18.S096 Problem Set 2 Spring 2018
Due Date: 2/23/2018
Where: On Stellar, prior to 11:59pm

Collaboration on homework is encouraged, but you will benefit from independent effort to solve the problems before discussing them with other people. **You must write your solution in your own words. List all your collaborators.**

1. **Moment-Generating Functions of linear transformations of random variables.**

   Suppose X is a random variable with density/pmf $f(x \mid \theta)$, indexed by the parameter θ and MGF:

   $$M_X(t) = E[e^{tX} \mid \theta] = \int_{\mathcal{X}} e^{tx} f(x \mid \theta)dx \text{ or}$$
   $$\left(\sum_{\mathcal{X}} e^{tx} f(x \mid \theta) \text{ if } X \text{ discrete}\right)$$

   (1a) If $Y = \mu + \sigma \times X$, where $\mu \in R$ and $\sigma \in R^+$ are known constants then the MGF of Y is
   $$M_Y(t) = e^{\mu t} M_X(\sigma t).$$

   (1b) Suppose $X \sim N(0, 1)$, i.e.,
   $$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}, \quad -\infty < x < +\infty$$
   Compute the moment-generating function of X:
   $$E[e^{tX}] = \int_{-\infty}^{+\infty} e^{tx} f(x)dx.$$

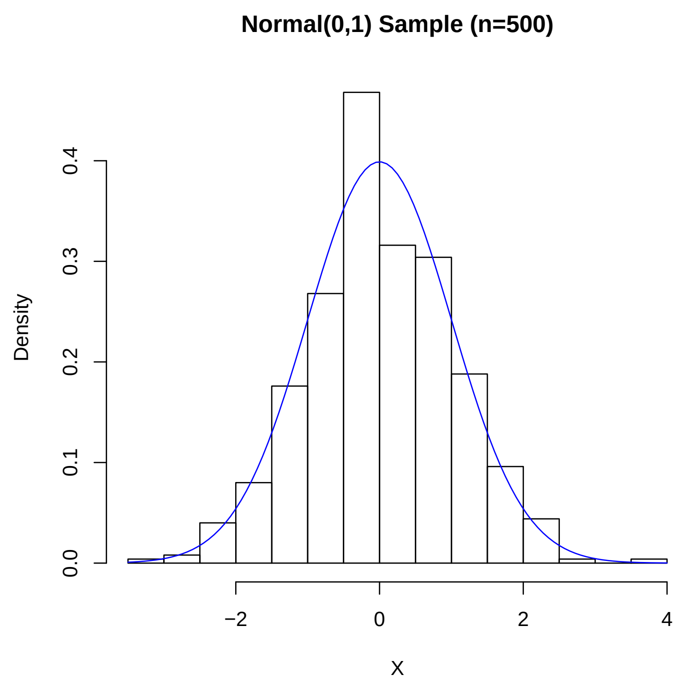
   (1c) Using the density of $X \sim N(0, 1)$, derive the density of $Y = \mu + \sigma \times X$. (Hint: use Jacobian in computing density of transformed random variable.)

   (1d) Apply part (a) to find the MGF of the linear transformation of X:
   $$Y = \mu + \sigma X.$$
   By the uniqueness of MGFs, recognize the MGF of Y as that of a $N(\mu, \sigma^2)$ distribution.

2. **Normal Q-Q Plots: Motivation and Computational Derivation**

   In R, generate an i.i.d. sample $n = 500$ from the $N(0, 1)$ distribution in a vector X. Plot the histogram of X (using the $freq = FALSE$ option) and super-pose the curve of its density function.

```
> set.seed(1);n=500; X=rnorm(n)
> hist(X, main="Normal(0,1) Sample (n=500)", freq=FALSE)
> f=function(x){dnorm(x)}
> X.density=curve(f,  add=TRUE, col='blue')
```



(2a) Apply the function $qqnorm()$ to X

```
> # The R function qqnorm() sorts the input vector
> # from smallest to largest values and plots these
> # on the vertical axis against horizontal values
> # equal to theoretical expected values for the
> #order statistics of a Normal(0,1) distribution.
> par(mfcol=c(1,1))
> qqnorm(X)
```

For a random sample of size n from a population,

$$x_1, x_2, \ldots, x_n$$

the sorted sample values constitute the n order statistics, ranging from the smallest to the largest. These are denoted by putting parentheses around the index:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

Notationally $x_{(j)} = x_{i_j}$, where i_j is the index of the jth smallest x_i. The $qqnorm$ plots the $x_{(j)}$ sorted values of X versus their expected values μ_j, i.e.,

$$\mu_{(j)} = E[x_{(j)}],$$

where $x_{(j)}$ is the jth order statistic from a simple random sample of size n of a $N(0,1)$ population. These expected values are called *theoretical quantiles*.

- In the plot from $qqnorm()$, add a straight line with intercept 0, and slope 1 to the plot. (Use the R function $abline()$ with arguments $abline(a = 0, b = 1)$.) Do the points in the plot follow the line?
- Repeat the exercise of generating a random sample ($n = 500$) from a $Normal(0, 1)$ distribution and constructing a normal qq plot 4 times. creating a normal qq plot for 4 samples in each panel of a 2-by-2 display. (In R use $par(mfcol = c(2, 2))$ to setup the display.)
  Comment on the degree of consistency of how the plots look.

(2b) Compute $Y = 10. + 5. * X$. Plot the histogram of Y (using the $freq = FALSE$ option) and super-pose a plot of the true density function. (The density function f should use the $mean =$ and $sd =$ arguments when using the R function $dnorm()$, i.e., $dnorm(x, mean = 10, sd = 5.)$

(2c) Compute the normal qq plot of Y using $qqnorm()$

The points fall close to a line. What specification of the line $abline(a = ?, b = ?)$ should fit the data?

(2d) The function $qqnorm()$ computes expected values of order statistics ("theoretical quantiles") from a $N(0, 1)$ sample using a theoretical formula. We now use R to compute numerical approximations of these theoretical values:

Generate $N^* = 1000$ samples of size $n = 500$ from the $N(0, 1)$ distribution. Order each sample from smallest to largest. These ordered values are the "order statistics" of each sample: $x_{(j)}, j = 1, 2, \ldots, n$.

Compute the average value of each order statistic over the $N^* = 1000$ samples.

```
> # Create matrix of of the 1000 samples of size 500
> nsamples=1000
> samplesize=500
> mat.samples=matrix(rnorm(nsamples*samplesize),
+                    nrow=samplesize, ncol=nsamples)
> # Use apply() to sort each column of mat.samples
> #     so that each row consists of random sample
> #     of the respective order statistic
> mat.samples.orderstats=apply(mat.samples,2,sort)
> # Use rowMeans() to approximate the expected value of order statistics
> mat.samples.orderstats.rowMeans=rowMeans(mat.samples.orderstats)
```

- Plot the sorted values of X against the row-means of the order statistics. Compare this plot to the plot generated by $qqnorm()$.
- For $Y = 10. + 5 * X$, plot the sorted sample Y against the same approximate expected values. Compare this plot to that produced using qqnorm().

3. Suppose $X \sim Gamma(\alpha, \beta)$ with density:

$$f(x \mid \alpha, \beta) = \frac{\beta^\alpha x^{\alpha-1}}{\Gamma(\alpha)} e^{-\beta x}, \ \ 0 < x < \infty$$

3(a) Compute the MGF of X.

3(b) If $X_1, X_2, \ldots, X_n$ denotes an i.i.d. sample from the $Gamma(\alpha, \beta)$ distribution prove that the distribution of $S_n = X_1 + X_2 + \cdots X_n$ is $Gamma(\alpha_*, \beta_*)$, where $\alpha_* = n \times \alpha$ and $\beta_* = \beta$.

(Hint: Compute the MGF of S_n and identify the underlying distribution using the uniqueness property of MGFs.)

3(c) In R, use the function $rgamma()$ to fill a 500×10 matrix $X1$ with random variates from an $Exponential(1) = Gamma(shape = 1, scale = 1)$ distribution. Define the vector $Y1$ to be the row-sums of $X1$.

```
> nrow0=500
> ncol0=10
> X1=matrix(rgamma(nrow0*ncol0, shape=1, scale=1), nrow=nrow0)
> Y1=rowSums(X1)
```

- Plot the histogram of $Y1$ (use the density scale)
- What is the probability distribution of the sample values in $Y1$?
- What are the mean and variance of this distribution?
- Add the probability density curve to the histogram plot (use the R function $dgamma()$ to compute values of a Gamma density function).

4. **Simulation Exercise: Comparing Method-of-Moments Estimates for Poisson Distribution**

Simulate $M = 1000$ random samples (sample size $n = 50$) from a $Poisson(\lambda)$ distribution with $\lambda = 5$.

```
> # Create matrix of of the 1000 samples of size 500
> nsamples=1000
> samplesize=50
> mat.samples=matrix(rpois(nsamples*samplesize, lambda=5), nrow=samplesize, ncol=nsampl
> # Use apply to compute MOM estimates
> mat.samples.mom1<-colMeans(mat.samples)
> mat.samples.mom2<-colMeans(mat.samples^2) - colMeans(mat.samples)^2
```

4(a) For each sample compute the two method-of-moments estimates of λ :

$$\hat{\lambda}_{MOM1} = \hat{\mu}_1 \text{ and } \hat{\lambda}_{MOM2} = \hat{\mu}_2 - (\hat{\mu}_1)^2.$$

Compare the two estimates using the simulated sampling distribution of the estimation error. Using the Mean-Squared-Error (equivalently RMSE) criterion:

$MSE(\hat{\lambda}) = \frac{1}{M}\sum_{j=1}^{M}(\hat{\lambda}_j - \lambda)^2.$

$RMSE(\hat{\lambda}) = \sqrt{MSE(\hat{\lambda})}.$

Which estimator is better?

4(b) Are the two MOM estimates dependent? Construct a scatterplot of the two estimates and compute their correlation over the $M = 1000$ pairs of simulated estimates.

4(c) Construct Normal Q-Q Plots of the simulated samples of each method-of-moments estimate. Which estimate has a simulation distribution which is closer to a Normal distribution? Explain why this is true.