HW7

Mia Murphy

4/7/2021

###Open libraries

```
library(ggplot2)
library(MASS)
```

###Read in data vector

##To illustrate, we will generate some fake data here:

```
nyc_squirrels1 <- readr::read_csv("https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_character(),
##   long = col_double(),
##   lat = col_double(),
##   date = col_double(),
##   hectare_squirrel_number = col_double(),
##   running = col_logical(),
##   chasing = col_logical(),
##   climbing = col_logical(),
##   eating = col_logical(),
##   foraging = col_logical(),
##   kuks = col_logical(),
##   quaas = col_logical(),
##   moans = col_logical(),
##   tail_flags = col_logical(),
##   tail_twitches = col_logical(),
##   approaches = col_logical(),
##   indifferent = col_logical(),
##   runs_from = col_logical(),
##   zip_codes = col_double(),
##   community_districts = col_double(),
##   borough_boundaries = col_double()
##   # ... with 2 more columns
## )
## i Use `spec()` for the full column specifications.
```

##In the third step of this exercise, you wil substitute in your own data for this fake data set. But for now, use the code chunks below to see how you fit different statistical distributions to a vector of observations, and then estimate the maximum likelihood parameters for each distribution.

##Plot histogram of data

##Plot a histogram of the data, using a modification of the code from lecture. Here we are switching from qplot to ggplot for more graphics options. We are also rescaling the y axis of the histogram from counts to density, so that the area under the histogram equals 1.0.

```
str(nyc_squirrels1)
```

```
## tibble [3,023 x 36] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ long                                       : num [1:3023] -74 -74 -74 -74 -74 ...
##  $ lat                                        : num [1:3023] 40.8 40.8 40.8 40.8 40.8 ...
##  $ unique_squirrel_id                         : chr [1:3023] "37F-PM-1014-03" "37E-PM-1006-03" "2E-AM-
##  $ hectare                                    : chr [1:3023] "37F" "37E" "02E" "05D" ...
##  $ shift                                      : chr [1:3023] "PM" "PM" "AM" "PM" ...
##  $ date                                       : num [1:3023] 10142018 10062018 10102018 10182018 1018
##  $ hectare_squirrel_number                    : num [1:3023] 3 3 3 5 1 2 2 3 9 14 ...
##  $ age                                        : chr [1:3023] NA "Adult" "Adult" "Juvenile" ...
##  $ primary_fur_color                          : chr [1:3023] NA "Gray" "Cinnamon" "Gray" ...
##  $ highlight_fur_color                        : chr [1:3023] NA "Cinnamon" NA NA ...
##  $ combination_of_primary_and_highlight_color: chr [1:3023] "+" "Gray+Cinnamon" "Cinnamon+" "Gray+"
##  $ color_notes                                : chr [1:3023] NA NA NA NA ...
##  $ location                                   : chr [1:3023] NA "Ground Plane" "Above Ground" "Above
##  $ above_ground_sighter_measurement           : chr [1:3023] NA "FALSE" "4" "3" ...
##  $ specific_location                          : chr [1:3023] NA NA NA NA ...
##  $ running                                    : logi [1:3023] FALSE TRUE FALSE FALSE FALSE FALSE ...
##  $ chasing                                    : logi [1:3023] FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ climbing                                   : logi [1:3023] FALSE FALSE TRUE TRUE FALSE FALSE ...
##  $ eating                                     : logi [1:3023] FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ foraging                                   : logi [1:3023] FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ other_activities                           : chr [1:3023] NA NA NA NA ...
##  $ kuks                                       : logi [1:3023] FALSE FALSE FALSE FALSE TRUE FALSE ...
##  $ quaas                                      : logi [1:3023] FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ moans                                      : logi [1:3023] FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ tail_flags                                 : logi [1:3023] FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ tail_twitches                              : logi [1:3023] FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ approaches                                 : logi [1:3023] FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ indifferent                                : logi [1:3023] FALSE FALSE TRUE FALSE FALSE FALSE ...
##  $ runs_from                                  : logi [1:3023] FALSE TRUE FALSE TRUE FALSE FALSE ...
##  $ other_interactions                         : chr [1:3023] NA "me" NA NA ...
##  $ lat_long                                   : chr [1:3023] "POINT (-73.9561344937861 40.794082388408
##  $ zip_codes                                  : num [1:3023] NA NA NA NA NA NA NA NA NA NA ...
##  $ community_districts                        : num [1:3023] 19 19 19 19 19 19 19 19 19 19 ...
##  $ borough_boundaries                         : num [1:3023] 4 4 4 4 4 4 4 4 4 4 ...
##  $ city_council_districts                     : num [1:3023] 19 19 19 19 19 19 19 19 19 19 ...
##  $ police_precincts                           : num [1:3023] 13 13 13 13 13 13 13 13 13 13 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   long = col_double(),
##   ..   lat = col_double(),
##   ..   unique_squirrel_id = col_character(),
##   ..   hectare = col_character(),
##   ..   shift = col_character(),
##   ..   date = col_double(),
##   ..   hectare_squirrel_number = col_double(),
```

```
##   ..   age = col_character(),
##   ..   primary_fur_color = col_character(),
##   ..   highlight_fur_color = col_character(),
##   ..   combination_of_primary_and_highlight_color = col_character(),
##   ..   color_notes = col_character(),
##   ..   location = col_character(),
##   ..   above_ground_sighter_measurement = col_character(),
##   ..   specific_location = col_character(),
##   ..   running = col_logical(),
##   ..   chasing = col_logical(),
##   ..   climbing = col_logical(),
##   ..   eating = col_logical(),
##   ..   foraging = col_logical(),
##   ..   other_activities = col_character(),
##   ..   kuks = col_logical(),
##   ..   quaas = col_logical(),
##   ..   moans = col_logical(),
##   ..   tail_flags = col_logical(),
##   ..   tail_twitches = col_logical(),
##   ..   approaches = col_logical(),
##   ..   indifferent = col_logical(),
##   ..   runs_from = col_logical(),
##   ..   other_interactions = col_character(),
##   ..   lat_long = col_character(),
##   ..   zip_codes = col_double(),
##   ..   community_districts = col_double(),
##   ..   borough_boundaries = col_double(),
##   ..   city_council_districts = col_double(),
##   ..   police_precincts = col_double()
##   .. )
```

```
summary(nyc_squirrels1)
```

```
##       long             lat        unique_squirrel_id   hectare
##  Min.   :-73.98   Min.   :40.76   Length:3023        Length:3023
##  1st Qu.:-73.97   1st Qu.:40.77   Class :character   Class :character
##  Median :-73.97   Median :40.78   Mode  :character   Mode  :character
##  Mean   :-73.97   Mean   :40.78
##  3rd Qu.:-73.96   3rd Qu.:40.79
##  Max.   :-73.95   Max.   :40.80
##
##     shift                date          hectare_squirrel_number
##  Length:3023        Min.   :10062018   Min.   : 1.000
##  Class :character   1st Qu.:10082018   1st Qu.: 2.000
##  Mode  :character   Median :10122018   Median : 3.000
##                     Mean   :10119487   Mean   : 4.124
##                     3rd Qu.:10142018   3rd Qu.: 6.000
##                     Max.   :10202018   Max.   :23.000
##
##      age             primary_fur_color  highlight_fur_color
##  Length:3023        Length:3023        Length:3023
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
```

```
##
##
##
##  combination_of_primary_and_highlight_color color_notes
##  Length:3023                                Length:3023
##  Class :character                           Class :character
##  Mode  :character                           Mode  :character
##
##
##
##
##    location         above_ground_sighter_measurement specific_location
##  Length:3023        Length:3023                      Length:3023
##  Class :character   Class :character                 Class :character
##  Mode  :character   Mode  :character                 Mode  :character
##
##
##
##
##   running          chasing          climbing         eating
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:2293      FALSE:2744      FALSE:2365      FALSE:2263
##  TRUE :730       TRUE :279       TRUE :658       TRUE :760
##
##
##
##
##   foraging       other_activities      kuks            quaas
##  Mode :logical   Length:3023        Mode :logical   Mode :logical
##  FALSE:1588      Class :character   FALSE:2921      FALSE:2973
##  TRUE :1435      Mode  :character   TRUE :102       TRUE :50
##
##
##
##
##    moans          tail_flags      tail_twitches   approaches
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:3020      FALSE:2868      FALSE:2589      FALSE:2845
##  TRUE :3         TRUE :155       TRUE :434       TRUE :178
##
##
##
##
##  indifferent     runs_from       other_interactions   lat_long
##  Mode :logical   Mode :logical   Length:3023        Length:3023
##  FALSE:1569      FALSE:2345      Class :character   Class :character
##  TRUE :1454      TRUE :678       Mode  :character   Mode  :character
##
##
##
##
##    zip_codes     community_districts borough_boundaries city_council_districts
##  Min.   :10090   Min.   :11          Min.   :4          Min.   :19.00
##  1st Qu.:12081   1st Qu.:19          1st Qu.:4          1st Qu.:19.00
```

```
##  Median :12420   Median :19          Median :4          Median :19.00
##  Mean   :11828   Mean   :19          Mean   :4          Mean   :19.07
##  3rd Qu.:12423   3rd Qu.:19          3rd Qu.:4          3rd Qu.:19.00
##  Max.   :12423   Max.   :23          Max.   :4          Max.   :51.00
##  NA's   :3014
##  police_precincts
##  Min.   :10
##  1st Qu.:13
##  Median :13
##  Mean   :13
##  3rd Qu.:13
##  Max.   :18
##
```

```
nyc_squirrels1 <- rnorm(n=3000,mean=0.2)
nyc_squirrels1 <- data.frame(1:3000, nyc_squirrels1)
names(nyc_squirrels1) <- list("ID","myVar")
nyc_squirrels1 <- nyc_squirrels1[nyc_squirrels1$myVar>0,]
str(nyc_squirrels1)
```

```
## 'data.frame':    1747 obs. of  2 variables:
##  $ ID   : int  2 5 6 7 8 11 12 13 14 16 ...
##  $ myVar: num  1.7185 0.8712 0.9538 0.0498 1.0133 ...
```

```
summary(nyc_squirrels1$myVar)
```

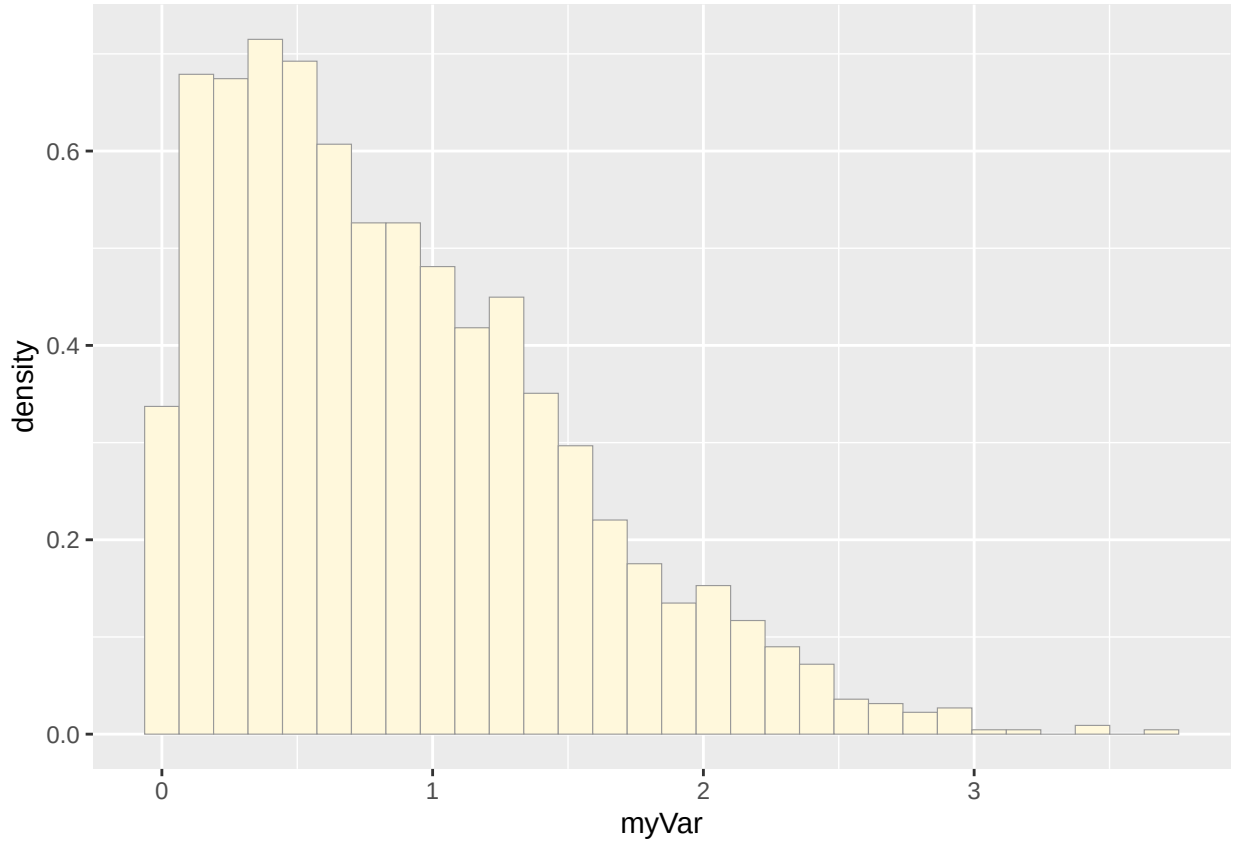
```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.000113 0.362097 0.740312 0.877820 1.270461 3.691887
```

###Add empirical density curve

##Now modify the code to add in a kernel density plot of the data. This is an empirical curve that is fitted to the data. It does not assume any particular probability distribution, but it smooths out the shape of the histogram:

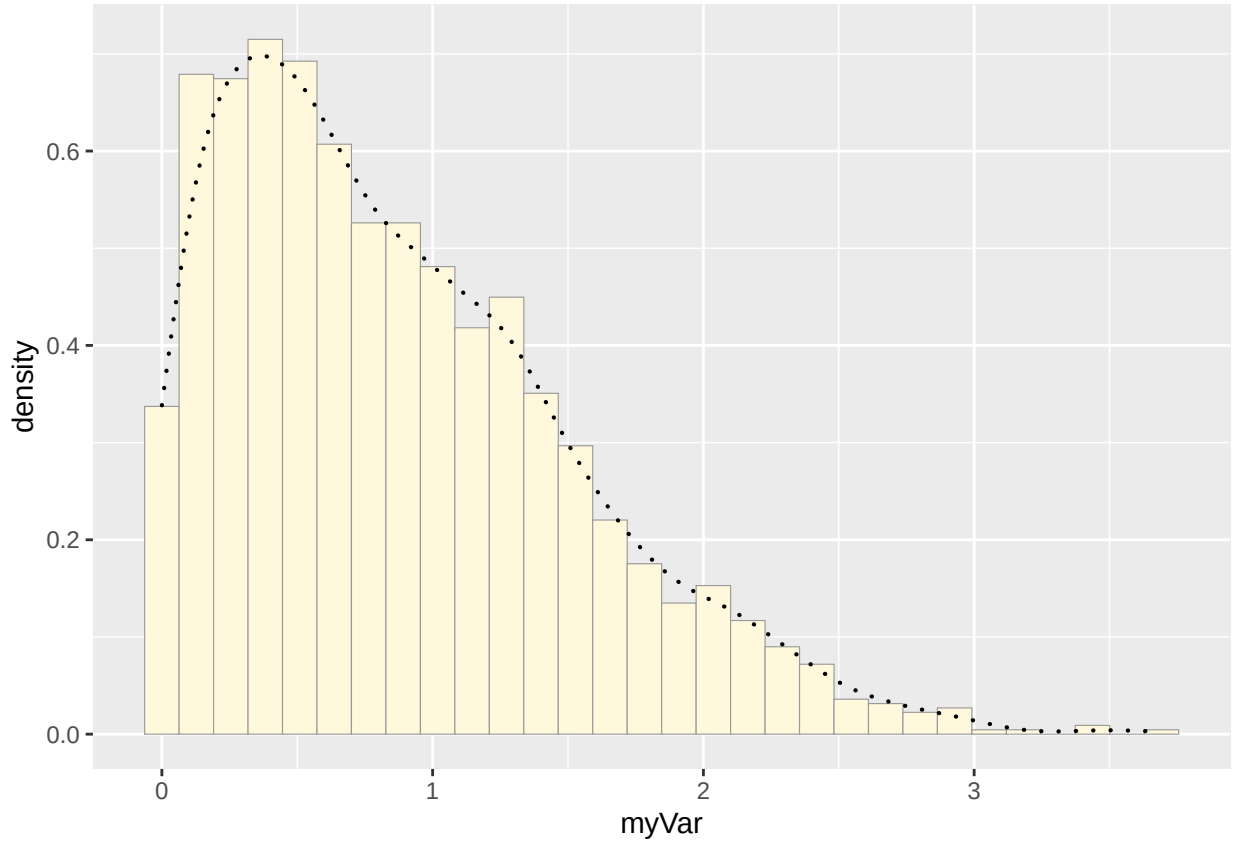
```
p1 <- ggplot(data=nyc_squirrels1, aes(x=myVar, y=..density..)) +
  geom_histogram(color="grey60", fill="cornsilk", size=0.2)
print(p1)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
p1 <-  p1 +  geom_density(linetype="dotted", size=0.75)
print(p1)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



###Get maximum likelihood parameters for normal

##Next, fit a normal distribution to your data and grab the maximum likelihood estimators of the two parameters of the normal, the mean and the variance:

```
normPars <- fitdistr(nyc_squirrels1$myVar, "normal")
print(normPars)
```

```
##       mean          sd
##   0.87781976   0.64225831
##  (0.01536609) (0.01086546)
```

```
str(normPars)
```

```
## List of 5
##  $ estimate: Named num [1:2] 0.878 0.642
##   ..- attr(*, "names")= chr [1:2] "mean" "sd"
##  $ sd      : Named num [1:2] 0.0154 0.0109
##   ..- attr(*, "names")= chr [1:2] "mean" "sd"
##  $ vcov    : num [1:2, 1:2] 0.000236 0 0 0.000118
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:2] "mean" "sd"
##   .. ..$ : chr [1:2] "mean" "sd"
##  $ n       : int 1747
##  $ loglik  : num -1705
##  - attr(*, "class")= chr "fitdistr"
```

```
normPars$estimate["mean"]
```

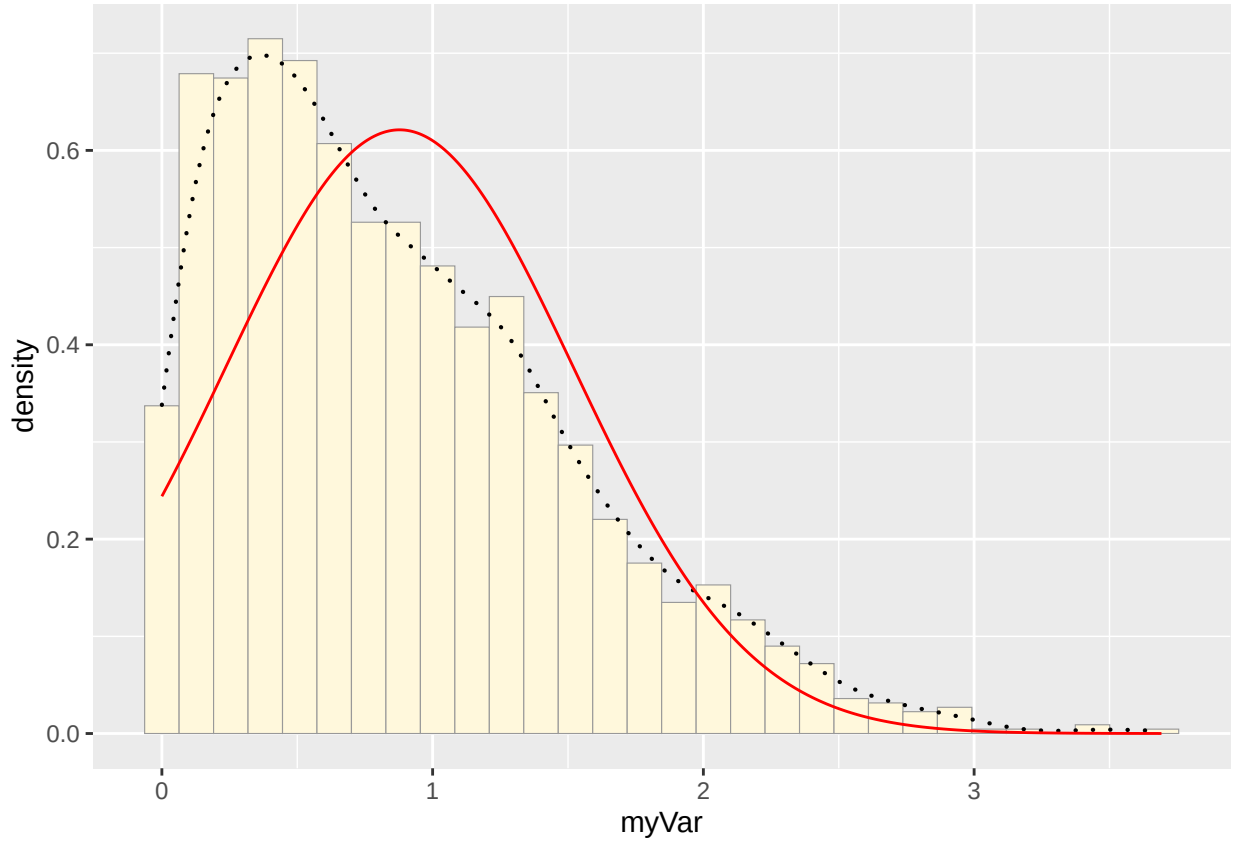
```
##      mean
## 0.8778198
```

###Plot normal probability density

##Now let's call the dnorm function inside ggplot's stat_function to generate the probability density for the normal distribution. Read about stat_function in the help system to see how you can use this to add a smooth function to any ggplot. Note that we first get the maximum likelihood parameters for a normal distribution fitted to thse data by calling fitdistr. Then we pass those parameters (meanML and sdML to stat_function:

```
meanML <- normPars$estimate["mean"]
sdML <- normPars$estimate["sd"]

xval <- seq(0,max(nyc_squirrels1$myVar),len=length(nyc_squirrels1$myVar))

 stat <- stat_function(aes(x = xval, y = ..y..), fun = dnorm, colour="red", n = length(nyc_squirrels1$my
 p1 + stat
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



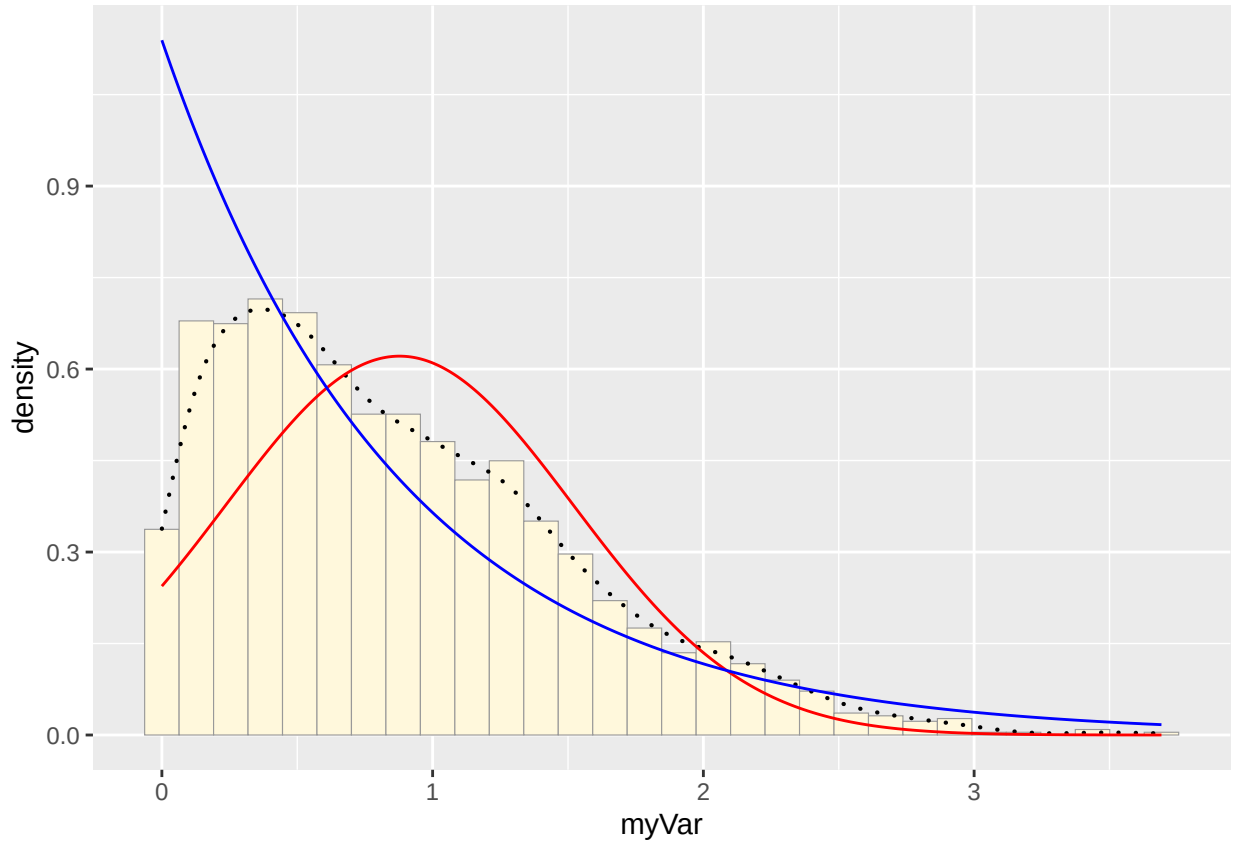
##Notice that the best-fitting normal distribution (red curve) for these data actually has a biased mean. That is because the data set has no negative values, so the normal distribution (which is symmetric) is not working well.

###Plot exponential probability density

##Now let's use the same template and add in the curve for the exponential:

```
expoPars <- fitdistr(nyc_squirrels1$myVar,"exponential")
rateML <- expoPars$estimate["rate"]

stat2 <- stat_function(aes(x = xval, y = ..y..), fun = dexp, colour="blue", n = length(nyc_squirrels1$my
 p1 + stat + stat2
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

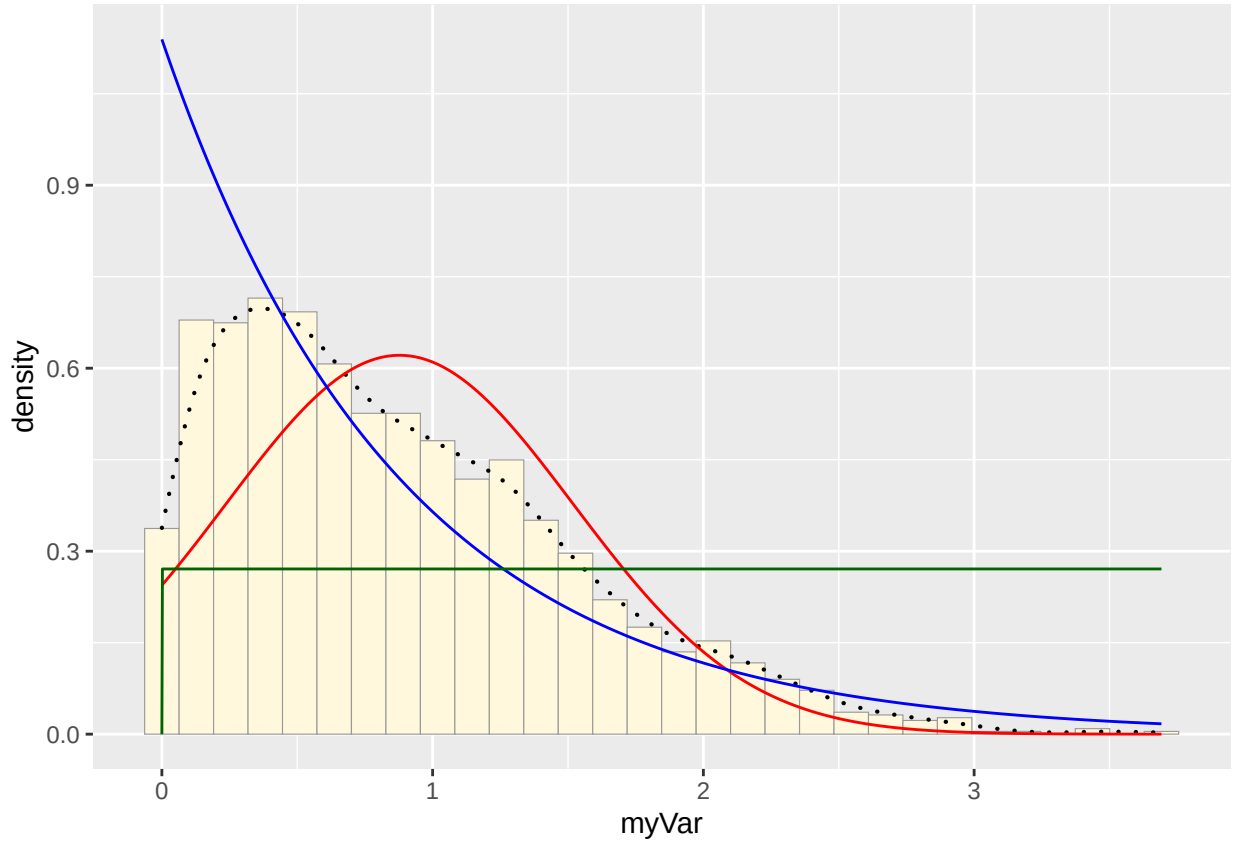


###Plot uniform probability density

##For the uniform, we don't need to use fitdistr because the maximum likelihood estimators of the two parameters are just the minimum and the maximum of the data:

```
stat3 <- stat_function(aes(x = xval, y = ..y..), fun = dunif, colour="darkgreen", n = length(nyc_squirre
 p1 + stat + stat2 + stat3
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



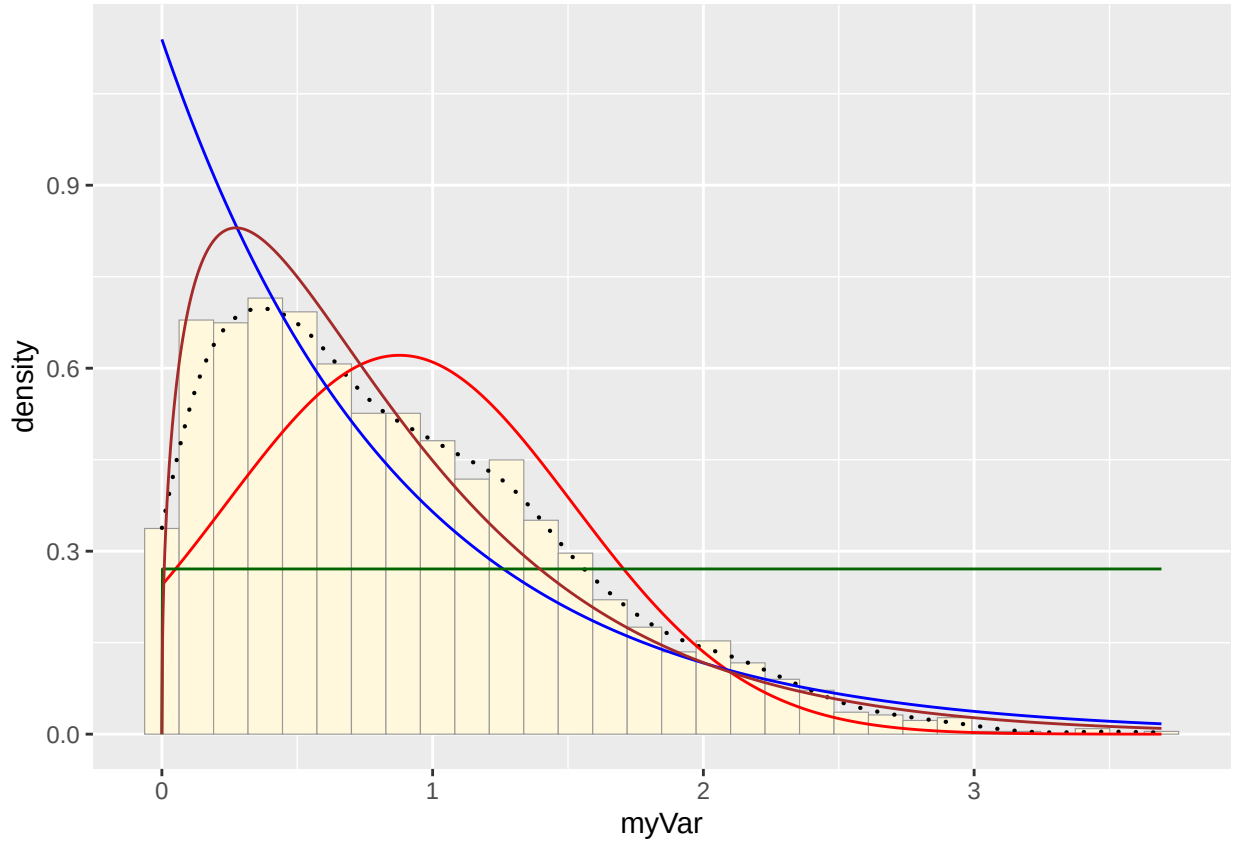
##Plot gamma probability density

```
gammaPars <- fitdistr(nyc_squirrels1$myVar,"gamma")
```

```
## Warning in densfun(x, parm[1], parm[2], ...): NaNs produced

## Warning in densfun(x, parm[1], parm[2], ...): NaNs produced

## Warning in densfun(x, parm[1], parm[2], ...): NaNs produced

## Warning in densfun(x, parm[1], parm[2], ...): NaNs produced

## Warning in densfun(x, parm[1], parm[2], ...): NaNs produced

## Warning in densfun(x, parm[1], parm[2], ...): NaNs produced

## Warning in densfun(x, parm[1], parm[2], ...): NaNs produced
```

```
shapeML <- gammaPars$estimate["shape"]
rateML <- gammaPars$estimate["rate"]

stat4 <- stat_function(aes(x = xval, y = ..y..), fun = dgamma, colour="brown", n = length(nyc_squirrels
 p1 + stat + stat2 + stat3 + stat4
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



##Plot beta probability density

##This one has to be shown in its own plot because the raw data must be rescaled so they are between 0 and 1, and then they can be compared to the beta.

```
pSpecial <- ggplot(data=nyc_squirrels1, aes(x=myVar/(max(myVar + 0.1)), y=..density..)) +
  geom_histogram(color="grey60", fill="cornsilk", size=0.2) +
  xlim(c(0,1)) +
  geom_density(size=0.75,linetype="dotted")

betaPars <- fitdistr(x=nyc_squirrels1$myVar/max(nyc_squirrels1$myVar + 0.1), start=list(shape1=1, shape
```

```
## Warning in densfun(x, parm[1], parm[2], ...): NaNs produced

## Warning in densfun(x, parm[1], parm[2], ...): NaNs produced

## Warning in densfun(x, parm[1], parm[2], ...): NaNs produced

## Warning in densfun(x, parm[1], parm[2], ...): NaNs produced
```

```
shape1ML <- betaPars$estimate["shape1"]
shape2ML <- betaPars$estimate["shape2"]

statSpecial <- stat_function(aes(x = xval, y = ..y..), fun = dbeta, colour="orchid", n = length(nyc_squ
pSpecial + statSpecial
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```