldsymbol{\theta}$ is a random *vector*, we will need a multivariate prior distribution. Consider the Dirichlet$(\alpha_1, \alpha_2, \ldots, \alpha_J)$ distribution, where $J \geq 2$ and each $\alpha_j > 0$ (pronounced "deer-ee-shlay").

If $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_J)$, its PDF is

$$f_{\boldsymbol{\theta}}(\theta_1, \ldots, \theta_J) = \frac{\Gamma(\sum_{j=1}^J \alpha_j)}{\prod_{j=1}^J \Gamma(\alpha_j)} \prod_{j=1}^J \theta_j^{\alpha_j - 1} \qquad \text{if } \boldsymbol{\theta} \in \Delta^{J-1},$$

and 0 otherwise. *Please don't be scared!*

Suppose that our sampling model is $Y_1, Y_2, \ldots Y_n | \boldsymbol{\theta} \overset{\text{iid}}{\sim} \text{Categorical}(\theta_1, \ldots, \theta_J)$, and we use a Dirichlet$(\alpha_1, \alpha_2, \ldots, \alpha_J)$ prior for $\boldsymbol{\theta}$.

(a) Obtain the posterior distribution $f_{\boldsymbol{\theta}|\mathbf{y}}(\boldsymbol{\theta}|\mathbf{y})$. You may either explicitly derive the marginal likelihood along the way, or you may not. You should be able to name the posterior exactly! *It may be useful to use indicator notation at some point!*

(b) Obtain the posterior predictive distribution $f_{Y^*|\mathbf{y}}(y^*|\mathbf{y}) = \Pr(Y^* = l|\mathbf{y})$, for $l = 1, 2, \ldots, J$. Interpret the value that you obtain. *It may be useful to use indicator notation at some point!*

3. Let's return to the set-up of Problem 1 where we had $Y_i|\theta \overset{\text{iid}}{\sim} \text{Poisson}(\theta),\ i = 1, \ldots, n$ and $\theta \sim \text{Gamma}(a, b)$.

(a) Find the posterior predictive distribution for a new, unobserved count Y^*. Some very helpful things:

- For z positive, $\Gamma(z+1) = z\Gamma(z)$. When z is a positive integer, $\Gamma(z) = (z-1)!$.
- When u, v are positive reals, $\binom{u}{v} = \frac{\Gamma(u+1)}{\Gamma(v+1)\Gamma(u-v+1)}$ (even though we're used to integer inputs into the Binomial coefficient).

(b) Compare the mean and variance of your posterior predictive distribution from (c) to the posterior mean and variance of θ. What might be a possible explanation?

4. I mentioned in class that the posterior mode is typically not a Bayes estimate for θ. However, we will prove here that in the case where θ is discrete (such that $\Pr(\theta = a)$ is well-defined), then under 0-1 loss, the posterior mode is the Bayes estimator! *Note that it's very rare to assume θ is discrete in real life...*

Recall that when using a to estimate θ, the loss that is incurred under 0-1 loss is

$$L(\theta, a) = \begin{cases} 0 & \theta = a \\ 1 & \theta \neq a \end{cases}$$

(a) First, re-write the 0-1 loss function using indicator notation.

(b) Prove that when θ is discrete, the Bayes estimator for θ using 0-1 loss is (any) posterior mode.

5. (R) A famous example of real-life Poisson data is the data from the Prussian Army. In particular, we have (real!) data on the number of soldiers killed by being kicked by a horse each year in each of three different cavalry corps, across a total of 20 years (1875-1894) . This data was shown by economist and statistican Ladislaus Josephovich Bortkiewicz to follow the Poisson distribution.

Download the `prussianHorses1.csv` dataset from the course website under Handouts. Load it into your RStudio using the code `read_csv(" /Downloads/prussianHorses1.csv")`. You may have to change the file path. Be sure to store the data in your `R` environment!

Try to be as reproducible in your code as possible! That means storing quantities as variables, and actually using those variables. Also, if you perform random sampling anywhere, please set a seed using the `set.seed()` function.

(a) Let Y_i represent the number of deaths by horsekick in observation i, where $i = 1, \ldots, 60$ corresponds to a specific year and cavalry corps combination. We will use the sampling model $Y_i|\theta \overset{\text{iid}}{\sim} \text{Poisson}(\theta)$. What is the interpretation of the unknown parameter, in context?

(b) I'd like my prior mean to be 0.5, with about 95% of the prior mass concentrated between 0 and 1.5. If I choose $\theta \sim \text{Gamma}(a, b)$, provide a choice of a and b that satisfies these criteria. You might want to play around in `R` to help you.

(c) Now, the researcher is interested in learning (or daresay, testing a hypothesis) if θ is less than 0.5, in light of the data. Using `R` code, provide support for/against the claim that θ is less than 0.5.

(d) As it turns out, there were 11 more cavalry corps than the three provided to you! Provide a histogram (or barplot to be more aesthetic) of 220 predictions for the number of soldiers killed by being kicked by a horse in the new cavalry corps (assuming that the deaths by horse are relatively similar across cavalry corps). Then, download and load in the remaining data from `prussianHorses2.csv"` on the website and visualize the counts. How do our predictions compare to the truth?

General rubric

Points	Criteria
5	The solution is correct *and* well-written. The author leaves no doubt as to why the solution is valid.
4.5	The solution is well-written, and is correct except for some minor arithmetic or calculation mistake.
4	The solution is technically correct, but author has omitted some key justification for why the solution is valid. Alternatively, the solution is well-written, but is missing a small, but essential component.
3	The solution is well-written, but either overlooks a significant component of the problem or makes a significant mistake. Alternatively, in a multi-part problem, a majority of the solutions are correct and well-written, but one part is missing or is significantly incorrect.
2	The solution is either correct but not adequately written, or it is adequately written but overlooks a significant component of the problem or makes a significant mistake.
1	The solution is rudimentary, but contains some relevant ideas. Alternatively, the solution briefly indicates the correct answer, but provides no further justification.
0	Either the solution is missing entirely, or the author makes no non-trivial progress toward a solution (i.e. just writes the statement of the problem and/or restates given information).
Notes:	For problems with multiple parts, the score represents a holistic review of the entire problem. Additionally, half-points may be used if the solution falls between two point values above.
Notes:	For problems with code, well-written means only having executed lines of code that are necessary to solving the problem (you're welcome to comment out code for yourself to keep), as well as presenting the solution for the reader to easily see.