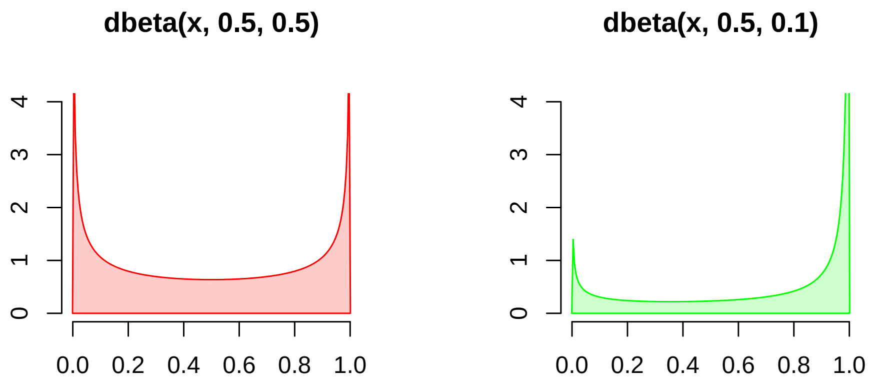
StatR 101: Fall 2012
Homework 8
Eli Gurarie
Due Wednesday, November 21

Instructions: Please submit a single document with all the R code, short answers and figures. Upload the completed homework assignment into the course webpage drop-box. Don't hesitate to discuss the problems amongst yourselves on the forum. This homework closely follows `lab8.r` and expands the exercises somewhat. Please refer to the latest version of `lab8.r`, in which I've numbered the exercises, and which I recently re-uploaded to the course web-site.

1. **Central Limit Theorem:** Carefully go through Part I of the computer lab.

   (a) Create a version of the `CLT(FUN, n, k)` function in the lab, which illustrates the Central Limit Theorem by sampling k random numbers for *any* specified distribution (with *any* parameter values) n times and summing them. Have it illustrate the distribution of the n sums and a qqnorm plot and line, but add the feature described in Exercise 2 where you also draw a normal density curve superposed on the histogram.

   (b) Run this function on random samples from the beta distributed random variable (see http://en.wikipedia.org/wiki/Beta_distribution) which is a distribution that takes two parameters (α and β, called `shape1` and `shape2` in the `rbeta()` function). Illustrate the central limit theorem for two random variables: $X \sim \text{Beta}(\alpha = 0.5, \beta = 0.5)$ and $Y \sim \text{Beta}(\alpha = 0.5, \beta = 0.1)$, summing each of these $n = 1, 2, 10$ and 20 times. Comment on the number of times you need to sum these random variables before the central limit theorem "kicks in", i.e. when the distribution begins to look normal.



2. **Sampling Distribution:** Carefully go through Part II of the computer lab. In this section, we sample a random variable X (in this case, waiting times between earthquakes) n times and look at the resulting distribution of $\overline{X_n}$ (the subscript n represents the size of the sample).

   (a) Tweak the `SampleMean()` function to illustrate (again) the normal approximation to the histogram of the sampling distributions, and make 4 separate plots showing the sample distribution for $\overline{X_5}$, $\overline{X_{10}}$, $\overline{X_{30}}$ and $\overline{X_{100}}$. At what point do you think the normal approximation is value?

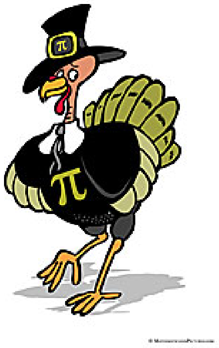
   (b) **Bonus Problem:** (*Optional but recommended!*) Perform a loop using the `SampleMean()` function that calculates the **sample standard deviation** of $\overline{X_n}$ (call it `X.bar.sd`) for values of n ranging from 1 to 100, and plot the result (i.e. `n <- 1:100` on the x-axis, and `X.bar.sd` on the y-axis.

   (c) **Bonus Problem:** (*Optional but recommended!*) According to the central limit theorem, what is the theoretical prediction for this curve in terms of σ_x, the standard deviation of the entire population? Draw the theoretical prediction on this plot. Approximately how large of a sample of waiting times would you need to estimate the mean waiting time to within a standard deviation of 5 minutes?

3. **Apocalyptic inference:** Carefully go through Part III of the lab. We determined in class that if we observe 10 earthquakes with a mean interval of 15 minutes, that we do not need to panic yet because under normal conditions there is a 13% chance that we would observe something that extreme anyway.

   (a) Do exercise 5, i.e. determine the probability that we would observe 30 or 100 earthquakes with a mean interval of 15 minutes or lower under the assumption that the world is not ending. This probability is called the p-value of a hypothesis test (we'll get into that more next week).

   (b) A typical "significance level" for rejecting a null hypothesis (e.g. that the world is NOT ending) is a p-value of 0.05. After how many earthquakes coming in at an average rate of one every 15 minutes should we start panicking?



Happy Thanksgiving!