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:

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

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

   $$M_Y(t) = e^{\mu t} M_X(\sigma t).$$

   Solution:

   $$\begin{aligned} M_Y(t) &= E[e^{t(\mu+\sigma X)} \mid \theta] \\ &= e^{t\mu} E[e^{\sigma t X} \mid \theta] = e^{t\mu} M_X(\sigma t) \end{aligned}$$

   (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:

   $$E[e^{tX}] = \int_{-\infty}^{+\infty} e^{tx} f(x)dx.$$

   Solution:

   $$\begin{aligned} M_X(t) = E[e^{tX}] &= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x^2-2xt)} dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}[(x-t)^2-t^2]} dx \quad \text{(completing square in exponent)} \\ &= e^{\frac{1}{2}t^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-t)^2} dx \\ &= e^{\frac{1}{2}t^2} \end{aligned}$$

   since the last integral is 1; change of variable to $y = x - t$ gives the integral of a $N(0,1)$ density.

   (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.)

   Solution:

   Given $f_X(x) = \frac{1}{\sqrt{2\pi} e^{-\frac{1}{2}x^2},}$ consider the change of variable to

   $$Y = g(X) = \mu + \sigma X.$$

Note that $g^{-1}(y) = \frac{y-\mu}{\sigma}$ and $\frac{d}{dy} g^{-1}(y) = \frac{1}{\sigma}$.
The density of Y is given by

$$\begin{aligned} f_Y(y) &= f_X(g^{-1}(y))|\tfrac{d}{dy} g^{-1}(y)| \\ &= f_X(\tfrac{y-\mu}{\sigma}) \times \tfrac{1}{\sigma} \\ &= \tfrac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\mu)^2} \end{aligned}$$

(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.

Solution:

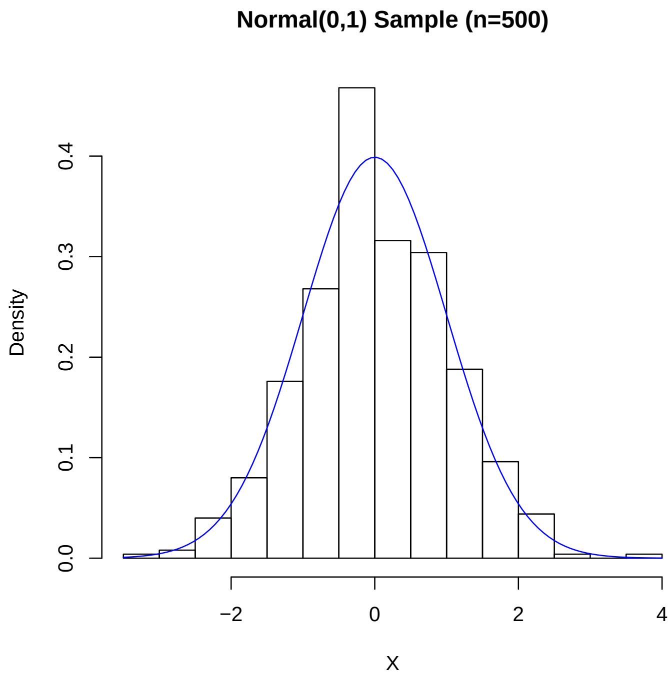
$M_Y(t) = e^{\mu t} M_X(\sigma t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$

This is the MGF of a $N(\mu, \sigma^2)$ distribution so it is the distribution of Y.

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')
```

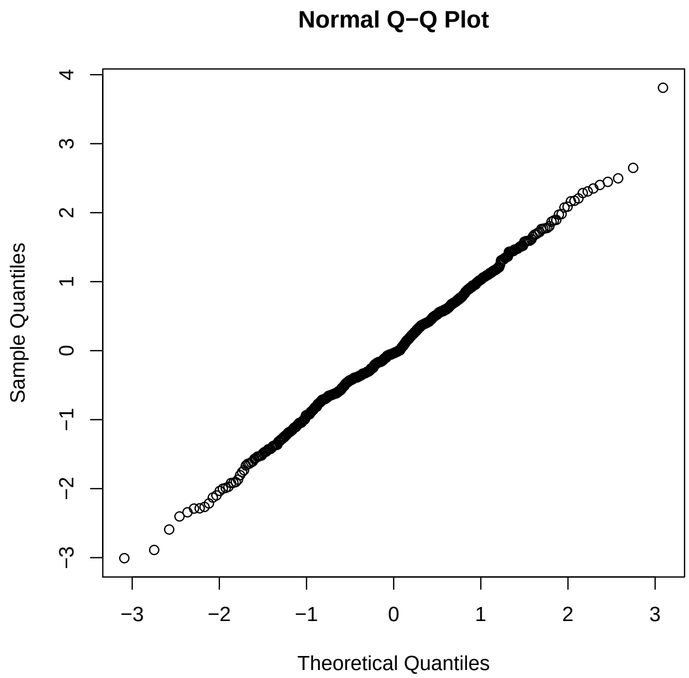


(2a) Apply the function *qqnorm*() to X

```
> # 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)
```

Solution:

```
> # 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

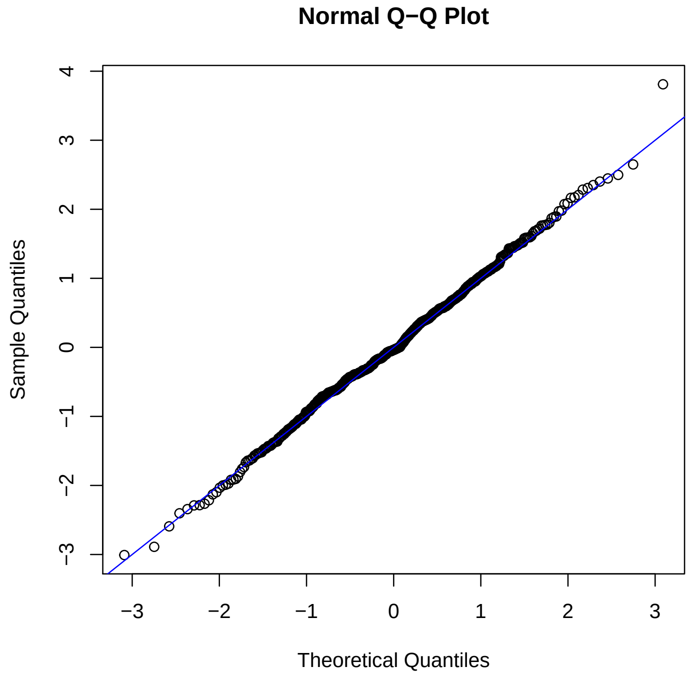
setup the display.)

Comment on the degree of consistency of how the plots look.

Solution:

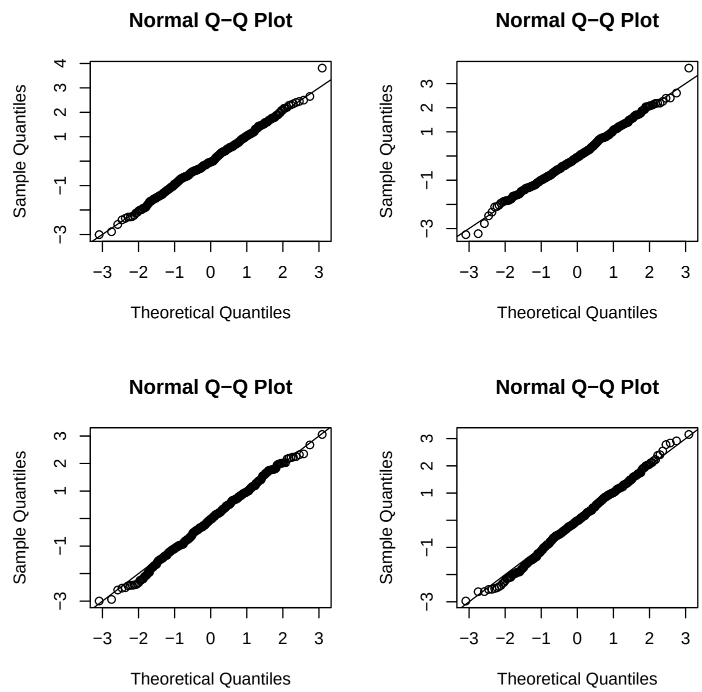
(a)

```
> par(mfcol=c(1,1))
> qqnorm(X)
> abline(a=0,b=1,col='blue')
> # Yes the points follow a straight line except possibly at the edges of the sampl
```



(b).

```
> set.seed(1);n=500;
> par(mfcol=c(2,2))
> for (i in 1:4){
+    X=rnorm(n)
+    qqnorm(X)
+    abline(a=0,b=1)
+ }
```

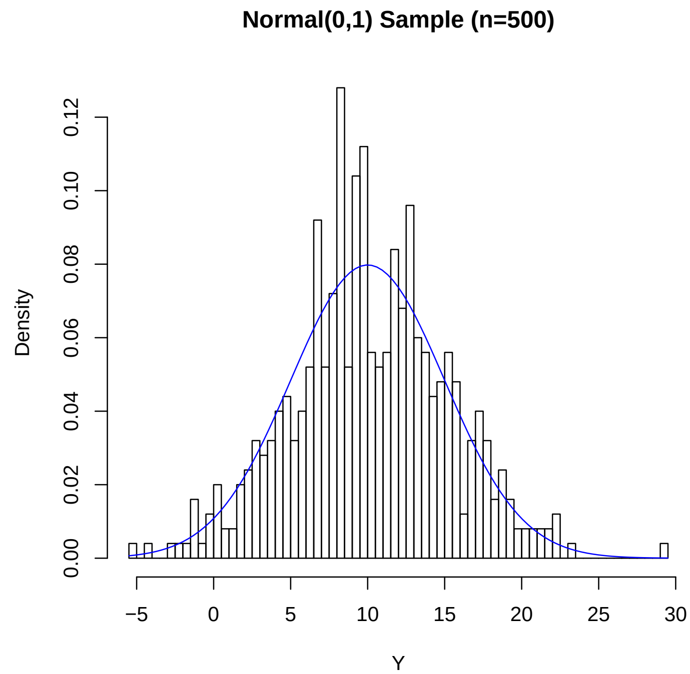


The qqnorm plots exhibit high consistency with the data being samples from a $N(0, 1)$ distribution. There appears to be greater variability about the line at the extremes of the data.

(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., $dnorm(x, mean = 10, sd = 5.)$

Solution:

```
> par(mfcol=c(1,1))
> set.seed(1);n=500; X=rnorm(n)
> Y= 10. + 5.*X
> hist(Y, main="Normal(0,1) Sample (n=500)", freq=FALSE,
+      breaks = 50)
> f=function(x){dnorm(x, mean=10., sd=5.)}
> X.density=curve(f,  add=TRUE, col='blue')
```

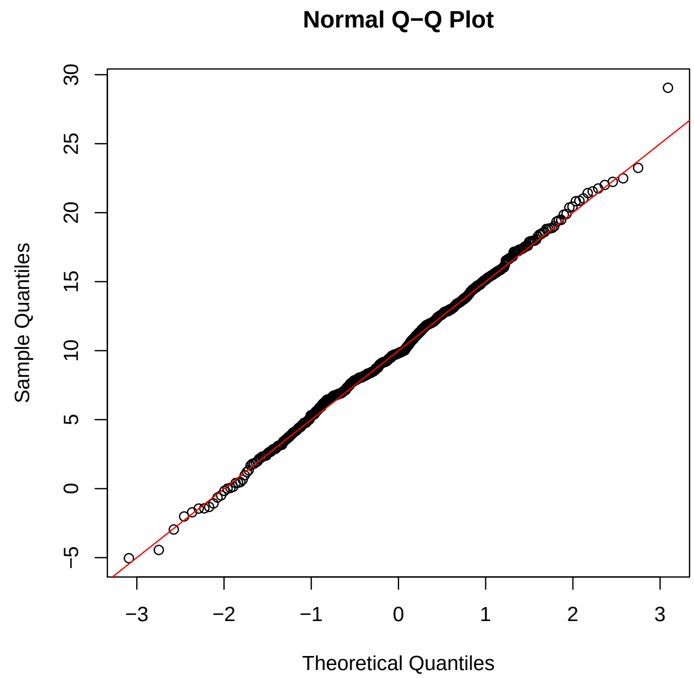


(2c) Compute the normal qq plot of Y using $qqnorm()$

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

Solution:

```
> qqnorm(Y)
> # The theoretical line of best fit for a N(mu,sigma^2)
> # sample should have intercept = mu and slope = sigma
> abline(a=10., b=5., col='red')
```



(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:

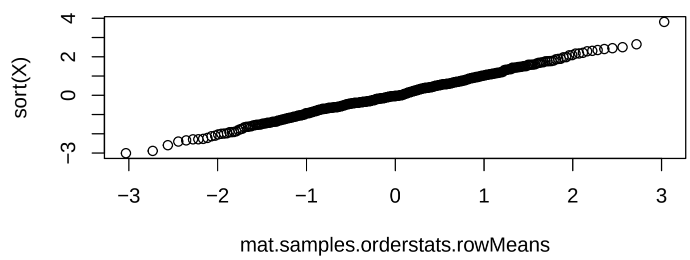
Generate $N^* = 1000$ samples of size $n = 500$ from the $N(0,1)$ distribution. Order each sample from smallest to largest. These ordered values are the "order statistics" of each sample: $x_{(j)}, j = 1, 2, \ldots, n$. Compute the average value of each order statistic over the $N^* = 1000$ samples.

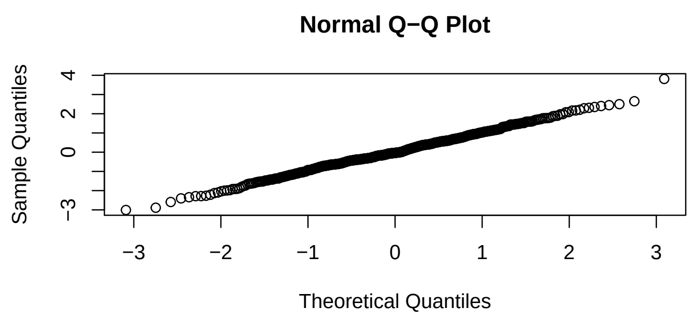
```
> # Create matrix of of the 1000 samples of size 500
> nsamples=1000
> samplesize=500
> mat.samples=matrix(rnorm(nsamples*samplesize),
+                    nrow=samplesize, ncol=nsamples)
> # Use apply() to sort each column of mat.samples
> #     so that each row consists of random sample
> #     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 $qqnorm()$.

Solution:

```
> par(mfcol=c(2,1))
> plot(mat.samples.orderstats.rowMeans, sort(X))
> # The R function qqnorm() creates this plot using
> #   theoretical expected values for the
> #   order statistics of a Normal(0,1) distribution.
> qqnorm(X)
```



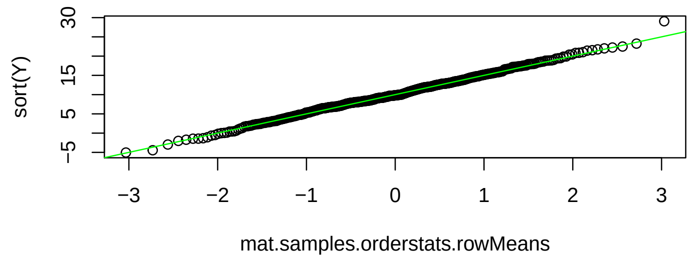


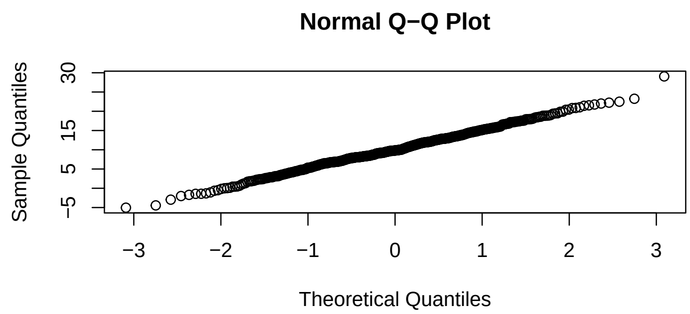
These plots look identical. The *Theoretical Quantiles* from the R function *qqnorm*() are comparable to the sample means of the order statistics.

- For $Y = 10. + 5 * X$, plot the sorted sample Y against the same approximate expected values. Compare this plot to that produced using qqnorm().

Solution:

```
> Y=10. + 5.*X
> par(mfcol=c(2,1))
> plot(mat.samples.orderstats.rowMeans, sort(Y))
> # Add the line Y= a + b X, where a=10. and b=5.
> abline(a=10.,b=5.,col='green')
> #
> qqnorm(Y)
>
```





The Normal Q-Q Plot always uses the *standard quantiles* corresponding to the $Normal(0, 1)$ distribution on the horizontal axis. Data following a $Normal(\mu, \sigma^2)$ distribution model will tend to fall close the line:

$$y = \mu + \sigma \times x.$$

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.

Solution:

$$M_X(t) = E[e^{tX}] = (1 - \frac{t}{\beta})^{-\alpha}, \;\; (t < \beta)$$

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.)

Solution:

The MGF of $S_n = X_1 + X_2 + \cdots X_n$ is:

$$\begin{aligned} M_{S_n}(t) &= E[e^{tS_n}] = E[e^{t(X_1+\cdots+X_n)}] \\ &= E[e^{tX_1}]E[e^{tX_n}]\cdots E[e^{tX_n}] \\ &= (1-\tfrac{t}{\beta})^{-n\alpha} \end{aligned}$$

This is the MGF of a $Gamma(\alpha_*, \beta)$ random variable with $\alpha_* = n\alpha$.

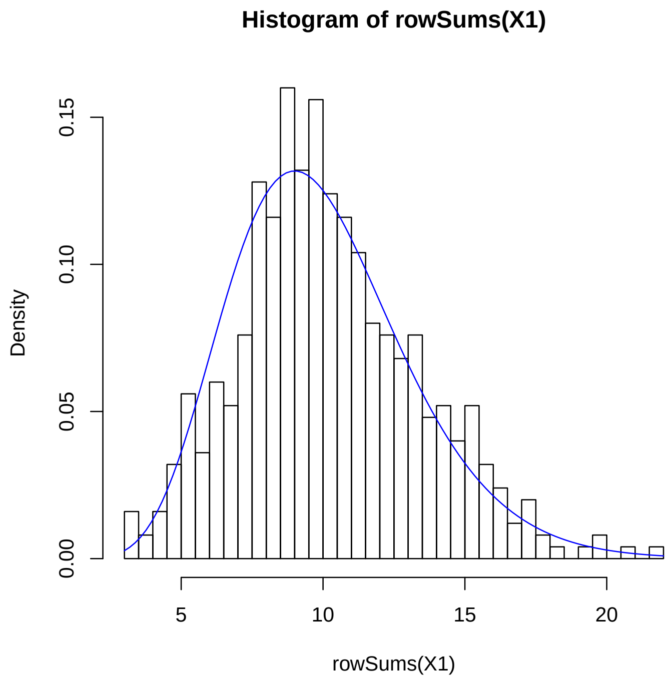
3(c) In R, use the function $rgamma()$ to fill a 500×10 matrix $X1$ with random variates from an $Exponential(1) = Gamma(shape = 1, scale = 1)$ distribution. Define the vector $Y1$ to be the row-sums of $X1$.

```
> 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).
  Solution:
  The row sums are $Gamma(\alpha, \beta)$ with $\alpha = 10$, and $\beta = 1$.
  The mean of a $Gamma(\alpha, \beta)$ distribution is α/β which equals 10. in this case.
  The variance is α/β^2 which is also 10. The standard deviation is $\sqrt{10} \approx 3.16$ which is apparent in the histogram.

Solution:

```
> par(mfcol=c(1,1))
> hist(rowSums(X1),nclass=50, freq=FALSE)
> f=function(x){dgamma(x, shape=10, scale=1.)}
> X.density=curve(f,  add=TRUE, col='blue')
```



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:

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

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

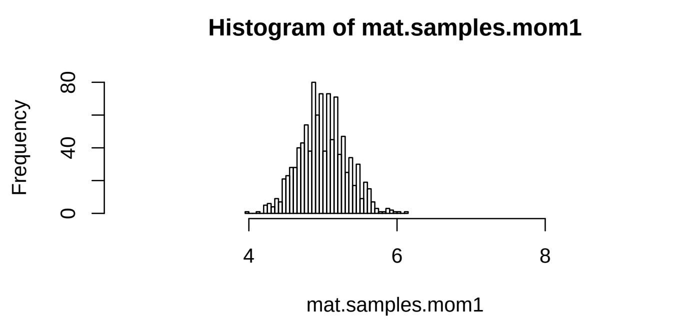
Which estimator is better?

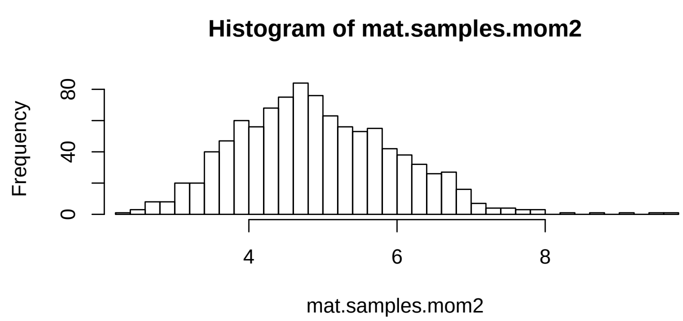
4(b) Are the two MOM estimates dependent? Construct a scatterplot of the two estimates and compute their correlation over the $M = 1000$ pairs of simulated estimates.

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.

Solution:

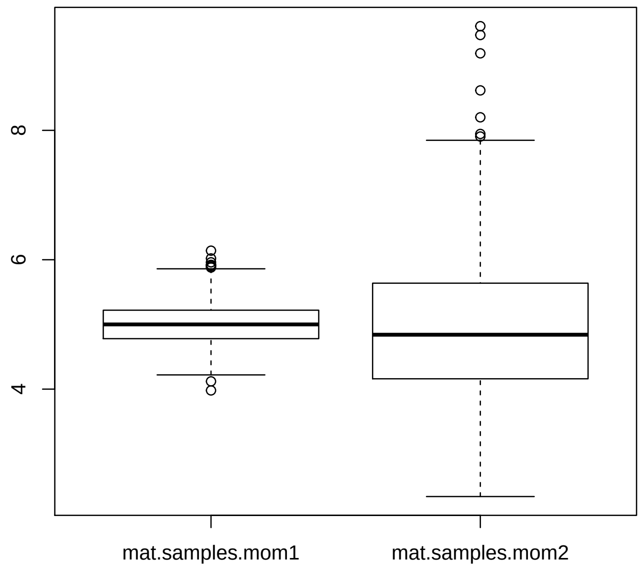
```
> # Plot separate histograms for each estimate
> par(mfcol=c(2,1))
> xlim0=c(min(c(mat.samples.mom1, mat.samples.mom2)),
+        max(c(mat.samples.mom1,mat.samples.mom2)))
> hist(mat.samples.mom1,breaks=50,xlim=xlim0)
> hist(mat.samples.mom2, breaks=50,xlim=xlim0)
```



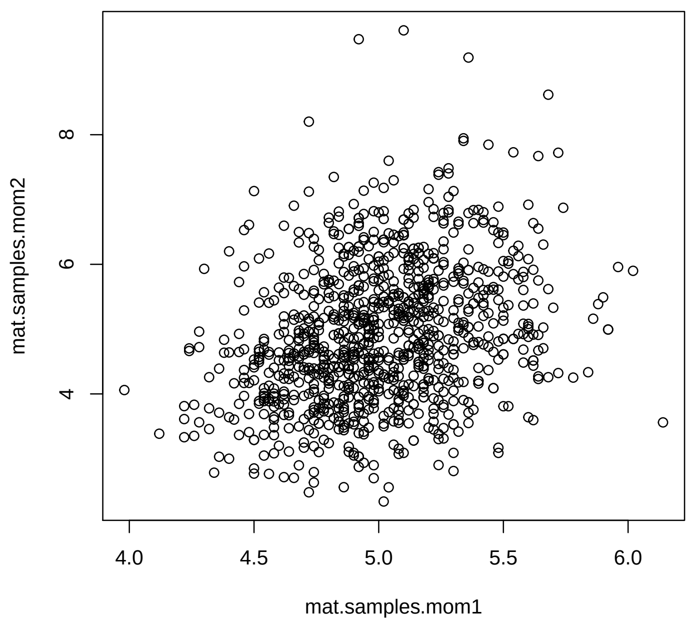



```
> # Compare the distributions with parallel boxplots
> boxplot.matrix(cbind(mat.samples.mom1, mat.samples.mom2))
> # Compute the MSE/RMSE
> MSE.mom1=mean( (mat.samples.mom1-5)^2)
> MSE.mom2=mean( (mat.samples.mom2-5)^2)
> print(sqrt(MSE.mom1))
[1] 0.3201437
```

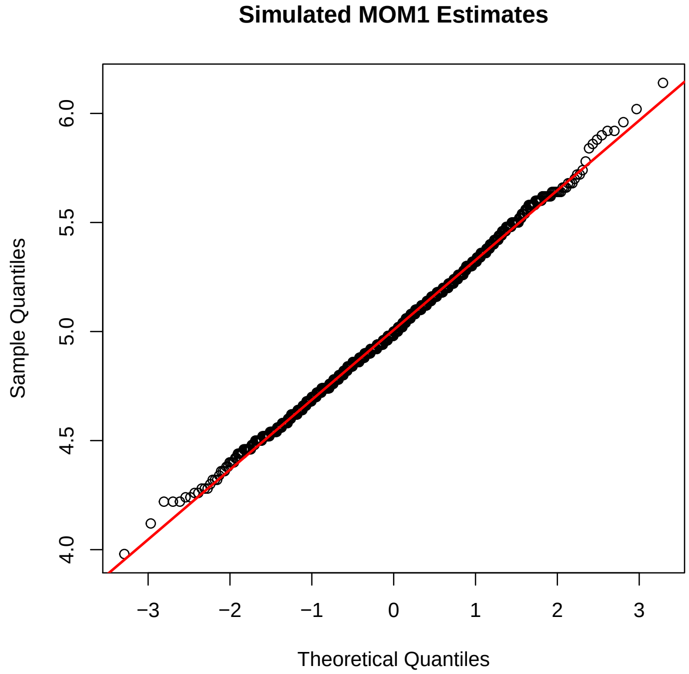
```
> print(sqrt(MSE.mom2))
[1] 1.074409
```



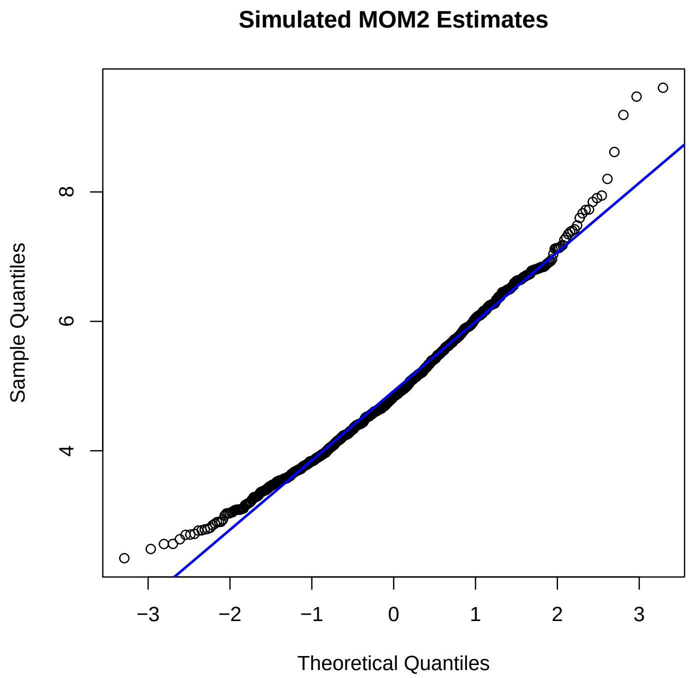

```
> # To evaluate whether the two estimates are dependent
> # examine their scatterplot and compute their correlation
> plot(mat.samples.mom1, mat.samples.mom2)
> cor(cbind(mat.samples.mom1,mat.samples.mom2))[1,2]
[1] 0.2881119
>
> #The two estimates appear to be modestly dependent
> # with this positive correlation.
```




```
> # Construct Normal Q-Q Plots of the simulated samples
> # each method-of-moments estimate
> # Which estimate has a simulation distribution which
> # is closer to a Normal distribution?
> # Explain why this is true.
>
> par(mfcol=c(1,1))
> qqnorm(mat.samples.mom1, main="Simulated MOM1 Estimates")
> abline(a=mean(mat.samples.mom1), b=sqrt(var(mat.samples.mom1)),col="red", lwd=2)
```




```
> qqnorm(mat.samples.mom2, main="Simulated MOM2 Estimates")
> abline(a=mean(mat.samples.mom2), b=sqrt(var(mat.samples.mom2)),col="blue", lwd=2)
```



The simulated distribution of $\hat{\lambda}_{MOM1}$ is closer to the normal distribution. By the Central Limit Theorem, this estimator as a sample mean converges to the Normal distribution as the sample size increases. The distribution of the sample variance $\hat{\lambda}_{MOM2}$ is skew relative to the normal distribution. This is evident from plotting the line on the normal QQ plot which matches the mean and variance of the estimate. The estimates tend to be less extreme on the low side (the variance is bounded below by zero), and more extreme on the high side (the sample variance will be extremely high if there are extreme values in the original sample).