b-worksheet - Describing Distributions with Numbers

1. The life span in days of 88 wildtype and 99 transgenic mosquitoes is given in `mosquitoes.csv`. Download these data using

```
# this command reads in a comma-separated-values file at the following url
# and stores it in a variable called mosquitoes

mosquitoes<-read.csv("http://math.arizona.edu/~jwatkins/mosquitoes.csv")

# you can also read in a local file by identifying the path to the file
```

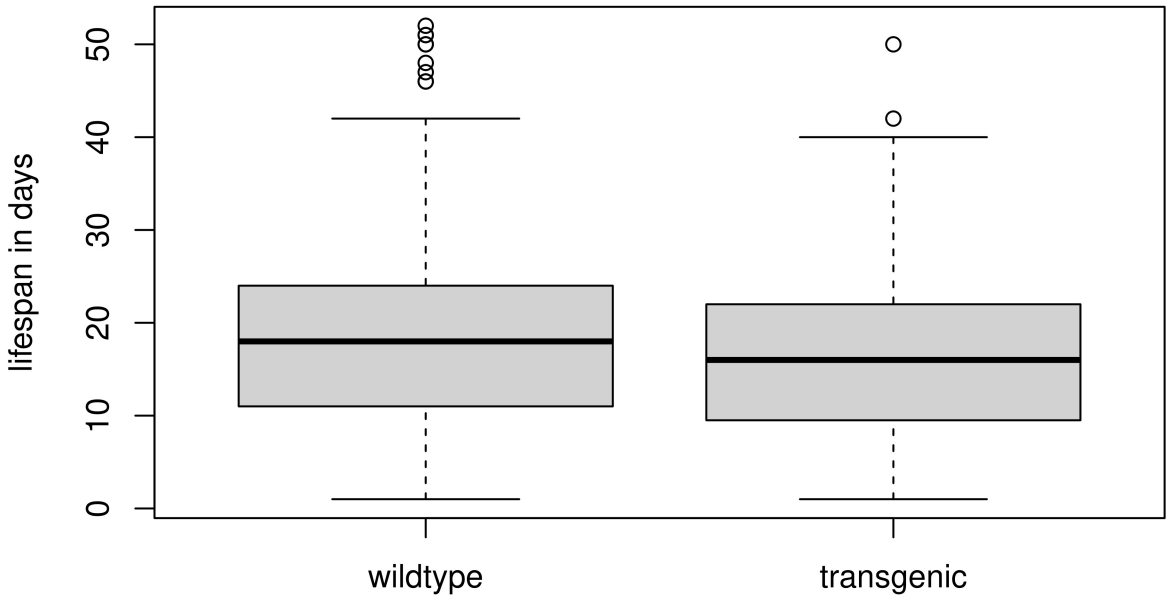
   a. Give the five number summary of the life span of both types of mosquitoes.

```
# Hint:  use the summary() function on the mosquitoes data frame
summary(mosquitoes)
```

```
##     wildtype       transgenic
##  Min.   : 1.00   Min.   : 1.00
##  1st Qu.:11.00   1st Qu.: 9.50
##  Median :18.00   Median :16.00
##  Mean   :20.78   Mean   :16.55
##  3rd Qu.:24.00   3rd Qu.:22.00
##  Max.   :52.00   Max.   :50.00
##  NA's   :11
```

   b. Give side by side box plots of the life span of both types of mosquitoes.

```
# Hint:  use the boxplot() function on the mosquitoes data frame
# set the parameter ylab = "lifespan in days"
boxplot(mosquitoes, ylab = "lifespan in days")
```



c. You can show the data for each column; we have already created two new vectors to do this: `wildtype` and `transgenic` . Place on one graph the empirical survival functions by using the command `par(new=TRUE)`. Be sure to give them the same limits for the values on each of the axes and use different colors for each mosquito type.

```
# na.omit(vector) provides a copy of the vector with the NAs removed
# we'll work with the data in vectors to avoid the issue of there
# being a different number of wildtype samples than transgenic samples
wildtype <- na.omit(mosquitoes$wildtype)
transgenic <- mosquitoes$transgenic

#Hint: What proportion of mosquitoes survive to a particular age?

# For an empirical CDF of a variable the x-values are the sorted
# values and the y-values should be equispaced between 0 and 1,
# the example code for y_wild shows one way to do this using
# the length() function.  So 1:10/10 = [0.1 0.2 ... 1.0]
# and 1:N/N = [1/N 2/N ... N/N]

x_wild <- sort(mosquitoes$wildtype,decreasing=TRUE)
y_wild <- (1:length(x_wild))/length(x_wild)

x_trans <- sort(mosquitoes$transgenic,decreasing=TRUE)
y_trans <- (1:length(x_trans))/length(x_trans)
```

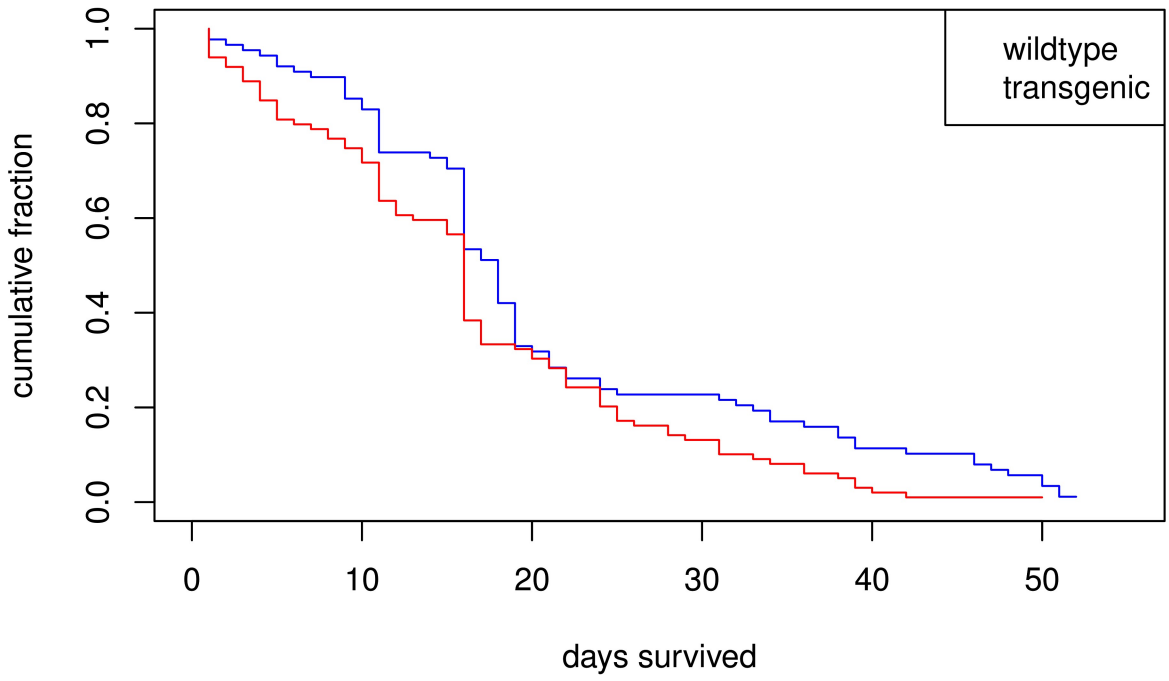
```r
# The following shows how to overlay two plots:
plot(x_wild,y_wild,xlim=c(0,55),ylim=c(0,1),
     xlab="days survived",ylab="cumulative fraction", type="s",col="blue")

# par() sets graphical parameters, new = TRUE tells R to dump the next
# plot on top of the first

par(new=TRUE)

# When overlaying plots make sure that xlim and ylim (the graph axes) are
# the same as well as xlab and ylab (the graph labels)

plot(x_trans,y_trans,xlim=c(0,55),ylim=c(0,1),
     xlab="days survived",ylab="cumulative fraction", type="s",col="red")
# We'll include a legend so the reader knows which line is which
legend("topright",legend=c("wildtype","transgenic"),col=c("blue","red"))
```



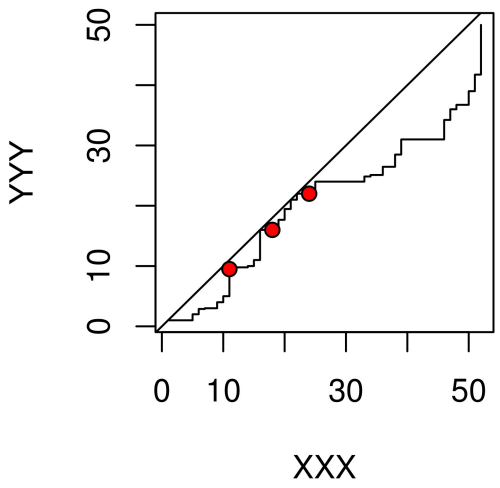
d. Give the Q-Q plot of the two types of mosquitoes. Indicate the median and the first and third quartiles on the graph.

```r
# notice that we made the plot square by setting fig.width and fig.height in
# the options for the chunk

 XXX = wildtype
 YYY = transgenic
```

```r
qqplot(XXX,YYY,type="s")
abline(a=0,b=1) # plot y = x

#Hint: Think back to part (a) to find the quartiles to plot.
 XXX = c(11.000,18.000,24.000) #are the 3 quartiles for wildtype
 YYY = c(9.50,16.000,22.000) #are the 3 quartiles for transgenic
# pch sets the character type to plot
# bg sets the color of the points to red
points(XXX,YYY,pch = 21,bg = "red")
```



e. One genotype of mosquito lives longer, on average, than the other. Explain how this can be seen in the boxplots, in the survival function and on the Q-Q plot.

Boxplots: We can see that the boxplot for the wild genotype of mosquitoes is higher than that of the transgenic genotype. This show that on average, the lifespans of these types of mosquitoes is greater, meaning that they live for a longer period of time

Survival function: Since the area of the survival function for transgenic is lower than the area of that of the wildtype mosquitoes, we can infer that on average, the lifespan of the wildtype mosquitoes is higher since the area of the survival function gives the average of the data that we plot.

Q-Q plot: Here, since the data that we have for both the mosquito types lies below the line with slope 1, the mosquitoes whose data is plotted on the x-axis(wildtype) live for a longer time than those that are on the y-axis (transgenic)