Mid_Project

Qi Huang

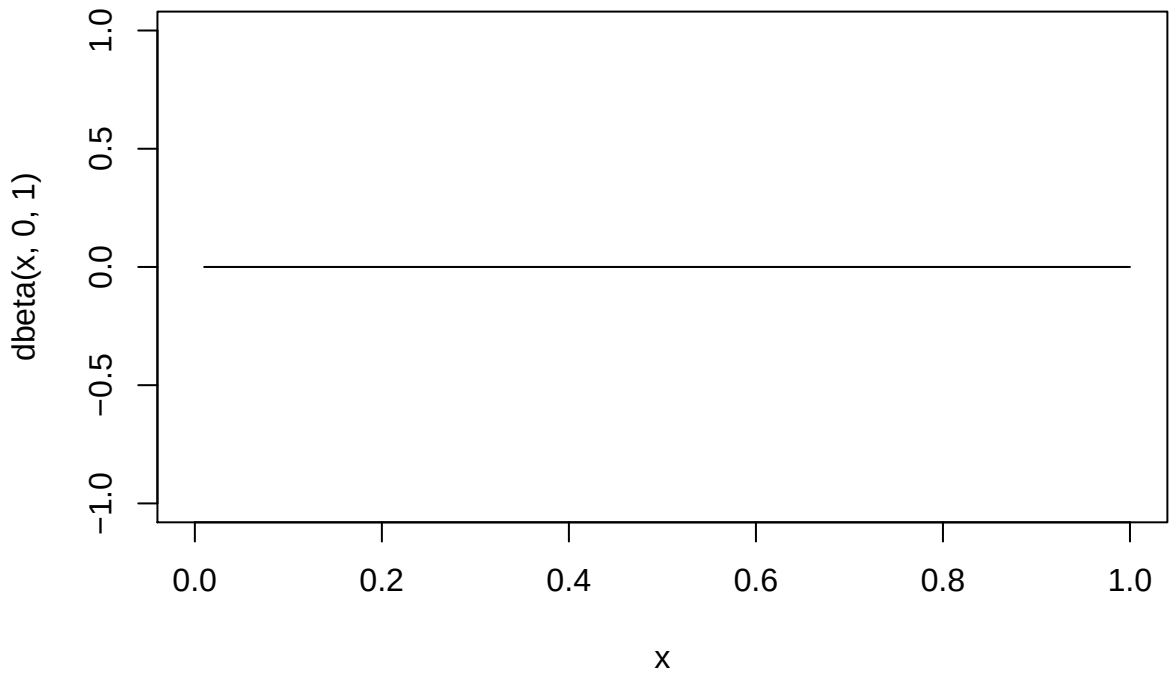
3/6/2020

Project1: Hand sanitizer 1).Prior distribution:

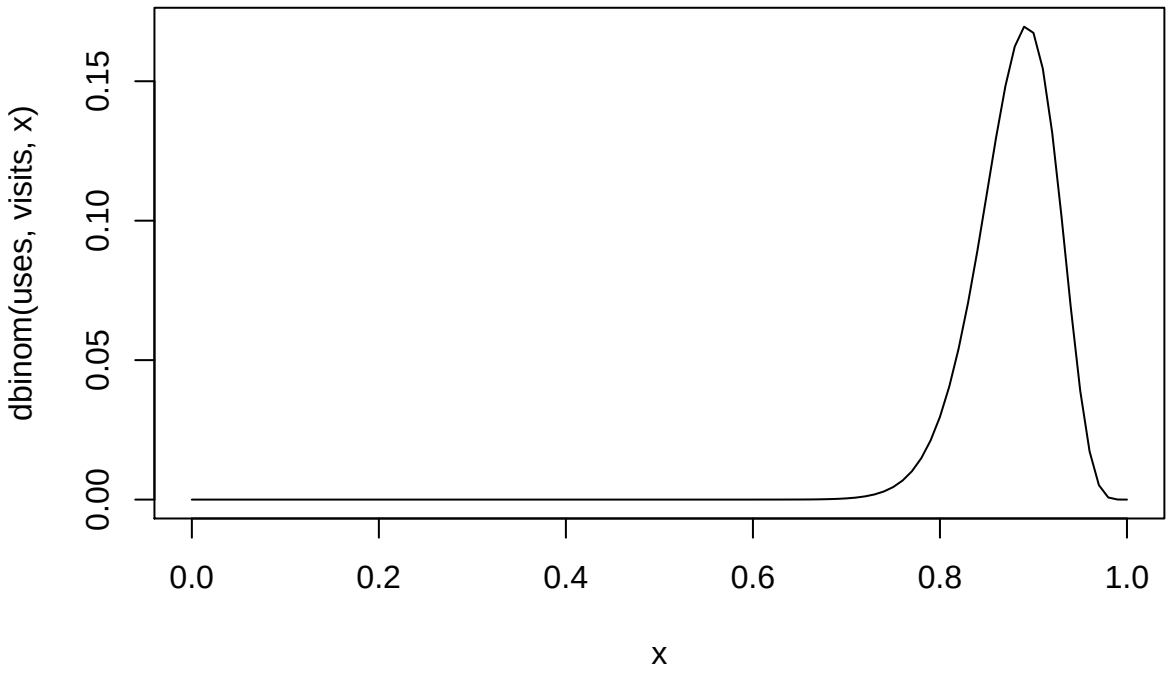
$$p\ U(0,1)$$

Calculating the Likelihood Function for a Proportion

```
n=100
curve(dbeta(x,0,1)) # plot the prior
```



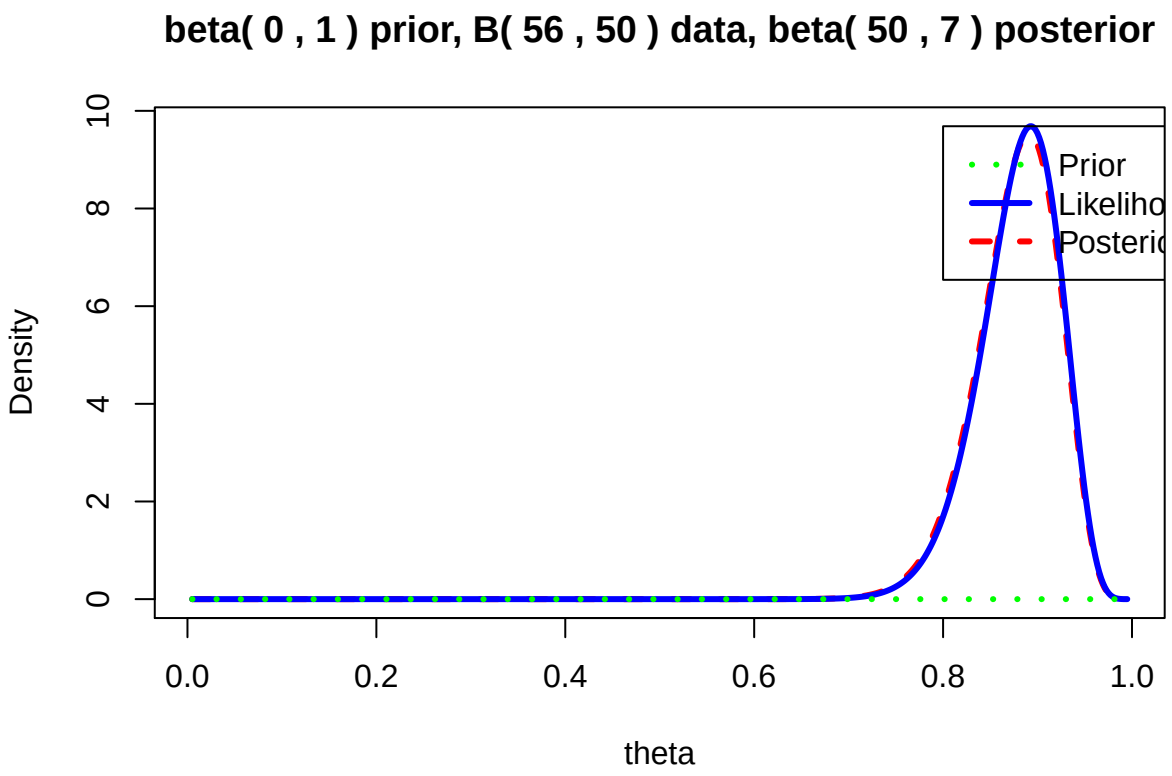

```
#Calculating the Likelihood Function for a Proportion
calcLikelihood <- function(uses,visits){
  curve(dbinom(uses,visits,x))
}
calcLikelihood(50,56)
```



Calculating the posterior Distribution for a Proportion

```
calcPosteriorForProportion <- function(successes, total, a, b)
  {
     # Adapted from triplot() in the LearnBayes package
     # Plot the prior, likelihood and posterior:
     likelihood_a = successes + 1; likelihood_b = total - successes + 1
     posterior_a = a + successes;  posterior_b = b + total - successes
     theta = seq(0.005, 0.995, length = 500)
     prior = dbeta(theta, a, b)
     likelihood = dbeta(theta, likelihood_a, likelihood_b)
     posterior  = dbeta(theta, posterior_a, posterior_b)
     m = max(c(prior, likelihood, posterior))
     plot(theta, posterior, type = "l", ylab = "Density", lty = 2, lwd = 3,
          main = paste("beta(", a, ",", b, ") prior, B(", total, ",", successes, ") data,",
          "beta(", posterior_a, ",", posterior_b, ") posterior"), ylim = c(0, m), col = "red")
     lines(theta, likelihood, lty = 1, lwd = 3, col = "blue")
     lines(theta, prior, lty = 3, lwd = 3, col = "green")
     legend(x=0.8,y=m, c("Prior", "Likelihood", "Posterior"), lty = c(3, 1, 2),
          lwd = c(3, 3, 3), col = c("green", "blue", "red"))
     # Print out summary statistics for the prior, likelihood and posterior:
     calcBetaMode <- function(aa, bb) { BetaMode <- (aa - 1)/(aa + bb - 2); return(BetaMode); }
     calcBetaMean <- function(aa, bb) { BetaMean <- (aa)/(aa + bb); return(BetaMean); }
     calcBetaSd   <- function(aa, bb) { BetaSd <- sqrt((aa * bb)/(((aa + bb)^2) * (aa + bb + 1))); retu
     prior_mode      <- calcBetaMode(a, b)
     likelihood_mode <- calcBetaMode(likelihood_a, likelihood_b)
     posterior_mode  <- calcBetaMode(posterior_a, posterior_b)
     prior_mean      <- calcBetaMean(a, b)
     likelihood_mean <- calcBetaMean(likelihood_a, likelihood_b)
     posterior_mean  <- calcBetaMean(posterior_a, posterior_b)
     prior_sd        <- calcBetaSd(a, b)
     likelihood_sd   <- calcBetaSd(likelihood_a, likelihood_b)
     posterior_sd    <- calcBetaSd(posterior_a, posterior_b)
```

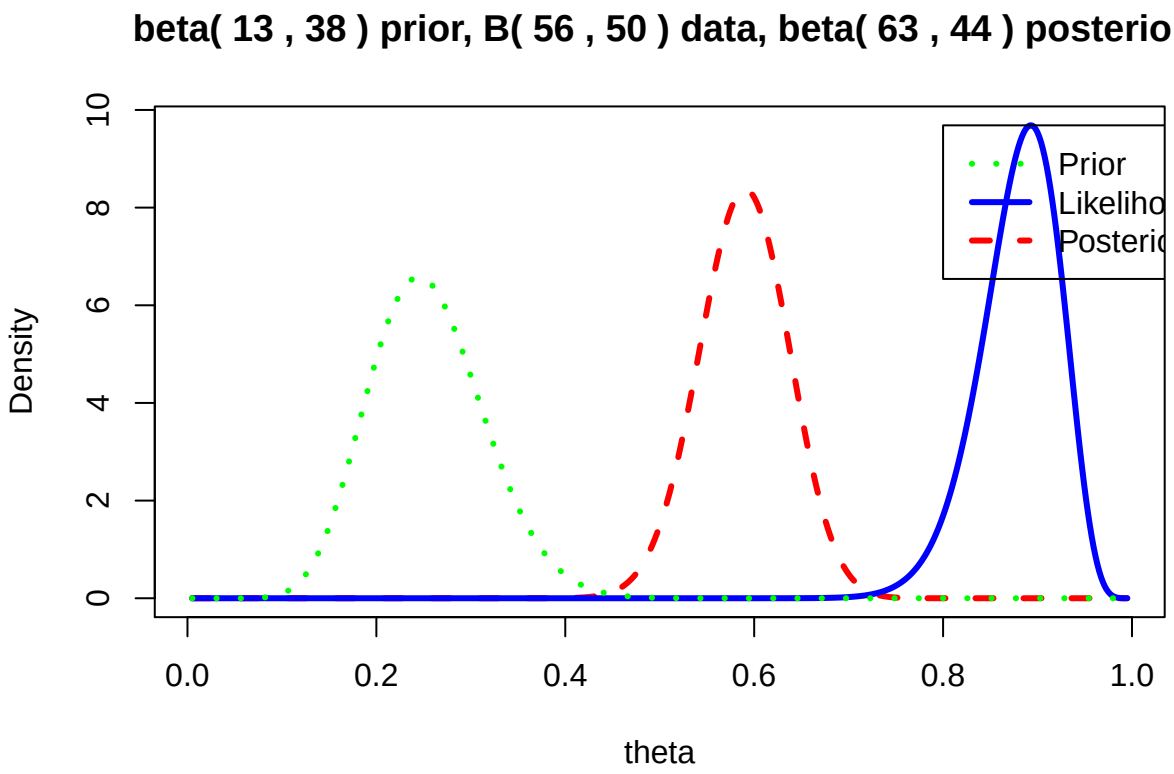
```
    print(paste("mode for prior=",prior_mode,", for likelihood=",likelihood_mode,", for posterior=",pos
    print(paste("mean for prior=",prior_mean,", for likelihood=",likelihood_mean,", for posterior=",pos
    print(paste("sd for prior=",prior_sd,", for likelihood=",likelihood_sd,", for posterior=",posterio
}
calcPosteriorForProportion(50,56,0,1)
```




```
## [1] "mode for prior= 1 , for likelihood= 0.892857142857143 , for posterior= 0.890909090909091"
## [1] "mean for prior= 0 , for likelihood= 0.879310344827586 , for posterior= 0.87719298245614"
## [1] "sd for prior= 0 , for likelihood= 0.0424111558530635 , for posterior= 0.0430968144246068"
```

2).

```
calcPosteriorForProportion(50,56,13,38)
```




```
## [1] "mode for prior= 0.244897959183673 , for likelihood= 0.892857142857143 , for posterior= 0.590476
## [1] "mean for prior= 0.254901960784314 , for likelihood= 0.879310344827586 , for posterior= 0.588785
## [1] "sd for prior= 0.0604354314016566 , for likelihood= 0.0424111558530635 , for posterior= 0.047347
```

Comparing the plots for the first part of the problem and second part of the problem, we can tell the relationship between prior distribution, likelihood, and posterior distribution. The prior distribution, combines with the likelihood, generating the posterior distribution. In other words, the probability of a nurse using hand sanitizers depends both on the probability a nurse using hand sanitizers before the education program and probability a nurse using hand sanitizers after the education program, which is the test result.
From the first plot, we see that when the prior distribution is uniform distribution, which does not tell any information, the posterior distribution is exactly the same as the evidence. Because we only get information form the test result. However, when the prior distribution tells some information, the posterior distribution will be determined by both the prior and the likelihood. Here, the prior distribution says the probability of a nurse using hand sanitizers would be most likely be around 0.25. However, after the education program, the mode of the probability density becomes 0.9. Therefore, the posterior distribution, as a weighted average of the prior and the likelihood, says the probability of a nurse using hand sanitizers is most likely to be around 0.6.

Project2: Irragation

```
#read the rotation data
rot <- read.delim("rot.txt",header = FALSE,sep = " ")
names(rot)[c(1,2,3,4,5)] <- "time"
time <- tidyr::pivot_longer(rot,time)
```

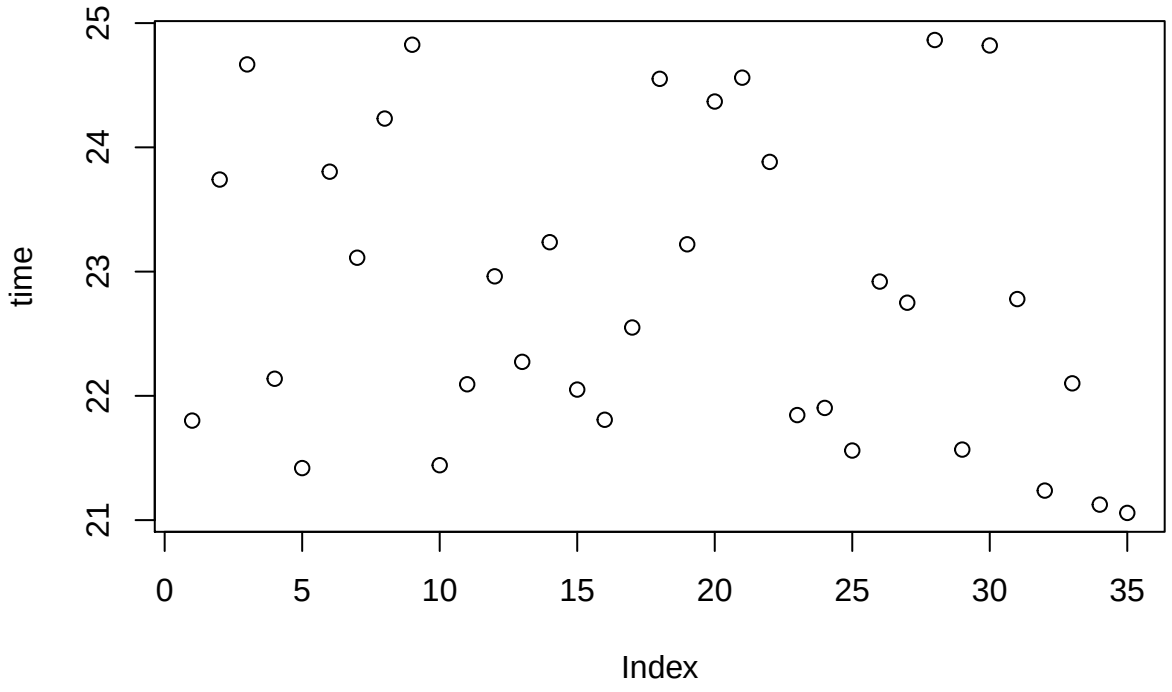
```
## Warning: Duplicate column names detected, adding .copy variable
```

```
as.data.frame(time)
```

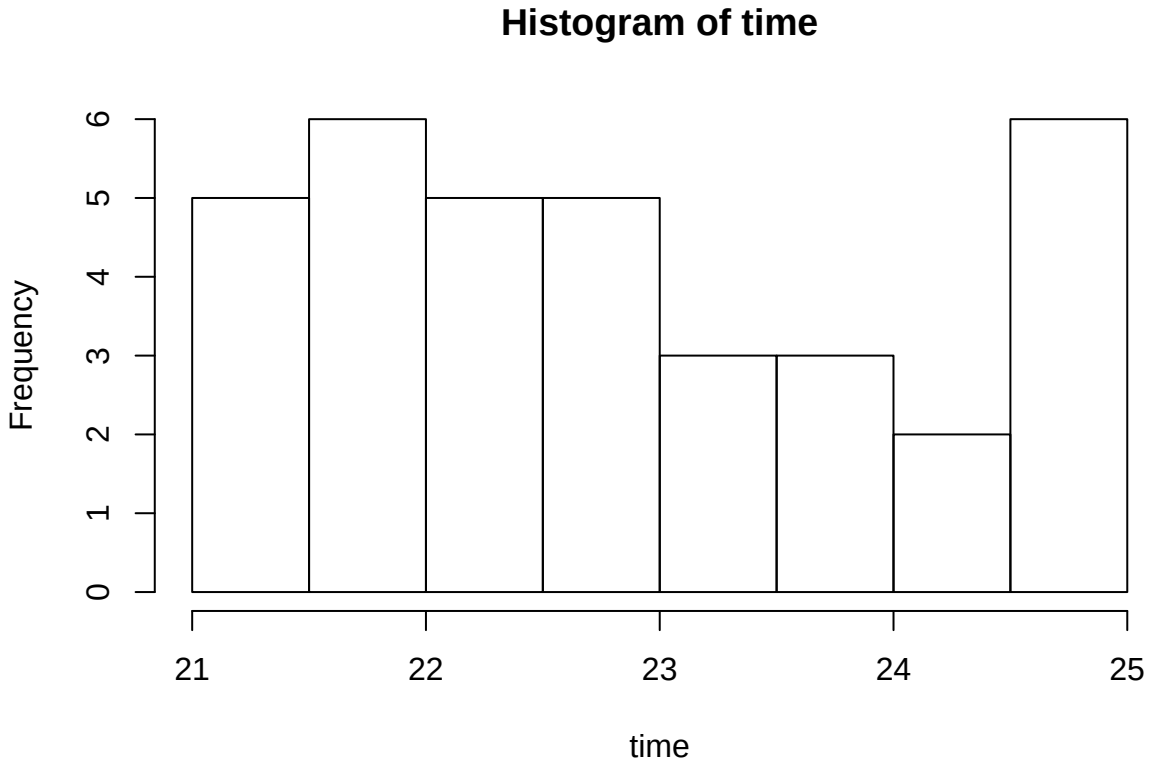
```
##    name .copy    value
## 1  time     1 21.80086
## 2  time     2 23.74087
```

```
## 3  time     3 24.66750
## 4  time     4 22.13760
## 5  time     5 21.41860
## 6  time     1 23.80423
## 7  time     2 23.11184
## 8  time     3 24.23174
## 9  time     4 24.82600
## 10 time     5 21.44181
## 11 time     1 22.09314
## 12 time     2 22.96205
## 13 time     3 22.27362
## 14 time     4 23.23669
## 15 time     5 22.05037
## 16 time     1 21.80750
## 17 time     2 22.55010
## 18 time     3 24.55148
## 19 time     4 23.21969
## 20 time     5 24.36872
## 21 time     1 24.56083
## 22 time     2 23.88280
## 23 time     3 21.84536
## 24 time     4 21.90287
## 25 time     5 21.55993
## 26 time     1 22.91966
## 27 time     2 22.74965
## 28 time     3 24.86386
## 29 time     4 21.56766
## 30 time     5 24.81992
## 31 time     1 22.77892
## 32 time     2 21.23745
## 33 time     3 22.10060
## 34 time     4 21.12459
## 35 time     5 21.05793
```

```
time <- time$value
plot(time)
```




```
hist(time)
```




```
mean(time)
```

```
## [1] 22.83618
```

```
var(time)
```

```
## [1] 1.50765
```

```r
#the mean of the sample is 23, the variance of the sample is 1.5


dis <- 3.14*1320*2
#distance is a fixed number, 8290
```

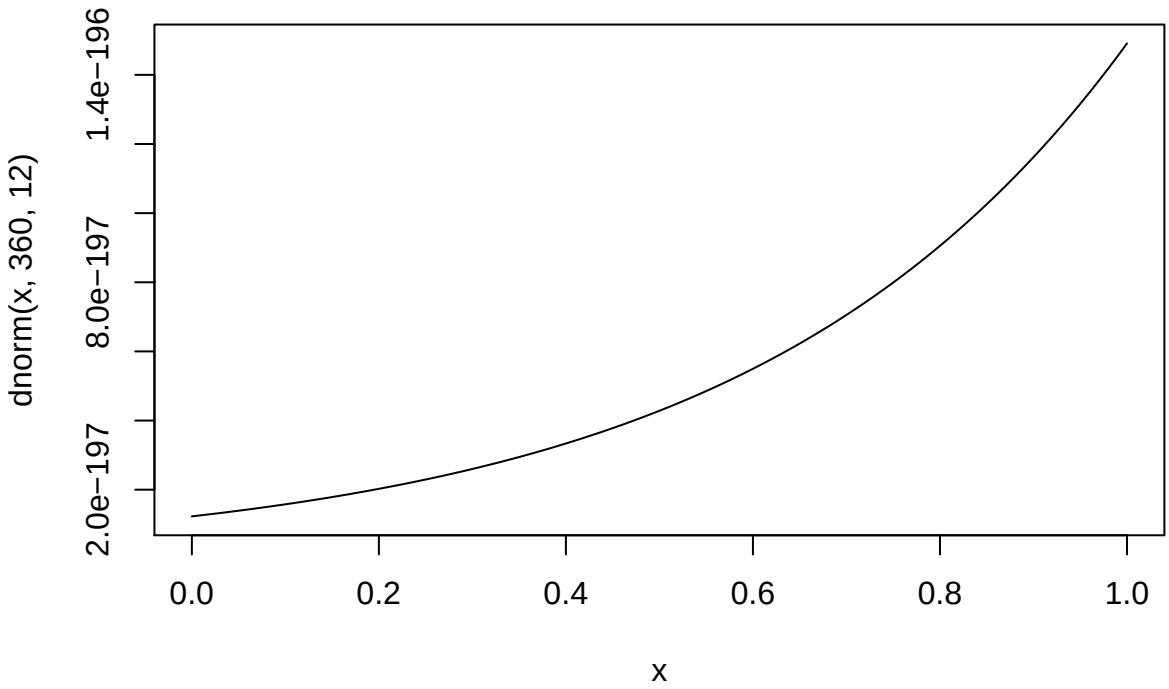
Central limit theorm: the distribution of the mean from each sample is a normal distribution with mean of 23 and variance of 1.5.

Delta method: because we care about the speed, which is the distance(8290) divided by time(x), we want to know the distribution of $8290/x$. So when we know the distribution of x, we use the delta method to generate the distribution of $8290/x$.

$$g(x) = 8290/x$$

The mean of the distribtion of $8290/x$ is $8290/\text{mean}(x)$, and the variance of the distribution of $8290/x$ is $(8290/(mean(x)^2))^2 * var(x)/n$ n = 35 newmean = 8290/23 = 360 newvar = (8290/ (1.5^2))*(1.5/35) = 147 mewsd = sqrt(var) = 12

```r
#distribution for the speed
curve(dnorm(x,360,12))
```




```r
lb <- qnorm(0.1,mean=360,sd = 12)
up <- qnorm(0.9,mean = 360, sd = 12)
lb
```

```
## [1] 344.6214
```

```r
up
```

```
## [1] 375.3786
```

The 90% confidence interval would be 344 to 375.

From the histogram of the orignal dataset, we can not determine whether it is a normal distribution or not, thus we can not generate a reasonable confidence interval. However, based on the central limit theorm,

once we know the mean and variance of the population exsit, the mean of the every sample, that is taken from the population, is nearly a normal distribution with a certain mean and variance. Thats the first step. Second, because we want to know the distribution of the speed, which is length/time, while we only know the distribution of time, we have to convert the time to speed and get a distribution of speed. Using the delta method, we can generate a normal distribution of speed with a known mean and variance that are calculated by the mean and variance of the distribution of time. Once we know the distribution of speed, which is a normal distribution, we can generate the 90% confidence interval of the speed. In this case, we can say 90% of the speed would fall within the range of 344 to 375.