STAT 206 Lab 4

Due Monday, October 30, 5:00 PM

General instructions for labs: You are encouraged to work in pairs to complete the lab. Labs must be completed as an R Markdown file. Be sure to include your lab partner (if you have one) and your own name in the file. Give the commands to answer each question in its own code block, which will also produce plots that will be automatically embedded in the output file. Each answer must be supported by written statements as well as any code used.

Agenda: Distributions as models, method of moments and maximum likelihood estimation.

The Beta is a random variable bounded between 0 and 1 and often used to model the distribution of proportions. The probability distribution function for the Beta with parameters α and β is

$$p(x|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)+\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

where $\Gamma()$ is the Gamma function, the generalized version of the factorial. Thankfully, for this assignment, you need not know what the Gamma function is; you need only know that the mean of a Beta is $\frac{\alpha}{\alpha+\beta}$ and its variance is $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

For this assignment you will test the fit of the Beta distribution to the on-base percentages (OBPs) of hitters in the 2014 Major League Baseball season; each plate appearance (PA) results in the batter reaching base or not, and this measure is the fraction of successful attempts. This set has been pre-processed to remove those players with an insufficient number of opportunities for success.

Part I

1. Load the file [http://faculty.ucr.edu/~jflegal/206/mlb-obp.csv] into a variable of your choice in R. How many players have been included? What is the minimum number of plate appearances required to appear on this list? Who had the most plate appearances? What are the minimum, maximum and mean OBP?

```
#the variable for the data will be mlb since this is baseball data
#we use nrow to find the ammount of players, min and max to find the minimum and mamimum values
#of specified attributes and mean to find the average of a certain feature

mlb = read.csv("http://faculty.ucr.edu/~jflegal/206/mlb-obp.csv", header=TRUE)
#head(mlb)
total_players = nrow(mlb)
total_players
```

```
## [1] 441
```

```
min_PA = min(mlb$PA)
min_PA
```

```
## [1] 103
```

```
max_PA = max(mlb$PA)
max_PA
```

```
## [1] 726
```

```r
min_OBP = min(mlb$OBP)
min_OBP
```

```
## [1] 0.168
```

```r
max_OBP = max(mlb$OBP)
max_OBP
```

```
## [1] 0.432
```

```r
avg_OBP = mean(mlb$OBP)
avg_OBP
```

```
## [1] 0.3119184
```

2. Plot the data as a histogram with the option `probability=TRUE`. Add a vertical line for the mean of the distribution. Does the mean coincide with the mode of the distribution?

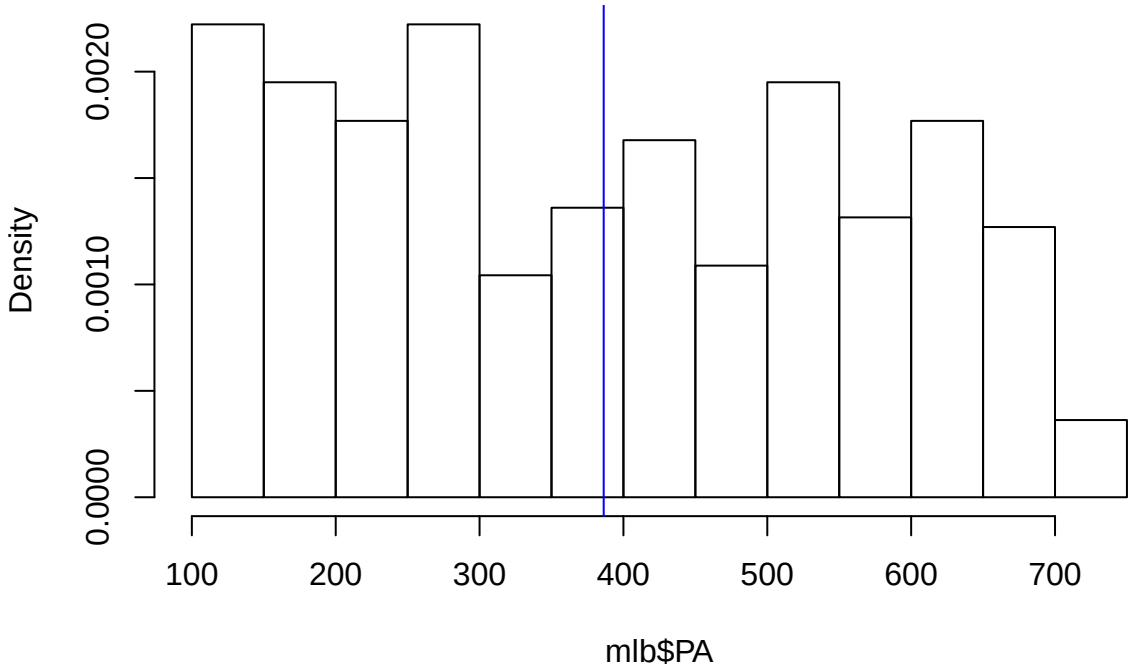
```r
avg_PA = mean(mlb$PA)
avg_PA
```

```
## [1] 386.2857
```

```r
hist_PA = hist(mlb$PA, probability = TRUE, main = "Plate appearances")
hist_PA
```

```
## $breaks
##  [1] 100 150 200 250 300 350 400 450 500 550 600 650 700 750
##
## $counts
##  [1] 49 43 39 49 23 30 37 24 43 29 39 28  8
##
## $density
##  [1] 0.0022222222 0.0019501134 0.0017687075 0.0022222222 0.0010430839
##  [6] 0.0013605442 0.0016780045 0.0010884354 0.0019501134 0.0013151927
## [11] 0.0017687075 0.0012698413 0.0003628118
##
## $mids
##  [1] 125 175 225 275 325 375 425 475 525 575 625 675 725
##
## $xname
## [1] "mlb$PA"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```

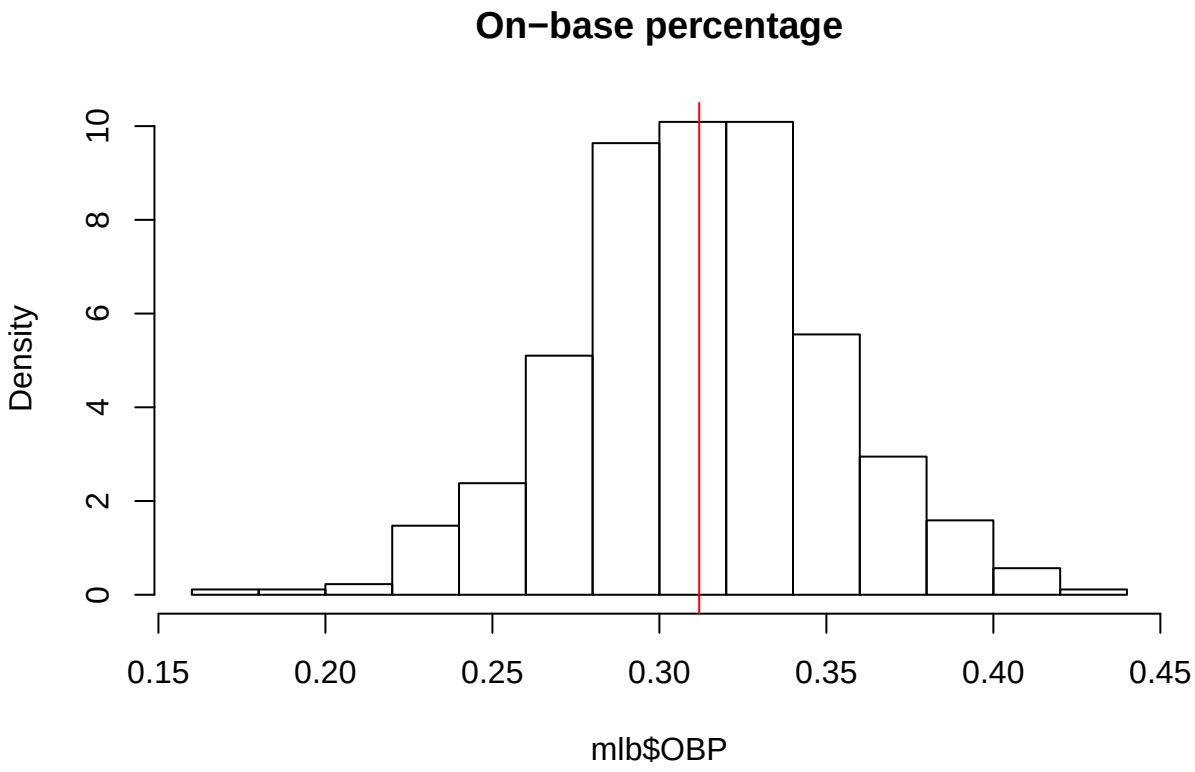
```r
abline(v=avg_PA, col = "blue")
```




```
hist_OBP = hist(mlb$OBP, probability = TRUE, main = "On-base percentage")
hist_OBP
```

```
## $breaks
##  [1] 0.16 0.18 0.20 0.22 0.24 0.26 0.28 0.30 0.32 0.34 0.36 0.38 0.40 0.42
## [15] 0.44
##
## $counts
##  [1]  1  1  2 13 21 45 85 89 89 49 26 14  5  1
##
## $density
##  [1]  0.1133787  0.1133787  0.2267574  1.4739229  2.3809524  5.1020408
##  [7]  9.6371882 10.0907029 10.0907029  5.5555556  2.9478458  1.5873016
## [13]  0.5668934  0.1133787
##
## $mids
##  [1] 0.17 0.19 0.21 0.23 0.25 0.27 0.29 0.31 0.33 0.35 0.37 0.39 0.41 0.43
##
## $xname
## [1] "mlb$OBP"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```

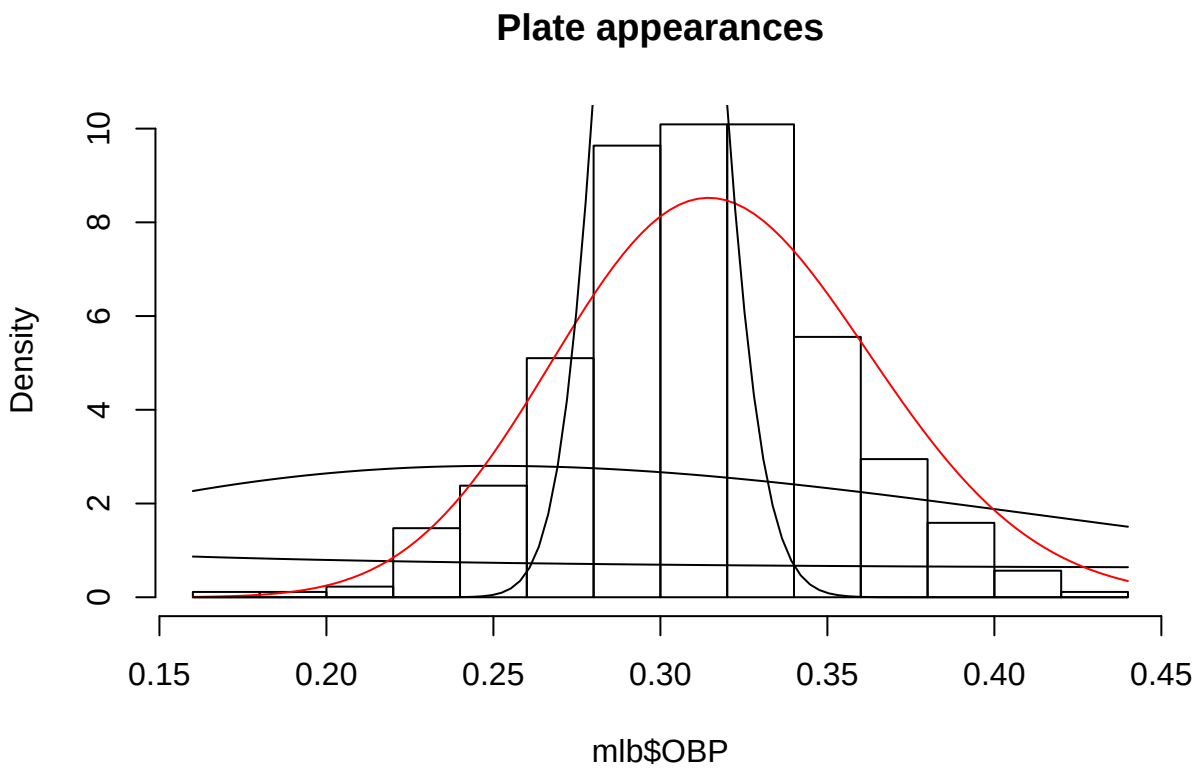
```r
abline(v=avg_OBP, col = "red")
```



3. Eyeball fit. Add a `curve()` to the plot using the density function `dbeta()`. Pick parameters α and β that matches the mean of the distribution but where their sum equals 1. Add three more `curve()`s to this plot where the sum of these parameters equals 10, 100 and 1000 respectively. Which of these is closest to the observed distribution?

```r
#The first step was just to place the histogram for On-base percentage which looks to be
#normally distirbuted. Then to add some curve lines. After playing with the numbers the line
#in red where alpha and beta sum to 100 seems to give the closest fit

hist(mlb$OBP, probability = TRUE, main = "Plate appearances")
curve(dbeta(x, .5 , .5), add = TRUE)
curve(dbeta(x, 3,7), add = TRUE)
curve(dbeta(x,31.5,67.5), add = TRUE, col = "red")
curve(dbeta(x,300,700), add = TRUE)
```



Part I

4. Method of moments fit. Find the calculation for the parameters from the mean and variance from [http://en.wikipedia.org/wiki/Beta_distribution] and solve for α and β. Create a new density histogram and add this `curve()` to the plot. How does it agree with the data?

```
#to solve for alpha and beta we need the variance equation which is (alpha*beta)/s0^2(s0+1)
#alpha = s0 * avg and beta = s0 - s0*avg
#we substitue into the top: numerator = (s0*avg)(s0-s0*avg) denom = s0*s0*(s0+1)
#we can then cancel out 2 factors of s0 and get variance = avg-avg^2/(s0+1)
#so s0 = ((avg-avg^2)/var)-1
#note: i'm sure there is a nice way to solve the equations with r, i just couldn't figure it out
#so i did it by hand on paper first
#the fitted line sits nicely with the data

avg = mean(mlb$OBP)
avg
```

```
## [1] 0.3119184
```

```
variance = var(mlb$OBP)
variance
```

```
## [1] 0.001500052
```

```
s0 = ((avg-avg^2)/variance)-1
s0
```

```
## [1] 142.0785
```

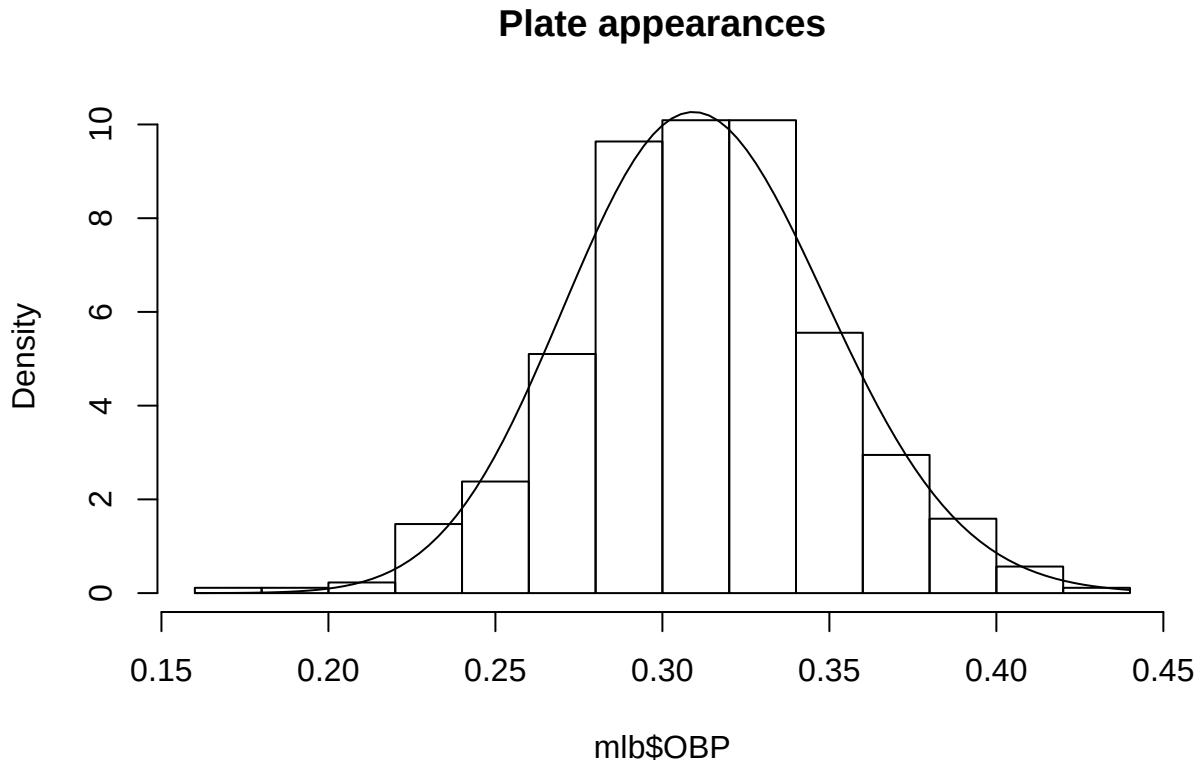
```
alpha = s0*avg
alpha
```

```
## [1] 44.3169
```

```
beta = s0 - (s0*avg)
beta
```

```
## [1] 97.76163
```

```
hist(mlb$OBP, probability = TRUE, main = "Plate appearances")
curve(dbeta(x, alpha , beta), add = TRUE)
```

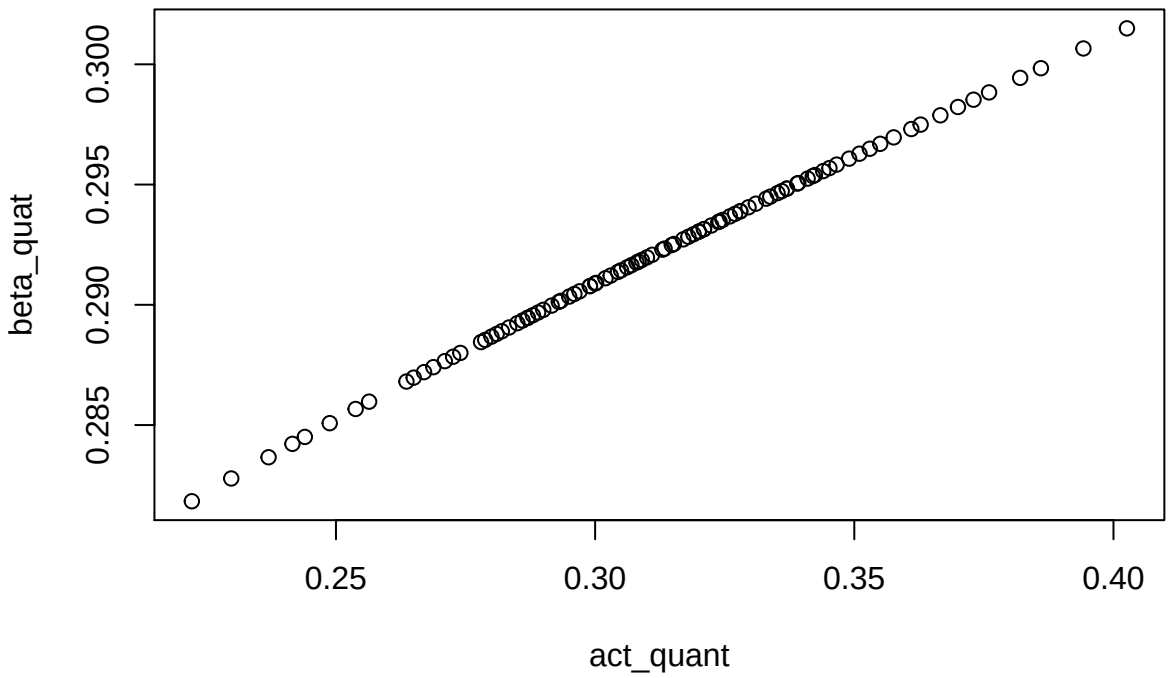


5. Calibration. For the previous part, find the 99 percentiles of the actual distribution using the `quantile()` function and plot them against the 99 percentiles of the beta distribution you just fit using `qbeta()`. How does the fit appear to you?

```
#first we put the 99 quantiles in a variable, then we get the beta quantiles and then plot
#with qqplot, the line appears straight meaning that they are highly correlated

act_quant = quantile(mlb$OBP, probs = seq(0.01 ,.99 , .01))
beta_quat = qbeta(act_quant, alpha, beta)

qqplot(act_quant, beta_quat)
```



6. Create a function for the log-likelihood of the distribution that calculates `-sum(dbeta(your.data.here, your.alpha, your.beta, log=TRUE))` and has one argument `p=c(your.alpha, your.beta)`. Use `nlm()` to find the minimum of the negative of the log-likelihood. Take the MOM fit for your starting position. How do these values compare?

```
log_like = function(input) {
  likelihood = sum(dbeta(mlb$OBP, input[1], input[2], log = TRUE)) *-1
  return(likelihood)

}

logg = log_like(c(alpha, beta))
test =nlm(log_like, c(alpha, beta))
```