Distribution Functions

{2.1} Consider a random variable Y with distribution function $F(.)$

a) Which if any of the distributions are discrete?

Distribution B is discret

b) Do any of the distributions appear to be symmetric about the mean of the distribution?

Distribution C and D appears symmetric

c) Which distribution appears to have a longer right tail, A or D?

Relative to one another distribution A would have a longer right tail, as values of y increase F(y) approaches 1

d) Which distributions if any are likely to be a normal distribution?

Distribution D as it is the only one that appears to be symmetrical

e) For which distribution is $Pr(Y \leq 2)$ the largest? For which distribution is it the smallest?

Largest-Distribution A

Smallest- Distribution B

f) Which distribution has the largest median? Which distribution has the smallest median?

Largest-Distribution B

Smallest-Distribution A

2.2 Consider a random variable Y with distribution function F(.) A plot of F(.) is given in figure 2.14.

a) Find the set of values of y for which $Pr(Y = y) > 0$

Based on the plot any $y \geq 1$ for $Pr(Y = y) > 0$

b) Find the value of y for which $Pr(Y = y)$ is the largest

Based on the plot the largest value for y which $Pr(Y = y)$ is approximately around y=4 because this is where the jump is the largest

c) Find the value of y among those with $Pr(Y = y) > 0$, for which $Pr(Y = y)$ is the smallest.

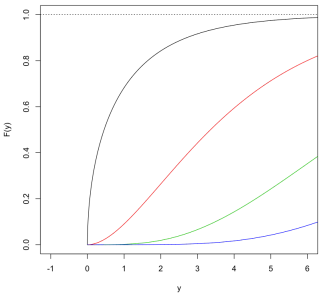
Based on the plot the smallest value for y which $Pr(Y = y)$ is approximately around y=2 because this is where the jump is the smallest

2.3 Plot the distribution function of the chi squared distribution with 1 df. On the same plot, add the distribution functions of the chi squared distributions with 4,8, and 12 df

Based on your results what can you conclude about how the chi-squared distribution changes as the df increases?

Because the mean of a chi sq distribution depends on the df, the distribution skew decreases as the df increases.

In [1]:
```
y<-seq(0,25,.001)
Fy<-pchisq(y, df=1)
df4<-pchisq(y, df=4)
df8<-pchisq(y, df=8)
df12<-pchisq(y, df=12)
plot(y,Fy, type="l", xlim=c(-1, 6), xlab="y", ylab="F(y)", col=1)
abline(a=1, b=0, lty=3)
lines(y, df4, col=2)
lines(y, df8, col=3)
lines(y, df12, col=4)
```



For exercises 2.4-2.8

Let $Y_1, Y_2, \ldots Y_n$ denote i.i.d. observations and let $\sigma^2 = Var(Y_1)$ The sample variance is defined typically by

$S^2 = (1/n-1)\sum(Y_j = \bar{Y})^2$

Because S^2 is an estimate of σ^2 this can now be written as

$S_0^2 = 1/n\sum(Y_j = \bar{Y})^2$=MLE of σ assuming that Y_j is normally distributed. Using the S^2 is preferred because it is an unbiased estimator of σ^2 Meaning it doesn't require the normal distribution assumption. S^2 applies only when estimating variance. But if we are interested in estimating the standard deviation of the distribution, σ S standard deviation is not an unbiased estimator of σ neither is S_0^2

Therefore you must use Monte Carlo simulation to study the properties of S and S_0 as estimators of σ. Here is another formula

$S_1^2 = (1/n-1.5)\sum(Y_j - \bar{Y})^2$

2.4 Suppose that $Y_1, Y_2, , Y_4, Y_5 (n = 5)$ are i.i.d. standard normal random variables.

Use Monte Carlo stimulation to find:

$E(S)$ = .93

$E(S_0)$=.84

$E(S_1)$=1.00

Based on these results, find the bias of each estimator of σ; note that here the true value of $\sigma = 1$. Based on these results which estimator has the smallest bias?

Bias is

$E(S)$ = -.06

$E(S_0)$=-.15

$E(S_1)$=.004

Based on the results the estimator S_1^2 has the smallest bias based on the magnitude

In [2]:
```
#simulate a matrix of random variables where each row is one observation of the vector or Y's
set.seed(35201)
y_mat<-matrix(rnorm(5*100000, mean=0, sd=1), 10000, 5)
#calculate the sample sd
E_S<-apply(y_mat,1, sd)
#E(S)
S<-mean(E_S)
#no need to resimulate
S0= 0.8944 * S
S1= 1.0690 * S

S
S0
S1

#to estimate bias take the expected minus true parameter sigma
S-1
S0-1
S1-1
```

0.939222164627378

0.840040304042727

1.00402849398667

-0.0607778353726224

-0.159959695957273

0.0040284939866666

2.5

Repeat the analysis in 2.4 but now assume that $Y_1 \ldots Y_5$ each have an exponential distribution with rate parameter 1 (the standard deviation of this distribution is also 1).

$E(S)$ = .87

$E(S_0)$=.78

$E(S_1)$=.92

Based on these results, find the bias of each estimator of σ; note that here the true value of $\sigma = 1$. Based on these results which estimator has the smallest bias?

Bias is

$E(S)$ = -.13

$E(S_0)$=-.22

$E(S_1)$=-.07

Based on the results the estimator S_1^2 has the smallest bias based on the magnitude

In [3]:
```
#exponential distribution with rate parameter 1
set.seed(35201)
y_mat<-matrix(rexp(1e6, rate=1), 10000, 5)
E_S<-apply(y_mat,1, sd)
#E(S)
S<-mean(E_S)
#no need to resimulate

S0= 0.8944 * S
S1= 1.0690 * S

S
S0
S1

#to estimate bias take the expected minus true parameter sigma
S-1
S0-1
S1-1
```

0.867961590444464

0.776304846493528

0.927850940185132

-0.132038409555536

-0.223695153506472

-0.0721490598148683

2.6

Repeat the analysis in 2.4 but now assume that $Y_1 \ldots Y_5$ each have a Poisson distribution with mean 1(sd also 1).

$E(S)$ = .93

$E(S_0)$=.83

$E(S_1)$=1.0

Based on these results, find the bias of each estimator of σ; note that here the true value of $\sigma = 1$. Based on these results which estimator has the smallest bias?

Bias is

$E(S)$ = -.06

$E(S_0)$=-.16

$E(S_1)$=-.003

Based on the results the estimator S_1^2 has the smallest bias based on the magnitude

In [4]:
```
#poisson distribution
set.seed(35201)
y_mat<-matrix(rpois(1e6, lambda=1), 10000, 5)
E_S<-apply(y_mat,1, sd)
#E(S)
S<-mean(E_S)
#no need to resimulate

S0= 0.8944 * S
S1= 1.0690 * S

S
S0
S1

#to estimate bias take the expected minus true parameter sigma
S-1
S0-1
S1-1
```

0.932253604719399

0.83380762406103

0.996579103445038

-0.067746395280601

-0.16619237593897

-0.00342089655496247

2.7

Repeat the analysis in 2.4 but now assume that $Y_1 \ldots Y_5$ each have a t-distribution with 6 df. The variance of the t-distribution is 3/2.

$E(S)$ = .91

$E(S_0)$=.82

$E(S_1)$=.98

Based on these results, find the bias of each estimator of σ; note that here the true value of $\sigma = 1$. Based on these results which estimator has the smallest bias?

Bias is

$E(S)$ = -.08

$E(S_0)$=-.18

$E(S_1)$=-.02

Based on the results the estimator S_1^2 has the smallest bias based on the magnitude

In [5]:
```
#tdistribution
set.seed(35201)
y_mat<-matrix(rt(1e6, df=6), 10000, 5)
y_mat<-y_mat *sqrt(2/3)
E_S<-apply(y_mat,1, sd)
#E(S)
S<-mean(E_S)
#no need to resimulate

S0= 0.8944 * S
S1= 1.0690 * S

S
S0
S1

#to estimate bias take the expected minus true parameter sigma
S-1
S0-1
S1-1
```

0.913431623498279

0.816973244056861

0.97645840551966

-0.0865683765017211

-0.183026755943139

-0.0235415944803399

2.8

Based on the results in 2.4-2.7 does one of the estimators consistently have smaller bias than the others? Or does the estimator with the smallest bias depend on the underlying distribution?

It would appear that S_1^2 consistently has the smallest bias suggesting that the underlying distribution does not influence the bias. This makes sense as S_1^2 is a more conservative estimate because we are subtracting the most relative to the others.

2.9 The dataset "software" contained the time between failures (in CPU minutes) of certain software systems

a) Plot the empirical distribution function for these data.

b) A commonly used distribution for modeling failure data is the exponential distribution with rate parameter λ Note that an estimate of λ is given by $1/\bar{Y}$ where $\bar{Y}$ is the sample mean of the data. For the software, find an estimate of λ

c) On the plot of empirical distribution function constructed in part (a) add a plot of the distribution function of the exponential distribution with the rate parameter taken to be the estimate found in part (b)

d) An alternative distribution for the failure data is the log-normal distribution. Estimate the μ and σ for the software data.

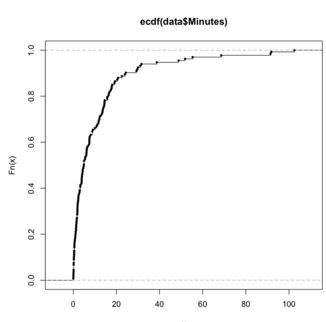
e) On the plot of empirical distribution function constructed in a and c add a plot of the distribution function of the log normal distribution with the parameters taken to be the estimates found in part d.

f) Based on these plots which of the distributions appears to be a better fit to the software failure data? Or do they fit the data equally well?

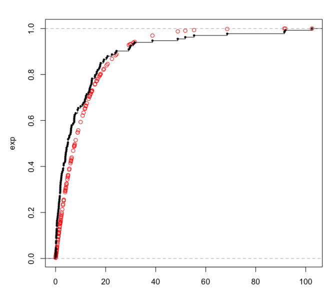
Based on the plots it would seem that all distributions fit the software failure data equally well

In [7]:
```
data_loc<-'/Users/Alexis/Documents/Spring2020/nonparametrics/data/software.csv'
#data_loc<-'/Users/aporter1350/Documents/Courses/Spring2020/nonparametrics/data/software.csv'
data<-read.csv(data_loc, header=FALSE)
colnames(data) <- c("Minutes")
```

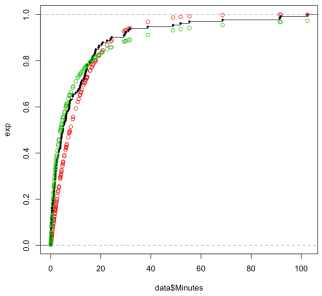
In [8]:
```
#a
plot(ecdf(data$Minutes), verticals=T, cex=0.5)
```



In [9]:
```
#b
y_bar<-mean(data$Minutes)
lambda<-1/y_bar
#c
exp<-pexp(data$Minutes, rate=lambda)
plot(data$Minutes,exp, col=2)
lines(ecdf(data$Minutes), verticals=T, cex=0.5)
```



In [10]:
```
#d
min_log<-log(data$Minutes)
mu=mean(min_log)
sd=sd(min_log)
#e
log<-plnorm(data$Minutes, meanlog=mu, sdlog=sd)
plot(data$Minutes,exp, col=2)
lines(ecdf(data$Minutes), verticals=T, cex=0.5)
points(data$Minutes, log, col=3)
```



In [ ]:

In [ ]: