https://dicook.github.io/Statistical_Thinking/tutorials/lab10/lab10help.pdf

SETU surveys are out, please fill them in. It's the first time we've run this subject so your feedback is really useful to improve things in the future.

Question One

Use Bayes' rule to get

$$p(Email = spam|heading = CTO) = \frac{p(heading = CTO|Email = spam)p(Email = spam)}{p(heading = CTO)}$$

where

$$p(heading = CTO) = \sum_{Email} p(heading = CTO|Email)p(Email)$$

with {spam, not spam} $\in$ Email.

Question Two

Let $p(\theta) \sim \mathcal{N}(\mu_0, \sigma_0^2)$, then if we observe $y_1, \ldots, y_n$ from a Gaussian likelihood with unknown mean θ and known variance σ^2 we have the posterior

$$p(\theta|y_{1:n}) \sim \mathcal{N}\left(\frac{\mu_0\sigma_0^{-2} + \sigma^{-2}\sum x_i}{\sigma_0^{-2} + n\sigma^{-2}}, (\sigma_0^{-2} + n\sigma^{-2})^{-1}\right).$$

Random number generation recap:

```
x <- rnorm(n, truemean, truesd)
```

What you need to do is generate n random numbers from the 'true' distribution, calculate your posterior mean and variance and plot it with the prior and the MLE for the normal distribution mean.

```
library(ggplot2)
library(tidyr)
support <- seq(-10, 10, 0.001)
prior <- dnorm(support, priormean, priorsd)
posterior <- dnorm(support, posteriormean, posteriorsd)
data.w = data.frame(x=support, prior=prior, posterior=posterior)
data.l = gather(data.w, distribution, density, -x)
ggplot(data.l, aes(x, density, colour = distribution)) +
  geom_line() + geom_vline(xintercept = MLE)
```

Question Three

Remember the denominator from Bayes' rule is constant with respect to θ, so

$$p(\theta|y < 3) \propto p(y < 3|\theta)p(\theta)$$

If θ has a $\mathscr{B}(\alpha, \beta)$ distribution then

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1},$$

and if $y|\theta$ has a binomial distribution then

$$p(Y = y|\theta) \propto \theta^{y}(1-\theta)^{n-y}.$$

We can always ignore the terms in the distribution that don't depend on the main variable (y for the likelihood, and θ for the prior).

Find $p(y < 3|\theta)$ and combine with $p(\theta)$ to get your un-normalised posterior.

If we had $p(\theta|y) \propto \theta + \theta^2$ your code to plot it is below, just switch the second line for your posterior.

```
theta <- seq(0, 1, 0.001)
ptheta <-  theta + theta^2
ggplot() + geom_line(aes(theta, ptheta))
```

Question Four

If $\theta \sim \mathscr{B}(\alpha, \beta)$ then $\mathbb{E}(\theta) = \alpha/(\alpha+\beta)$ and $\text{Var}(\theta) = \alpha\beta((\alpha+\beta)^2(\alpha+\beta+1))^{-1}$. Look up the Method of Moments estimator for the Beta distribution to find α and β.

The Beta prior is conjugate to the Binomial likelihood and has a really easy posterior update rule you can derive.

Question Five

Gamma/Poisson make another conjugate pair and we will get another easy parameter update from prior to posterior.

Remember that

$$\begin{aligned}
\text{Gamma}(\theta|\alpha, \beta) &= \frac{\beta^{\alpha}}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta} \\
\mathbb{E}(\theta) &= \frac{\alpha}{\beta} \\
\text{Var}(\theta) &= \frac{\alpha}{\beta^2} \\
\text{Poisson}(x_{1:n}|\theta) &= \frac{e^{-\theta}\theta^{\sum x_i}}{\prod x_i!}
\end{aligned}$$

Why Bayes works with proportions

Bayesian statistics usually doesn't use the whole density function. In this lab we just said that the posterior was proportional ($\propto$) to part of the product of each density. We can always split a probability density into the kernel, the part that depends on the random variable, and a constant, which does not depend on the variable. Saying something is proportional to the kernel instead of hanging onto the constant makes working in the Bayesian paradigm much easier.

Bayes' Rule gives us the posterior:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \tag{1}$$

Working in this form is hard, the maths can get messy and we can almost never know what the marginal distribution $p(y)$ is.

Let's use $p(y|\theta) = \mathcal{N}(0, \theta^{-1})$ for the likelihood and use its conjugate prior, Gamma(α, β), for $p(\theta)$. Note that θ is the inverse of the variance, which is known as the precision.

Then we have

$$\begin{aligned} p(\theta) &= \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}\exp\{-\beta\theta\} \\ &= A\theta^{\alpha-1}\exp\{-\beta\theta\} \\ &= Ap(\theta)^* \qquad (2) \\ p(y|\theta) &= \prod_{i=1}^{n}\left(\sqrt{\frac{\theta}{2\pi}}\right)\exp\left\{\frac{y_i^2}{-2\theta^{-1}}\right\} \\ &= \left(\sqrt{\frac{\theta}{2\pi}}\right)^n\exp\left\{\frac{\sum_{i=1}^n y_i^2}{-2\theta^{-1}}\right\} \\ &= B\theta^{n/2}\exp\left\{\frac{\sum_{i=1}^n y_i^2}{-2\theta^{-1}}\right\} \\ &= Bp(y|\theta)^* \qquad (3) \end{aligned}$$

We have seperated each distribution into its kernel (the * functions in (2) and (3)), which is multiplied by a constant (A or B).

Putting this into Bayes' Rule in (1) we get

$$p(\theta|y) = \frac{Ap(y|\theta)^*Bp(\theta)^*}{p(y)}$$

and we could write

$$p(\theta|y) = Cp(y|\theta)^*p(\theta)^* = Cp(\theta|y)^*,$$

with

$$C = \frac{AB}{p(y)}. \tag{4}$$

Lets look at the kernel of the posterior,

$$\begin{aligned} p(\theta|y)^* &= p(y|\theta)^* p(\theta)^* \\ &= \theta^{\alpha-1} \exp\left\{-\beta\theta\right\} \theta^{n/2} \exp\left\{ \frac{\sum_{i=1}^n y_i^2}{-2\theta^{-1}} \right\} \\ &= \theta^{\alpha+n/2-1} \exp\left\{ -\theta \left(\beta + 1/2 \sum_{i=1}^n y_i^2 \right) \right\} \end{aligned} \tag{5}$$

You might notice that a Gamma$(\alpha+n/2, \beta+1/2\sum_{i=1}^n y_i^2)$ distribution looks like

$$\frac{(\beta + 1/2\sum_{i=1}^n y_i^2)^{\alpha+n/2}}{\Gamma(\alpha+n/2)} \theta^{\alpha+n/2-1} \exp\left\{ -\theta \left(\beta + 1/2 \sum_{i=1}^n y_i^2 \right) \right\}. \tag{6}$$

This looks pretty messy, but it has exactly the same kernel as our $p(\theta|y)$ in (5), and as a probability distribution must integrate to 1.

So we have two identical kernels from our Posterior in (5) and this new Gamma distribution in (6).

If we know what the kernel integrates to we can use the fact that a probability distribution must integrate to 1 to find the full density function. If

$$\int_{\theta\in\Theta} p(\theta|y)^* d\theta = D,$$

then

$$p(\theta|y) = \frac{p(\theta|y)^*}{D}$$

as

$$\int_{\theta\in\Theta} p(\theta|y) d\theta = \frac{\int_{\theta\in\Theta} p(\theta|y)^* d\theta}{D} = 1.$$

The constant is always the inverse of the integral of the kernel. As our kernels are the same, their integrals are the same and hence the constants are the same. Therefore we know that our unknown constant from the posterior, $AB/p(y)$ in (4), must be the same as the known constant from the Gamma distribution in (6),

$$\frac{(\beta + 1/2\sum_{i=1}^n y_i^2)^{\alpha+n/2}}{\Gamma(\alpha+n/2)}.$$

So both distributions are identical, and we can say that $p(\theta|y)$ is Gamma($\alpha + n/2, \beta + 1/2 \sum_{i=1}^{n} y_i^2$).

We didn't need to hold onto the constant terms in the likelihood and prior to find the posterior. Instead, we just said that the posterior is propotional to the product of the kernels of the likelihood and prior. This approach is much easier than holding onto both constants, performing the integration $p(y) = \int_{\theta \in \Theta} p(y|\theta)p(\theta)d\theta$, and then putting everything together in (1). However, we do need to be able to look at the product of kernels and say *'This looks like a Whatever-Distribution kernel'* and use the known constant from that distribution.

If you use a conjugate pair for the likelihood and prior, then the posterior will be from the same distibutional family (eg. a different gamma) as the prior and it's easy to recognize the kernel. It is no coincidence you saw a Gaussian, Beta and Gamma posterior in this lab.

A lot of the time the model will be too complex to use the easy conjugate pair setup and the kernel comes from a distribution you can't recognize. That's where Bayesian Statistics begins to get interesting.