**Fitting Distributions to Data**

**Problem:** Suppose we have data for a continuous random variable X. Can we find (estimate) the distribution for X?

Part 1: What type of distribution? Exponential, normal, uniform, gamma, beta, ….

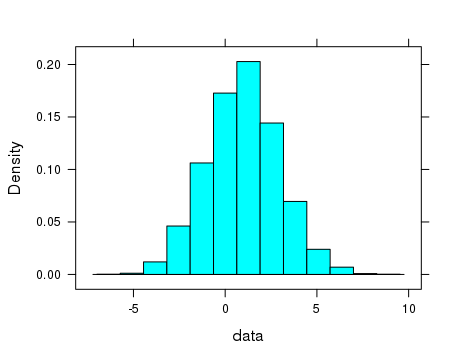
Part 2: What are the values of the parameters for the distribution?

**R-command**: **fitdistr(data, “type”)** where **data** is the data and **type** is the type of

distribution.

**Example:**

* data<-rtype(5000,p1,p2)
* histogram(data)



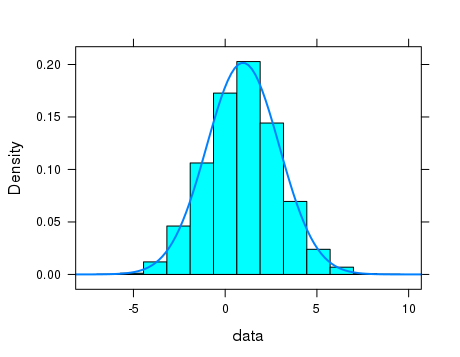
> fitdistr(data,"normal")

mean sd

0.97086512 1.97853505

(0.02798071) (0.01978535)

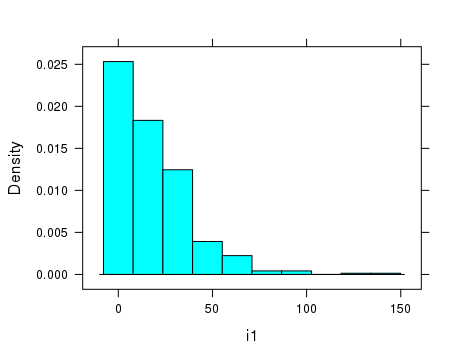
* histogram(data,fit="normal")



**Example**

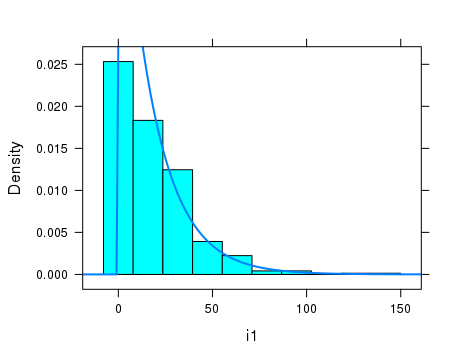
The variable i1 in the data frame HELPrct contains data for the average number of alcoholic drinks consumed per day for each participant. If X is the average number of drinks per day for a randomly selected participant, what is the distribution of X?

* histogram(~i1, data=HELPrct)



It looks like the type of distribution is .

* histogram(~i1,data=HELPrct,fit="exponential")



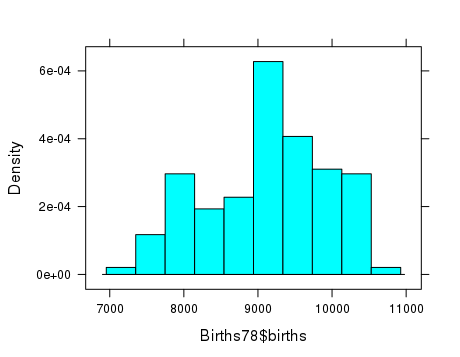
> fitdistr(HELPrct$i1,"exponential")

rate

0.055843195

(0.002623742)

Example Births78 contains the number of births in the US on each day in 1978.

* histogram(Births78$births)
* 

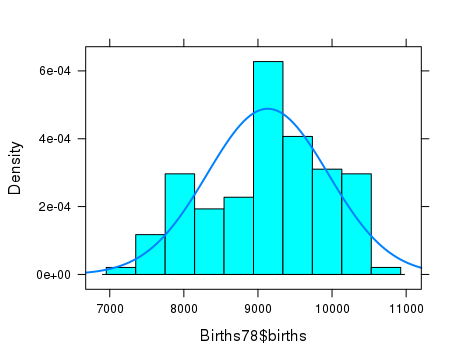
> fitdistr(Births78$births, "normal")

mean sd

9132.16164 816.76093

( 42.75122) ( 30.22968)

* histogram(Births78$births, fit = "normal")



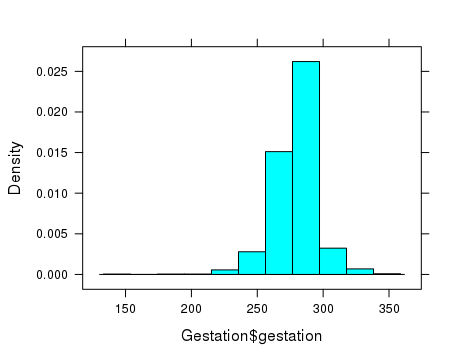
The dataframe **Gestation** contains data on 1236 human births during 1961-62. Variables include:

gestation: gestation period in days

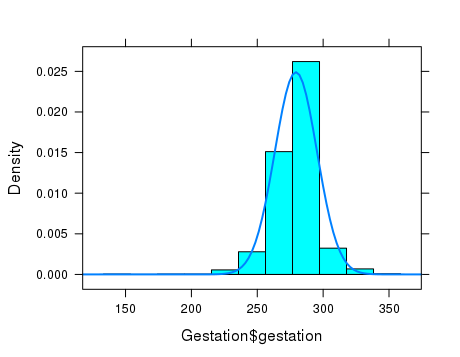
ht: height of mother in inches

age: age of mother in years

* histogram(Gestation$gestation)



* histogram(Gestation$gestation, fit = "normal")



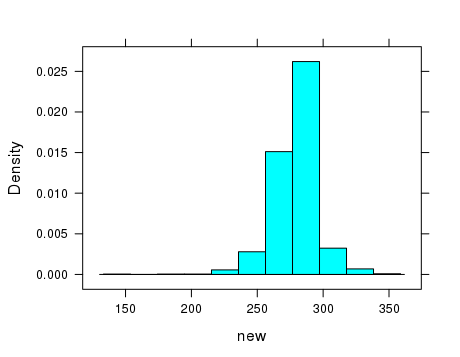
> fitdistr(Gestation$gestation, "normal")

Error in fitdistr(Gestation$gestation, "normal") :

'x' contains missing or infinite values

> new<-na.omit(Gestation$gestation)

> histogram(new)



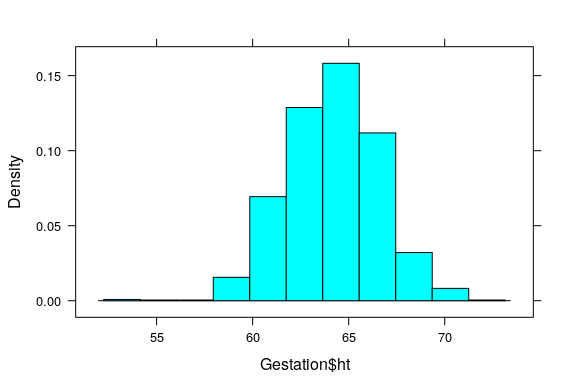
> fitdistr(new, "normal")

mean sd

279.3385119 16.0211393

( 0.4581210) ( 0.3239404)

> histogram(Gestation$ht)



> new<-na.omit(Gestation$ht)

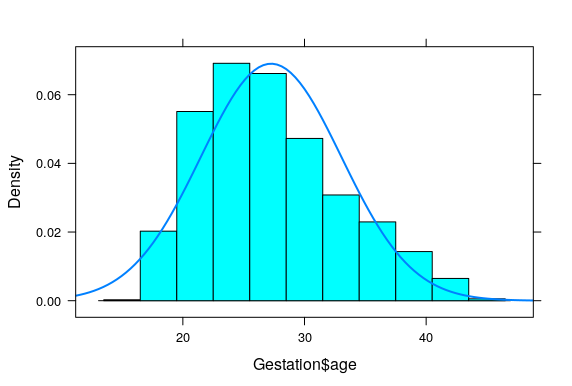
> fitdistr(new,"normal")

mean sd

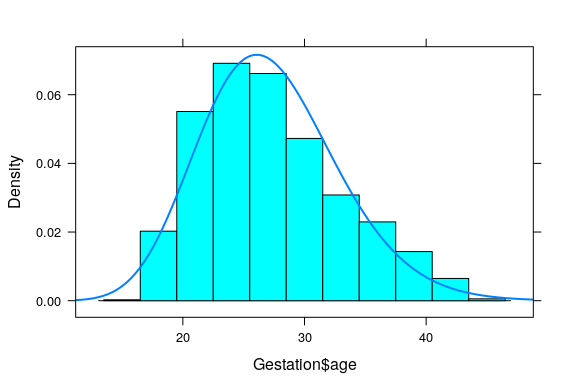
64.04777595 2.53236529

( 0.07268035) ( 0.05139277)

histogram(Gestation$age,fit = "normal")



> histogram(Gestation$age,fit = "gamma")



> fitdistr(new,"gamma")

shape rate

23.0981793 0.8474759

( 0.9232297) ( 0.0342433)

**How are the estimates of the parameters computed from the data?**

**1. Maximum Likelihood**

**Example:** exponential distribution , parameter = 

Data : x1, x2, x3

Likelihood function

Find the value of  that maximizes the value of the likelihood function

The general formula for estimating 

**Exanple**

> fitdistr(HELPrct$i1,"exponential")

rate

0.055843195

(0.002623742)

> 1/mean(HELPrct$i1)

[1] 0.0558432

**2. Method of Moments**

Express the parameters as functions of the mean, variance, etc. Use the mean, variance, etc., of the data to estimate the parameters.

**Example**: exponential distribution

 so that 

estimate 

**Example**: uniform distribution on [a,b]

Solve the two equations above for a and b in terms of μ and σ.

> sim<-runif(1000,1,4)

> mean(sim)-sqrt(3)\*sd(sim)

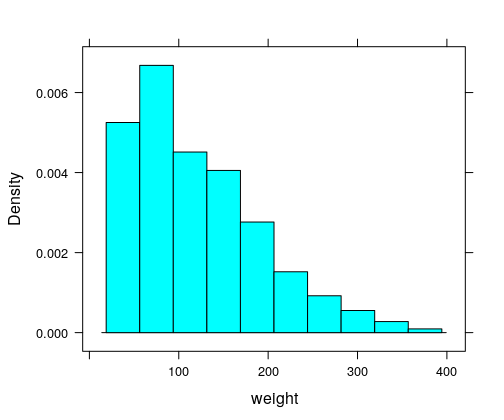
[1] 0.9430267

> mean(sim)+sqrt(3)\*sd(sim)

[1] 3.978281

**Exercises 8. NOTE: WHEN A JUDGEMENT IS REQUIRED, BE FLEXIBLE IN GRADING.**

1. **(10 pts**)The column **weight** in the dataframe **ChickWeight** contains the weights (gm) of baby chicks at different times under different feeds.
2. Plot a histogram of the values in the column **weight.**



1. Of the distributions: uniform, exponential, normal, and gamma, which fits the data the best?

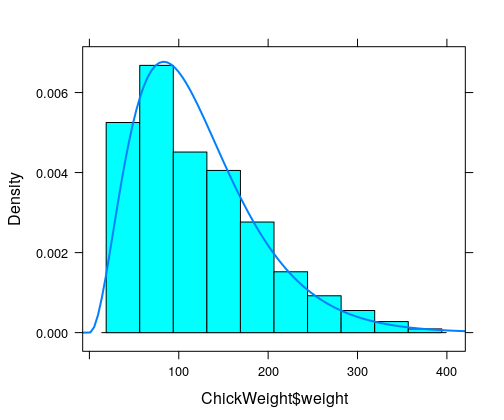
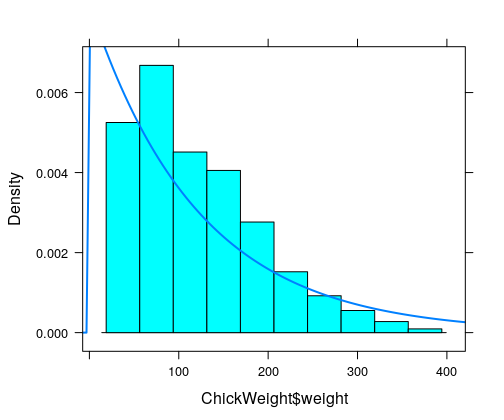
**Either Exponential or gamma is a good answer**

1. Use fitdistr in R to fit the type of distribution in (b) to the data. What is the estimate of the parameter(s) for the distribution?

**Estimated value is rate = .0082 if exponential**

**Estimated values are shape = 3.15, rate = .026 if gamma**

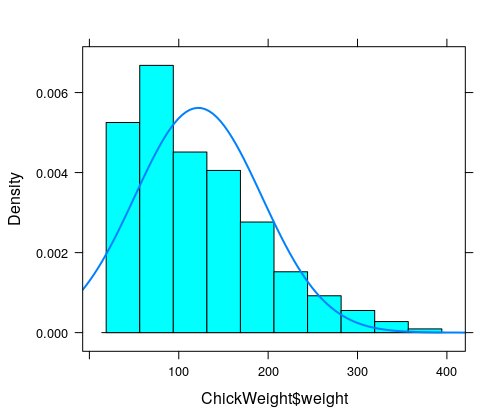
1. Plot the histogram with the fitted density function.



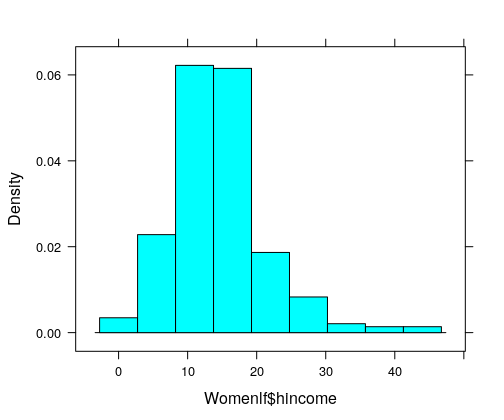
**Exponential**  **Gamma**

1. Use fitdistr in R to fit one of the other two types of distributions from (b) and plot the histogram with the density curve for that type of distribution. How does the fit look?

**With normal fit, Mean = 121.8 SD = 71.0**

**Normal is not as good as either the exponential or gamma.**

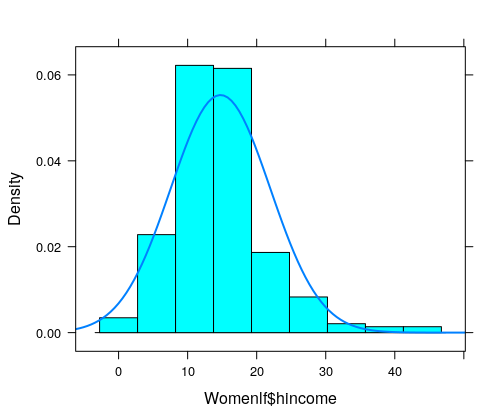
1. **(10 pts). Load the package: carData**. **T**he column **hincome** in the dataframe **Womenlf** contains household income for 263 Canadian women.
2. Plot a histogram of this data.



1. Use R to fit a normal distribution to this data. What are the estimates for the mean and standard deviation of the resulting normal distribution?

**Mean = 14.76 SD = 7.21**

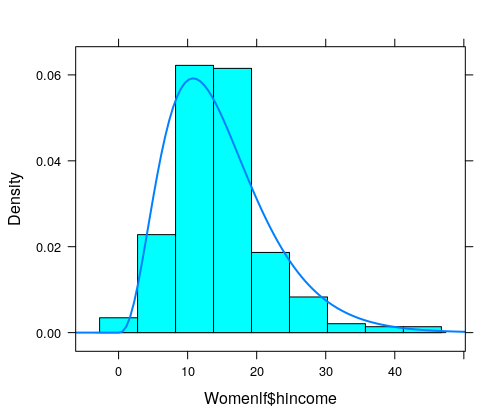
1. Plot the histogram with the fitted normal density function.



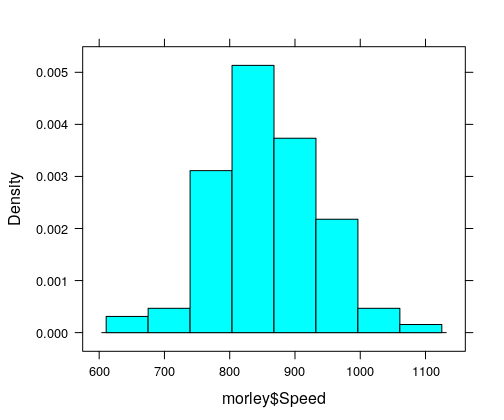
1. Use R to fit a gamma distribution to this data. What are the estimates for the shape and rate of the fitted gamma distribution?

**Shape = 3.73. rate = .25**

1. Plot the histogram with the fitted gamma density function.



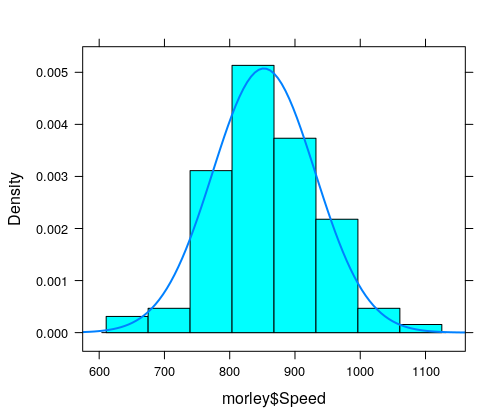
1. Compare the quality of the fits with normal vs gamma. Does one clearly fit better than the other? **They both seem to fit reasonably well. Given that the data is skewed a bit to the right, maybe the gamma fits better.**
2. **(6 pts).** The dataframe **morley** contains 100 measurements that Morley made of the speed of light. These values are contained in the column **Speed**.
3. Plot the histogram of this data,



1. Which kind of distribution best fits the data? **normal**
2. Find the estimated parameters for this distribution.

**Mean = 852.4 SD = 78.6**

1. Plot the histogram with the fitted density curve.



1. **(2 pts)** The gamma distribution has two parameters,  and . The mean and variance of the gamma distribution are given by  and . Suppose a data set has . Use the method of moments to compute estimates of  and 

** = 6  = 36**