Visualization case

# http://www.grenswetenschap.nl/images/artikelfoto/OVLemminglente111.jpg Lemmings in Rotterdam

# Student: Jeroen van der Steen Student number: 0867254 Minor: Data Science

# Research

The case is about lemmings in Rotterdam. The data for the case is taken from Abbey Waldron’s Github account[[1]](#footnote-1):

The data contains three variables, explained below:

|  |  |  |
| --- | --- | --- |
| X | **sightings** | Lemming sightings per hour in Rotterdam |
| Y | **weights** | Lemming weights in kg |
| Z | **lifetimes** | Lemming lifetimes in years |

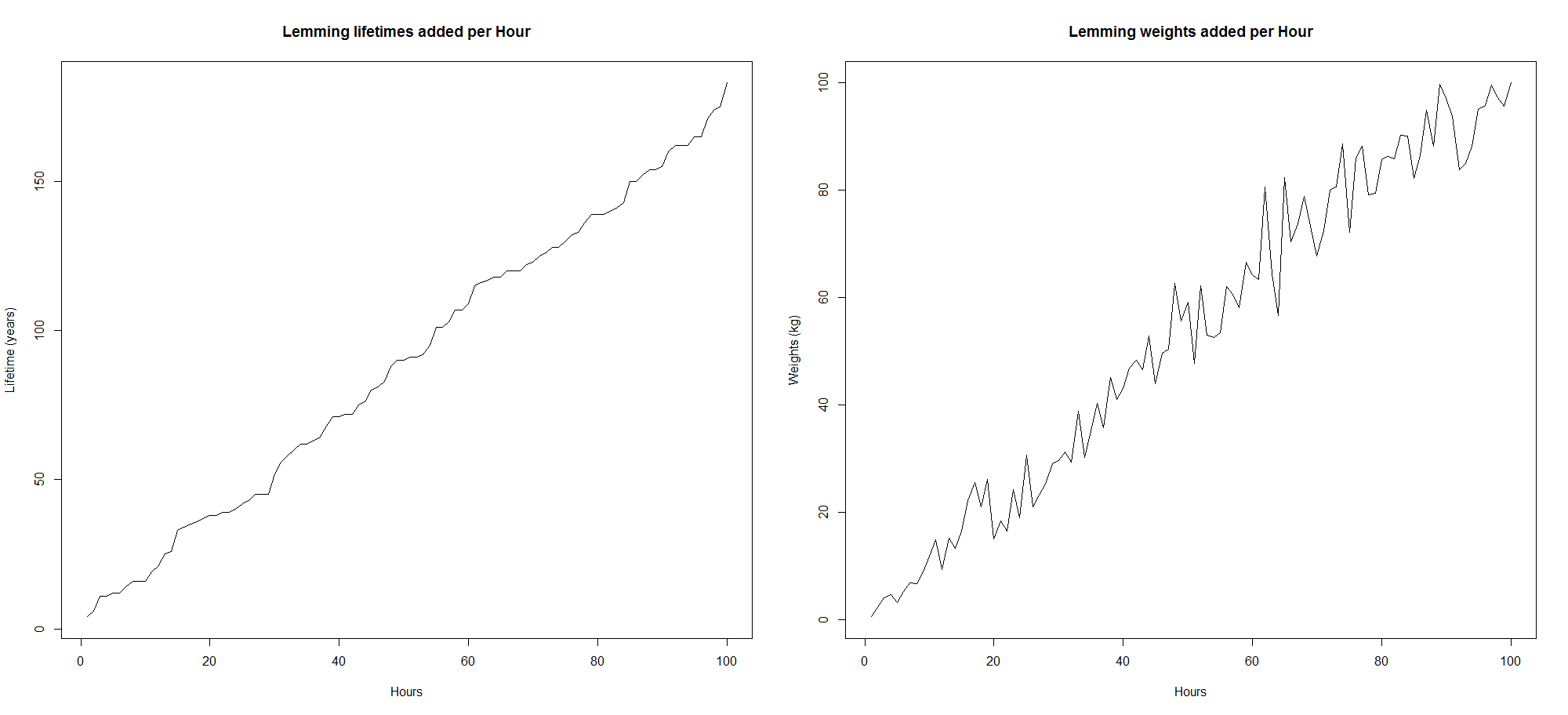
When you look carefully to the data, something is notable. If you order the data based on the weights and lifetimes, you can see the order of measurement per hour (up to 100 hours).

The number of sightings are probability counted by several lemming spotters, after which the average number of sightings where taken. The weights are also a mean of something. I assume that both the weights and lifetimes where calculated by taking the cumulative sum.

The number of lemming spotters that counted the lemmings is unknown.

The maximum of lemming sightings is known and is about 50.

On figure “Lemming lifetimes added per Hour” and “Lemming weights added per Hour”, on the next page, shows how probability the lemming research was done.



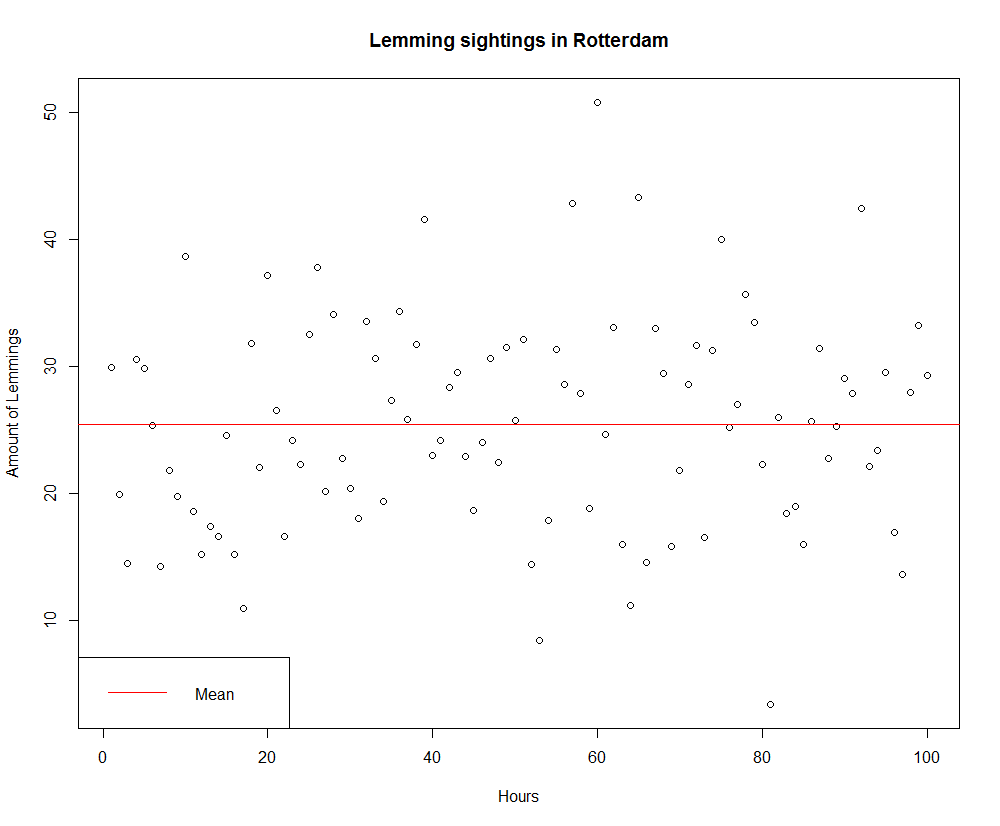
# Lemming sightings in Rotterdam

I computed the mean, standard deviation and standard error of the mean of lemming sightings in Rotterdam.  
You can see the results in this table below:

|  |  |
| --- | --- |
|  | **Lemming sightings** |
| **mean** | 25.47741 per hour |
| **standard deviation** | 8.373527 |
| **standard error of the mean (sem)** | 0.8373527 |

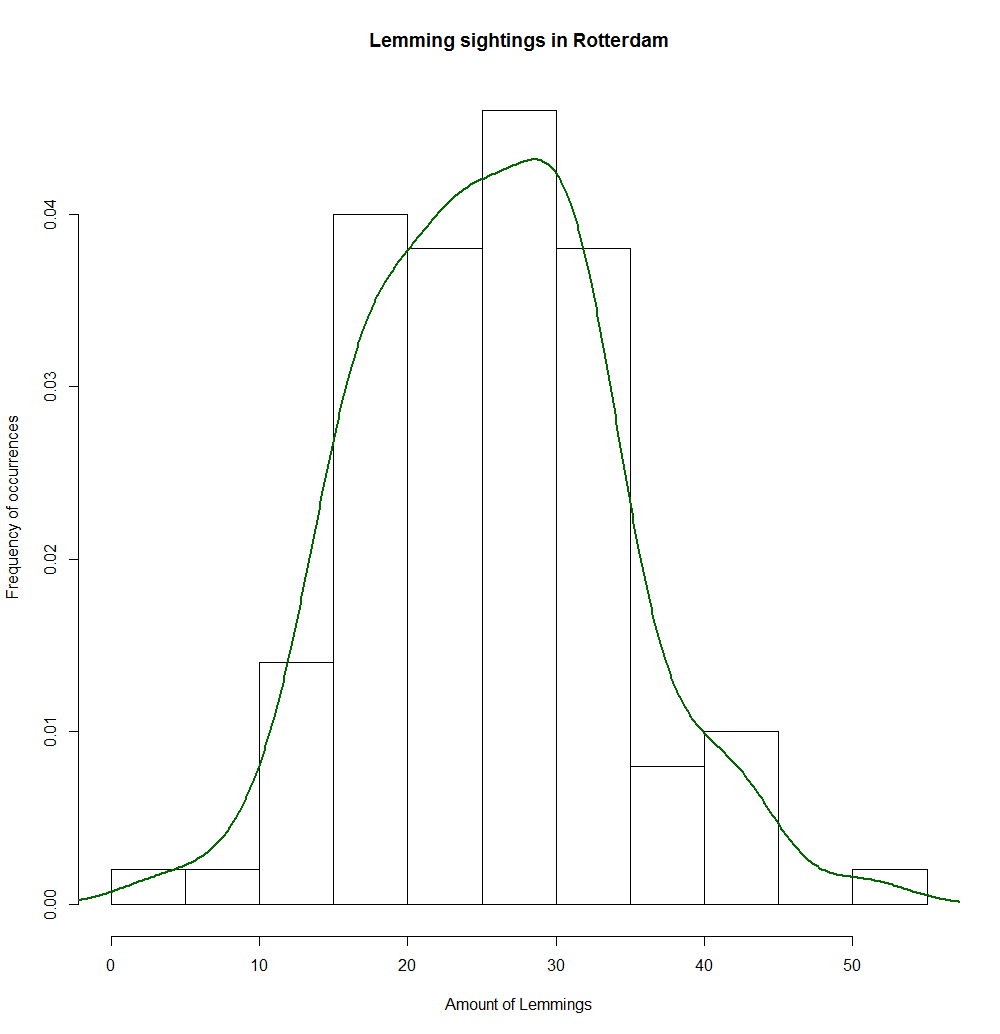
If you observe the plot on figure: “Lemming sightings in Rotterdam I” (see below), you can see that spotters of the lemmings saw each time a different mean of lemmings between each hour. They spotted the lemmings in 100 hours.

Notice the red line that shows the mean of lemmings per hour.

 *Figure: Lemming sightings in Rotterdam I*

If you observe the plot on figure: “Lemming sightings in Rotterdam II” (see below), you can see the frequency of the number of sightings per hour more clearly.

You can see that the spotters saw 25 lemmings per hour the most, so the half of the population in Rotterdam.

 *Figure: Lemming sightings in Rotterdam II*

## Lemming weights vs lifetimes

I computed the mean, standard deviation and standard error of the mean of lemming sightings in Rotterdam on three ways. You can see the results of the first two methods this table below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Method I** | | **Method II** | |
|  | **Lemming weights (kg)** | **Lemming lifetimes (years)** | **Lemming weights (~~kg~~)** | **Lemming lifetimes (years)** |
| **mean** | 52.98623 | 89.51 | 25.45842 | 14.46949 |
| **standard deviation** | 30.25457 | 50.30623 | 3.743661 | 2.1581 |
| **standard error on the mean (sem)** | 3.025457 | 5.030623 | 0.3743661 | 0.21581 |

I found fits for the model “Lemming weights vs lifetimes”.  
Again for both methods (incorrect and correct), you can see the parameters in this table below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Method I** | | **Method II** | |
| **Parameters** | **Values** | **Function** | **Values** | **Function** |
| **intercept** | 2.769 | **slope\*lifetime + intercept**  1.637\*lifetime + 2.769 | 11.8776 | **slope\*lifetime + intercept**  0.1056 \* lifetime + 11.8776 |
| **slope of weights** | 1.637 | 0.1056 |

On figure “Lemming weights vs lifetimes (method I)”, you can see a strong linear relationship between the weights and lifetimes. Notice also figure “Lemming weights vs lifetimes (method I)” for my second attempt.

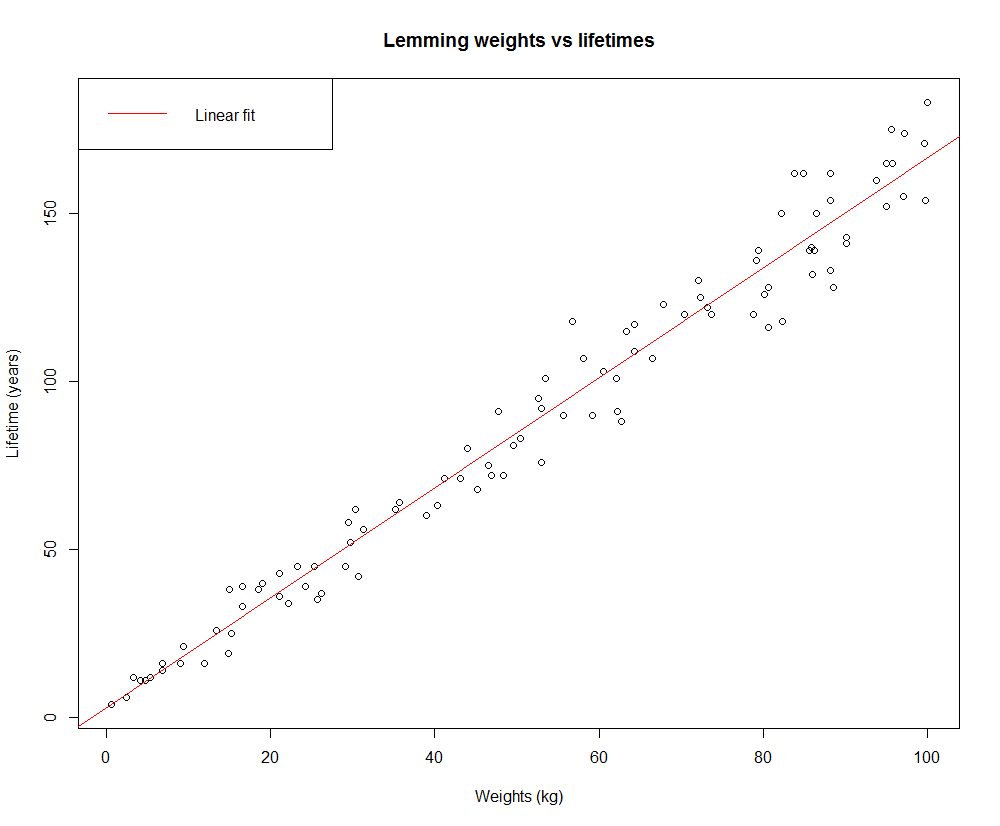
***Warning***

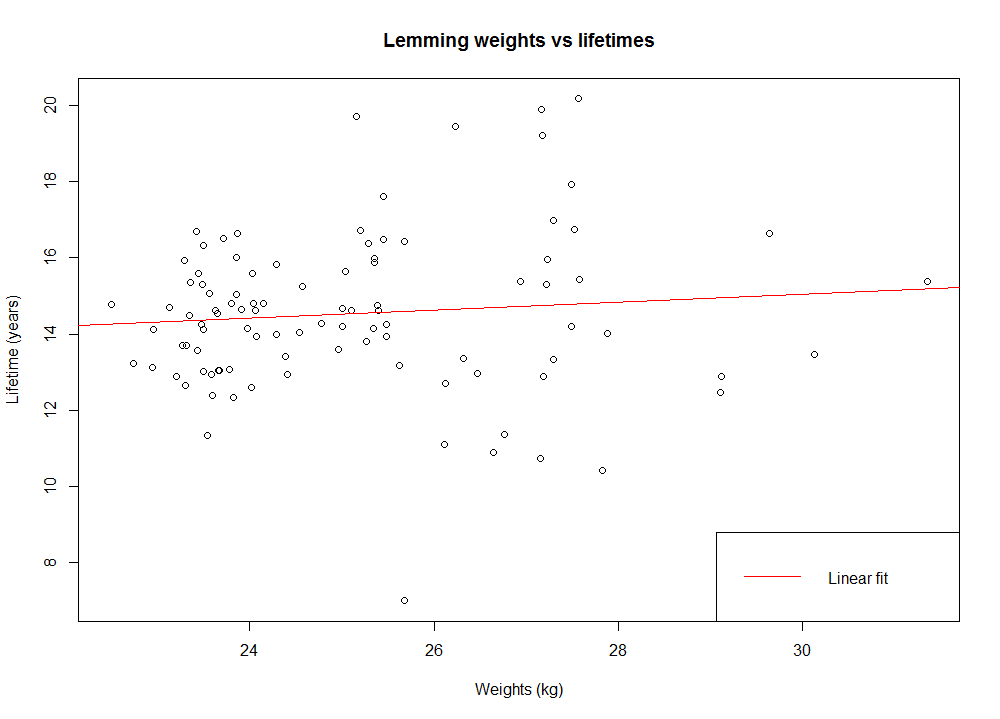
If you look at the facts about lemmings, see discussion, I think these results are ***not*** right, because:

* According to the article from “Vervloed” lemmings are not that fat;
* According to the article “Norway Lemming” lemmings don’t have that longevity;
* Personally I don’t trust the unit of measurement “kg/kilograms” either.

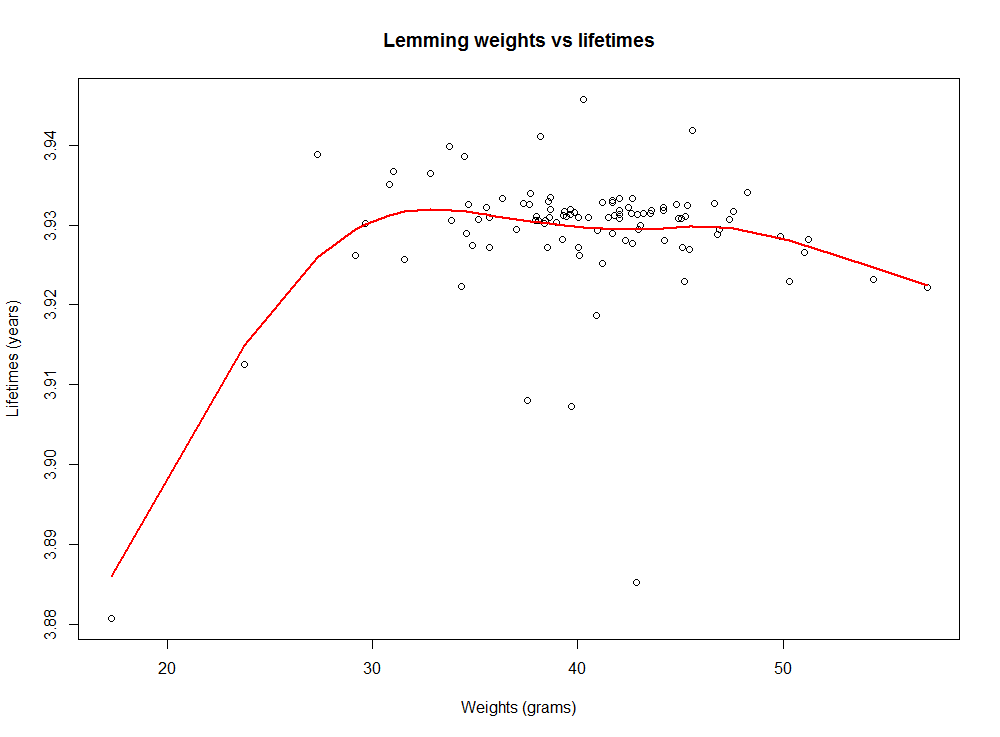
So, to find the real relationship between the lemming weights and their lifetimes, I had to come to the conclusion that the data haves to be normalized. Since the weights and lifetimes where calculated by taking the cumulative sum, and the sightings where not. I thought I also had to do the same for the sightings, take their cumulative sums.

See figure “Lemming weights vs lifetimes (method III)” for a more logical plot.

 *Figure: Lemming weights vs lifetimes (method I)*

****

*Figure: Lemming weights vs lifetimes (method II)*



*Figure: Lemming weights vs lifetimes (method III)*

This is my final attempt; see the tables on this page below for the adjusted results.

|  |  |  |
| --- | --- | --- |
|  | **Method III** | |
|  | **Lemming weights (grams)** | **Lemming lifetimes (years)** |
| **mean** | 52.98623 | 3.92922 |
| **standard deviation** | 30.25457 | 0.008588741 |
| **standard error on the mean (sem)** | 3.025457 | 0.0008588741 |

|  |  |  |
| --- | --- | --- |
|  | **Method III** | |
| **Parameters** | **Values** | **Function** |
| **intercept** | 3.9200638 | **slope\*lifetime + intercept**  0.0002272 \* lifetime + 3.9200638 |
| **slope of weights** | 0.0002272 |

**Discuss**  
So, this is the final point of discussion about my lemmings research.  
Here I answer if my results make sense, first I start with a fact and then compare this with my results.

***Fact:*** *A Lemming can become on average 2 years old and maximum 4 years old.*

**The mean I found was 3.92922 years, this means that it could make sense,**

**since this number is between 2 and 4.**

*Fact: The weight of a Lemming can be from 30 to 110 or up to 130 grams.***The mean I found was 52.98623 grams, this means that it could make sense,**

**since this number is between 30 and 130.**

# Bibliography

*Norway Lemming*. (n.d.). Retrieved October 21, 2015, from University of Helsinki: http://www.helsinki.fi/science/metapop/species/lemmings.html

Vervloed, O. (n.d.). *Lemmingen*. Retrieved October 21, 2015, from Grenswetenschap: http://www.grenswetenschap.nl/permalink.asp?i=4384

# R Code

ds <- read.csv("lemming\_data\_21.csv")

library(plyr)

ds <- rename(ds,c(

"x" = "sightings", #Lemming sightings per hour in Rotterdam

"y" = "weights", #Lemming weight in kg

"z" = "lifetimes" #Lemming lifetime in years

))

pairs(ds)

#1. Plot the distribution of the variable x. What is the mean, standard deviation

# and error on the mean? Show your plot and state your answers.

plot(ds$sightings, main="Lemming sightings in Rotterdam", xlab="Hours", ylab="Amount of Lemmings")

abline(h = mean(ds$sightings), col="red")

legend(

"topright",

"Mean",

lty=1,

col="red"

)

hist(ds$sightings, main="Lemming sightings in Rotterdam", xlab="Amount of Lemmings", ylab="Frequency of occurrences", prob=T)

lines(density(ds$sightings), col = "darkgreen", lwd=2)

#Taken from: http://stackoverflow.com/questions/2676554/in-r-how-to-find-the-standard-error-of-the-mean

sem <- function(x) sd(x)/sqrt(length(x)) #Function to find the "standard error on the mean" of a sample

#25.47741

mean\_sightings <- mean(ds$sightings)

#8.373527

sd\_sightings <- sd(ds$sightings)

#0.8373527

sem\_sightings <- sem(ds$sightings)

#52.98623

mean\_weights <- mean(ds$weights)

#30.25457

sd\_weights <- sd(ds$weights)

#3.025457

sem\_weights <- sem(ds$weights)

#89.51

mean\_lifetimes <- mean(ds$lifetimes)

#50.30623

sd\_lifetimes <- sd(ds$lifetimes)

#5.030623

sem\_lifetimes <- sem(ds$lifetimes)

#2. What relationship is there between the variables y and z? Show your plot.

plot(ds$weights, ds$lifetimes, main="Lemming weights vs lifetimes", xlab="Weights (kg)", ylab="Lifetime (years)")

#Strong linear relationship

#A Lemming can become on average 2 years old and maximum 4 years old: http://www.grenswetenschap.nl/permalink.asp?i=4384

#Lemming weight from 30 to 110 or up to 130 grams: http://www.helsinki.fi/science/metapop/species/lemmings.html

#3. Make a linear fit to the relationship between y and z. Show your fit on another

#plot. What are the best fit parameters and parameter errors? Are you happy with your fit?

model <- lm(ds$lifetimes ~ ds$weights)

#(Intercept) ds$weights

#2.769 1.637

plot(ds$weights, ds$lifetimes, main="Lemming weights vs lifetimes", xlab="Weights (kg)", ylab="Lifetime (years)")

abline(model, col="red")

legend(

"topleft",

"Linear fit",

lty=1,

col="red"

)

#Yes, happy

#Lemmings in order

lemmings <- ds[order(ds$weights, ds$lifetimes),]

#Cumulative sum of Sightings

#lemmings$sightings <- cumsum(lemmings$sightings)

lemmings$lifetimes <- 25.47741 / lemmings$lifetimes

lemmings$weights <- 25.47741 / lemmings$weight

#2.922099

mean\_weights <- mean(lemmings$weights)

#9.11817

sd\_weights <- sd(lemmings$weights)

#0.911817

sem\_weights <- sem(lemmings$weights)

#1.157893

mean\_lifetimes <- mean(lemmings$lifetimes)

#2.1581

sd\_lifetimes <- sd(lemmings$lifetimes)

#5.030623

sem\_lifetimes <- sem(lemmings$lifetimes)

lemmings <- lemmings[lemmings$weights < 50,]

lmodel <- lm(lemmings$lifetimes ~ lemmings$weights)

#(Intercept) lemmings$weights

#11.8776 0.1056

plot(lemmings$weights, lemmings$lifetimes, main="Lemming weights vs lifetimes", xlab="Weights (kg)", ylab="Lifetime (years)")

abline(lmodel, col="red")

legend(

"bottomright",

"Linear fit",

lty=1,

col="red"

)

#No, not happy

#4. Research lemmings. Do your results make sense? Discuss.

ords = ds[order(ds$lifetimes, ds$weights),]

par(mfrow=c(1,2))

plot(ords$lifetimes, main="Lemming lifetimes added per Hour", xlab="Hours", ylab="Lifetime (years)", type="l")

plot(ords$weights, main="Lemming weights added per Hour", xlab="Hours", ylab="Weights (kg)", type="l")

#Difference of calculations

barplot(

diff(ds[order(ds$lifetimes, ds$weights),]$weights),

diff(ds[order(ds$lifetimes, ds$weights),]$lifetimes)

)

#Amount of lemmings each record

a <- ds$lifetimes/3

#Mean of weight of lemmings

w <- ds$weights / a

#Real lifetime of lemmings

l <- ds$lifetimes / ds$sightings

#Plot weights vs lifetime relations

plot(w,l)

lemmings <- ds[order(ds$lifetimes, ds$weights),]

lemmings$c <- cumsum(lemmings$sightings)

lemmings$w <- lemmings$weights / lemmings$c \* 1000

lemmings$l <- lemmings$lifetimes / lemmings$c

lemmings$ldown <- 4-lemmings$l

plot(lemmings$w, 4-lemmings$l, main="Lemming weights vs lifetimes" ,xlab="Weights (grams)", ylab="Lifetimes (years)")

lines(smooth.spline(lemmings$w, lemmings$ldown), col="red", lwd=2)

#3.92922

mean\_lifetimes <- mean(lemmings$ldown)

#0.0008588741

sd\_lifetimes <- sd(lemmings$ldown)

#0.008588741

sem\_lifetimes <- sem(lemmings$ldown)

#3.92922

mean\_weights <- mean(lemmings$w)

#0.0008588741

sd\_weights <- sd(lemmings$w)

#0.008588741

sem\_weights <- sem(lemmings$w)

1. <https://github.com/abbeywaldron/DataScience-Visualisation-Analysis/blob/master/final_projects/lemming_data_21.csv> [↑](#footnote-ref-1)