STA3030F Assignment 3

Kiara Beilinsohn

4/22/2019

Introduction

The data for this project represents proportions of infant birds surviving to adulthood at various locations. The beta distribution is useful for modelling this, as it often models data which lies between 0 and 1. The pdf of the beta distribution is shown below, noting that β and α are greater than 0:

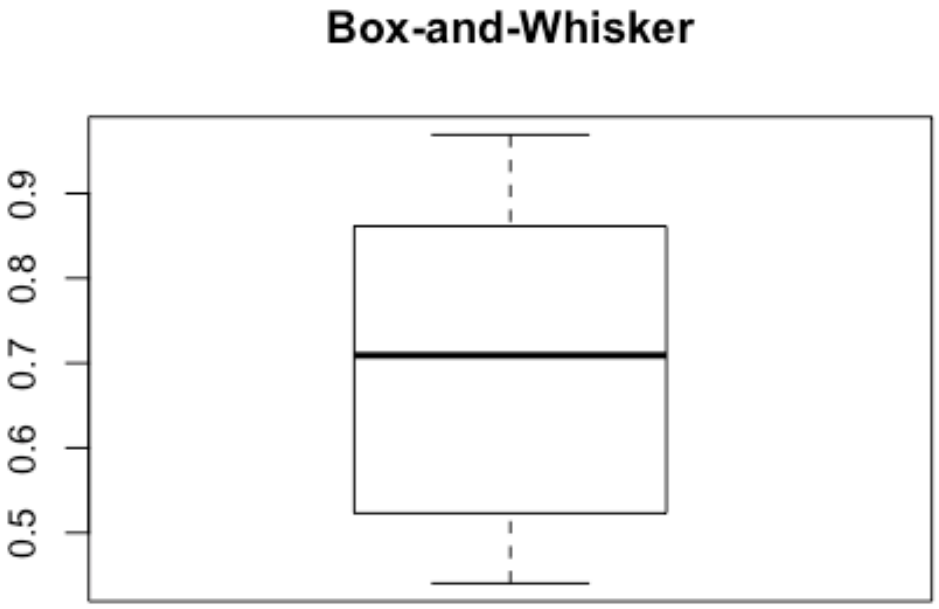
$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

For this project we will be estimating both the β and α parameters using R and excel methods. The aim of the assignment is to fit a beta distribution to the data and perform a goodness-of-fit test on the distribution.

Question 1:

data = c(0.658760658, 0.928971401
,0.842943952,0.508333565,0.657546099,0.635335118,
0.440458789,0.488153204,0.910936447,0.44179225,0.832158101,0.759574661,0.50641
8231,0.541490007,0.879809129,0.822131406,0.829105149,0.889004285,0.537892284,0.
969078343)

The initial data was analysed. Below the data is analysed through a box and wisker plot, the data seems to be relatively symetrically distrbuted with a small skewness to the left. It has a first quatile equal to 0.5305, a median equal to 0.7082, a mean equal to 0.7040 and a third quatile equal to 0.8522.The standard deviation is equal to 0.189. The 5 number summary can also be seen below:

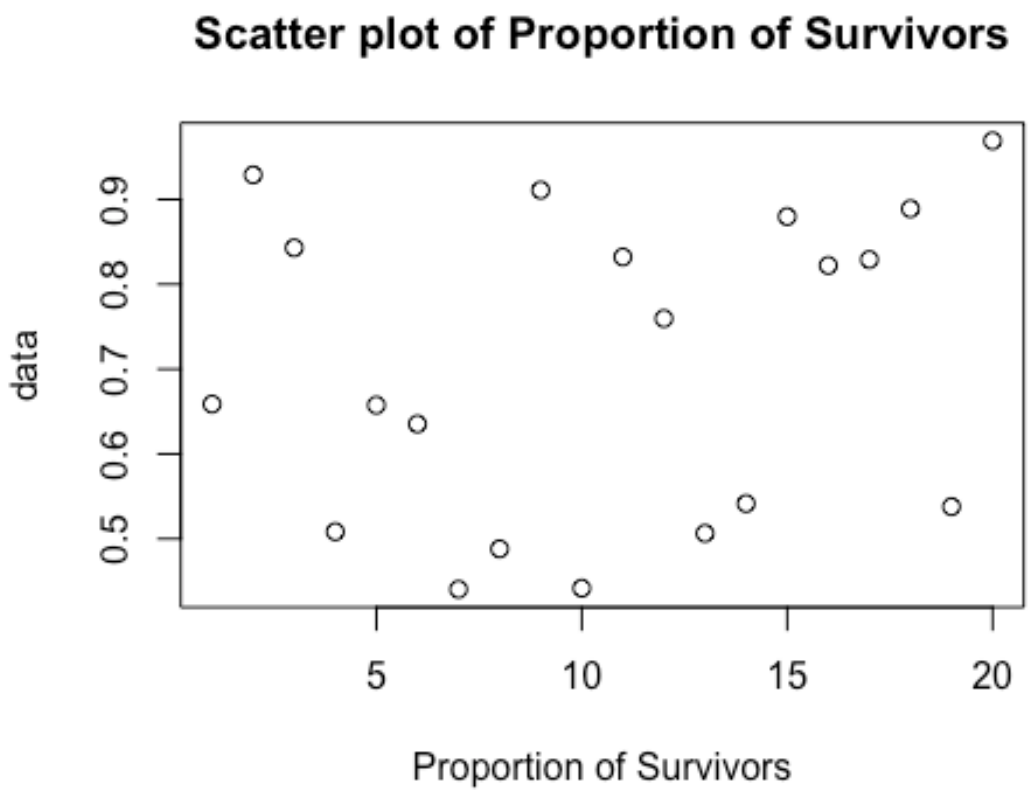


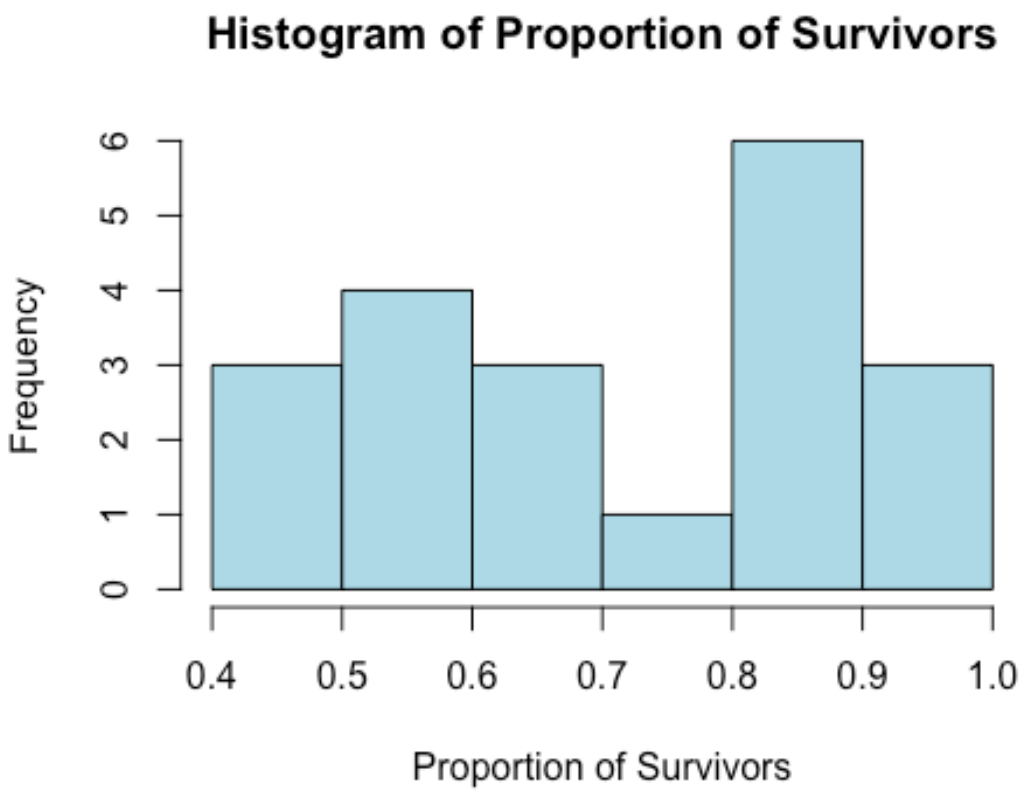

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4405  0.5305  0.7092  0.7040  0.8522  0.9691

## [1] "Standard deviation: "

## [1] 0.1809773
```

A plot of the distribution can be shown below and the data points seem to have no linear relationship or dependence on each other.The histogram indicates that the data does not appear to have a specific distribution, it is not normally distributed.


Scatter plot of Proportion of Survivors
data
0.5
0.6
0.7
0.8
0.9
5
10
15
20
Proportion of Survivors



Ouestion 2:

We will also be interested in being able to obtain its maximum likelihood estimators (MLEs) and showing that the values obtained, are indeed, good estimators.

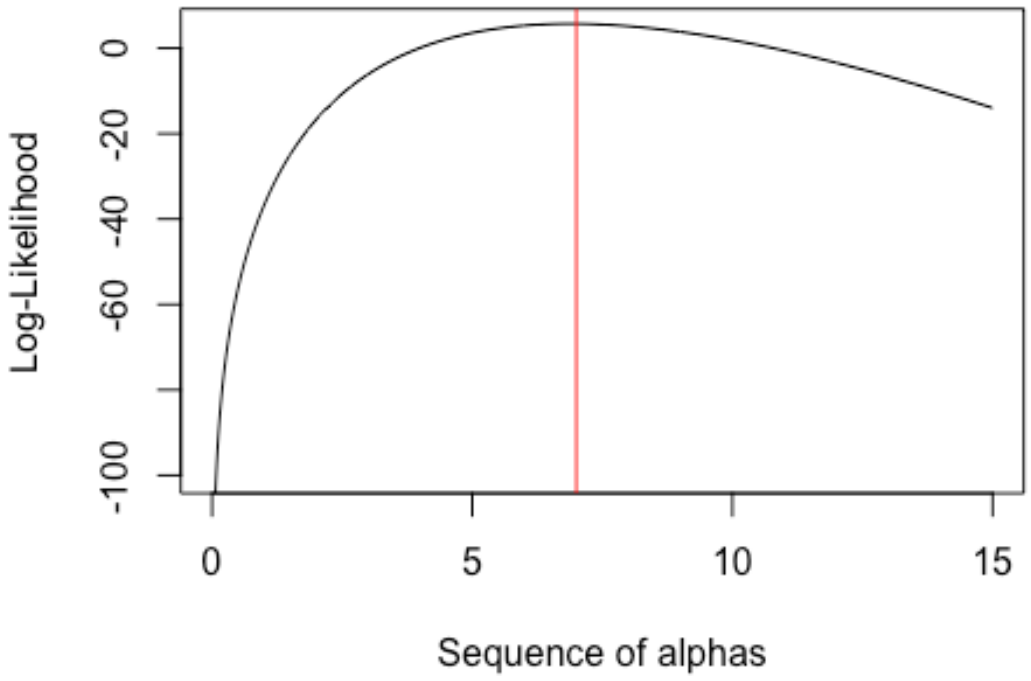
The first step is to derive the likelihood function of the distribution. This is done by taking the product of f(x) n times. Doing so provides the following equation:

$$L(\alpha,\beta|x) = \Pi_{i=1}^{n} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x_i^{\alpha-1}(1-x_i)^{\beta-1}$$

The next step is to derive the natural logarithm of the likelihood function as it is easier to work with. This is due to the fact that the logarithm is a stricly increasing funtion and therefore the logarithm of a function achieves its maximum at the same point(s) as the original funtion. The log-likelihood function obtained is:

$$LogL(\alpha,\beta|x) = nlog(\Gamma(\alpha+\beta)) - nlog(\Gamma(\alpha)) - nlog(\Gamma(\beta)) + (\alpha-1)\sum_{i=1}^{n} log(x_i) + (\beta-1)\sum_{i=1}^{n} log(1-x_i)$$

This function was derived in R and the value of β was set to 3, while the value of α was a generated sequence of numbers from 0.0001 to 15. Thereafter a graph was plotted for the value of α against the log likelihood with a fixed value of 3 for β. This can be seen plotted below:



Through analysing the above likelihood plot it can be seen that the value of α that maximises the function is 7 (the MLE), yeilding a likelihood of 5.7031822 (this is where the red line intercepts the function).

Question 3:

The following question was performed using excel solver to obtain maximum likelihood estimates of both α and β. The detailed method done using excel solver can be seen in the appendix. Below is the simplified steps followed:

1. The data (xi) was inputted into column B.
2. Parameter α was stored in B25, and β in. The total number of obsevations, n (20), was also stored, B28. The log -liklihood was stored in B31.
3. $ln(x_i)$ and $ln(x_i - 1)]$ needed to be calculated for each data point and the totals were calculated. In addition the values for α, β, x, (x-1) , $\Gamma(x)$ and $ln\Gamma(x)$ were stored.
4. I then used the relevant values along with the log-likelihood formula to obtain a value and stored it in B31.
5. Thereafter I used the Solver tool in Excel to maximise the estimates of α and β.

6. Since there is no definite greater than, ">", option in solver, I used ">=0.00001" as a constraint, since they should all be ">0".

7. Using the GRG Nonlinear solving method (due to the fact that the log-likelihood function is not linear), the estimates we obtain are as follows:

$$\hat{\alpha} = 4{,}036087093$$

$$\hat{\beta} = 1{,}667578959$$

$$log - likelihood = 8{,}14$$

Question 4:

The chi-squared goodness of fit test can be used to check the if of the distribution obtained from question 3 is a good fit for the data.In this case four bins are specifically used.

The hypotheses for this test are:

H_0: There is no difference between the expected data values and the observed data values, the distribution fits the data well.

H_1: There is a difference between the expected data values and the observed data values, the distribution does not fit the data well.

Below the test statistic was calculated by splitting the sorted data into four bins each 0.25 apart. Then using the qbeta() function the data was grouped into the various quatiles and the value of each quantile was calculated. these being 0.6094588, 0.7137633, 0.8044926, 1. Then the observed values were calculated by counting the number of values within each quantile. Next the expected value was calculated for each quantile that being 5 (20/4). Therafter for the test statistic was calculated using the following formula (this was across the bins):

$$\sum_{i=1}^{4} \frac{(Observed - Expected)^2}{Expected} \sim \chi_1^2$$

The test statistic yeiled is 8.

Next, the pvalue was calculated using a X^2 statistic with k-d-1 degrees of freedom, thus with four (k) bins and two estimated parameters (d). The degrees of freedom is 1. This yeilded a pvalue of 0.0046777. Since the pvalue is less then 0.005, we reject H_0 at the 5% significance level and conclude that the data does not fit the beta distribution well.

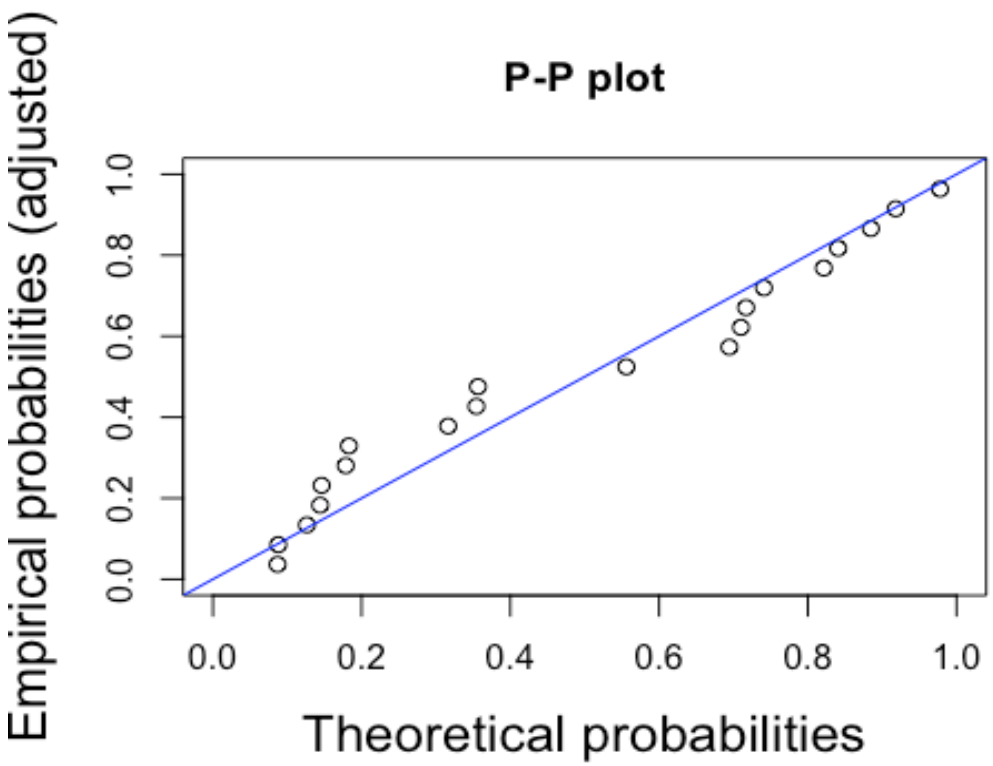
Question 5:

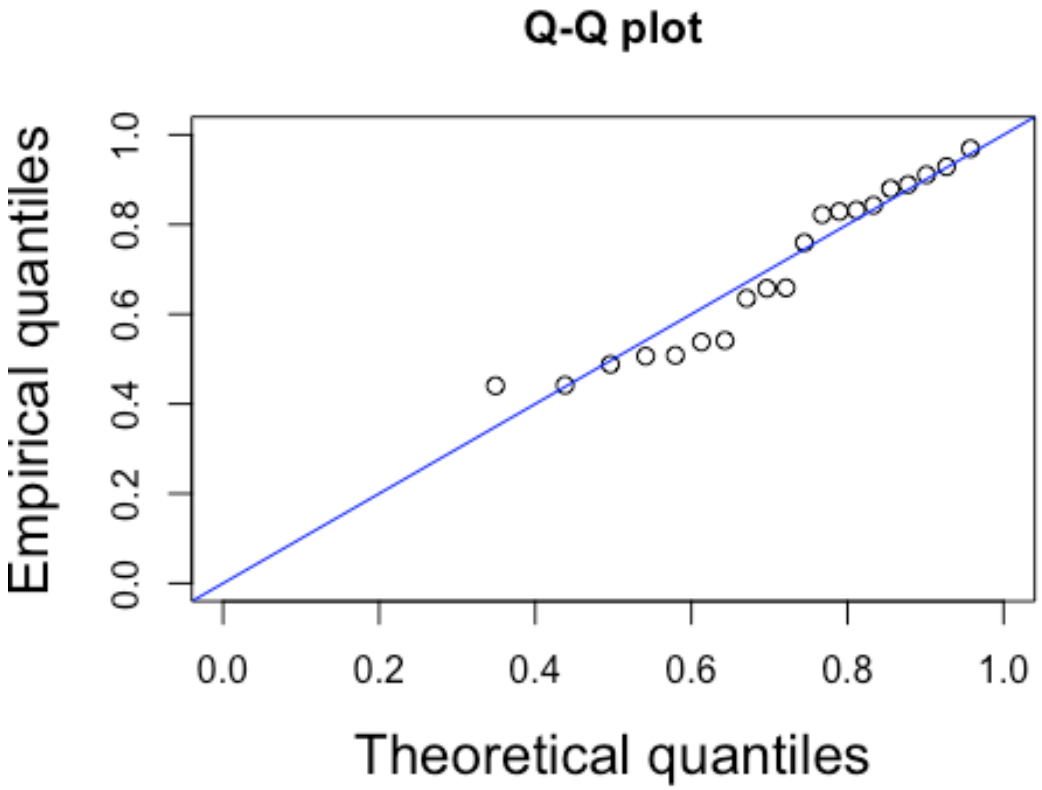
To determine whether the values obtained are suitable estimates for the parameters, we look at the Q-Q and P-P plots. In order to generate the P-P plot, the data was sorted and ranked from smallest to largest and then each data point was assigneed a rank from 1 to 20. The rank is notated by k.
k is then divided by n to get the theoretical probabilities. due to the fact that k cannot be equal to n, we adjust the probabilities with the following formula:

$$\frac{k - 0.25}{n + 0.5}$$

The cumulative probabilities using the pbeta() function was then calculated. The sample values were placed into the pbeta function with the estimates for α and β that were calculated in question 3. This generates the culmulative probabilities of observing these values.
The QQ-Plot is made by plotting the observed data against the theoretical quantiles. The theoretical quantiles are generated using the qbeta function in R with the same values as above for α and β, by inputting the adjusted probabilities into the function.

In order to see whether the values obtained are good estimates for the parameters, we look at the Q-Q Plot. This plots $x_{(k)}$ values directly against the predicted values, $F^{-1}(\frac{k-0.25}{n+0.5})$, where k is the different ranks.


P-P plot
Empirical probabilities (adjusted)
Theoretical probabilities
0.0
0.2
0.4
0.6
0.8
1.0


Q-Q plot
Empirical quantiles
Theoretical quantiles
0.0
0.2
0.4
0.6
0.8
1.0

A table of all the values used for calculating the PP and QQ plot is summarised below.

k	Adj k/n	xk	fxk	finv
1	0.0366	0.4405	0.0870	0.3491
2	0.0854	0.4418	0.0880	0.4382
3	0.1341	0.4882	0.1265	0.4962
4	0.1829	0.5064	0.1443	0.5415
5	0.2317	0.5083	0.1462	0.5795
6	0.2805	0.5379	0.1787	0.6129
7	0.3293	0.5415	0.1829	0.6430
8	0.3780	0.6353	0.3164	0.6708
9	0.4268	0.6575	0.3543	0.6967
10	0.4756	0.6588	0.3565	0.7213
11	0.5244	0.7596	0.5560	0.7448
12	0.5732	0.8221	0.6945	0.7675
13	0.6220	0.8291	0.7101	0.7897
14	0.6707	0.8322	0.7170	0.8115
15	0.7195	0.8429	0.7411	0.8333
16	0.7683	0.8798	0.8215	0.8552
17	0.8171	0.8890	0.8408	0.8777
18	0.8659	0.9109	0.8849	0.9013
19	0.9146	0.9290	0.9182	0.9270
20	0.9634	0.9691	0.9779	0.9576

Through analysing the above P-P plot and Q-Q plot it can be said that the model is a relatively bad fit, and can be improved since there is a lot of deviation from the straight line through the origin, a better model can be implemented. We can conclude that the values for the estimators are average and that distribution used to fit the data is valid in this context, but can be significantly improved.

Conclusion

Through the use of the maximum liklihood estimation procedure we are able to obtain the estimated parameters (α and β). We can confirm the fit of these parameters by viewing the P-P nad Q-Q plot, with the data provided, when using the estimations, does not closely approximate the straight line through the origin. Therefore showing that the distribution using the estimators is an average fit for the data. We can also make conclusions about the fit of the distribution via the chi-squared goodness of fit test, which in this case reveals that it is not a good fit.

Appendix:

Code:

```
# Loading in the data
data = c(0.658760658,   0.928971401
,0.842943952,0.508333565,0.657546099,0.635335118,
0.440458789,0.488153204,0.910936447,0.44179225,0.832158101,0.759574661,0.5064
18231,0.541490007,0.879809129,0.822131406,0.829105149,0.889004285,0.537892284
,0.969078343)

# Question 1: Give brief descriptive statistics or graphical displays of the
data
hist(data, col = "lightblue")
boxplot(data,main="Box-and-Whisker")
summary(data)
print('Standard deviation: ')
sd(data)
plot(data)

# Question 2
# alphas and beta, n
b <- 3
a <- seq(from = 0.00001, to = 15, length = 1000)
n <- length(data)

# loglikelihood function
x <- as.matrix(data)
loglike <- function(a)
{
  return(n*(log(gamma(a+b)) - log(gamma(a)) - log(gamma(b))) + (a-
1)*sum(log(x)) + (b-1)*sum(log(1-x)))
}

llofa <- loglike(a = a)

plot(a,llofa, type = "l", xlab = "Sequence of alphas", ylab = "Log-
Likelihood", ylim = c(-100, 5))

# estimate max
amax <- 7

abline(v = amax, col = "red")

# max loglikelihood
loglike(amax)
```

```r
# Question 4
# vector of probs
prob.bins <- seq(0.25, 1, by = 0.25)

# theretically quarter of data should lie between each quantile
# thereotical quantiles
ther.quants <- qbeta(prob.bins, shape1 = amax, shape2 = b)

# what is observed in data
# observed bins
sorted.data <- sort(data)
bin1 <- sorted.data[1:7]
bin2 <- sorted.data[8:10]
bin3 <- sorted.data[11]
bin4 <- sorted.data[12:20]

# for the test stat - (O-E)^2/E the sum of them
s1 <- (length(bin1)- 5)^2/5
s2 <- (length(bin2)- 5)^2/5
s3 <- (length(bin3)- 5)^2/5
s4 <- (length(bin4)- 5)^2/5

test.stat <- s1 + s2 + s3 + s4

# find p-value
pvalue <- (1 - pchisq(q = test.stat, df =1 )) # check with someone dof 17
print(pvalue)

# therefore since > 0.05 we fail to reject H0


# Question 5
# P-P  plot


#functions for P-P plot
seq <- seq(from = 0.5, to = 1, length= 100)

F.xk <- function(seq, a,b)
  {
   pbeta(seq,a,b)
  }

k.adj <- function(n){
  k = 1:n
  (k-0.25) / (n + 0.5)
}

Cum.probs <- F.xk(sorted.data, 4.036087093,1.667578959)
```

```r
Emp.probs <- k.adj(20)

plot(Emp.probs ~ Cum.probs, ylab = "Empirical probabilities (adjusted)", xlab
= "Theoretical probabilities", main = "P-P plot", cex.lab = 1.5, xlim =
c(0,1), ylim = c(0,1))
abline(0,1, col = "blue")


# Q-Q Plot
#function for Q-Q plot
F.inv <- function(seq,a , b){
  qbeta(seq,a,b)
}

Theor.quant <- F.inv(Emp.probs, 4.036087093,1.667578959)
plot(sorted.data ~ Theor.quant, ylab = "Empirical quantiles", xlab =
"Theoretical quantiles", main = "Q-Q plot", cex.lab = 1.5, xlim = c(0,1),
ylim = c(0,1))
abline(0, 1, col = "blue")
```

Excel solver workings:

	A	B	C	D	E	F
1		data (xi)	ln(xi)	ln(1-xi)		
2		0,66	-0,417395	-1,0751712		
3		0,93	-0,0736773	-2,6446727		
4		0,84	-0,1708548	-1,8511525		
5		0,51	-0,6766174	-0,7099548		
6		0,66	-0,4192404	-1,0716182		
7		0,64	-0,4536027	-1,0087765		
8		0,44	-0,8199384	-0,5806381		
9		0,49	-0,717126	-0,6697299		
10		0,91	-0,0932821	-2,4184051		
11		0,44	-0,8169155	-0,5830241		
12		0,83	-0,1837328	-1,7847328		
13		0,76	-0,2749967	-1,4253457		
14		0,51	-0,6803924	-0,7060667		
15		0,54	-0,6134307	-0,7797732		
16		0,88	-0,1280503	-2,1186742		
17		0,82	-0,195855	-1,7267102		
18		0,83	-0,1874083	-1,7667068		
19		0,89	-0,1176532	-2,1982637		
20		0,54	-0,620097	-0,7719573		
21		0,97	-0,0314098	-3,4762985		
22	Totals	14,08	(7,69)	(29,37)		
23						
24		x	x-1	Gamma(x)	ln(gamma(x))	
25	a	4,03608709	3,03608709	6,2793928	1,837273289	
26	b	1,66757896	0,66757896	0,90289531	-0,102148671	
27	a+b	5,70366605	4,70366605	72,9678274	4,290018624	
28	n	20				
29						
30						
31	log-likelihood	8,14				
32						