Comparisons of CDFs and ECDFs of F for Exponential Data

In the JSM paper we look only at results for a nominal significance level $\alpha = 5\%$. Consider again the case where $k = 3$ groups and $r = 5$ observations on each group. If the null hypothesis is true, then for data that are normal and homoscedastic the F-statistic is distributed as $F(2, 12)$.

**When the null hypothesis is true:** Here, for exponential data with $\mu_1 = \mu_2 = \mu_3 = 10$, we compare the cumulative distribution function of $F(2, 12)$ with the empirical cumulative distribution function (ECDF) of the simulated values of the F-statistic. Because the data are not normal, we can’t expect the F-statistic to have exactly an F distribution. The question is whether log and rank transformations result in F-statistics that are more nearly F-distributed than the original data.

The bold lines of code make the required graphs. We make graphs first for original data, then for log-transformed data, and finally for rank-transformed data. Because the critical value of the nominal 5% level test in this case is $q_f(0.95, 2, 3*(r-1)) \approx 3.885$, the value of the cdf above 3.885 is 0.95. The ecdf values above this point correspond to the value of $\text{mean}(\text{rej})$ used as our main criterion in the paper, by which we have judged that transformations are helpful even in the case where the null hypothesis is true. The graphs below show it is not at the value 3.885 alone that the transformations give more satisfactory values.

```r
r <- 5;  m <- 20000
mu1 <- 10;  mu2 <- 10;  mu3 <- 10
mu <- c(rep(mu1,r), rep(mu2,r), rep(mu3,r))
x <- rexp(3*r*m, rate=1/mu)
DTA <- matrix(x, m, byrow=T)
#DTA <- log(DTA)  # Activate this line for log transf.
#DTA <- t(apply(DTA, 1, rank))  # Activate this line for rank transf.
m1 <- rowMeans(DTA[,1:r])
m2 <- rowMeans(DTA[,r+1:(2*r)])
m3 <- rowMeans(DTA[(2*r+1):(3*r)])
v1 <- rowSums((DTA[,1:r] - m1)^2)/(r-1)
v2 <- rowSums((DTA[,r+1:(2*r)] - m2)^2)/(r-1)
v3 <- rowSums((DTA[(2*r+1):(3*r)] - m3)^2)/(r-1)
g <- (m1 + m2 + m3)/3
MSF <- r * rowSums((cbind(m1,m2,m3) - g)^2)/2
MSE <- rowMeans(cbind(v1, v2, v3))
F.rat <- MSF/MSE
rej <- (F.rat > qf(0.95, 2, 3*(r-1)))
mean(rej)
#cut <- seq(0,max(F.rat)+1,.5)
#xx <- seq(0.0001,max(F.rat)+.5)
#hist(F.rat, breaks=cut, prob=T, xlim=c(0,5), ylim=c(0,1))
#lines(xx,df(xx,2,3*(r-1)),col="red")
sF <- sort(F.rat);  rF <- rank(sF)
plot(sF, (rF-.5)/m, type="s", xlim=c(0,5), lwd=2, col="green",
     xlab="F", ylab="cdf", main="No Transformation: ECDF (Green) and CDF")
     # ecdf
lines(xx,pf(xx,2,3*(r-1)))  # cdf
```
The height of the ECDF at 3.885 for this run was $1 - 0.037 = 0.963$. Of course, the height of the CDF there is 0.95.

For this run with log-transformed data the height of the (green) ECDF is $1 - 0.046 = 0.954$. 
For this run with rank-transformed data, the height of the ECDF at 3.885 is $1 - 0.057 = 0.943$. Notice that the granularity of the ranks can be seen here in the ECDF—especially for values below 2. This is typical of a sample distribution based on rank-transformed data. But in this case the granularity is less evident in the vicinity of reasonable critical values.

The bold lines of code that have been suppressed (with #s) can be used to compare histograms of the simulated values of the F-statistic with the pdf of $F(2, 12)$. The histogram, if smoothed, gives an empirical estimate of the PDF. (This is sometimes called “density estimation.”) While comparisons of histograms with PDFs may be more familiar in elementary texts, it is often more accurate to compare ECDFs with CDFs.

**Exercise.** Make histograms based on simulated data and draw corresponding PDF curves through the data. Do this for original, log-transformed, and rank-transformed exponential data with $\mu_1 = \mu_2 = \mu_3 = 10$.

*When the null hypothesis is not true,* it is not straightforward to make similar comparative graphs. For normal, homoscedastic data with different group population means, the power of the F-test is computed using a noncentral F-distribution, where the noncentrality parameter depends on (i) the pattern of differences among the means, (ii) the number of replications per group, and (iii) the common variance. If the data are exponential, there is no common variance without transformation; with transformation the pattern of the differences among the group population means is altered.