2001 First Year Exam
In Class Portion

1. Please use a **separate page for each problem** so different people can grade different problems.
2. If you use extra paper, please indicate your **code and the problem number on top of each page.**
3. This is a closed book exam. No notes are permitted.
4. It is to your advantage to show your work and explain your answers. Don't erase. Draw a line through work you don't want graded.
5. Put your code on each page of your exam.
6. You should attempt all questions, and credit will be given for partial answers.
7. You have three hours to finish.
8. Tear off the last pages with the take-home problem and submit it by tomorrow 5pm to my office.

Problem 1.

For each of the following statements, tell if they are true or false.
If true, give a proof. If false, give a counterexample.

1a. If X and Y are independent random variables, then $X + Y$ is independent of $X - Y$.

1b. If X and Y are independent random variables, then X^2 is independent of Y^2.

1c. If $E(XY) = E(X) \cdot E(Y)$, then X and Y are independent.

1d. If $E(X) = E(Y)$ and $\mathrm{Var}(X) = \mathrm{Var}(Y) = 0$, then X and Y are independent.

Problem 2.

Let $X_1, \ldots, X_n \sim N(\mu, 4)$, i.i.d.

2a. Find a minimal sufficient statistic for μ.

2b. Find a likelihood ratio test for testing $\mu = 0$ against $\mu = 1$.
Express the test in terms of the minimal sufficient statistic.

2c. Let $\alpha = Pr\{\text{reject } \mu = 0 | \mu = 0\}$ and $\beta = Pr\{\text{accept } \mu = 0 | \mu = 1\}$ for the test in part b).
Find n and a likelihood ratio test such that $\alpha \approx 0.01$ and $\beta \approx 0.10$.

Problem 3.

Consider a normal sampling model

$$y_i \overset{ind}{\sim} N(\mu_i, 1), \quad i = 1, 2$$

with a normal prior for (μ_1, μ_2)

$$\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \sim N\left\{ \begin{pmatrix} m \\ m \end{pmatrix}, \varphi \begin{bmatrix} a & b \\ b & a \end{bmatrix} \right\} \tag{1}$$

and fixed hyperparameters m, φ, a and b.

3a. Let $d = \mu_1 - \mu_2$. Find the posterior distribution $p(d \,|\, y_1, y_2)$.

Now consider an alternative prior parameterization:

$$\begin{gathered} \mu_i = u + \tau_i, \quad i = 1, 2 \\ u \sim N(v, \sigma) \text{ and } \tau_i \overset{ind}{\sim} N(0, \rho), \quad i = 1, 2 \end{gathered} \tag{2}$$

and fixed hyperparameters v, σ and ρ.

3b. Under model (2), find the implied prior $p(\mu_1, \mu_2)$.

3c. Can the hyperparameters (v, σ, ρ) in model (2) be fixed such that the posterior $p(d \,|\, y_1, y_2)$ under (2) matches the posterior distribution which you found in part **3a.** for model (1)?

If yes, describe how. If no, give a proof.

Problem 4.

A single observation X is to be drawn from pdf

$$f(x \mid \theta) = 2(1-\theta)x + \theta, \qquad 0 < x < 1$$

where $\theta \in [0, 2]$ is unknown.

4a. Sketch the likelihood function.

4b. Find the Maximum Likelihood Estimator of θ, as a function of x:

4c. Find the Posterior Mean Estimator of θ, as a function of x, for a uniform prior on θ.

4d. Is this an Exponential Family? If not, explain why. If so, put it in canonical form

$$f(x \mid \theta) = h(x) \cdot \exp[\eta(\theta) \cdot T(X) - B(\theta)]$$

4e. For a large number n of observations, if we *only* observe the number Y_n of observations X_i, $1 \leq i \leq n$ which exceed $1/2$ (we do not get to observe the X_i themselves), then what is the MLE for θ? (*Hint:* Recall the constraint $0 \leq \theta \leq 2$).

Problem 5.

A genomic technique aims to measure whether or not a particular gene is actively "expressed" in a biological sample. In one experiment, indexed i, the interest lies in a binary variable $x_i = 0$ (if the gene is unexpressed in sample i) or $x_i = 1$ (if the gene is expressed). A set of n independent biological samples therefore provides n independent observations on $X = \{x_1, \ldots, x_n\}$. Assume the x_i to be independent Bernoulli random quantities, and that the population probability of the gene being expressed is θ.

5a. Write down the likelihood function for θ based on the observed data X. Derive the formula for the MLE of θ.

5b. An investigator familiar with the gene in question from prior studies indicates that, based on these prior studies, he expects the gene to be expressed in about 25% of similar samples, and that he views it as very unlikely indeed that it expresses in more than 40% of samples. Describe how you could use this opinion to specify informative prior distributions for θ.

5c. The prior information above is used to determine a beta prior, $Beta(15, 45)$. What is the posterior distribution, and what are the posterior mean and standard deviation?

5d. Derive the asymptotic normal approximation to the posterior distribution for θ assuming a uniform prior. Use this to evaluate the endpoints of an approximate 95% posterior interval for θ in the case $n = 25$ and $\bar{x} = 0.172$.

5e. What is the asymptotic normal approximation to the sampling distribution of $\bar{x}$? Describe how you can use this to compute an approximate 95% confidence interval for θ, clearly stating any additional assumptions or approximations.

The genomic lab method is actually not perfect. Due to instrument noise and other experimental features, there is about a 5% error rate when the gene is not expressed, and about a 1% error rate when it is expressed. That is, the technique cannot directly measure x_i, but instead it measures the binary outcome y_i such that $Pr(y_i = 0|x_i = 0) = 0.95$ and $Pr(y_i = 1|x_i = 1) = 0.99$, whatever the value of θ may be.

5f. Show that $Pr(y = 1|\theta) = (94\theta + 5)/100$.

5g. Write down the likelihood function for θ based on data $Y = \{y_1, \ldots, y_n\}$.

5h. Describe a numerical strategy for computing the MLE of θ in this model.

5i. Under a specified prior $p(\theta)$, indicate how you might develop a computational strategy to explore the posterior and compute the posterior mean and a 95% posterior interval for θ.

Take home portion of First Year Exam, 2001

Problem 6 – Diversity in Insect Fossils

A biological reseracher hypothesizes that evolutionary lineages with greater degrees of *modularity* should exhibit higher rates of diversification.

Modularity describes the degree to which an organism's traits are independent from one another. Thus, highly modular organisms can be thought of as a union of coherently linked components, with the components functioning fairly independently of one another. Because of this independence between components, it is expected that modular organisms more easily diversify over large (macroevolutionary) time scales.

This application focuses on two groups or *taxa* (singular form is *taxon*) of insects. Here, a taxon is a collection of similar insect species – both extinct and current. A given taxon may be further refined into orders of species. A given order may be refined into families of species. For a given order or taxon, it is believed that the number of unique families is a good measure of diversity. Hence if one taxon has a higher rate of diversification, then that taxa should consistantly produce more unique families over large (macroevlolutionary) time scales.

In order to investigate this hypothesis, data from fossil records have been collected on two different taxa of insects. The first taxon, *Holometabola*, exhibits a high degree of modularity. It is hypothesized that this modularity facilitates the formation of new families of species due to evolution over macroevolutionary time periods. The second taxon, *Hemimetabola*, is far less modular. Hence it is hypothesized that evolution will, on average, produce new families at a lower rate relative to that of the Holometabola taxon.

The data in the file

`www.stat.duke.edu/~pm/fye/insectfossils.csv`

gives the number of unique families detected in the Holometabola and Hemimetabola taxa over 53 different epochs (time periods) of the fossil record. The first column gives the epoch name, the second column gives the time of the epoch in *millions of years ago*. Columns 3 – 12 give the number of families recorded for each order in the Holometabola taxa. Columns 13 – 37 give the number of families recorded for each order within the Hemimetabola taxa. The first row gives the column names. The names of columns 3 – 12 end in O to remind you they correspond to the HOlometabola taxon; the names of columns 13 – 37 end in E to remind you they correspond to the HEmimetabola taxon.

The data file looks like

```
epoch    mya     trichopO ... raphidioO  hemipE ... protorthopE
present  0       43       ...  2         138    ...   0
holo     0.01    21       ...  2          82    ...   0
plei     0.83    21       ...  2          82    ...   0
 :        :       :       ...  :          :     ...   :
prag   248.75     1       ...  0          19    ...   1
loch   251.25     1       ...  0          12    ...   1
```

And can be read into Splus with read.table using the following commands:

```
> insect_read.table("insectfossils.csv",header=T,sep=",")
```

Use the above dataset to address the following question:

6a. Does the modular Holometabola lineage exhibit greater rate of family-level diversification than the less modular Hemimetabola lineage?

6b. Address this question using aggregated taxa level data. If time permits, you may condsider the order level data as well.

6c. Try to carry out a graphical analysis which suggests an answer, as well as a more quantitative analysis.

Summarize your findings in a **2 page** document.
A good graphical exploratory data analysis is preferable to overly complex incomplete formal inference.

Supplementary documentation may be added to an appendix of no more than 5 pages. However grading will be limited to your 2 page summary. Dave or Peter will be present Tuesday morning (from 9am until 11am) to answer any questions you might have. The exam is due in Peter's office by 5pm on Tuesday May 15. Late answers will not be accepted.

The following facts might help you in your analysis:

- The Homometabola and Hemimetabola taxa only became distinct about 300 million years ago. Hence only epochs from no more than 250 million years ago are included in this dataset.
- The actual rate of diversification is of interest rather than the total number of families produced. You should consider a response that captures this. One possibility is the relative change in families from one epoch to the next. If y_{ij} is the number of families recorded at epoch i for order j then
$$r_{ij} = (y_{i+1,j} - y_{ij})/y_{ij}$$
gives the observed rate at which new families were created between epochs i and $i+1$. You will have to decide how best to deal with zero divisions here.
- When a given $y_{ij} = 0$, it means that no species were found that belonged to order j at epoch i. This does not necessarily mean that all the species within that order have gone extinct.
- The environmental conditions may vary substantially from one epoch to the next. Any analysis you carry out should account for this.
- A simple – but thoughtful – analysis is almost always preferable to a rushed, complicated analysis.
- You should focus on exploratory data analysis and simple models/tests. There is no need to build a sophisticated model.