Worksheet 01 - R Intro and Displaying/Describing Distributions

##Shawn Kim

NOTE: *While I wish I could foresee every error you may find yourself facing when trying to knit to a PDF, I sadly do not have that superpower. Regardless, my first suggestion is to ensure the package* ***tinytex*** *is installed by entering* `tinytex::install_tinytex()` *into the console. You may also need to install the* `formatR` *package. To do so, enter* `install.package("formatR")` *into the console window and run it. If you have additional errors, I encourage you to post to Teams so that a fellow student or myself may provide instructions on how to remedy them.*

1. In D2L access Appendix A, which is included as part of this assignment and work through the pieces in the R console. If you are using R studio, click on console underneath the box that you can type in here. Using your outputs, answer the following questions.

a. Give exp(1) to 4 decimal places. **2.7183**
b. Give mean(x) and sd(x) for the 50 normal random variables. **mean(x) = -0.1965967, sd(x) = 1.040841**
c. In sorting the values of x, what is the smallest value for x? the largest? **smallest: -2.238445258, largest: 1.706856970**
d. Explain what the following commands do in a plot.
   - "h" **histogram like vertical lines**
   - "l" **lines**
   - "s" **steps**
   - "b" **both lines and points**
e. Describe seq(25,1,-1) and rep(0,25). **seq(25,1,-1) starts from 25 and counts downward in steps of -1 all the way to 1, rep(0,25) repeats 0 twenty-five times**
f. Give the top row of matrix A. **1 17 12 10 22 25 4 6 4 10**
g. Give the second, third, and fourth columns of the matrix A. **2nd col: 17, 13, 2; 3rd col: 12, 16, 5; 4th col: 10, 6, 3**
h. Describe what barplot(data) shows. **barplot(data) shows the sum of every number in each column and there are 10 columns, 3 rows**
i. What does abline do? **abline draws a line of best fit through the scatterplot**
j. Describe table(fair) and table(biased). What do you think that this command is doing? **table(fair) was data for an evenly weighted or "fair" six sided dice rolled 1000 times, table(biased) was data for an unevenly weight six sided dice that did not have equal odds for every side**
k. Describe curve(dnorm(x),-3,3). Do you think that the contents of the variable x affect the result of this command? **curve(dnorm(x), -3, 3) shows the curve of dnorm(x) over the interval [-3,3]. The contents of variable x affect the result of the command because if x contained a different set of values, the dnorm of x could change, and thefore its curve could change as well**

2. The life span in days of 88 wildtype and 99 transgenic mosquitoes is given in `mosquitoes.csv`. Download these data from D2L.

```
# We begin with an RMD file that includes many hints for the parts that use R.
# As the semester continues, we will reduce/eliminate these helpful comments.
# To use the useful comments you will sometimes (most often) need to uncomment
# the lines and fill in text to the places that are indicated as FILL IN

# 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
# mosquitoes<-read.csv('mosquitoes.csv')
```

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

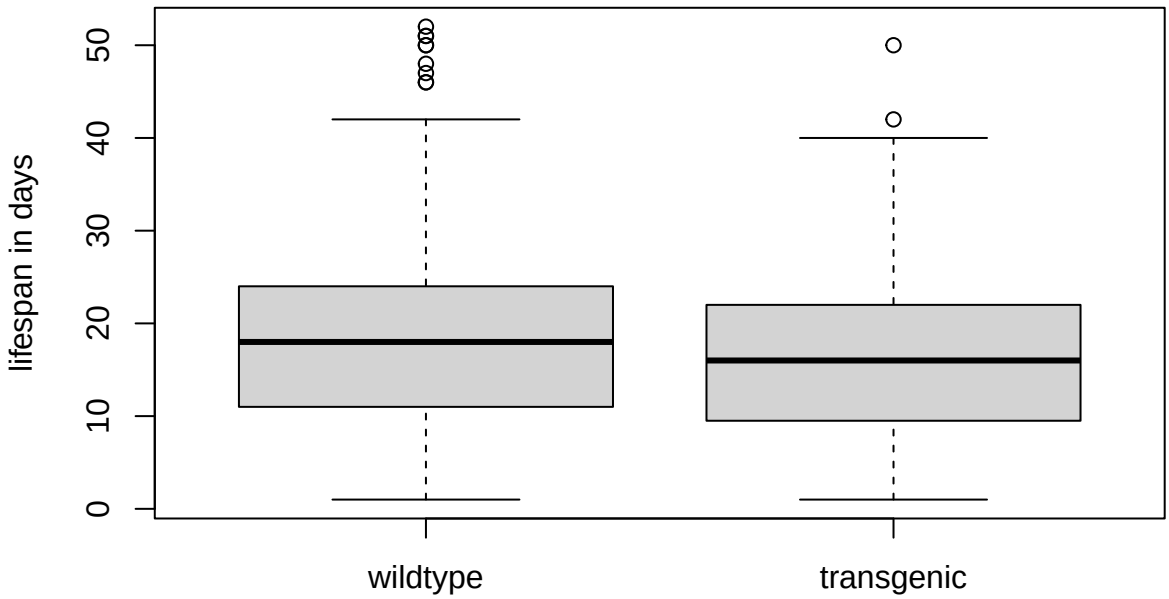
```
# NOTE: use the summary() function on the mosquitoes data frame

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.

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

# boxplot(mosquitoes, ylab='lifespan in days')
```

```
boxplot(mosquitoes, ylab = "lifespan in days")
```

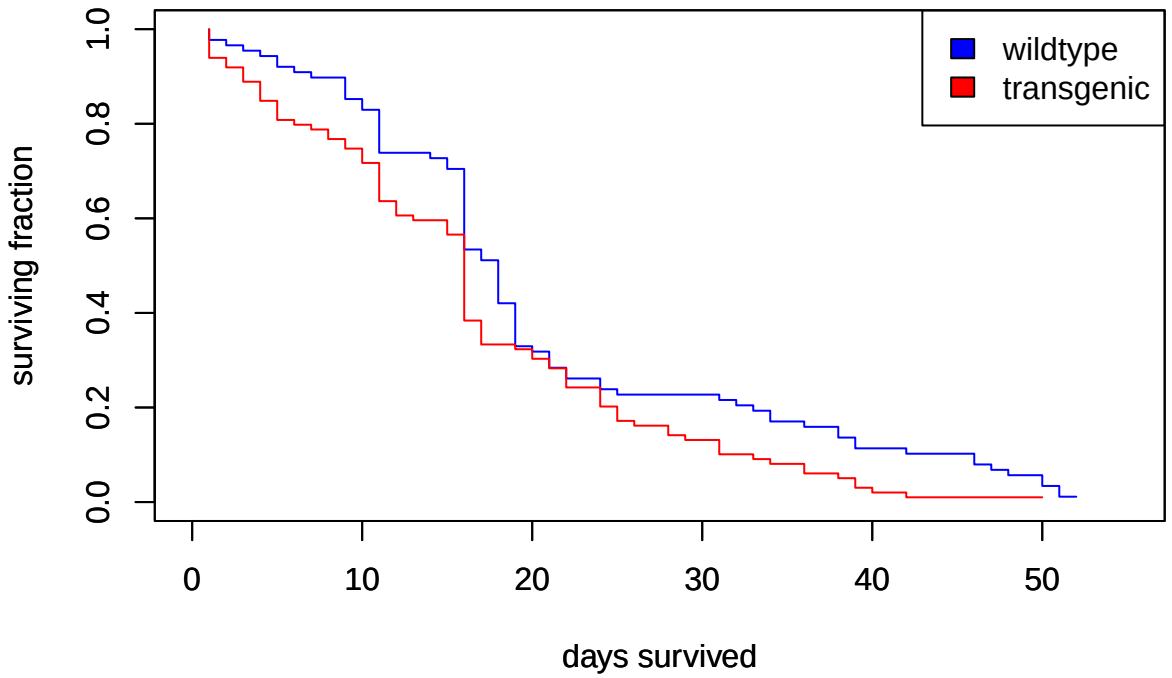


c. Graph the empirical survival functions for both types of mosquitoes in one graph. We start by creating two new vectors to do this: `wildtype` and `transgenic` . To overlay the empirical survival functions use the command `par(new=TRUE)`. Be sure to give the same limits for the values on each of the axes and use different colors for each mosquito type.

NOTE: To do these tasks, enter your code in the R chunk below by filling in the blanks denoted with `FILL IN` and un-commenting all non-NOTE lines.

```
# NOTE: na.omit(vector) provides a copy of the vector with the NAs removed
# NOTE: We'll work with the data in vectors to avoid the issue of there being a
# different number of wildtype samples than transgenic samples. The following
# creates the vectors, one called wildtype and the other transgenic.

wildtype <- na.omit(mosquitoes$wildtype)
transgenic <- mosquitoes$transgenic

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

# NOTE: 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.

# NOTE: So 1:10/10 = [0.1 0.2 ... 1.0] and 1:N/N = [1/N 2/N ... N/N]

# NOTE: This will be for the wildtype mosquitoes.

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

# NOTE: This will be for the transgenic type mosquitoes.

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

# NOTE: The following shows how to overlay two plots:
```

```r
# NOTE: First we create a plot of the wildtype mosquitoes by plotting the
# x-values against the y-values for the wildtype that we created.

plot(x_wild, y_wild, xlim = c(0, 55), ylim = c(0, 1), xlab = "days survived", ylab = "surviving fractio
    type = "s", col = "blue")

# NOTE: par() sets graphical parameters NOTE: new = TRUE tells R to dump the
# next plot on top of the first NOTE: You will need to know how to overlay
# plots in the future

par(new = TRUE)

# NOTE: 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 = "surviving fracti
    type = "s", col = "red")

# NOTE: We'll include a legend so the reader knows which line is which NOTE:
# You will need to know how to add a legend to plots in the future

legend("topright", legend = c("wildtype", "transgenic"), fill = c("blue", "red"))
```



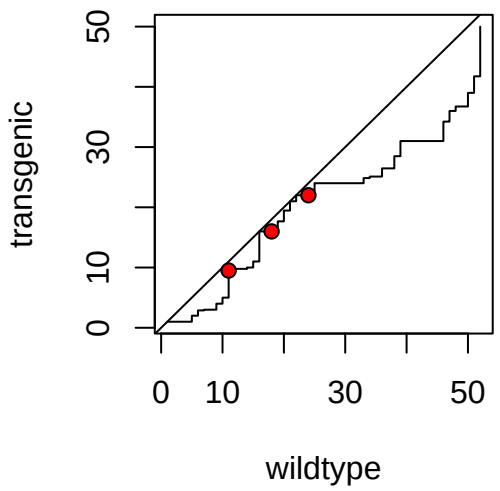
b

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

NOTE: To do these tasks, enter your code in the R chunk below by filling in the blanks denoted with `FILL IN` and un-commenting all non-NOTE lines.

```r
# NOTE: Notice the nifty option in the chuck opener NOTE: we made the plot
# square by wisely choosing fig.width and fig.height

# NOTE: What is a qqplot? This type of plot matches up the various quantiles of
```

```
# the two data sets. For example, one of the points on the qqplot will be
# (median of wildtype data, median of transgenic type data).

# NOTE: Let's make a qqplot!  NOTE: We want the wildtype data on the x-axis
# (first missing entry in qqplot) and the transgenic data on the y-axis (second
# missing entry in qqplot)

qqplot(wildtype, transgenic, type = "s")

# NOTE: we also want to add the diagonal line y=x to the qqplot NOTE: You will
# need to know how to add a line to plots in the future

abline(a = 0, b = 1)

# NOTE: Now we indicate the requested values NOTE: Use part (a) to find the
# quartiles to plot.

XXX = c(11, 18, 24)  #are the 3 quartiles for wildtype (these will just be values)
YYY = c(9.5, 16, 22)  #are the 3 quartiles for transgenic (these will just be values)

# NOTE: Finally, we can add the points to the plot to indicate the quantiles
# NOTE: pch sets the character type (large filled circle) to plot NOTE: bg sets
# the color (red) of the points

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: **For the boxplots, you can see that the wildtype has higher mean, median, and quartiles than the transgenic types which means wildtypes live longer on average compared to the trangenic type**

Survival function: **for the survival function, you can see that the plot for wildtype (in red) is consistently under the plot for the transgenic type (in blue) over the entire x axis. That means that generally, the transgenic type has a larger fraction of its population that lived at least or less than a specific lifespan than the wildtype as we go up in lifespan, meaning that on average, the transgenic type population had mosquitoes with less lifespan than that of the wildtype**

Q-Q plot:**for the qq plot, we can see that the data is skewed below the diagonal line which tells us that the wildtype population lives longer on average (since the wildtype is the x-axis), the 1st and 3 quartile as well as the median are also below the diagonal line which means that these values for the wildtype are higher than that for the transgenic type**