STAT 206 Lab 3

Xin Feng(Vanessa)

10/21/2019

Due Monday, October 21, 5:00 PM

General instructions for labs: Labs must be completed as a pdf 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: Writing functions to automate repetitive tasks; fitting statistical models.

The ***gamma*** distributions are a family of probability distributions defined by the density functions,

$$f(x) = \frac{x^{a-1}e^{-x/s}}{s^a\Gamma(a)}$$

where the ***gamma function*** $\Gamma(a) = \int_0^\infty u^{a-1}e^{-u}du$ is chosen so that the total probability of all non-negative x is 1. The parameter a is called the ***shape***, and s is the ***scale***. When $a = 1$, this becomes the exponential distributions we saw in the first lab. The gamma probability density function is called `dgamma()` in R. You can prove (as a calculus exercise) that the expectation value of this distribution is as, and the variance as^2. If the mean and variance are known, μ and σ^2, then we can solve for the parameters,

$$\begin{aligned} a &= \frac{a^2s^2}{as^2} = \frac{\mu^2}{\sigma^2} \\ s &= \frac{as^2}{as} = \frac{\sigma^2}{\mu} \end{aligned}$$

In this lab, you will fit a gamma distribution to data, and estimate the uncertainty in the fit.

Our data today are measurements of the weight of the hearts of 144 cats.

Part I

1. The data is contained in a data frame called `cats`, in the R package `MASS`. (This package is part of the standard R installation.) This records the sex of each cat, its weight in kilograms, and the weight of its heart in grams. Load the data as follows:

```r
library(MASS)
data(cats)
```

Run `summary(cats)` and explain the results.

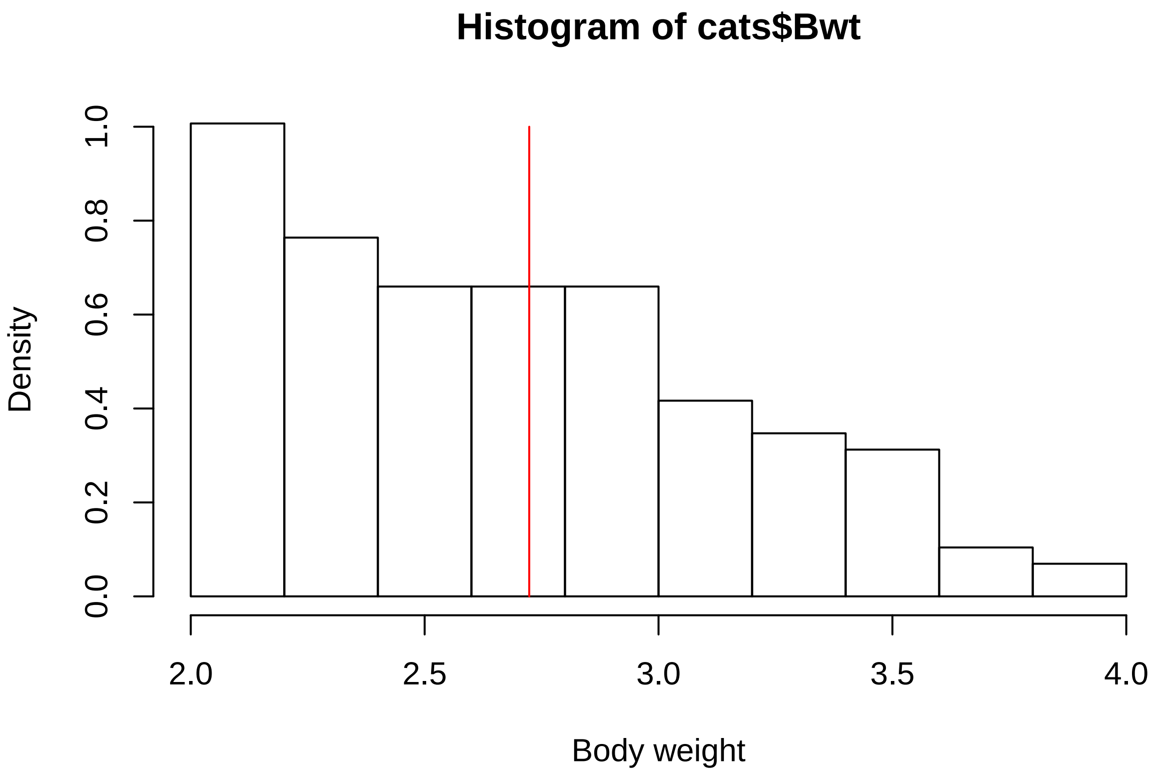
```r
# View(cats)
summary(cats)
```

```
##  Sex         Bwt             Hwt       
##  F:47   Min.   :2.000   Min.   : 6.30  
##  M:97   1st Qu.:2.300   1st Qu.: 8.95  
##         Median :2.700   Median :10.10  
##         Mean   :2.724   Mean   :10.63  
##         3rd Qu.:3.025   3rd Qu.:12.12  
##          Max.   :3.900   Max.   :20.50
```

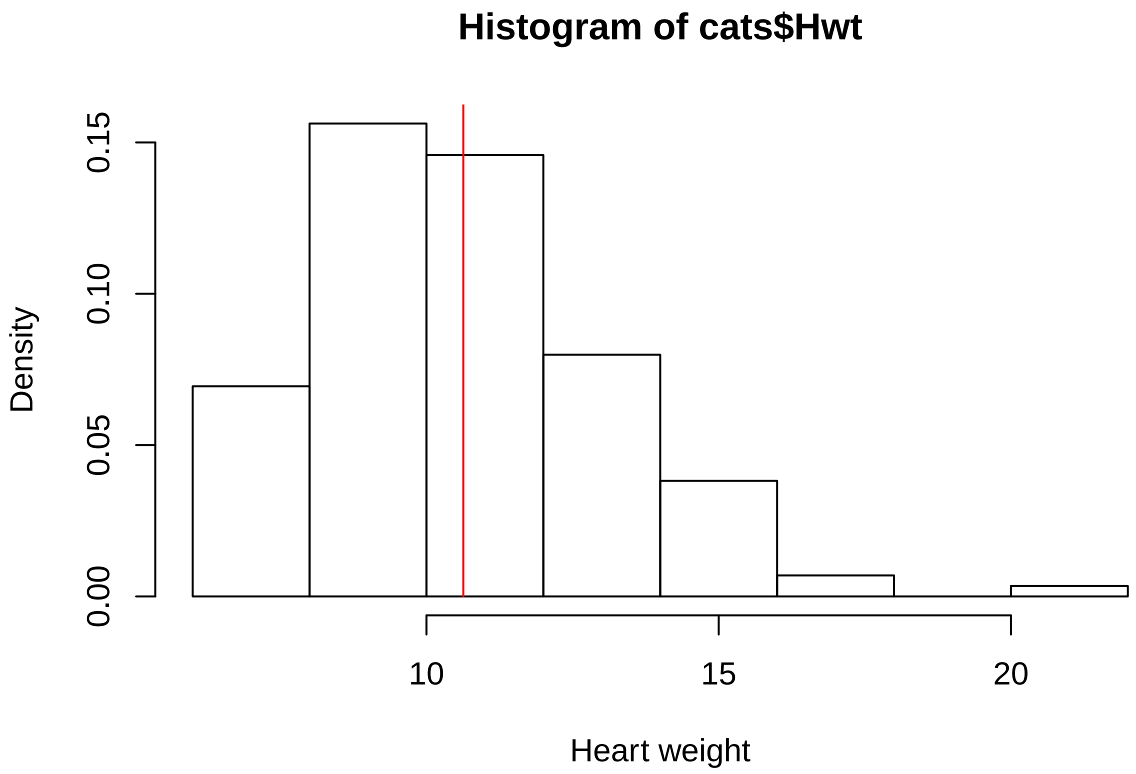
```
# Sex: number(Female)=47  number(Male)=97
# Body weight(kg): minimal=2,mean=2.724, max=3.9, 1st quantile=2.3, 3rd quantile=3.025,
# Median=2.7
# Heart weight(g): minimal=6.3,mean=10.63, max=20.5, 1st quantile=8.95, 3rd quantile=12.12,
# Median=10.10
```

2. Plot a histogram of these weights using the `probability=TRUE` option. Add a vertical line with your calculated mean using `abline(v=yourmeanvaluehere)`. Does this calculated mean look correct?

```
Bwt_mean <- mean(cats$Bwt)
Hwt_mean <- mean(cats$Hwt)
hist(cats$Bwt, probability=TRUE, xlab="Body weight")
segments(Bwt_mean, 0, Bwt_mean, 1, col=2)
```



```
hist(cats$Hwt, probability=TRUE, xlab="Heart weight")
segments(Hwt_mean, 0, Hwt_mean, 1, col=2)
```



```
# The calculated mean seems correct.
```

3. Define two variables, `fake.mean <- 10` and `fake.var <- 8`. Write an expression for a using these placeholder values. Does it equal what you expected given the solutions above? Once it does, write another such expression for s and confirm.

```
fake.mean <- 10
fake.var <- 8
a <- fake.mean^2/fake.var
cat("From the value of fake.mean and fake.var, the value of a is: ", a, "\n")
```

```
## From the value of fake.mean and fake.var, the value of a is:  12.5
```

```
s <- fake.var/fake.mean
```

4. Calculate the mean, standard deviation, and variance of the heart weights using R's existing functions for these tasks. Plug the mean and variance of the cats' hearts into your formulas from the previous question and get estimates of a and s. What are they? Do not report them to more significant digits than is reasonable.

```
cats_mean_Hwt <- mean(cats$Hwt)
cats_var_Hwt <- var(cats$Hwt)
cat("Mean of heart weight: ", cats_mean_Hwt, "\n")
```

```
## Mean of heart weight:  10.63056
```

```
cat("Standard deviation of heart weight: ", sd(cats$Hwt), "\n")
```

```
## Standard deviation of heart weight:  2.434636
```

```
cat("Varience of heart weight: ", cats_var_Hwt, "\n")
```

```
## Varience of heart weight:  5.927451
```

```
a_hat <- cats_mean_Hwt^2/cats_var_Hwt
s_hat <- cats_var_Hwt/cats_mean_Hwt
cat("Estimates of a: ", a_hat, ",\nEstimates of s: ", s_hat, "\n")
```

```
## Estimates of a:  19.06531 ,
## Estimates of s:  0.5575862
```

5. Write a function, `cat.stats()`, which takes as input a vector of numbers and returns the mean and variances of these cat hearts. (You can use the existing mean and variance functions within this function.) Confirm that you are returning the values from above.

```
cat.stats <- function(value_x=cats$Hwt){
  cats_mean_Hwt <- mean(value_x)
  cats_var_Hwt <- var(value_x)
  return(c(cats_mean_Hwt, cats_var_Hwt))
}
cat("Mean of the cats' hearts weight: ", cat.stats()[1], "\n")
```

```
## Mean of the cats' hearts weight:  10.63056
```

```
cat("Varience of the cats' hearts weight: ", cat.stats()[2], "\n")
```

```
## Varience of the cats' hearts weight:  5.927451
```

Part II

6. Now, use your existing function as a template for a new function, `gamma.cat()`, that calculates the mean and variances and returns the estimate of a and s. What estimates does it give on the cats' hearts weight? Should it agree with your previous calculation?

```
gamma.cat <- function(value_x=cats$Hwt){
  cats_feature <- cat.stats(value_x)
  a_hat <- cats_feature[1]^2/cats_feature[2]
  s_hat <- cats_feature[2]/cats_feature[1]
  return(c(a_hat, s_hat))
}
cats_para <- gamma.cat()
cat("Estimate of a: ", cats_para[1], "\n")
```

```
## Estimate of a:  19.06531
```

```
cat("Estimate of s: ", cats_para[2], "\n")
```

```
## Estimate of s:  0.5575862
```

```
cats_Hwt_hat <- cats_para[1]*cats_para[2]
cat("Estimate of the cats' hearts weight: ", cats_Hwt_hat)
```

```
## Estimate of the cats' hearts weight:  10.63056
```

```
# According the results, the estimate of cats' hearts weight equal to the previous
# calculation.
```

7. Estimate the a and s separately for all the male cats and all the female cats, using `gamma.cat()`. Give the commands you used and the results.

```
# Male
Index_male <- which(cats$Sex=="M")
```

```r
cats_Hwt_male <- cats$Hwt[Index_male]
cats_para_male <- gamma.cat(cats_Hwt_male)
cat("Male-estimate of a: ", cats_para_male[1], ", estimates of s: ",
    cats_para_male[2], "\n")
```

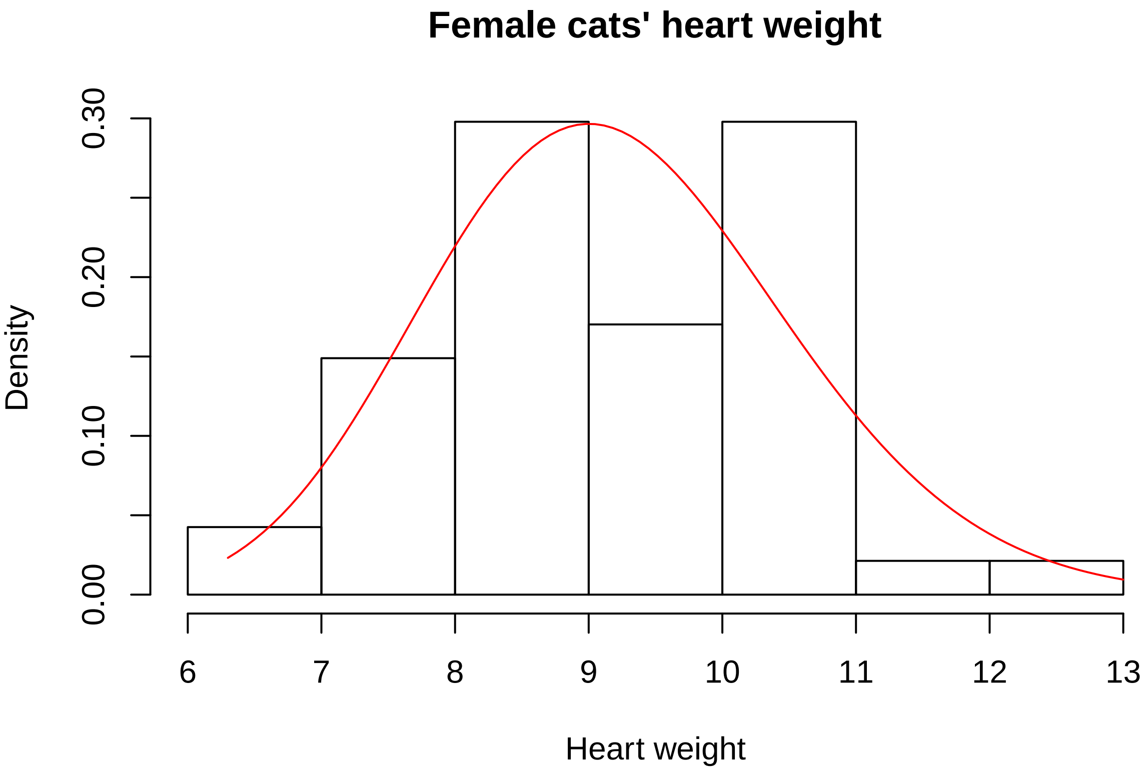
```
## Male-estimate of a:  19.83576 , estimates of s:  0.5708216
```

```r
# Female
Index_female <- which(cats$Sex == "F")
cats_Hwt_female <- cats$Hwt[Index_female]
cats_para_female <- gamma.cat(cats_Hwt_female)
cat("Female-estimate of a: ",cats_para_female[1],", estimates of s: ",
    cats_para_female[2],"\n")
```

```
## Female-estimate of a:  45.93998 , estimates of s:  0.2003076
```

8. Now, produce a histogram for the female cats. On top of this, add the shape of the gamma PDF using `curve()` with its first argument as `dgamma()`, the known PDF for the Gamma distribution. Is this distribution consistent with the empirical probability density of the histogram?

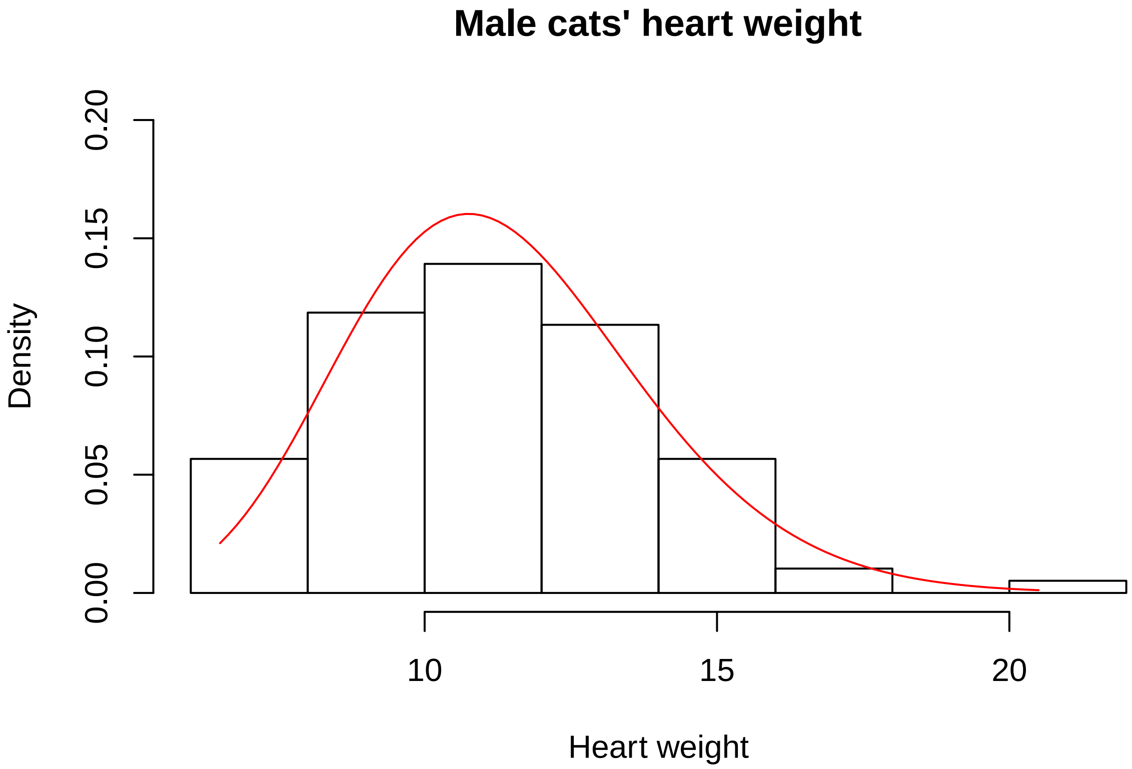
```r
hist(cats_Hwt_female, main="Female cats' heart weight", xlab="Heart weight",
     probability=T)
curve(dgamma(x,shape=cats_para_female[1],scale=cats_para_female[2]),
      from=min(cats_Hwt_female), to=max(cats_Hwt_female), add=T, col=2)
```



```r
# According to the picture, the distribution is partly consistent to the
# empirical probability density.
# But there are still some difference. I consider that may due to the too
# small sample space.
```

9. Repeat the previous step for male cats. How do the distributions compare?

```
hist(cats$Hwt[Index_male], main="Male cats' heart weight", xlab="Heart weight", ylim=c(0,0.2),
     probability = T)
curve(dgamma(x, shape=cats_para_male[1], scale=cats_para_male[2]), from=min(cats_Hwt_male),
      to=max(cats_Hwt_male), add=T, col=2)
```



```
# The distribution is greatly consistent to the empirical probability density.
```