STAT 509 Homework 3b

Nathan Bickel

2022-10-11

Question

4. You are in charge of monitoring errors in the production of a certain part. Let X represent the difference in mm between the length of part and its desired length. You have collected data on these differences in 50 randomly selected parts. You wish to use this data to construct a model for the error in these parts so you can make predictions and informed decisions.

a. Fit the following models to the data: normal, exponential, gamma, and Weibull. You must include the maximum likelihood estimates of the model parameters for each model. Which model do you prefer? Do you believe the model fits the data well and thus can be used to model the population? You must include supporting evidence for your conclusions.

b. A part is defective if the error is at least 1 mm. Based on your model, what is the estimated probability that a randomly produced part is defective?

c. A shipment consists of a batch of five parts. The shipment will be rejected if it contains a defective part. Assuming each part is independent of the other parts, what is the estimated probability that a shipment is rejected?

d. Based on your answer in (c), do you believe the production quality of the parts is adequate?

e. You wish to use your sample data to make inferences about the average population level error in the length of a part. Assuming you use the sample mean as your point estimate, what is your point estimate and its estimated standard error?

f. Using your point estimate, its standard error, and the appropriate quantile from a t distribution, construct $\underline{\text{and}}$ interpret a 99% confidence interval for the true mean error in the length of a part.

g. Do the assumptions needed for the confidence interval seem justified? Provide supporting evidence.

h. Provide the appropriate R function which verifies your calculations in (f).

i. Based on your interval, are you confident that the mean error does not exceed 1?

Solution

a. We have our dataset of error in mm in a CSv. We import it and observe the following list:

```
filepath = "C:\\Users\\natha\\OneDrive\\Desktop\\data.csv"
data = read.csv(filepath)[,-1]
n=length(data)
library(MASS)
print(data)
```

```
##  [1] 0.36030287 0.57271827 0.21157539 1.87067467 0.99410107 0.06651121
##  [7] 0.32217295 0.71184883 0.60575354 0.57404848 0.62162387 0.73499768
## [13] 0.64769200 1.51412899 0.41617430 0.67435920 0.91147283 0.36391466
## [19] 0.16749452 1.69829395 1.31453272 0.26718627 0.57405491 0.69663555
## [25] 1.08126191 1.25569431 0.16834188 0.95706865 0.39048939 0.48708892
## [31] 0.92049122 0.13202693 0.68732270 0.50938196 0.62094825 0.61557079
## [37] 0.70061933 0.85632851 0.94564820 0.40730127 1.19022912 0.56232161
## [43] 1.37164883 0.83289765 1.74159892 1.16789639 1.05483672 0.70924422
## [49] 0.40549461 0.45256887
```

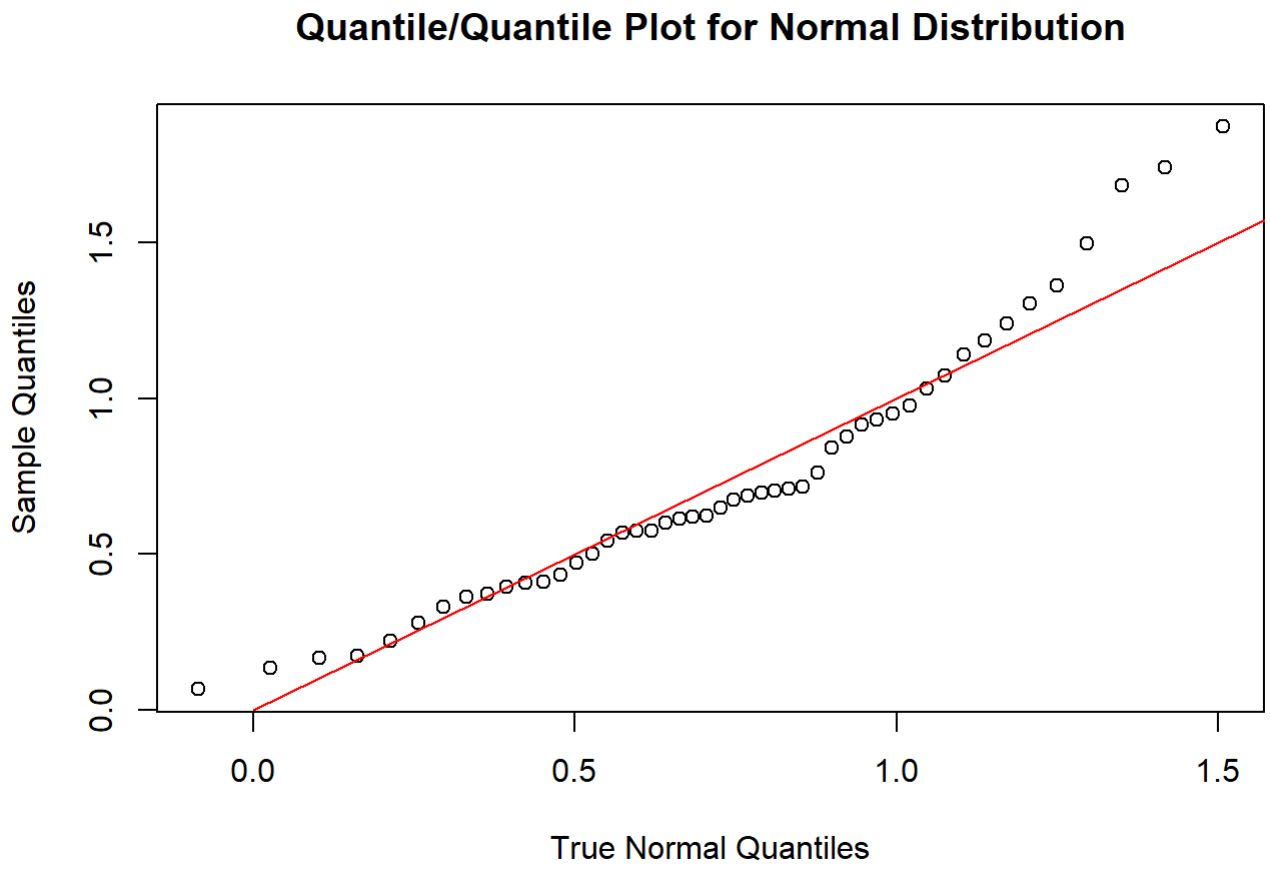
We now trying to fit the data to the four distributions listed.

Normal Distribution

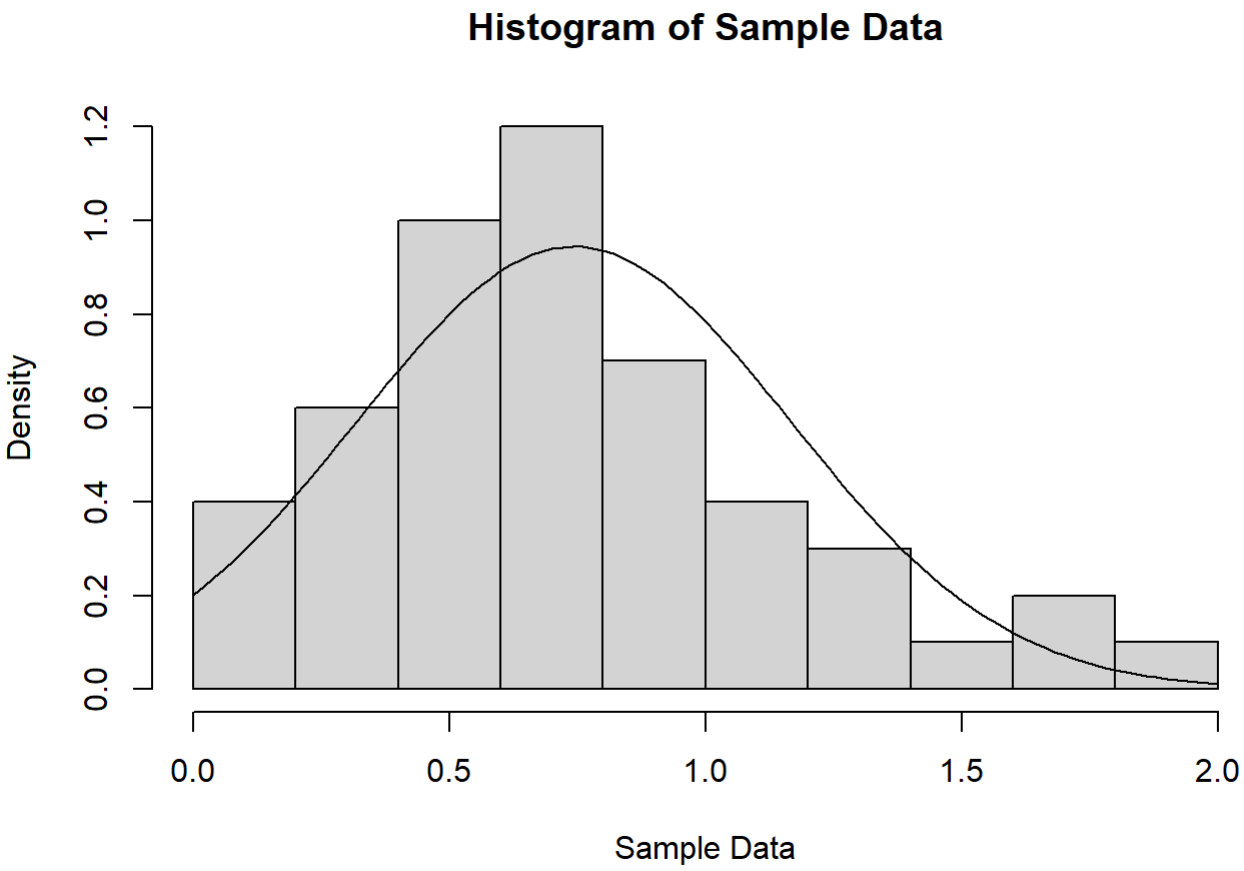
```
dist=fitdistr(data,"normal",lower=c(0,0))
print(dist)
```

```
##      mean         sd
##   0.7423318   0.4223910
##  (0.0597351) (0.0422391)
```

```
qqplot(qnorm(seq(.025,.975,by=1/n),dist$estimate[1],dist$estimate[2]),data,xlab="True Normal Qua
ntiles",ylab="Sample Quantiles",main="Quantile/Quantile Plot for Normal Distribution")
lines(c(0,100),c(0,100),col="red")
```




```
hist(data,freq=F,xlab="Sample Data",main="Histogram of Sample Data")
y=function(x){dnorm(x,dist$estimate[1],dist$estimate[2])}
curve(y,add=T)
```



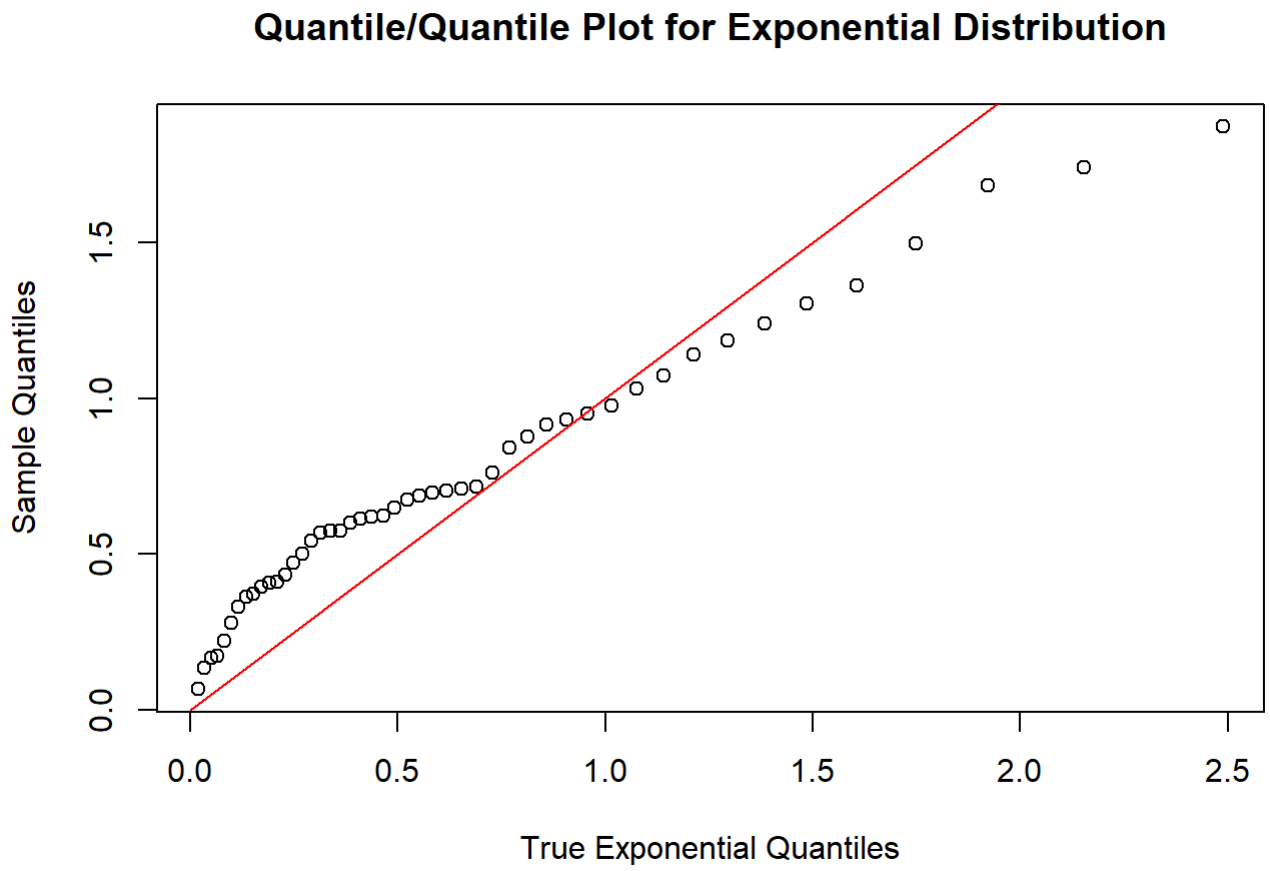
This is not an appropriate model to use. The points are consistently above the line at the beginning, then switch to below, and then switch back to above. Additionally, the model predicts that a significant proportion of the data should be negative, which isn't possible by our definition.

Exponential Distribution

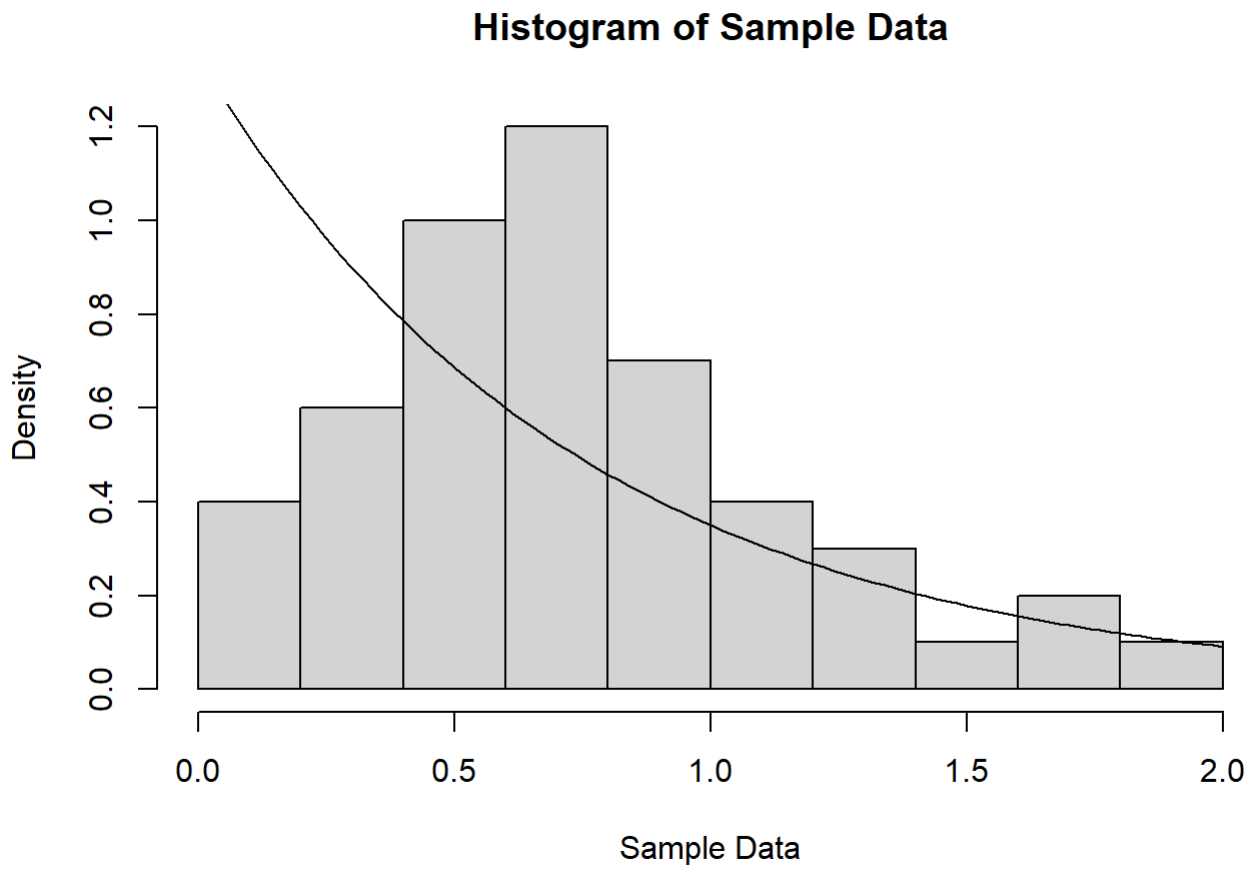
```
dist=fitdistr(data,"exponential",lower=c(0,0))
print(dist)
```

```
##      rate
##   1.3471065
##  (0.1905096)
```

```
qqplot(qexp(seq(.025,.975,by=1/n),dist$estimate[1]),data,xlab="True Exponential Quantiles",ylab=
"Sample Quantiles",main="Quantile/Quantile Plot for Exponential Distribution")
lines(c(0,100),c(0,100),col="red")
```




```
hist(data,freq=F,xlab="Sample Data",main="Histogram of Sample Data")
y=function(x){dexp(x,dist$estimate[1])}
curve(y,add=T)
```



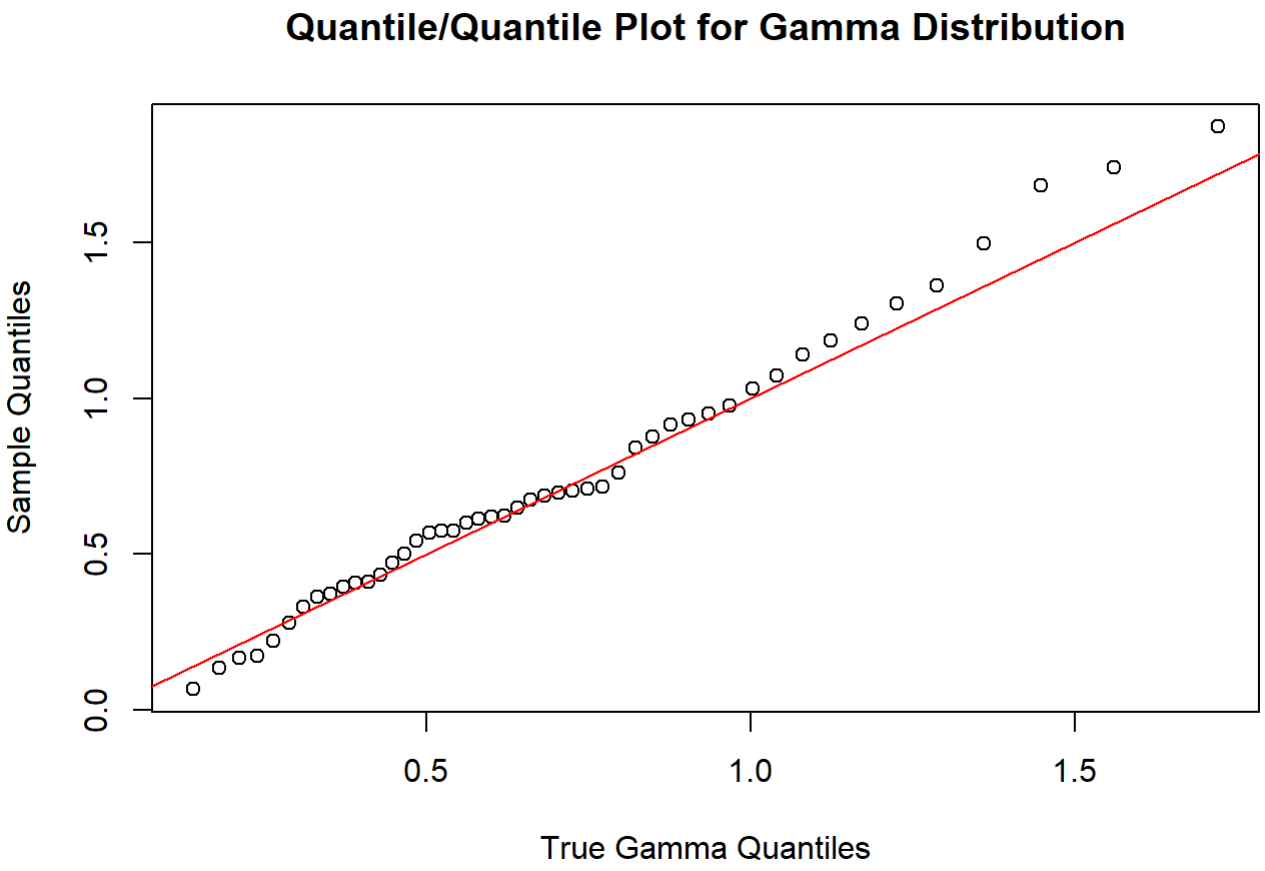
This clearly fits the data terribly. The peak is not at the low end of the data but in the middle, which is not consistent with an exponential distribution.

Gamma Distribution

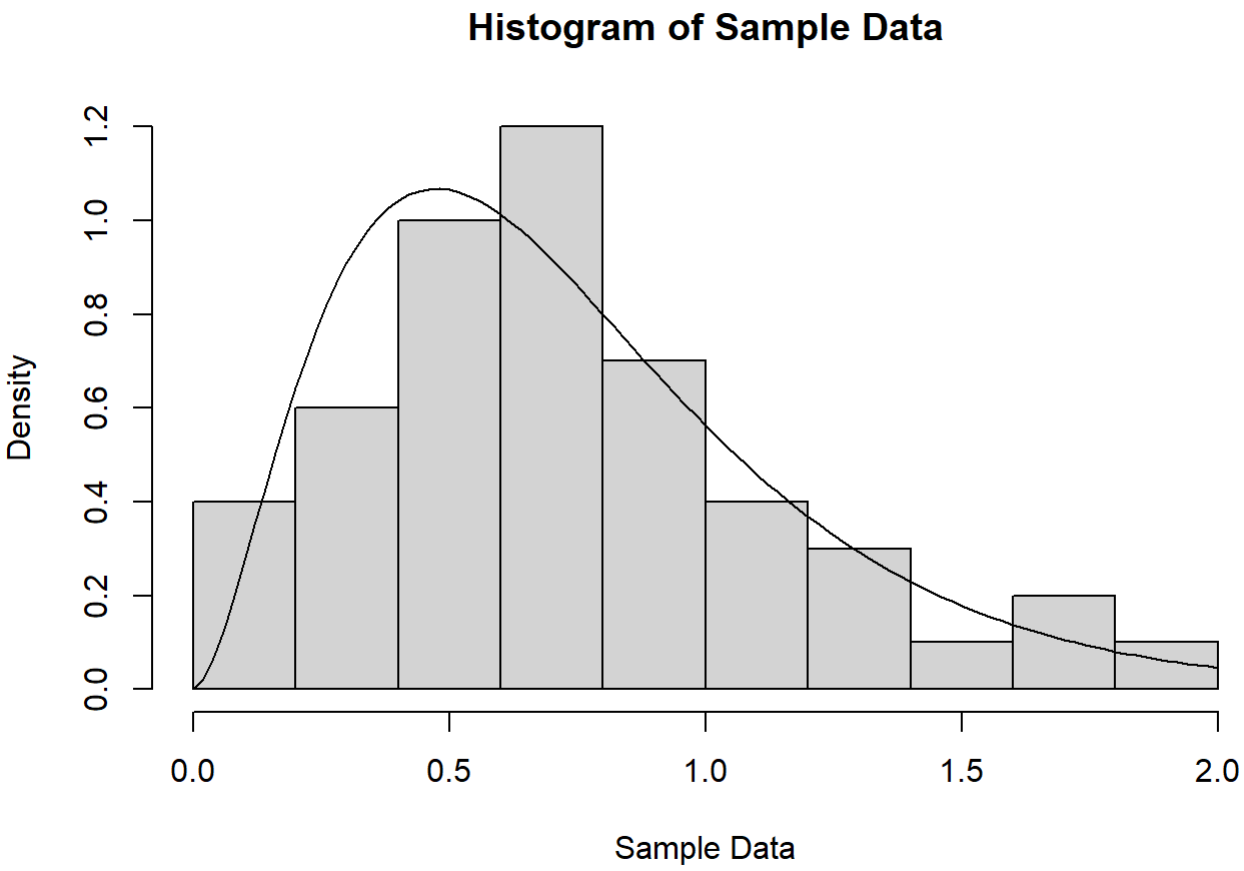
```
dist=fitdistr(data,"gamma",lower=c(0,0))
print(dist)
```

```
##      shape       rate
##   2.7713953   3.7333663
##  (0.5243536) (0.7742925)
```

```
qqplot(qgamma(seq(.025,.975,by=1/n),dist$estimate[1],dist$estimate[2]),data,xlab="True Gamma Quantiles",ylab="Sample Quantiles",main="Quantile/Quantile Plot for Gamma Distribution")
lines(c(0,100),c(0,100),col="red")
```




```
hist(data,freq=F,xlab="Sample Data",main="Histogram of Sample Data")
y=function(x){dgamma(x,dist$estimate[1],dist$estimate[2])}
curve(y,add=T)
```



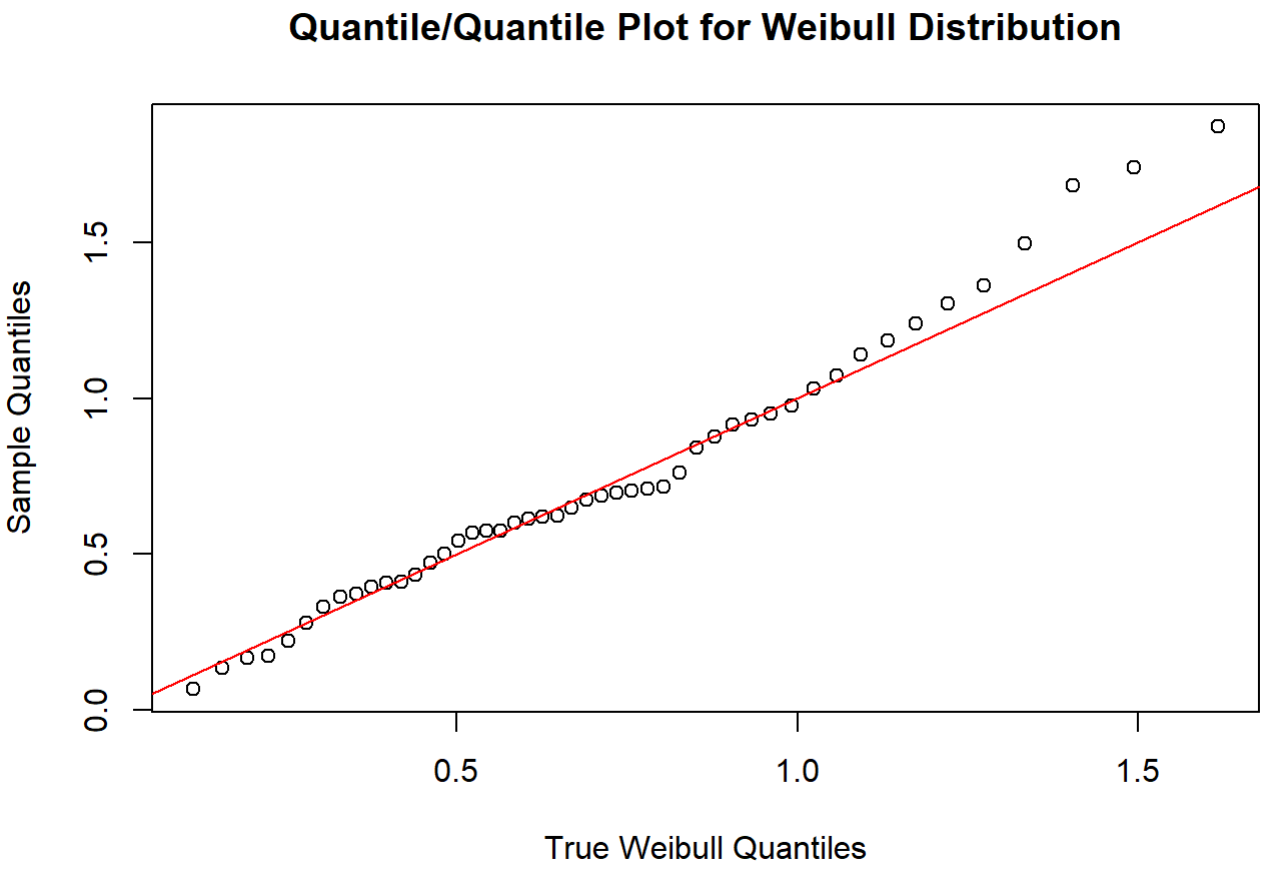
This is a better fit than we've seen so far. The points largely cluster fairly closely around the line of best fit. However, they seem to deviate more and more from the model as the values get largely, and in a way that suggests it is deviating systematically.

Weibull Distribution

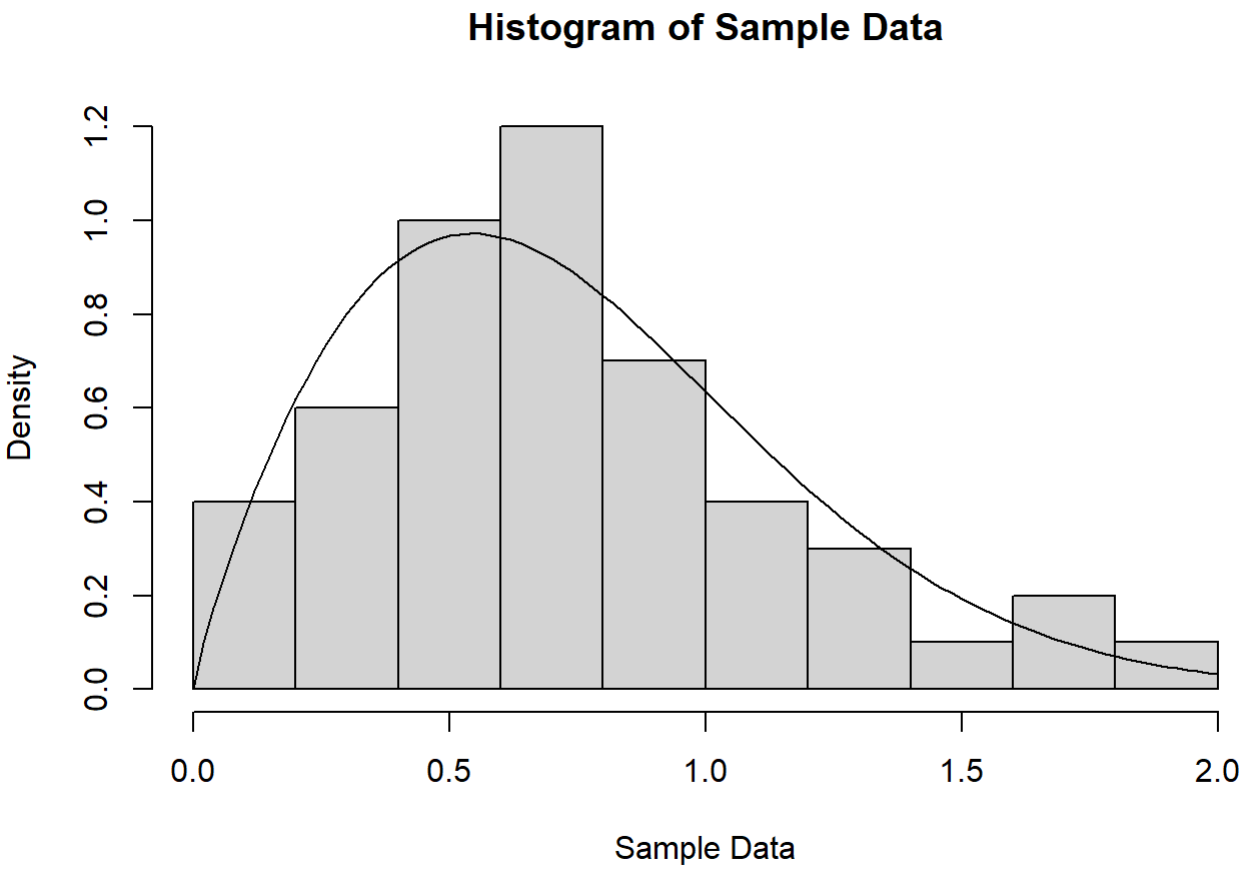
```
wdist=fitdistr(data,"weibull",lower=c(0,0))
print(wdist)
```

```
##      shape        scale
##   1.83290223   0.83632565
##  (0.20053280) (0.06802313)
```

```
qqplot(qweibull(seq(.025,.975,by=1/n),wdist$estimate[1],wdist$estimate[2]),data,xlab="True Weibu
ll Quantiles",ylab="Sample Quantiles",main="Quantile/Quantile Plot for Weibull Distribution")
lines(c(0,100),c(0,100),col="red")
```




```
hist(data,freq=F,xlab="Sample Data",main="Histogram of Sample Data")
y=function(x){dweibull(x,wdist$estimate[1],wdist$estimate[2])}
curve(y,add=T)
```



This looks quite similar to the gamma distribution. However, it seems to stay consistent with the line of best fit longer than the gamma distribution, so it may be a better description of the data (although neither seem totally ideal for the high end of the spectrum).

Overall, the Weibull distribution seems most preferable here. It seems to fit most of the data well as the points are mostly clustered around the line of best fit, but there are somewhat more large observations than the Weibull distribution predicts. It seems mostly appropriate to use this distribution, but to take its predictions with a grain of salt.

b. Our model is a Weibull distribution with shape ≈ 1.833 and scale ≈ 0.836. We are interested in the probability a part's error is greater than 1mm, which we can find with:

```
p=pweibull(1,wdist$estimate[1],wdist$estimate[2],lower.tail=F)
print(p)
```

```
## [1] 0.2496626
```

c. We can model this as a binomial distribution with 5 Bernoulli trials and probability of success $p \approx 1 - 0.250 = 0.750$. The probability of 5 successes can be found with:

```
dbinom(5,5,1-p)
```

```
## [1] 0.2378389
```

d. Based on the fact that less than 1 in 4 shipments of a mere 5 parts will have no defective parts, the production quality seems vastly lacking.

e. Our point estimate for the mean is simply the sample mean, and our standard error is the sample standard deviation scaled by the square root of the number of observations:

```
avg=mean(data)
ste=sd(data)/sqrt(n)
paste("Point Estimate for Mean:", round(avg,3))
```

```
## [1] "Point Estimate for Mean: 0.742"
```

```
paste("Standard Error:", round(ste,4))
```

```
## [1] "Standard Error: 0.0603"
```

f. We construct a confidence interval from the t distribution with 49 degrees of freedom with

$$\left(\overline{X} - t_{n-1,\alpha/2}\frac{s_x}{\sqrt{n}}, \overline{X} + t_{n-1,\alpha/2}\frac{s_x}{\sqrt{n}}\right).$$

Substituting our values, we obtain:

```
lower=avg-qt(.995,n-1)*ste
upper=avg+qt(.995,n-1)*ste
paste("(",round(lower,3),",",round(upper,3),")",sep="")
```

```
## [1] "(0.581,0.904)"
```

So we are 99% confident that the true mean error of a part is between 0.581 and 0.904 mm.

g. The data are from a random sample, and we have from the central limit theorem that the sampling distribution is approximately normal, so it seems we can true this procedure.

h. We use the `t.test` function on data:

```
t.test(data,conf.level=0.99)
```

```
##
##  One Sample t-test
##
## data:  data
## t = 12.302, df = 49, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 99 percent confidence interval:
##  0.5806193 0.9040443
## sample estimates:
## mean of x
## 0.7423318
```

This matches our confidence interval from (f).

i. We can conclude that the mean error does not exceed 1, because 1 is greater than the upper-bound of our 99% confidence interval.