STAT 206 Lab 4_Lihua_Xu

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?

```
mlb_obp <- read.csv("http://faculty.ucr.edu/~jflegal/206/mlb-obp.csv")
column_N <- dim(mlb_obp)[1]
column_N
```

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

```
#There are 441 players are included.

min_PA <- min(mlb_obp$PA)
min_PA
```

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

```
#The minimum number required to appear on this list is 103.

max_PA <- max(mlb_obp$PA)
position_row <- which(mlb_obp$PA==max_PA, arr.ind=TRUE)
max_PA_name <- mlb_obp$Name[position_row]
max_PA_name
```

```
## [1] Ian Kinsler
## 441 Levels: A.J. Ellis A.J. Pierzynski A.J. Pollock ... Zack Cozart
```

```
#Ian Kinsler had the most plate appearances

min(mlb_obp$OBP)
```

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

```
max(mlb_obp$OBP)
```

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

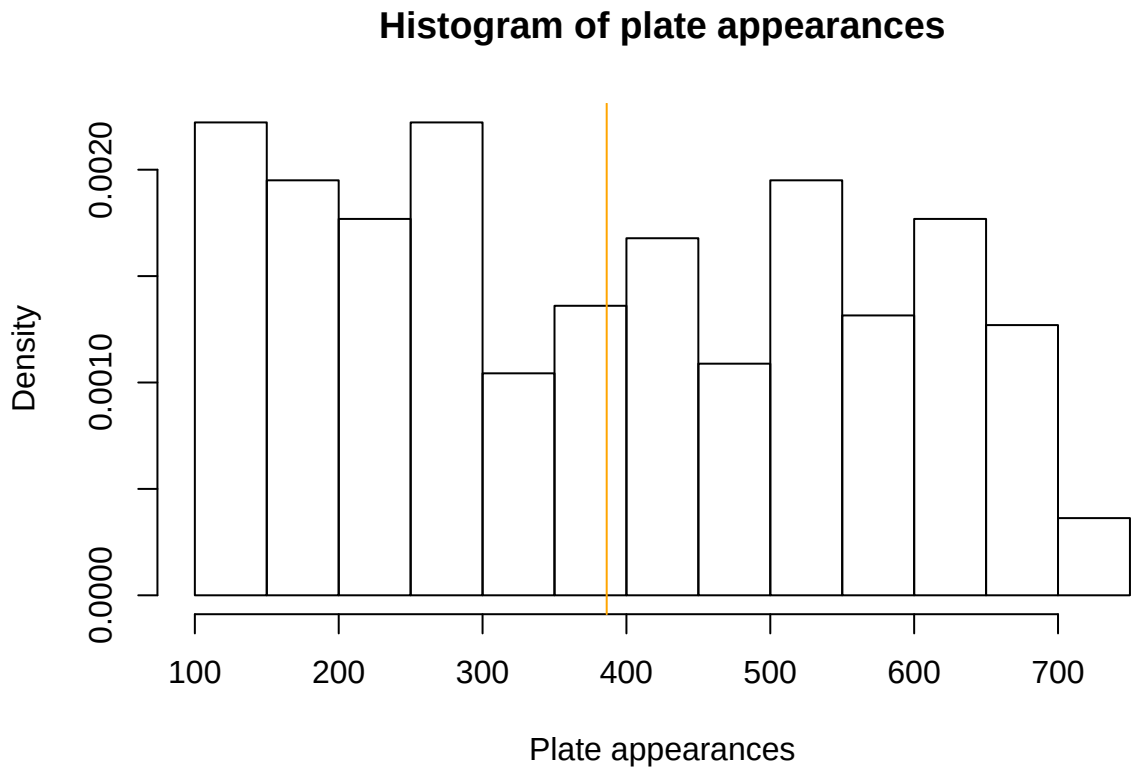
```
mean(mlb_obp$OBP)
```

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

```
#The minimum OBP is 0.168.
#The maxnimum OBP is 0.432.
#The mean of OBP is 0.3119184.
```

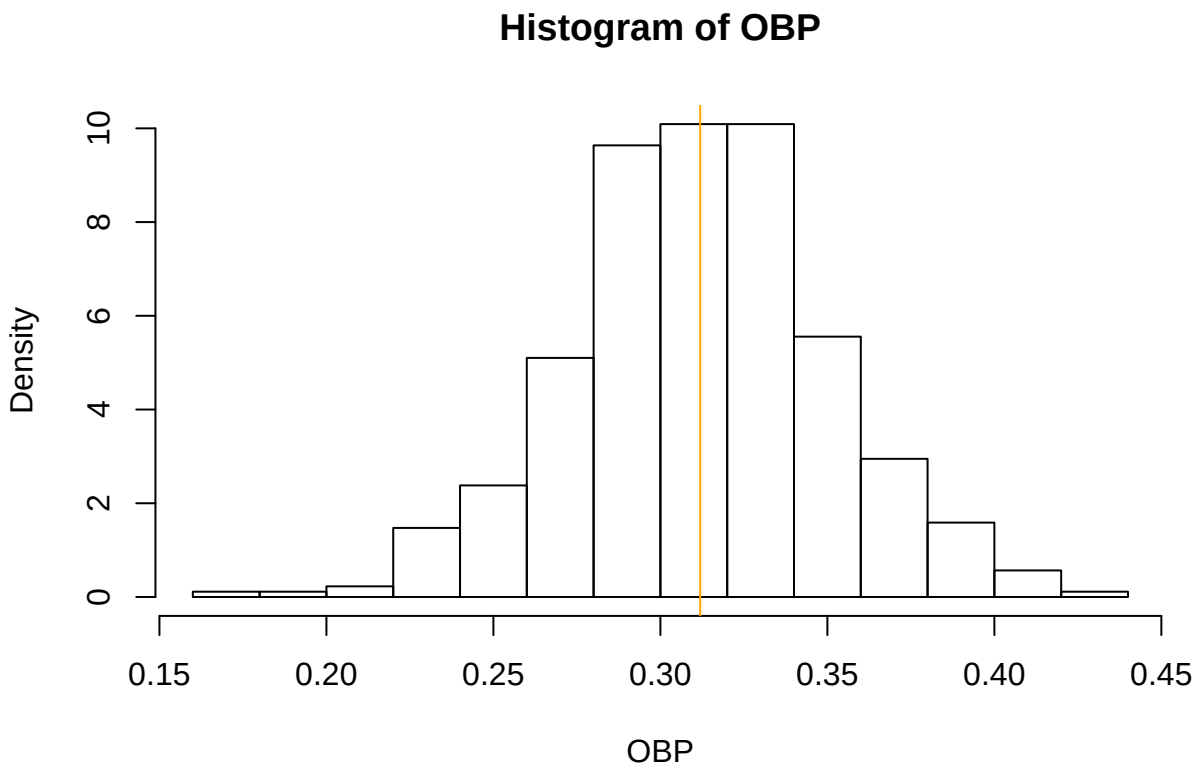
2. Plot the data as a histogram with the option `probability=TRUE`. Add a vertical line for the mean of the distribution. Does the mean coincide with the mode of the distribution?

```
#Plot the plate appearances
hist(mlb_obp$PA,probability=TRUE,xlab="Plate appearances",main="Histogram of plate appearances")
abline(v=mean(mlb_obp$PA),col="orange")
```




```
#For plate appearances, the mean doesn't coincide with the mode of the distribution .
```

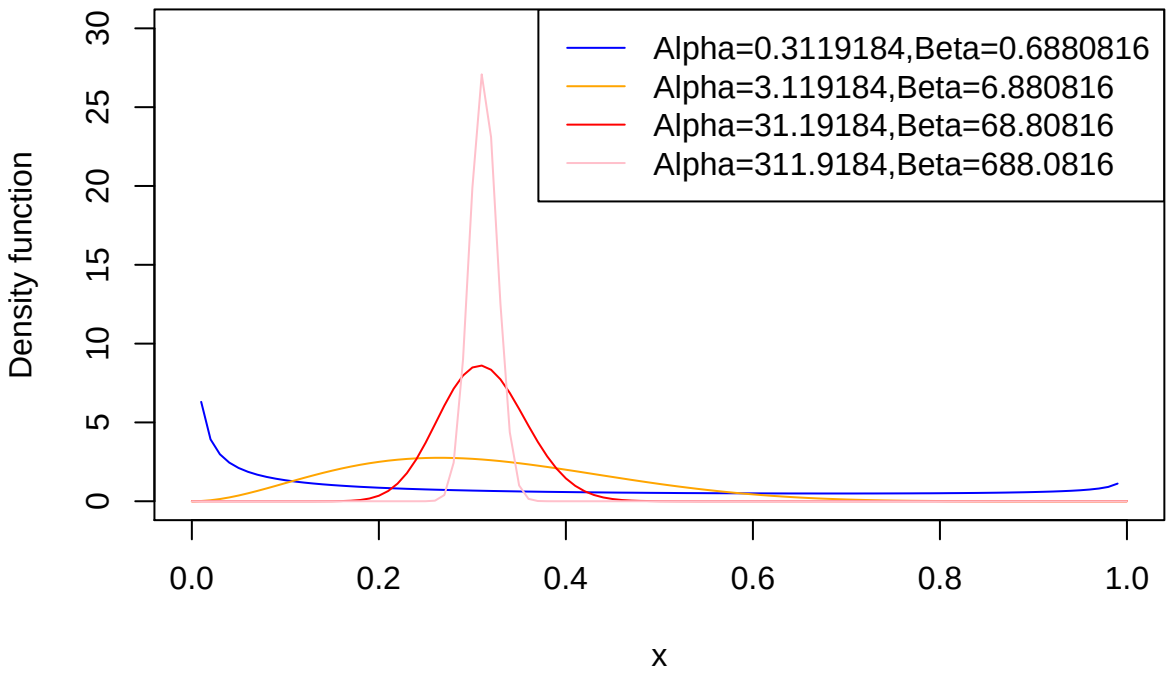
```r
#Plot the OBP
hist(mlb_obp$OBP,probability=TRUE,xlab="OBP",main="Histogram of OBP")
abline(v=mean(mlb_obp$OBP),col="orange")
```




```r
#For OBP, the mean coincide with the mode of the distribution.
```

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
##For OBP data:
#The sum of these parameters equals 1.
curve(dbeta(x,0.3119184,0.6880816),col="blue",xlab="x",ylab="Density function",ylim=c(0,30))
#The sum of these parameters equals 10.
curve(dbeta(x,3.119184,6.880816),col="orange",add=TRUE)
#The sum of these parameters equals 100.
curve(dbeta(x,31.19184,68.80816),col="red",add=TRUE)
#The sum of these parameters equals 1000.
curve(dbeta(x,311.9184,688.0816),col="pink",add=TRUE)
legend('topright',legend=c('Alpha=0.3119184,Beta=0.6880816','Alpha=3.119184,Beta=6.880816','Alpha=31.19
       col=c('blue','orange','red','pink'),lwd=1)
```




```
##When the sum of these parameters equals 100,
##the trend of the curve is closest to the observed distribution.
```

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?

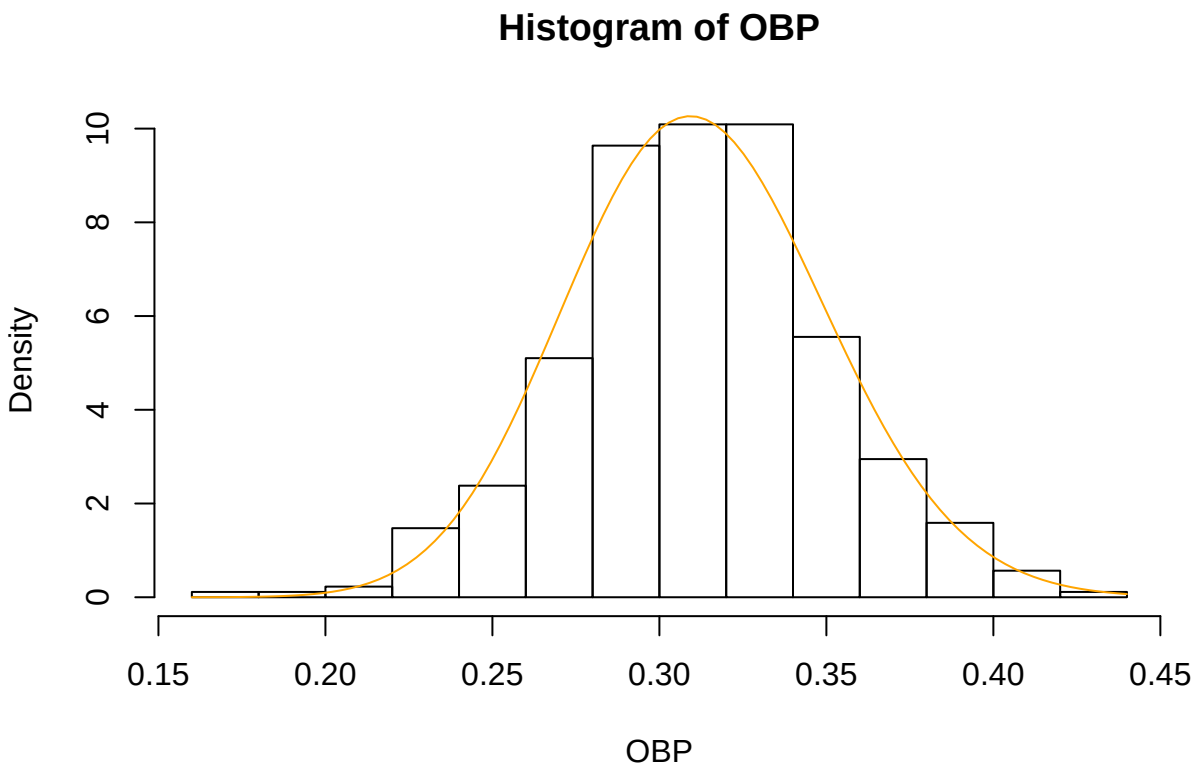
Solution: For the mean: $mean = \frac{\alpha}{\alpha+\beta}$ For the variance: $var = \frac{\alpha\beta}{(1+\alpha+\beta)(\alpha+\beta)}$

```
#The var of OBP
var(mlb_obp$OBP)
```

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

As we know, for the OBP: The mean is equal to 0.3119184. The var is equal to 0.001500052. So after doing math calculations for the above function: $\alpha = 44.32$ $\beta = 97.76$

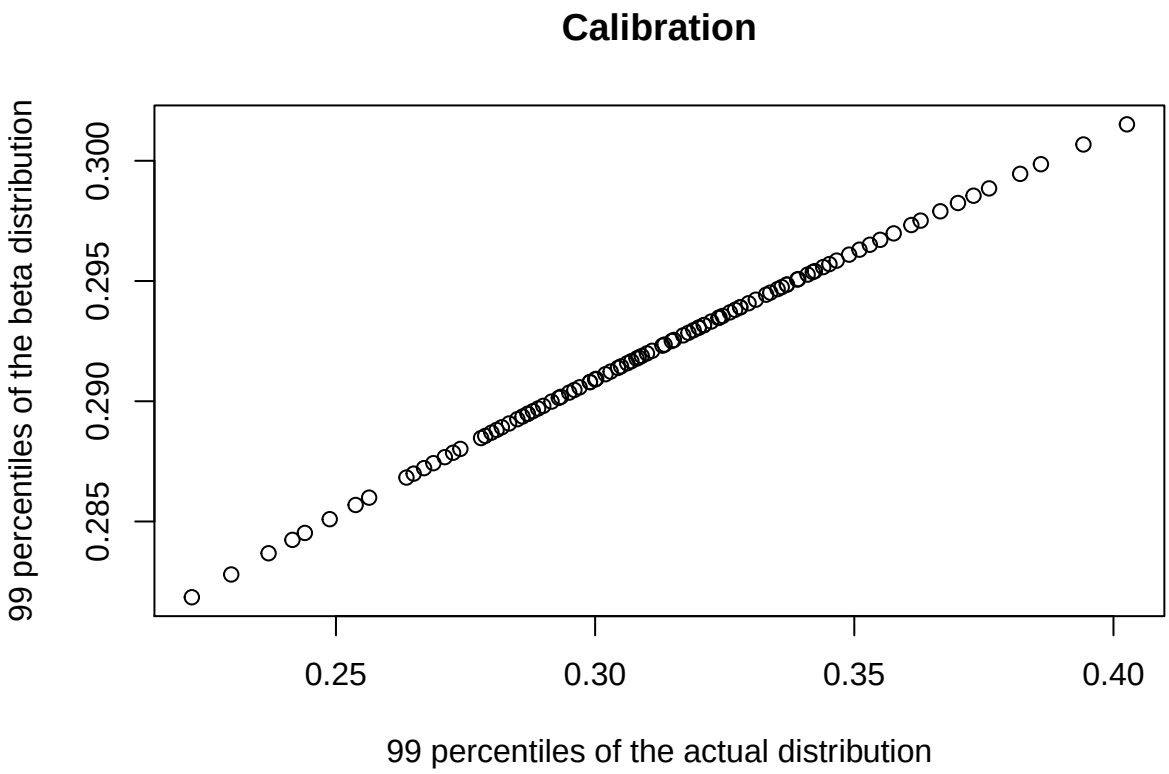
```
hist(mlb_obp$OBP,probability=TRUE,xlab="OBP",main="Histogram of OBP")
curve(dbeta(x,44.32,97.76),col="orange",add=TRUE)
```




```
#The curve matches the data very well.
```

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?

```
Prob_Quantile <- quantile(mlb_obp$OBP,probs = seq(0.01,0.99,0.01))
qbeta_fit <- qbeta(Prob_Quantile,44.32,97.76)
plot(Prob_Quantile,qbeta_fit,main="Calibration",xlab="99 percentiles of the actual distribution",
     ylab="99 percentiles of the beta distribution")
```




```
##The fit looks like a line, and "99 percentiles of the beta distribution" is in
##an linear increasing tendency when "99 percentiles of the actual distribution" is increasing.
```

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?

```
beta.loglikelihood <- function(p,vector=mlb_obp$OBP) -sum(dbeta(vector, p[1], p[2], log=TRUE))
nlm(beta.loglikelihood,c(44.32,97.76))
```

```
## $minimum
## [1] -805.8799
##
## $estimate
## [1] 43.73915 96.49892
##
## $gradient
## [1] 2.770748e-06 3.886602e-06
##
## $code
## [1] 1
##
## $iterations
## [1] 4
```

```
##Compared with the MOM fit alpha and beta values with the "nlm" minimum
##estimated alpha and beta values, they are almost the same.
```