```
## Loading required package:  ggplot2
## rstan (Version 2.9.0, packaged:  2016-01-05 16:17:47 UTC, GitRev:
05c3d0058b6a)
## For execution on a local, multicore CPU with excess RAM we recommend
calling
## rstan_options(auto_write = TRUE)
## options(mc.cores = parallel::detectCores())
```

STAT 544 Mid-term Exam
Thursday 10 March 8:00-9:20

Instructor: Jarad Niemi

INSTRUCTIONS

Please check to make sure you have 5 pages with writing on the front and back (some pages are marked 'intentionally left blank'). Feel free to remove the last page, i.e. the one with R code.

On the following pages you will find short answer questions related to the topics we covered in class for a total of 50 points. Please read the directions carefully.

You are allowed to use a calculator and one $8\frac{1}{2} \times 11$ sheet of paper with writing on both front and back. A non-exhaustive list of items you are not allowed to use are **cell phones, laptops, PDAs, and textbooks**. Cheating will not be tolerated. Anyone caught cheating will receive an automatic F on the exam. In addition the incident will be reported, and dealt with according to University's Academic Dishonesty regulations. Please refrain from talking to your peers, exchanging papers, writing utensils or other objects, or walking around the room. All of these activities can be considered cheating. **If you have any questions, please raise your hand.**

You will be given only the time allotted for the course; no extra time will be given.

Good Luck!

Please print your name below:

Student Name: _____

(intentionally left blank)

Diagnostic testing

1. Details of two diagnostic tests for a fairly prevalent disease is given below.

Test	Sensitivity	Specificity
CDC	0.90	0.80
MD	0.70	0.60

Recall that sensitivity is the probability of a positive test result when the individual has the disease and specificity is the probability of a negative test result when the individual does not have the disease. Suppose the prevalence of this disease is one case per ten individuals.

(a) Determine the marginal probability of a positive CDC test. Show all your work and define all your notation. (5 pts)

Notation

- D the individual has the disease
- D^c the individual does not have the disease
- $+$ the CDC test was positive
- $-$ the CDC test was negative

$$P(+) = P(+|D)P(D) + P(+|D^c)P(D^c) = P(+|D)P(D) + [1 - P(-|D^c)][1 - P(D)]$$
$$= 0.9 \times 0.1 + [0.2][0.9] = 0.27$$

(b) Calculate the probability of having Zika for an individual with a positive CDC test. (5 pts)

$$P(D|+) = \frac{P(+|D)P(D)}{P(+)} = \frac{0.9 \times 0.1}{0.27} = \frac{1}{3}$$

The table below provides the cost for each combination of true disease state and diagnostic test result.

Truth	Test result	Cost ($)	Probability CDC	Probability MD
Disease	Positive	0	0.09	0.07
Disease	Negative	10	0.01	0.03
No disease	Positive	1	0.18	0.36
No disease	Negative	0	0.72	0.54

(c) Complete the table above by calculating the probabilities for each truth-test result combination for each of the two diagnostic tests. Define your notation and show your work here. (5 pts)

Let CDC indicate the CDC test and MD indicate the MD test. Otherwise the notation is the same as it was before. Note that the prevalence is independent of the test being used.

$$P(D,+|CDC) = P(+|D,CDC)P(D) = 0.9 \times 0.1 = 0.09$$
$$P(D,-|CDC) = P(-|D,CDC)P(D) = 0.1 \times 0.1 = 0.01$$
$$P(D^c,+|CDC) = P(+|D^c,CDC)P(D) = 0.2 \times 0.9 = 0.18$$
$$P(D^c,-|CDC) = P(-|D^c,CDC)P(D) = 0.0 \times 0.9 = 0.72$$
$$P(D,+|MD) = P(+|D,MD)P(D) = 0.7 \times 0.1 = 0.07$$
$$P(D,-|MD) = P(-|D,MD)P(D) = 0.3 \times 0.1 = 0.03$$
$$P(D^c,+|MD) = P(+|D^c,MD)P(D) = 0.4 \times 0.9 = 0.36$$
$$P(D^c,-|MD) = P(-|D^c,MD)P(D) = 0.6 \times 0.9 = 0.54$$

(d) The cost for the CDC test is $50 while the cost for the MD test is $10. Determine the optimal diagnostic test to use. (5 pts)

The easiest approach is to realize that the cost to administer the test is so different between these two that it would take a lot for the CDC test to be preferred over the MD test. If the CDC test was perfect (which it is not) such that you never incurred a penalty and the MD test was always wrong (which it is not), then the cost for using the MD test would be $10 (to administer) and at most $10 if the individual has the disease and the test said they didn't. Thus, even in this scenario you would prefer the MD test (although you may question your loss function).

Formally, the optimal diagnostic test is the one that will minimize expected costs. The expected costs for the CDC test is $50 + 10 \times 0.01 + 1 \times 0.18 = 50.28$ while the expected costs for the MD test is $10 + 10 \times 0.03 + 1 \times 0.36 = 10.66$. Thus the optimal test is the MD test.

Prior-posterior

2. Suppose Y has the probability density function

$$p(y|\theta) = (\theta+1)y^{\theta}\mathrm{I}(0<y<1) \text{ for } \theta > 0.$$

(a) Derive Jeffreys prior for this model. (If you are having trouble, I will give you the answer and mark off 2 points. You can still obtain the remaining 6 points by deriving the answer now that you know it.) (8 pts)

$$\begin{aligned}
\log L(\theta) &= \theta \log(y) + \log(\theta+1) \\
\frac{\partial}{\partial\theta} \log L(\theta) &= \log(y) + \frac{1}{\theta+1} \\
\frac{\partial^2}{\partial\theta^2} \log L(\theta) &= -\frac{1}{(\theta+1)^2} \\
\mathcal{I}(\theta) &= -E\left[\frac{\partial^2}{\partial\theta^2} \log L(\theta)\right] = \frac{1}{(\theta+1)^2} \\
p(\beta) &\propto \sqrt{|\mathcal{I}(\beta)|} = \frac{1}{\theta+1}
\end{aligned}$$

(b) Is Jeffreys prior proper? (2 pt)

No, this prior is not proper because $\int_0^\infty \frac{1}{\theta+1} = \int_1^\infty \frac{1}{t} dt$ diverges.

(c) Derive the posterior for one observation under Jeffreys prior. (4 pts)
The posterior is

$$p(\theta|y) \propto p(y|\theta)p(\theta) = (\theta+1)y^\theta \frac{1}{\theta+1}\mathrm{I}(0<y<1) = y^\theta \mathrm{I}(0<y<1)$$

(d) When is this posterior proper? (Hint: $y^\theta = e^{\theta \log y}$.) (3 pts)

$$p(y) = \int p(y|\theta)p(\theta)d\theta = \int_0^\infty e^{\theta \log y} d\theta = \int_0^\infty \frac{1}{\log y} e^{\theta \log y} d\theta = \frac{1}{\log y}(0-1) = -\frac{1}{\log y}$$

which is finite for any $0 < y < 1$. So the posterior is always proper.

(e) Find a 95% HPD interval for θ for a single observation y. (3 pts)
Since the posterior is monotonically decreasing, a 95% HPD will be an interval of the form $(0, c)$ where c is obtained from

$$0.95 = \frac{\int_0^c y^\theta d\theta}{\int_0^c y^\theta d\theta} = \frac{\frac{1}{\log y}(e^{c \log y} - 1)}{\frac{1}{\log y}(0-1)} = 1 - e^{c \log y}.$$

Solving for c we obtain

$$c = \frac{\log 0.05}{\log y} = \log_y(0.05)$$

Pilot training

3. Airplane pilots were randomly assigned to a group where each group used a different teaching method. The pilots in the group where randomly assigned to a training scenario. The result of an individual pilot's success on the training scenario was binary: a 1 if the pilot successfully met the scenario's objectives and a 0 otherwise. Each group-scenario combination had from 5 to 8 pilots and the response p is the proportion of pilots in that group-scenario combination to successfully satisfy the scenario's objectives.

   (a) Using statistical notation, write down the model (including priors) encoded by Stan. (5 pts)

   Let p_i be the proportion of successes in the ith group-scenario combination. Let $s[i]$ and $g[i]$ indicate the group and scenario, respectively for the ith combination. The model is

$$
\begin{aligned}
p_i &\overset{ind}{\sim} N(\eta + \gamma[g[i]] + \delta[s[i]], \sigma^2) \\
\gamma_g &\overset{ind}{\sim} N(0, \sigma_\gamma^2) \\
\delta_g &\overset{ind}{\sim} N(0, \sigma_\delta^2) \\
p(\eta, \sigma, \sigma_\gamma, \sigma_\delta) &\propto Unif(\sigma; 0, 1) Unif(\sigma_\gamma; 0, 1) Unif(\sigma_\delta; 0, 1)
\end{aligned}
$$

   (b) Are the priors on the standard deviations reasonable? Why or why not? (3 pts)

   Yes. The range of the data is only 0 to 1, so these standard deviations are certainly no larger than 1.

   (c) Are you concerned about convergence of **this** Markov chain? Why or why not? (2 pts)

   There is no concern about convergence of the Markov chain since all potential scale reduction factors (Rhats) are close to one and the effective samples sizes are large.

(d) Provide a 95% credible interval for the overall proportion of successes. (1 pt)

(0.13,.76)

(e) Are the mean proportion of successes for groups or scenarios more variable? Explain your answer. (2 pts)

Scenarios appear more variable than groups since the estimated standard deviation is larger, i.e. (0.21,.75) compared to (0.00,0.24).

(f) Suggest a test statistic for which you could calculate a posterior predictive pvalue that could identify heavy-tails in the scenario means and defend this suggestion. (3 pts)

This is a difficult question. I gave full credit for anyone who provided a test statistical that looked at the tails of the data, e.g. how many zeros or ones there were.

(g) Using statistical notation, write down a model (without priors) that accommodates the discreteness of the response. (4 pts)

Let y_i be the number of successes in combination i and n_i be the number of attempts. Then a model that seems reasonable is

$$y_i \overset{ind}{\sim} Bin(n_i, \theta_i)$$
$$\text{logit}(\theta_i) = \eta + \gamma[g[i]] + \delta[s[i]]$$

with the hierarchical structure given in part a).

Stan code

```
pilots_stan = "
data {
  int<lower=0> N;
  int<lower=0> n_groups;
  int<lower=0> n_scenarios;
  int<lower=1,upper=n_groups> group_id[N];
  int<lower=1,upper=n_scenarios> scenario_id[N];
  vector[N] p;
}
parameters {
  vector[n_groups] gamma;
  vector[n_scenarios] delta;
  real eta;
  real<lower=0,upper=1> sigma_gamma;
  real<lower=0,upper=1> sigma_delta;
  real<lower=0,upper=1> sigma;
}
transformed parameters {
  vector[N] mu;

  for (i in 1:N)
    mu[i] <- eta + gamma[group_id[i]] + delta[scenario_id[i]];
}
model {
  gamma ~ normal(0, sigma_gamma);
  delta ~ normal(0, sigma_delta);
  p ~ normal(mu, sigma);
}
"
library(rstan)
m = stan_model(model_code = pilots_stan)
r = sampling(m,
             data = dat,
             pars = c('mu'), include = FALSE, # removes mu from the output
             iter = 20000,
             chains = 4)
```

Stan results

r

```
## Inference for Stan model: a5814b234db6af24770e1cd8dc27fa6d.
## 4 chains, each with iter=20000; warmup=10000; thin=1;
## post-warmup draws per chain=10000, total post-warmup draws=40000.
##
##              mean se_mean   sd  2.5%   25%   50%   75% 97.5% n_eff Rhat
## gamma[1]    -0.01    0.00 0.06 -0.15 -0.04 -0.01  0.01  0.10  6357    1
## gamma[2]    -0.01    0.00 0.06 -0.14 -0.03  0.00  0.02  0.11  6023    1
## gamma[3]     0.00    0.00 0.06 -0.12 -0.02  0.00  0.03  0.13  5959    1
## gamma[4]    -0.01    0.00 0.06 -0.15 -0.03  0.00  0.01  0.10  6434    1
## gamma[5]     0.03    0.00 0.06 -0.08 -0.01  0.01  0.05  0.18  5039    1
## delta[1]    -0.09    0.00 0.17 -0.45 -0.20 -0.09  0.02  0.25  3409    1
## delta[2]    -0.28    0.00 0.18 -0.64 -0.39 -0.28 -0.17  0.06  3387    1
## delta[3]     0.00    0.00 0.18 -0.35 -0.11  0.00  0.11  0.34  2762    1
## delta[4]    -0.22    0.00 0.17 -0.58 -0.33 -0.22 -0.12  0.12  4312    1
## delta[5]    -0.02    0.00 0.17 -0.37 -0.12 -0.02  0.09  0.33  3342    1
## delta[6]     0.49    0.00 0.18  0.15  0.37  0.48  0.60  0.84  4012    1
## delta[7]    -0.31    0.00 0.18 -0.67 -0.41 -0.30 -0.20  0.03  4222    1
## delta[8]     0.44    0.00 0.18  0.10  0.33  0.44  0.55  0.80  3467    1
## eta          0.44    0.00 0.16  0.12  0.35  0.44  0.53  0.76  2831    1
## sigma_gamma  0.07    0.00 0.07  0.01  0.02  0.05  0.08  0.25  3048    1
## sigma_delta  0.39    0.00 0.14  0.21  0.30  0.37  0.46  0.74  7796    1
## sigma        0.23    0.00 0.03  0.18  0.21  0.22  0.25  0.30 10118    1
## lp__        51.60    0.11 4.97 41.38 48.43 51.69 54.99 61.06  2127    1
##
## Samples were drawn using NUTS(diag_e) at Thu Mar 10 09:51:40 2016.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```