Statistical Computing 1, Stat 590 Homework 5b Name: _____
Fall 2015 Prof. Erik B. Erhardt

Part I. (80 points) Do all calculations in LaTeX + R + knitr. Insert computer text output and graphics to support what you are saying. For this assignment, all R code should well commented and be visible (`echo=TRUE`) in the document where you have written it.

(10pts) **1. Inverse CDF method**

Suppose $X \sim \text{Exponential}(\lambda)$ with density

$$f(x|\lambda) \;=\; \lambda e^{-\lambda x}, \quad x \geq 0,$$

and 0 otherwise. Show by the inverse cdf method that if $u \sim \text{Uniform}(0,1)$, that

$$x \;=\; -\frac{1}{\lambda}\log_e(u) \;\sim\; \text{Exponential}(\lambda).$$

(40pts) **2. Importance sampling, Beta**

Suppose $X \sim \text{Beta}(\alpha, \beta)$ and we wish to estimate the moment generating function

$$\text{M}_X(t) \;=\; \text{E}[e^{tX}] \;=\; \int e^{tX} f(x|\alpha,\beta)\,\text{d}x,$$

where $f(x|\alpha,\beta)$ is the $\text{Beta}(\alpha,\beta)$ density function. In the notes we discussed two methods. The first was a crude MC method based on sampling $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Beta}(\alpha,\beta)$. The second is a simple importance sample based on sampling $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Uniform}(0,1)$.

(a) (20 pts) Write a script that

- specifies n, α, and β,
- computes the two estimates for $t = -1$ to 1 in increments of size 0.025 or smaller,
- computes an estimated standard error for each MC estimate, as a function of t,
- computes a point-wise approximate 95% CI of the form EST $\pm$ 1.96SE for each method, and plots the estimate and CI bounds as a function of t, one plot for each estimate labeled appropriately, and
- computes the ratio of the standard error obtained by crude MC and that obtained by IS as a function of t and plot the results.

Choosing, say $n = 2000$ or so, run your script for the following scenarios:

1. $\alpha = \beta = 0.25$
2. $\alpha = \beta = 3$
3. $\alpha = 0.25$, $\beta = 3$
4. $\alpha = 3$, $\beta = 0.25$

(b) (20 pts)

1. Discuss the efficiency of the IS method here.
2. Relative to the crude MC method, are there some situations where IS works better than others?

3. Try to formulate whether any choices of α, β, and t result in a fairly good performance by IS, and why.
4. Can you think of an alternative IS distribution then the Uniform that might be better suited (and why) to estimate $\mathrm{M}_X(t)$?

This is an open-ended question. The uniform distribution was used for convenience, so we want to know does it ever work well, and if not, what might work better.

(30$^{\text{pts}}$) **3. Control variates**

For the multinomial goodness-of-fit problem $\underset{\sim}{X} = \{X_1, X_2, \ldots, X_k\}$ has a Multinomial$(m, \underset{\sim}{\theta})$ distribution, where $\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_k)$ such that $\theta_i > 0$ and $\sum_i^k \theta_i = 1$. The likelihood ratio statistic for testing $H_0 : \theta_1 = \theta_{01}, \ldots, \theta_k = \theta_{0k}$, versus $H_1 :$ not H_0, is

$$G^2 \;=\; 2\sum_{i=1}^{k} x_i \log_e\left(\frac{x_i}{m\theta_{0i}}\right),$$

where $0\log_e(0) \equiv 0$.

For a given value of m, k, and $\underset{\sim}{\theta}_0 = (\theta_{01}, \ldots, \theta_{0k})$ below you will be asked to generate n Multinomial samples, and from each compute G^2 and the Pearson statistic P. That is, get $G_1^2, \ldots, G_n^2$ and $P_1, \ldots, P_n$.

The crude MC estimate of $\mathrm{E}(G^2)$ is

$$\hat{\mu}_{\mathrm{C}} \;=\; \frac{1}{n}\sum_{i=1}^{n} G_i^2,$$

while the estimate based on using P as a control variate is

$$\hat{\mu}_{\mathrm{CV}} \;=\; \frac{1}{n}\sum_{i=1}^{n}\{G_i^2 - P_i\} + (k-1).$$

Based on the set of n samples, also compute $\widehat{\mathrm{Var}}[\hat{\mu}_{\mathrm{C}}]$, $\widehat{\mathrm{Var}}[\hat{\mu}_{\mathrm{CV}}]$, and $\mathrm{Corr}[G^2, P]$. Tabulate the two estimates, the standard error of the two estimates, and the correlation for the following scenarios:

1. $\theta_{0i} = k^{-1}$ for $i = 1, \ldots, k$ (equal cell probabilities) with $k = 10, 25, 50$ and $m = 200$
2. $\theta_{01} = 0.9 + 0.1/k$ and $\theta_{0i} = 0.1/k$ for $i = 2, \ldots, k$ with $k = 10, 25, 50$ and $m = 200$

Consider two choices for n

1. $n = 2000$
2. n selected so that the "error" in $\hat{\mu}_{\mathrm{C}}$ is small, for example

$$2\sqrt{\widehat{\mathrm{Var}}[\hat{\mu}_{\mathrm{C}}]} \;\le\; c\mathrm{E}[G^2] \;=\; c\hat{\mu}_{\mathrm{C}}$$

where c is small, say $c = 0.01$. In this case, $\hat{\mu}_{\mathrm{C}}$ should be within $2\sqrt{\widehat{\mathrm{Var}}[\hat{\mu}_{\mathrm{C}}]}$ of $\mathrm{E}(G^2)$ with probability 0.95, that is, with high probability relative to the very small error in $\hat{\mu}_{\mathrm{C}}$. Noting that

$$\widehat{\mathrm{Var}}[\hat{\mu}_{\mathrm{C}}] \;=\; \frac{\mathrm{Var}[G^2]}{n},$$

you might choose n assuming, for simplicity, that

$$\mathrm{E}[G^2] \;=\; \mathrm{E}[\chi^2_{k-1}] \;=\; k-1,$$

but

$$\mathrm{Var}[G^2] \;\doteq\; 2\mathrm{Var}[\chi^2_{k-1}] \;=\; 2\times 2(k-1) \;=\; 4(k-1).$$

Remark: You created code in earlier problems to generate multiple multinomial samples and functions to compute P and G^2 from multiple samples. Use this code here. If possible, you might think of automating things by writing a script that loops over all combinations of k, m, n, and $\underset{\sim}{\theta}$.