STAT 4880/5088 Exam 2

Name: ____

Tuesday, November 5, 2024

Write directly on this exam. You may "show work" by handing in and R script, .Rmd file, or knit Markdown document.

You may use R, the internet, and any reference material. You are not allowed to communicate with anyone - no email, messaging, internet forums, AI, etc. If you happen to find exact copies of the exam questions on a "homework help" website, please bring that to the instructor's attention.

Honor Pledge (10 points)

The work I have submitted represents my own effort. While working on this exam, I did not use AI or communicate in any form with individuals other than the instructor.

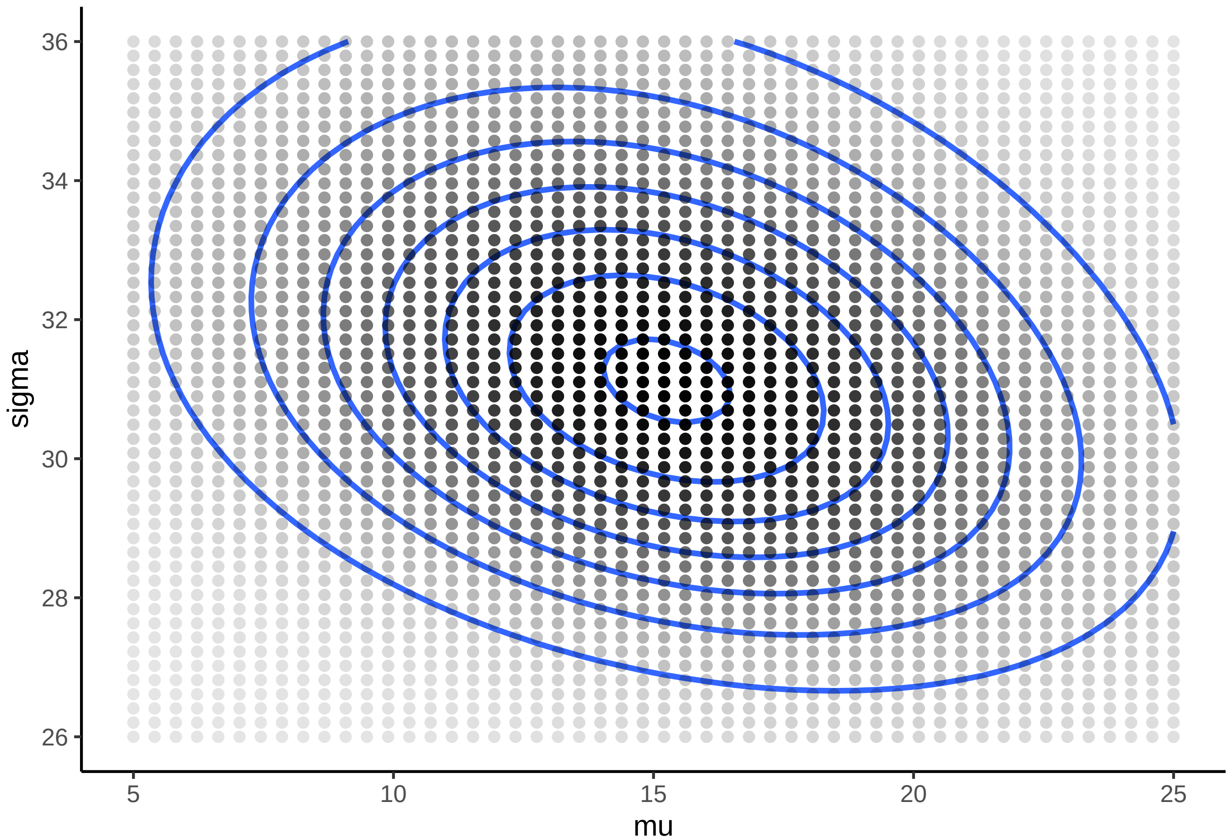
Signed:

____

Problem 1 (10 points)

Consider a normal data model with unknown mean μ and standard deviation σ. Here is a grid approximation to the posterior, shown with the grid shaded and approximate contours.

a. Estimate (by eye) the values of μ and σ at the mode of the posterior.
b. The prior for this model assumes that μ and σ are independent. How can you tell from the picture that this is not a conjugate prior model?



Solution

a. $\mu = 15$, $\sigma = 31$, plus or minus 1 at most.
b. The parameters μ and σ are no longer independent, they are negatively correlated. The downward slanted contours show that larger μ values correspond to smaller σ.

Problem 2 (10 points)

Suppose you are running a Metropolis MCMC to sample from a posterior which is normal with mean 10 and sd 2, and your proposal function is symmetric.

a. If the current value is $\theta = 11$, what is the acceptance probability for a proposed value of $\theta' = 12$?
b. If the current value is $\theta = 11$, what is the acceptance probability for a proposed value of $\theta' = 10$?
c. If the current value is $\theta = 11$, what is the acceptance probability for a proposed value of $\theta' = 16$?

Solution a. 0.69; b. 1; c. 0.013.

```
dnorm(c(12,10,16),10,2)/dnorm(11,10,2)
```

```
## [1] 0.68728928 1.13314845 0.01258814
```

Problem 3 (10 points)

a. By default, the stan sampler discards half of every chain. Why?
b. By default, the stan sampler runs four chains. Why not run just one chain for four times as long?

Solution

a. Chains need time to converge to the target distribution. The early values may not be representative, so we discard them.

b. If the chains aren't mixing well, the starting position will have a long-term effect on the chain. Starting four chains in different places allows us to detect that.

Problem 4 (10 points)

MCMC samplers (like stan) don't always provide an accurate sample of the target posterior distribution.

What are some diagnostics would indicate that your sampler was unsuccessful?

Solution A traceplot can show that chains are mixing slowly or not in agreement with each other. Low effective sample size indicates the chains aren't mixing. An R-hat value below one indicates the chains have not reached agreement.

Problem 5 (10 points)

For parameter π, suppose you have a Beta(2,2) prior model and a Beta(220,240) posterior.

a. What is the posterior probability of the hypothesis $\pi < 0.5$?

b. What are the posterior odds for the hypothesis $\pi < 0.5$?

c. Compute the Bayes factor for this hypothesis.

Solution The posterior probability is 0.825. The posterior odds are 4.7 to 1. The Bayes factor is also 4.7 since the prior odds are 1 to 1.

```r
postp <- pbeta(0.5, 220, 240)
postp
```

```
## [1] 0.8247185
```

```r
postp/(1-postp)
```

```
## [1] 4.705108
```

The final four problems are all about a model we could call the "Normal-Uniform" model.

Problem 6 (10 points)

In the Normal-Normal model, the normal prior can be vague but always has a peak somewhere.

a. The prior in the Normal-Normal model is $\mu \sim N(\theta, \tau^2)$. If you want to make the prior less informative, what can you do?

An alternative would be to choose a uniform prior on some interval that surely covers the possible parameter values. This model looks like:

$$Y \sim N(\mu, \sigma^2); \quad \mu \sim \text{uniform}(a, b)$$

where a and b are chosen constants and for simplicity we assume σ is known.

b. Is the Normal-Uniform model a conjugate prior model?

Solution a) You can increase the standard deviation τ. b) No, the posterior distribution will not be uniform.

Problem 7 (10 points)

Continue with the Normal-Uniform model. Let's use data from https://turtlegraphics.org/bayes/data/normtemp.csv on human body temperature, let $\sigma = 0.7$, and choose $(a, b) = (98, 99)$

Code this in stan by changing the Normal-Normal model to have `mu ~ uniform(98,99)` in the model.

If you run this, it should fail with a whole stream of errors that say:

```
Chain 1: Rejecting initial value:
Chain 1:   Log probability evaluates to log(0), i.e. negative infinity.
Chain 1:   Stan can't start sampling from this initial value.
```

a. What's going wrong?

You can fix the errors by changing the parameter `mu` to have type `real<lower=98, upper=99> mu;`

b. Why did this fix the errors?

c. Hand in your working stan program.

Solution

a. The initial value for the chain is not in the interval [98,99] and so the acceptance probability has division by zero.
b. stan now picks an initial value in the range [98,99]

```
nu_prog <- "
  data {  vector[130] y; }
  parameters { real<lower=98, upper=99>  mu; }
  model {
    y ~ normal(mu, 0.7);
    mu ~ uniform(98,99);
  }
"
```

Problem 8 (10 points)

You should now have stan code to generate samples from the posterior of the Normal-Uniform model.

Use your model to estimate the middle 95% credible interval for the temperature parameter μ.

Solution About [98.13,98.37].

```
normtemp <- read.csv("https://turtlegraphics.org/bayes/data/normtemp.csv")
temp_mod <- stan(model_code = nu_prog, data = list(y=normtemp$temperature))
```

```
## Trying to compile a simple C file
```

```
mu <- extract(temp_mod)$mu
quantile(mu, c(0.025, 0.975))
```

```
##     2.5%    97.5%
## 98.12474 98.36973
```

Problem 9 (10 points)

Use your samples from the Normal-Uniform posterior to generate a sample from the posterior predictive distribution $Y'|Y$.

What is the posterior probability that a healthy person has body temperature over 98.6? That is, estimate $P(Y' > 98.6|Y)$.

Solution

```r
y_prime <- rnorm(4000, mu, 0.7)
mean(y_prime > 98.6)
```

```
## [1] 0.3115
```