PROBLEM SET # 2

EC/ACM/CS 112: Bayesian Statistics
Caltech

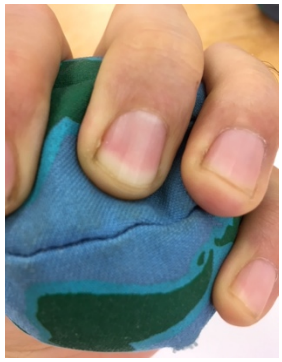
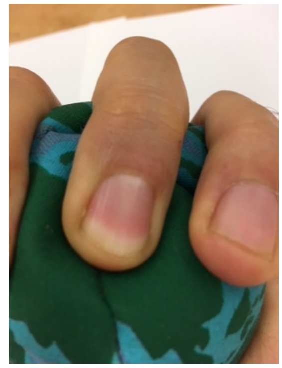
Submission instructions	>> Create a pdf of the R Notebook with your solutions (details below) >> Submit in Canvas
Additional files included in the problem set package	>> dataset problem set 2 >> solutions template (.Rmd file)

PART 1
IMPACT OF MORE DATA ON THE BALL TOSSING PROBLEM

BACKGROUND

- Remember the application of the Binomial Model (BM) that we carried out in Lecture 2.
- We have two juggling balls with maps of the Earth on their surface

- We are interested in applying the Binomial Model to estimate the percentage of surface in each of the balls that represents land (in green) or water (in dark or light blue).
- To do this, we tossed each ball in the air 100 times and recorded whether, after catching it, the base of the middle finger touches water or land

- We then estimated a bivariate binomial model with following assumptions:
  ++ $Prior(p_1) = Beta(5,5)$, where p_1 denotes the probability of a water landing for ball 1
  ++ $Prior(p_2) = Beta(5,5)$, where p_2 denotes the probability of a water landing for ball 2
  ++ The two priors are independent
  ++ The tosses across balls and across trials are independent

DATASET

- The dataset for this problem set is in the file "PS2_data.csv" included with this package
- The dataset is similar to the one used in class, except for two differences:
  ++ There is a new variable, called *tosser*, which indicates the name of person tossing the ball
  ++ There are twice as many observations for each ball: the initial 100 observations were generated by Prof. Rangel (indicated by ar); the next 100 were generated by a student (indicated by nh).

GOAL

- To explore the role that additional data and variation across samples has in the estimates of a simple binomial model

TO DO

STEP 1 (1 point): Estimate and summarize the model using only the original data collected by ar

- Plot the joint posterior density as a heat map
- Plot the marginal posterior densities for p_1 and p_2 in a single plot (add a figure legend identifying the two curves)
- Compute the mean and standard deviation of the posterior marginal distributions
- Compute the posterior probability that $p_1 < p_2$

STEP 2 (0.5 points). Repeat step 1 using <u>only</u> the new data collected by nh.

STEP 3 (0.5 points). Repeat step 1 using <u>all</u> of the data collected.

POTENTIALLY USEFUL MATERIALS	
R functions	colSums() rowSums() legend() levelplot() – in lattice package

PART 2
MODEL CHECKING: ROLE OF INDEPENDENT PRIORS

BACKGROUND

- A potential concern with the basic model used in class is that, given that the two juggling balls come from the same manufacturer, it is likely that the values of p_1 and p_2 are correlated.
- Thus, the assumption that the joint priors are uncorrelated is a potential concern

GOAL

- Explore the role that the degree of correlation in the joint priors has on the posterior distribution.

TO DO

STEP 1 (2 points). Build a function that takes as inputs the five parameters that describe a bivariate normal distribution and returns a prior matrix based on the associated bivariate normal distribution, truncated to [0,1] X [0,1]

- Start with a refresher or the bivariate distribution (e.g., here http://mathworld.wolfram.com/BivariateNormalDistribution.html)
- Recall that the bivariate normal distribution is described by the following objects:
  ++ The means for p_1 and p_2 denoted, respectively, μ_1 and μ_2
  ++ A covariance matrix with parameters $\begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$,
  where σ_i^2 denotes the variance for ball-*I*, and ρ denotes the correlation coefficient.
- Make sure that you understand the impact of the different parameters on the shape of the distribution.

- Build a function that takes as input a vector (p_1, p_2) and the parameters that describe the bivariate normal and returns the bivariate normal density at that vector.

- Build a function that takes as inputs the five parameters that describe a bivariate normal distribution and returns a prior matrix based on the associated bivariate normal distribution, truncated to [0,1] X [0,1]
- The prior matrix should have resolution 100 x 100
- Note that summing over all entries of the prior matrix you should get that sum(prior matrix) * $\text{gridSize}^2 \approx 1$ (see Lecture 2 for a related discussion in the case of one parameter).
- In order to test your work, plot the resulting prior matrix as a heat map for each of the following parameter combinations:
  a) $\mu_1 = \mu_2 = 0.5$, $\sigma_1 = \sigma_2 = 1$, and $\rho = 0$.
  b) $\mu_1 = \mu_2 = 0.5$, $\sigma_1 = \sigma_2 = 1$, and $\rho = 0.25$.
  c) $\mu_1 = \mu_2 = 0.5$, $\sigma_1 = \sigma_2 = 1$, and $\rho = 0.5$

STEP 2 (1 point). Re-estimate the model in Part 1 using all of the available data for each of the following priors:
a) $\mu_1 = \mu_2 = 0.5$, $\sigma_1 = \sigma_2 = 1$, and $\rho = 0$.
b) $\mu_1 = \mu_2 = 0.5$, $\sigma_1 = \sigma_2 = 1$, and $\rho = 0.25$.
c) $\mu_1 = \mu_2 = 0.5$, $\sigma_1 = \sigma_2 = 1$, and $\rho = 0.5$

In each case,
- Plot the joint posterior density as a heat map
- Plot the marginal posterior densities for p_1 and p_2 in a single plot (add a figure legend identifying the two curves)
- Compute the mean and standard deviation of the posterior marginal distributions
- Compute the posterior probability that $p_1 < p_2$

STEP 3 (1 point). What do the results suggest about the concern regarding the uncorrelated priors, given the amount of available data?

PART 3
MODEL CHECKING: ROLE OF MEASUREMENT ERROR

BACKGROUND
- Another potential concern with the basic model is that it ignores the problem of measurement error.
- This is a natural concern, since as the following image shows, sometimes it is hard to determine the location of the ball landing using our simple measurement method

GOAL

- Use synthetic data to explore the impact that ignoring this issue could have on the quality of our statistical model

A SIMPLE MODEL OF BALL TOSSING WITH MEASUREMENT ERROR

- In order to explore this issue we need to modify our simple model to allow for the possibility of measurement error.
- The idea of the modified model is simple.
- Every time we toss a ball, with probability $1 - \theta$ it lands in a region where there is no measurement error (e.g., as in the pictures shown in PART 1), but with probability θ it lands in a region where it is very hard to tell (as in the picture just above).
- When the latter case occurs, our measurement is *Water* with probability 50% and *Land* with probability 50%, irrespective of the true location of the landing.
- Details on how to implement the model in code are given below.

TO DO

STEP 1 (3 points). Use synthetic data to understand the impact that ignoring measurement error can have on the quality of our statistical model.

- In order to facilitate replication of your simulation results, initialize the random number generator seed using the command **set.seed(123)**
- Carry out 10,000 simulations of the following steps:
  ++ Randomly select the parameters for each simulation step by choosing $p_{True} \sim Unif(0,1)$ and θ_{True} ~*Unif*({*0.05, 0.15, 0.25*}).
  ++ In each step, build a dataset of 100 tosses of a ball with probability of a water landing given by p_{True} under the assumption that there are no errors. Call it dataNoError.
  ++ In each step, build a closely related dataset, call it dataWithError, by starting with dataNoError and then selecting uniformly at random a fraction θ_{True} of the entries in which

measurements are determined by flipping a coin.
++ The idea is to be able to compare two closely related datasets: the one that would have occurred without any measurement error and the one that we actually observe, but is otherwise identical.
++ For each step, compute the posterior distribution separately for the dataNoError and dataWithError, under the assumption that $prior(p) \sim Beta(5,5)$. Note that in both cases the analyst computes the posterior distribution as if there was no measurement error.
++ For each simulation step, compute the mean posterior in both cases.

- Use a scatter plot to compare the two mean posteriors. Each point in the plot is the result of simulation step.
- The plot should include the following features:
  ++ A 45-degree line.
  ++ Points for different values of θ_{True} should be displayed in different colors.
  ++ The points should be semi-transparent to facilitate seeing the relative density of points at different locations (see lecture code file *L2_example_2.R* for a related example).
  ++ A legend describing the color code used in the plot

STEP 2 (1 point).

- Provide a brief qualitative description of the resulting pattern.
- Do you have an intuition for why the pattern looks like this?
  (Hint: What happens if $\theta_{True} = 1$?)