Leptospirosis in Kazakhstan: Exercise in R

Otar Chokoshvili

30 July, 2024

Table of Contents

# 1 Install and load packages in R.

## 1.1 To conduct this excercise in R you need to install the following packages:

* “tidyverse”
* “knitr”,
* “readxl”,
* “lubridate”,
* “gtsummary”,
* “flextable”

### 1.1.1 Install the required packages, general reccomendations

Packages are installed only once in R, but you need to loed them every time you start new session in R.

**To install package using RStudio you need to use following menu.**

* To install package you should write the package name in quotation marks (’ ’ or ” “), in parenthesis in the *install.packages()* function.
* Make sure the function includes the statement *dependencies=TRUE*.
* Here is the example of code to install *tidyverse* package: *install.packages(‘gtsummary’, dependencies=TRUE)*.
* This is the way you install packages one by one.
* After installation packages you need to load them using *library()* function.
* You should pass the name on package between parenthesis.
* Your code to load package *tidyverse* will look like *library(tidyverse)*

### 1.1.2 Use this code to install and load the required packages simultaneously

knitr::opts\_chunk$set(comment=NA, prompt=TRUE, out.width=1050, fig.height=8, fig.width=8)  
## istall and load packages that needed  
ipak <- function(pkg){  
 new.pkg <- pkg[!(pkg %in% installed.packages()[, "Package"])]  
 if (length(new.pkg))  
 install.packages(new.pkg, dependencies = TRUE)  
 sapply(pkg, require, character.only = TRUE)  
}  
# indicaete package names  
packages <- c("tidyverse", "dplyr","readxl","lubridate","gtsummary", "knitr", "Rcmdr", "flextable", "cardx")  
  
# use created function ipak and pass packages you wont to install and load  
ipak(packages)

## tidyverse dplyr readxl lubridate gtsummary knitr Rcmdr flextable   
## TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE   
## cardx   
## TRUE

# 2 Introduction

## 2.1 Learning Objectives

After completing this exercise, the student should be able to:

* List the components of descriptive epidemiology.
* Given data from a surveillance system or field investigation, use descriptive statistics, tables. graphs, and maps to summarize the descriptive epidemiology.
* Apply computer programs (R and its libraries, including R Commander) to compile the raw data into summary tables, graphs.
* Identify and interpret patterns in the summary statistics, tables and graphs.
* Compile descriptive epidemiologic findings into a working hypothesis.
* Develop a One Health approach for further investigation of this outbreak.

## 2.2 Leptospirosis as a disease

Leptospirosis is a zoonotic disease with epidemic potential, especially after a heavy rainfall, caused by a bacterium called Leptospira. Leptospirosis is caused by any of 10 pathogenic spirochete species of the genus Leptospira. Clinical leptospirosis disease in humans features high fever, severe headache, chills, and muscle aches. Less frequent features may include jaundice (yellow skin and eyes), red eyes, abdominal pain, diarrhea, or a rash. In some cases, infections may occur without symptoms. The diagnosis should be confirmed with laboratory testing of blood or urine. Two reliable serologic tests are available to diagnose Leptospirosis: 1) the microagglutination test (MAT) and 2) the IgM ELISA tests. These become positive 1 to 2 weeks after onset of symptoms. Early diagnosis can be made using darkfield microscopy of fresh blood to visualize the active leptospiral spirochetes, but this requires a well-practiced microscopist. Culture requires specialized media and needs 3 months to complete. For details of other confirmatory tests see the accompanying reference.

Leptospirosis is a zoonosis affecting over 160 mammalian species. Leptospires have been found in domestic livestock (cattle, pigs, horses, sheep, goats, and dogs), pests (rats, mice, and other rodents), and many wild mammals. Leptospira species have many serovars which are adapted to specific non-human mammals. Mammals infected with an adapted strain maintain the infection for months or years without disease while excreting leptospires in their urine. Hence, these mammals are considered to be maintenance or reservoir hosts. Humans are exclusively accidental hosts. Excretion from the urine usually ceases within two weeks after the acute disease ends. The infection is not spread from person-to-person except in a few rare incidents.

Humans become infected through direct contact with infected mammals or through water, food, or soil (usually muddy or wet) containing urine from infected mammals. Exposure may happen from swallowing contaminated water or food or through contact of broken skin or mucosal surfaces with water or soil. Outbreaks and seasonal increases typically follow heavy rains that wash contaminated soil or water into streams, ponds, and other collections of surface water. Pathogenic leptospires do not multiply in water but can survive for months in aquatic environments. The incubation period between a person’s exposure to a contaminated source and becoming sick is usually from 5 to 14 days (average 9-10 days). Rarely, the onset of infection may occur from 2 days to 4 weeks after exposure.

## 2.3 Outbreak of Leptospirosis in Kenkolat village in East Kazakhstan

An outbreak of suspected leptospirosis occurred in Kenkolat village in East Kazakhstan, during August 2004. Kenkolat had a population of 319 people living in approximately 70 houses. Villagers raised livestock, particularly cattle (bovines). Kenkolat lay alongside a permanent stream in hilly steppe land. The climate classification is Dfb (Humid continental, no dry season, with warm summers). Monthly rainfall for the area averages 52 mm in July, 30 mm in August and 540 mm annually. The mean daily temperature is 25C in July and 23C in August.

An epidemiologic team travelled to Kenkolat from August 23 to 27, 2004. The outbreak investigation team collected clinical and epidemiological data from leptospirosis cases reported from the local public health authorities.

**Question 1.** List the variables you will require to measure the values of descriptive epidemiology of this outbreak?

**Answer 1.** Descriptive epidemiology should include information on the disease (clinical features), time, place, and person. Variables to assess might include:

Table 1. List of variables for Leptospirosis outbreak.

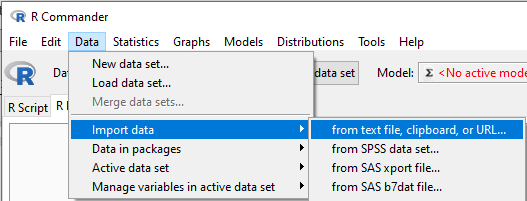
| Disease | Time | Place | Person |
| --- | --- | --- | --- |
| Fever | Onset | Home | Age |
| Headache | Hospitalization | Workplace or school | Sex |
| Chills | Death | Possible sources | Occupation |
| Muscle aches | Possible causative factors |  | Contact with mammals |
| Vomiting | Possible controlling factors |  |  |
| Diarrhea |  |  |  |
| Jaundice |  |  |  |
| Red eyes |  |  |  |
| Rash |  |  |  |
| Lab tests including confirmation |  |  |  |

# 3 Data management steps.

## 3.1 Import data into R.

**Using Rcmdr point-and-click:**

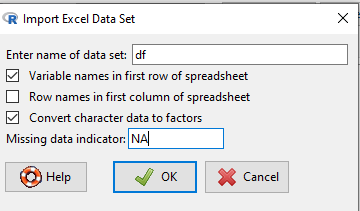
Data -> Import data -> From text file…



And the dialog box opens.

In the dialogue window that opens:

* Enter name for data set: leptodb
* Convert character variables to factors: Uncheck
* Missing: NA
* Field separator: Select commas [,]
* Decimal-point character: [.]
* Click OK



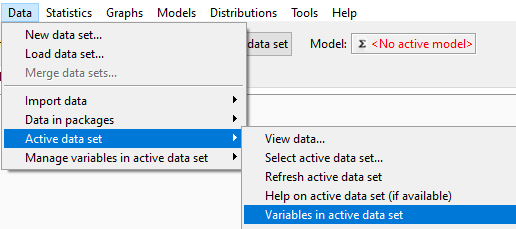
**Import data using R code.**

> ## Import data from excel file.  
> df <- read\_xlsx("data/LeptospDBclean.xlsx", sheet = "LeptoDB", na = "NA")

### 3.1.1 Check the names of variables in the data set.

**Using Rcmdr point-and-click:**

Data -> Active data sets -> Variables in active data set



**Using R code :**

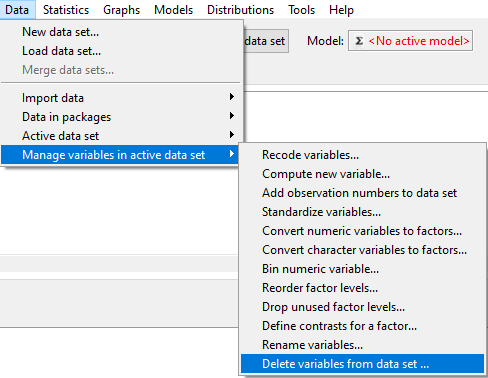
> ## Check names of variables in the data set  
> names(df)

[1] "id" "status" "sex"   
 [4] "dob" "address" "occupation"   
 [7] "profsn" "drinkwaterspring" "drinkwaterwell"   
[10] "homeusedwaterspring" "homeusedwaterwell" "homeusedwaterriver"   
[13] "swiminwaterpond" "swiminwatertype" "swimdate"   
[16] "ownedgoat" "owndesheep" "ownedcow"   
[19] "ownedhoarse" "owneddog" "ownedcat"   
[22] "ownedpig" "contactgoataug6" "contactsheepaug6"   
[25] "contactcowaug6" "contacthoarseaug6" "contactdogaug6"   
[28] "contactcataug6" "contactpigaug6" "wildanimaonctactaug6"   
[31] "fishingaug6" "fishingaug6place" "ratinyard"   
[34] "ratinhome" "contactwithillpersons" "dateofsymptoms"   
[37] "dateofhospitalization" "datediagnosis" "outcome"   
[40] "headache" "fever" "chils"   
[43] "myalgia" "myalgiaw" "calfmusc"   
[46] "eyered" "lappet" "mregdt"   
[49] "liveren" "livrtnd" "mcpbldlp"

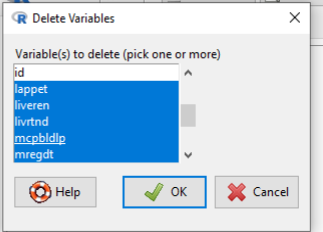
### 3.1.2 Delete variables form data set.

**Using Rcmdr point-and-click:**

Data -> Manage variables in active data set -> Delete variables in active data set



* Then you need to select variable names in the window to delete.



* Click Ok.
* Your selected variables are deleted.

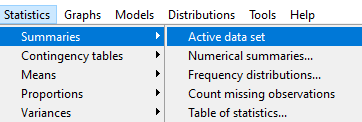
**Using R code:**

> ## Delete variables  
> df <- df %>%   
+ select(-c(calfmusc, eyered, lappet, mregdt, liveren, livrtnd, mcpbldlp))

### 3.1.3 View the summary of your data set.

**Using Rcmdr point-and-click:**

Statistics -> Summaries -> Active data set.



**Using R code:**

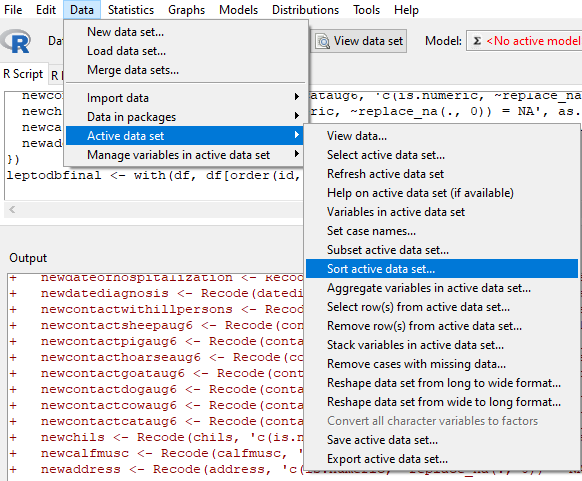
> ## see the summary statistics of data set.   
> summary(df)

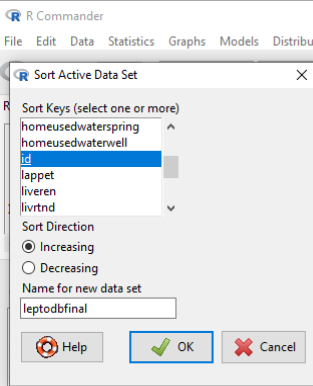
id status sex   
 Min. : 1.00 Length:150 Length:150   
 1st Qu.: 38.25 Class :character Class :character   
 Median : 75.50 Mode :character Mode :character   
 Mean : 75.66   
 3rd Qu.:112.75   
 Max. :151.00   
   
 dob address occupation   
 Min. :1920-01-01 00:00:00 Length:150 Length:150   
 1st Qu.:1957-02-19 12:00:00 Class :character Class :character   
 Median :1976-09-19 00:00:00 Mode :character Mode :character   
 Mean :1971-09-19 20:19:12   
 3rd Qu.:1986-08-07 18:00:00   
 Max. :2001-08-16 00:00:00   
   
 profsn drinkwaterspring drinkwaterwell homeusedwaterspring  
 Length:150 Length:150 Length:150 Length:150   
 Class :character Class :character Class :character Class :character   
 Mode :character Mode :character Mode :character Mode :character   
   
   
   
   
 homeusedwaterwell homeusedwaterriver swiminwaterpond swiminwatertype   
 Length:150 Length:150 Length:150 Length:150   
 Class :character Class :character Class :character Class :character   
 Mode :character Mode :character Mode :character Mode :character   
   
   
   
   
 swimdate ownedgoat owndesheep   
 Min. :2004-07-15 00:00:00 Min. :0.0000 Min. :0.0000   
 1st Qu.:2004-07-20 00:00:00 1st Qu.:1.0000 1st Qu.:1.0000   
 Median :2004-07-25 00:00:00 Median :1.0000 Median :1.0000   
 Mean :2004-07-24 08:00:00 Mean :0.8367 Mean :0.8639   
 3rd Qu.:2004-07-29 00:00:00 3rd Qu.:1.0000 3rd Qu.:1.0000   
 Max. :2004-08-02 00:00:00 Max. :1.0000 Max. :1.0000   
 NA's :147 NA's :3 NA's :3   
 ownedcow ownedhoarse owneddog ownedcat   
 Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. :0.0000   
 1st Qu.:1.0000 1st Qu.:1.0000 1st Qu.:1.0000 1st Qu.:0.0000   
 Median :1.0000 Median :1.0000 Median :1.0000 Median :1.0000   
 Mean :0.9796 Mean :0.8844 Mean :0.9252 Mean :0.6395   
 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000   
 Max. :1.0000 Max. :1.0000 Max. :1.0000 Max. :1.0000   
 NA's :3 NA's :3 NA's :3 NA's :3   
 ownedpig contactgoataug6 contactsheepaug6 contactcowaug6   
 Min. :0 Min. :0.0000 Min. :0.0000 Min. :0.0000   
 1st Qu.:0 1st Qu.:1.0000 1st Qu.:1.0000 1st Qu.:1.0000   
 Median :0 Median :1.0000 Median :1.0000 Median :1.0000   
 Mean :0 Mean :0.7687 Mean :0.7891 Mean :0.9116   
 3rd Qu.:0 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000   
 Max. :0 Max. :1.0000 Max. :1.0000 Max. :1.0000   
 NA's :3 NA's :3 NA's :3 NA's :3   
 contacthoarseaug6 contactdogaug6 contactcataug6 contactpigaug6  
 Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. :0   
 1st Qu.:1.0000 1st Qu.:1.0000 1st Qu.:0.0000 1st Qu.:0   
 Median :1.0000 Median :1.0000 Median :1.0000 Median :0   
 Mean :0.8027 Mean :0.8435 Mean :0.5918 Mean :0   
 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:0   
 Max. :1.0000 Max. :1.0000 Max. :1.0000 Max. :0   
 NA's :3 NA's :3 NA's :3 NA's :3   
 wildanimaonctactaug6 fishingaug6 fishingaug6place ratinyard   
 Min. :0.00000 Min. :0.00000 Length:150 Min. :0.0000   
 1st Qu.:0.00000 1st Qu.:0.00000 Class :character 1st Qu.:0.0000   
 Median :0.00000 Median :0.00000 Mode :character Median :0.0000   
 Mean :0.06122 Mean :0.08844 Mean :0.4082   
 3rd Qu.:0.00000 3rd Qu.:0.00000 3rd Qu.:1.0000   
 Max. :1.00000 Max. :1.00000 Max. :1.0000   
 NA's :3 NA's :3 NA's :3   
 ratinhome contactwithillpersons dateofsymptoms   
 Min. :0.0000 Min. :0.0000 Min. :2004-07-20 00:00:00.00   
 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:2004-08-05 00:00:00.00   
 Median :0.0000 Median :0.0000 Median :2004-08-07 00:00:00.00   
 Mean :0.2993 Mean :0.2993 Mean :2004-08-07 01:39:18.61   
 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:2004-08-10 00:00:00.00   
 Max. :1.0000 Max. :1.0000 Max. :2004-08-17 00:00:00.00   
 NA's :3 NA's :3 NA's :121   
 dateofhospitalization datediagnosis outcome   
 Min. :2004-08-06 00:00:00.00 Min. :2004-08-16 00:00:00 Min. :1.000   
 1st Qu.:2004-08-09 00:00:00.00 1st Qu.:2004-08-16 00:00:00 1st Qu.:1.000   
 Median :2004-08-11 00:00:00.00 Median :2004-08-17 00:00:00 Median :1.000   
 Mean :2004-08-10 06:37:14.48 Mean :2004-08-18 08:00:00 Mean :1.138   
 3rd Qu.:2004-08-12 00:00:00.00 3rd Qu.:2004-08-19 00:00:00 3rd Qu.:1.000   
 Max. :2004-08-17 00:00:00.00 Max. :2004-08-24 00:00:00 Max. :3.000   
 NA's :121 NA's :123 NA's :121   
 headache fever chils myalgia   
 Min. :1.000 Min. :37.30 Min. :1 Min. :0.0000   
 1st Qu.:1.000 1st Qu.:38.00 1st Qu.:1 1st Qu.:1.0000   
 Median :1.000 Median :38.70 Median :1 Median :1.0000   
 Mean :1.586 Mean :38.62 Mean :1 Mean :0.7586   
 3rd Qu.:2.000 3rd Qu.:39.20 3rd Qu.:1 3rd Qu.:1.0000   
 Max. :4.000 Max. :40.20 Max. :1 Max. :1.0000   
 NA's :121 NA's :121 NA's :121 NA's :121   
 myalgiaw   
 Length:150   
 Class :character   
 Mode :character

### 3.1.4 Rename the data set and sort it by id.

**Using Rcmdr point-and-click:**

Statistics -> Active data set -> Sort active data set.



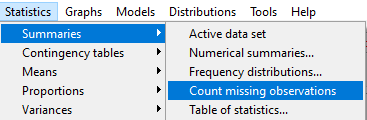


**Using R code:**

> ## Rename the data set and sort it by id.  
>   
> leptodbfinal <- df[order(df$id),]

### 3.1.5 Count missing observations fro all variables in data set.

**Using Rcmdr point-and-click:**



**Using R code:**

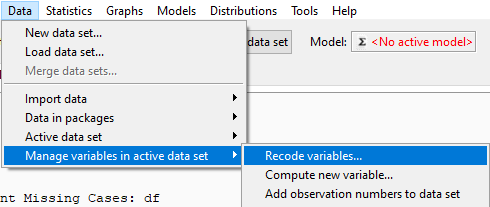
> # check numver of missing values in for each variables.  
> colSums(is.na(leptodbfinal))

id status sex   
 0 0 0   
 dob address occupation   
 0 0 0   
 profsn drinkwaterspring drinkwaterwell   
 149 3 3   
 homeusedwaterspring homeusedwaterwell homeusedwaterriver   
 3 3 3   
 swiminwaterpond swiminwatertype swimdate   
 3 93 147   
 ownedgoat owndesheep ownedcow   
 3 3 3   
 ownedhoarse owneddog ownedcat   
 3 3 3   
 ownedpig contactgoataug6 contactsheepaug6   
 3 3 3   
 contactcowaug6 contacthoarseaug6 contactdogaug6   
 3 3 3   
 contactcataug6 contactpigaug6 wildanimaonctactaug6   
 3 3 3   
 fishingaug6 fishingaug6place ratinyard   
 3 137 3   
 ratinhome contactwithillpersons dateofsymptoms   
 3 3 121   
dateofhospitalization datediagnosis outcome   
 121 123 121   
 headache fever chils   
 121 121 121   
 myalgia myalgiaw   
 121 145

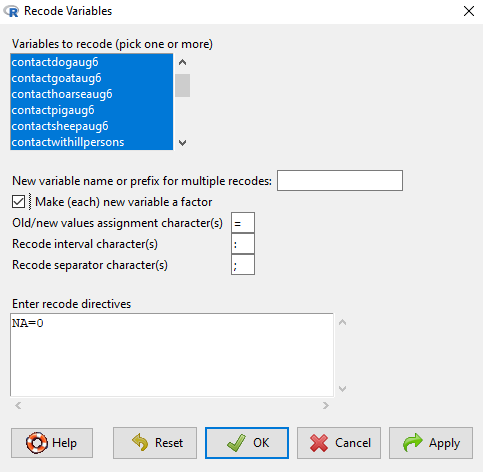
### 3.1.6 Recode variables with missing values (NA) to 0:

**Using Rcmdr point-and-click:**

Data -> Manage variables in active data set -> Recode variables



* in the recode variables window select variables you wont to recede and in the code write **NA=0**



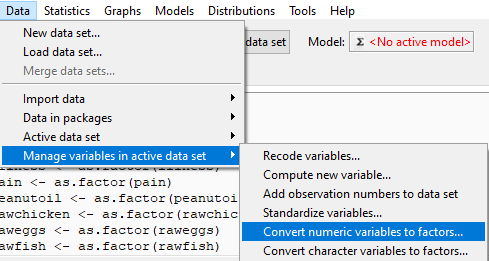
**Using R code:**

> ## replace all NA values to 0 for sleected variables  
> leptodbfinal <- leptodbfinal %>%   
+ #mutate\_if(is.numeric, ~replace\_na(., 0))   
+ mutate\_at(c("contactwithillpersons", "contactsheepaug6", "contactpigaug6", "contacthoarseaug6",  
+ "contactgoataug6", "contactdogaug6", "contactcowaug6", "contactcataug6"), ~replace\_na(., 0))

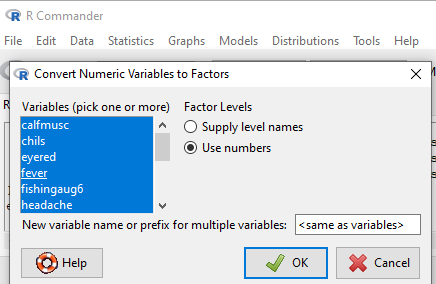
### 3.1.7 Recode variables from numeric to factor:

**Using Rcmdr point-and-click:**

Data -> Manage variables in active data set -> Convert numeric variables to factors



* In the dialogue window that opens up, select the above variables (hold the “shift” button to select more than one).
* Check that the factor level is set to “use names”.
* Click OK



**Using R code:**

> leptodbfinal <- leptodbfinal %>%   
+ mutate\_if(is.numeric, as.factor)

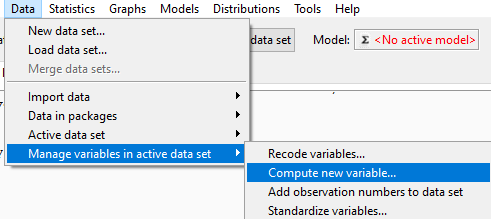
### 3.1.8 Convert variable to date

Some variables in the data set are shown as character but they are actually date variables.

This is the case for variables “dateonset”, “datediagnosis”

**Using Rcmdr point-and-click:**

Data -> Manage variables in active data set -> Compute new variable



-When you click it the new dialogue window pops up

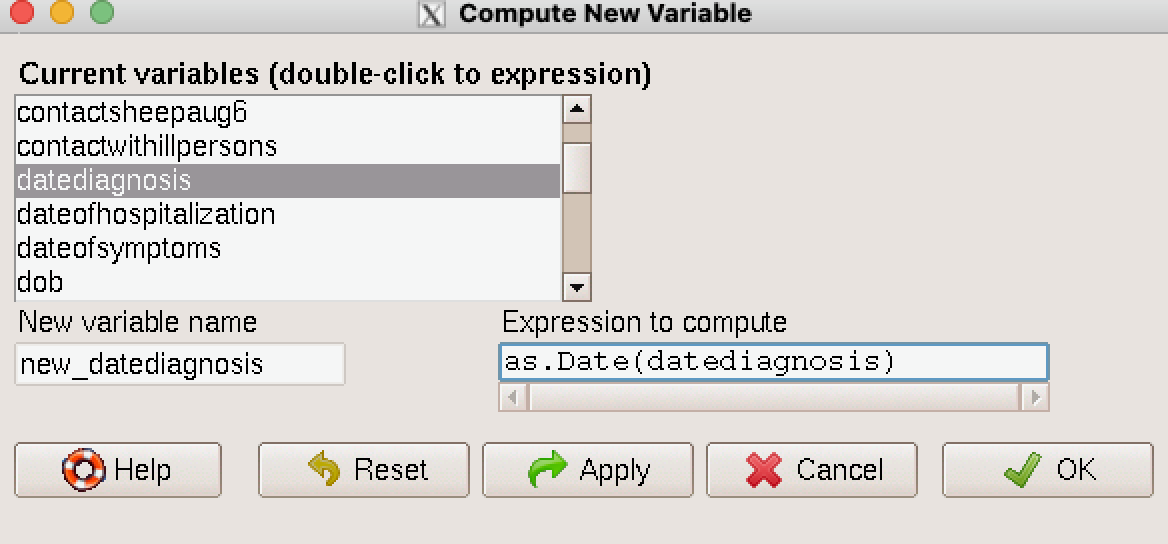
-Type the new variable name: new\_datediagnosis

-In expression to compute type: as.Date(datediagnosis)

-Click OK

You have now created a new date type variable called new\_datediagnosis

You can repeat this same step to create the date variable “new\_dateonset” from variable “dateofsymptoms”



**Using R code:**

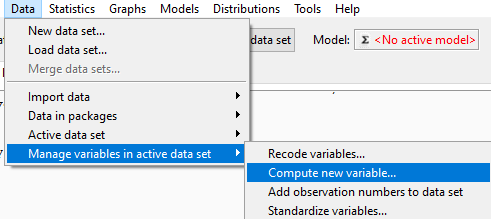
> ## Recode dates using dplyr package  
> leptodbfinal <- leptodbfinal %>%   
+ mutate(new\_datediagnosis = as.Date(datediagnosis),  
+ new\_dateonset = as.Date(dateofsymptoms))

### 3.1.9 Calculate difference between two dates

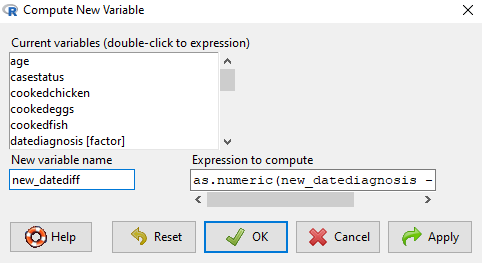
**Using Rcmdr point-and-click:**

* Calculate difference between date of diagnosis and date of onset to detect the delay between diagnosis and onset of disease.

Data -> Manage variables in active data set -> Compute new variable



* In the new variable calcuation window we need to write formula to calculate difference:
* new\_datediff <- as.numeric(new\_datediagnosis - new\_dateonset)



**Using R code:**

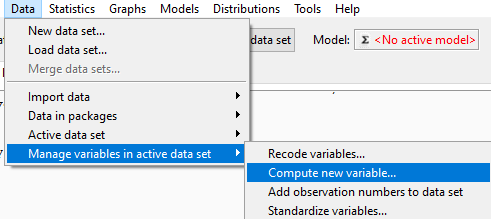
> # calcualate differnce between date of onset and date of diagnosis  
> leptodbfinal$new\_datediff <- as.numeric(leptodbfinal$new\_datediagnosis - leptodbfinal$new\_dateonset)

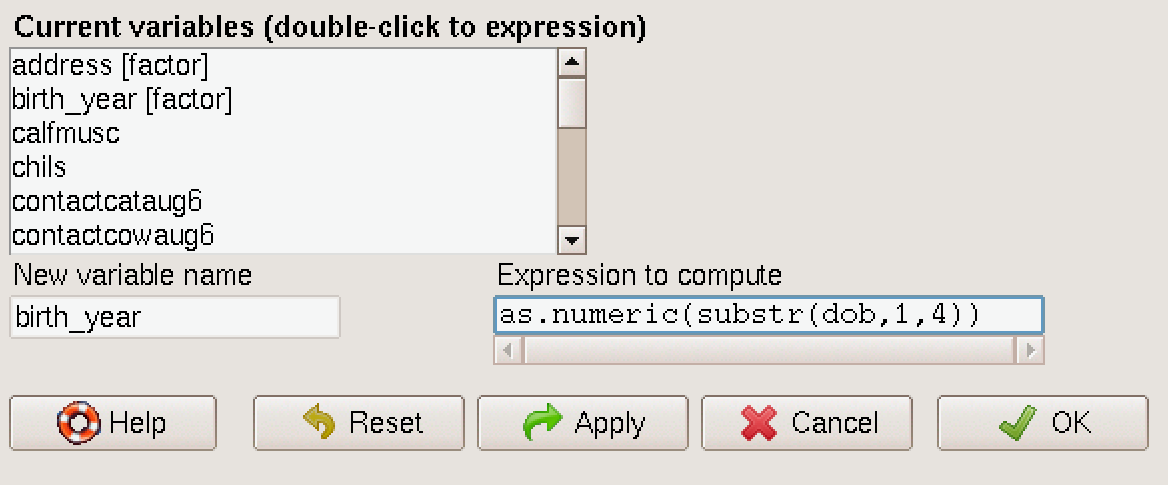
### 3.1.10 Create new categorical variable from numeric variable by recoding it.

#### 3.1.10.1 First calculate year of birth from date of birth variable.

**Using Rcmdr point-and-click:**

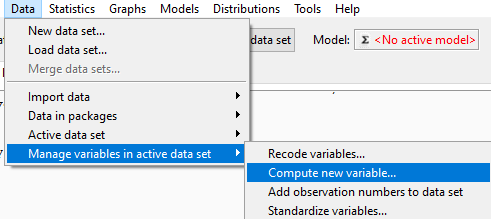
Data -> Manage variables in active data set -> Compute new variable

 - In the new window calculate variable birth\_year

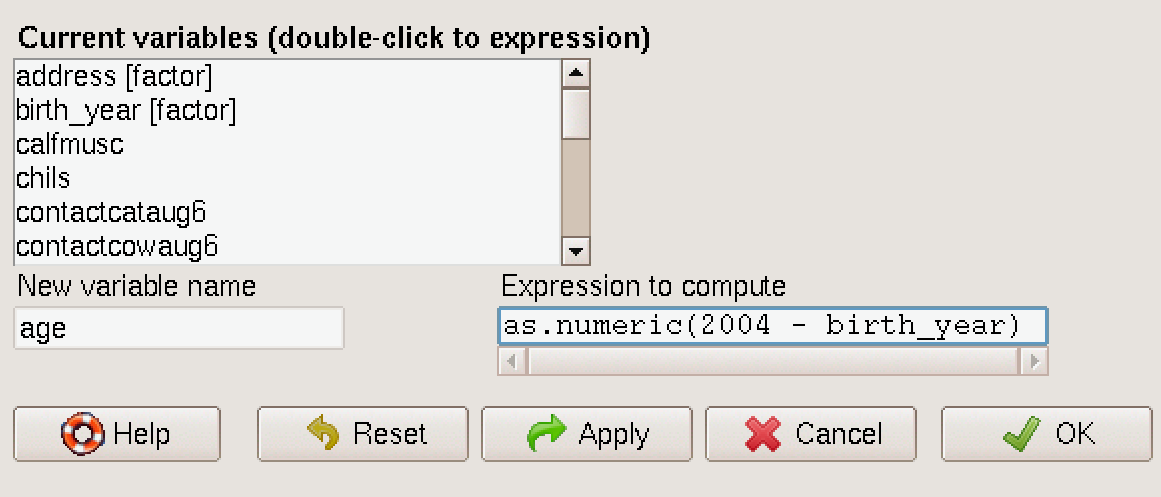


#### 3.1.10.2 Calculate age from year of diagnosis **2004** and year of birth.

Data -> Manage variables in active data set -> Compute new variable

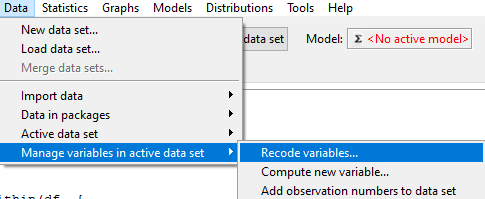


* In the new window calculate variable age



#### 3.1.10.3 Calcualte age groups by recoding age into age group catgories.

Data -> Manage variables in active data set -> Recode variables



* Type the new variable name: new\_age.
* Select checkbox for make new variable a factor.
* Recode age in 3 categoreis, <15, 15-39 and >= 40.
* In recode directives:

**0:14 = "<15"** **15:39 = "15 - 39"** **40:100 = "40+**

* Click OK
* You have now created a new numeric variable called “new\_age”

**Using R code:**

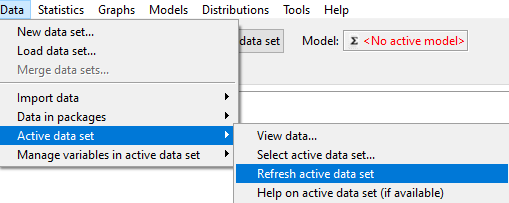
> ## Substract year from date of birth  
> leptodbfinal$birth\_year <- as.numeric(substr(leptodbfinal$dob,1,4))  
>   
> ## Calculate age from 2004 using birth\_year  
> leptodbfinal$age <- as.numeric(2004 - leptodbfinal$birth\_year)  
>   
> ## Calculate new\_age as age groups  
> leptodbfinal$age\_group <- cut(leptodbfinal$age, breaks = c(0,14,39,100), labels = c("<15", "10-39", "40+"))  
>   
> ## Calcualte year of birth, age and age groups using dplyr package in one code.  
> leptodbfinal <- leptodbfinal %>%   
+ # calculate age variabel from date of birth variable (DOB)  
+ mutate(age = as.numeric(2004 - year(dob)),  
+ age\_group = cut(age, breaks = c(0,14,39,100), labels = c("<15", "10-39", "40+")))

## 3.2 Refresh active data set

In Rcmdr after you create new variables you need to refresh.

**Using Rcmdr point-and-click:**

Data -> Active data set -> Refresh active data set

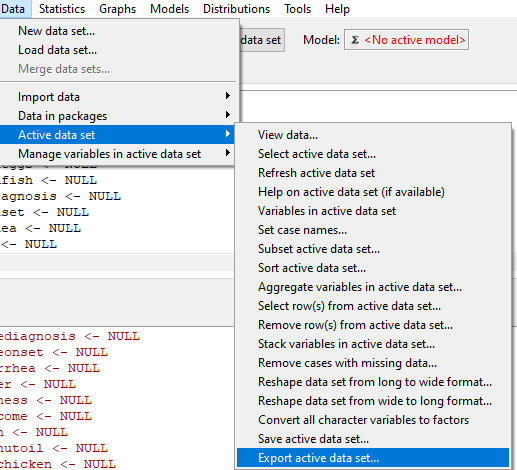


**This is only for R commander, R code does not require to refresh active data set !!!**

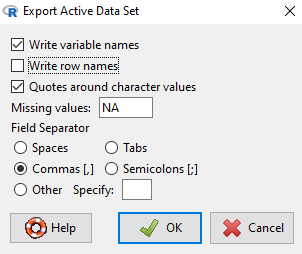
## 3.3 Save and export data into csv format

**Using Rcmdr point-and-click:**

Data -> Active data set -> Export active data set



– Open the window and indicate several paramaters: - In the dialogue window that opens up, uncheck write wrote names - Type in missing value: NA - Select field separator commas[,] - In the next dialogue window that opens up, type in the new file name “leptospirosis\_clean.csv” - Click OK



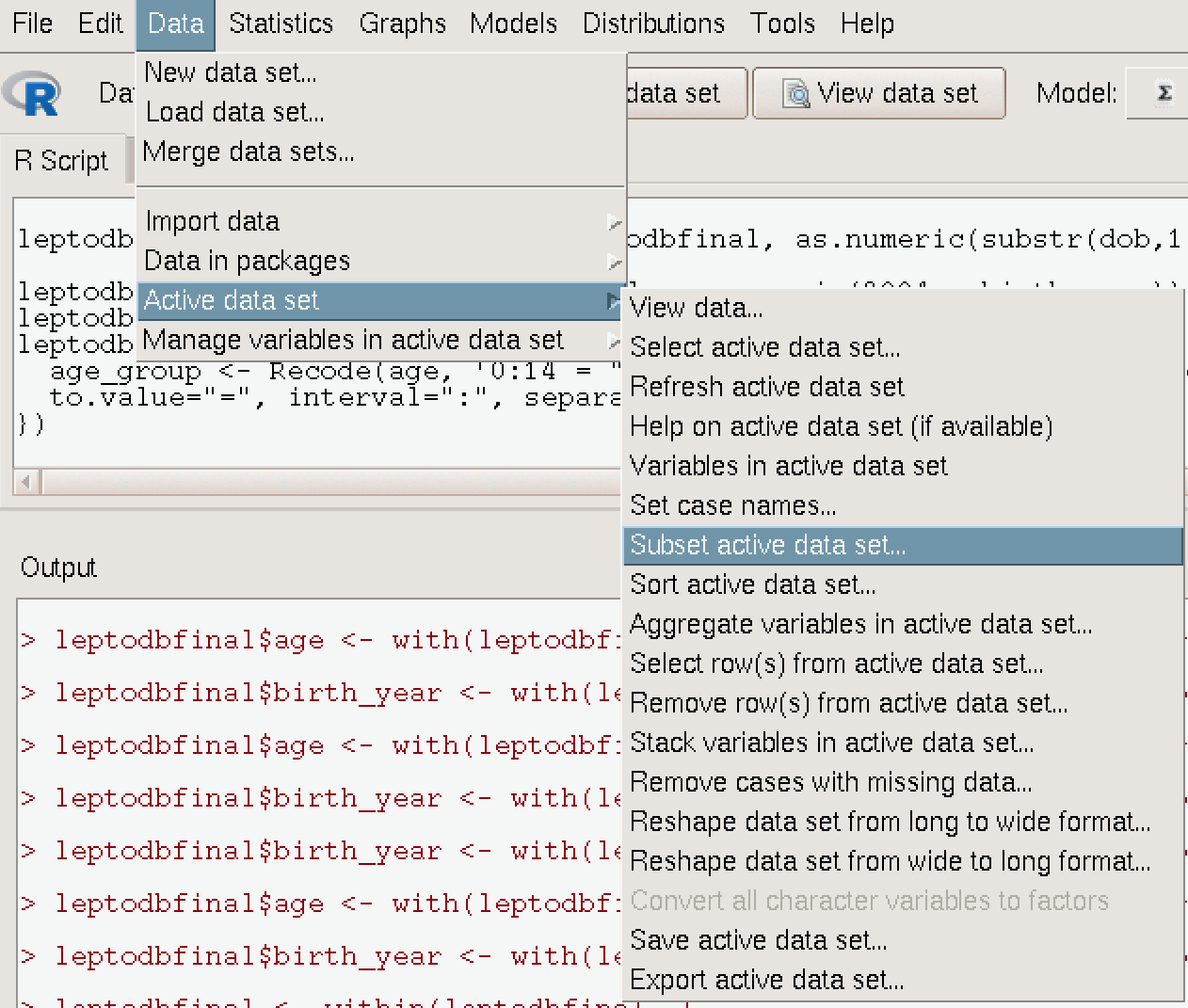
**Using R code:**

> ## export data into csv  
> write.table(leptodbfinal, "output/leptospirosis\_clean.csv", sep=",", col.names=TRUE, row.names=FALSE, na="NA")

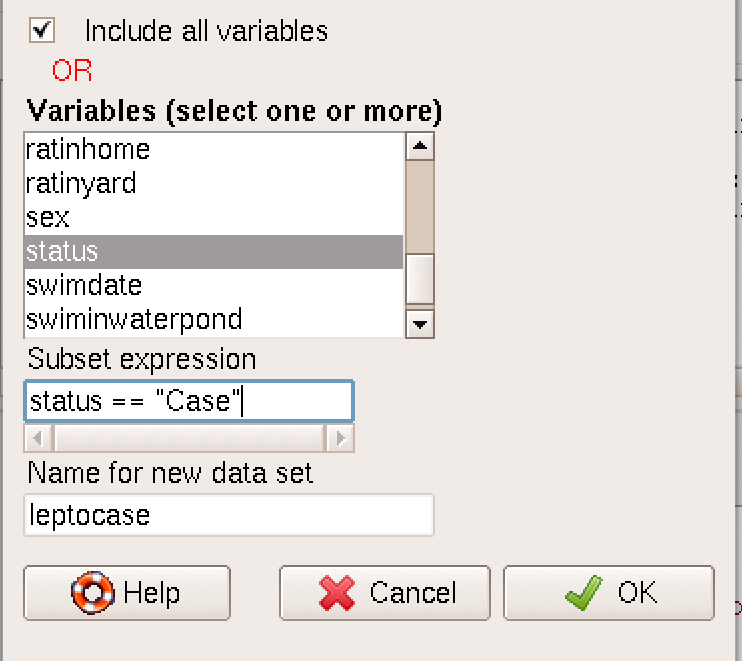
## 3.4 Subset data for Cases only

**Using Rcmdr point-and-click:**

Data -> Active data set -> Refresh active data set -> Subset active data set



* Name your new data file as **leptocase**



**Using R code:**

> ## Select only confirmed cases.  
> leptocase <- leptodbfinal %>% filter(status == "Case")

# 4 Leptospirosis data analysis.

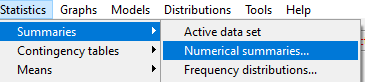
## 4.1 Univariate data analysis

### 4.1.1 Continious data analysis of numeric variables

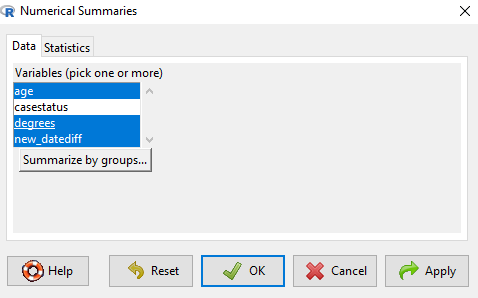
* Do summaries of the numeric variables for age, new\_datediff
* Using Rcmdr point-and-click:

**Using Rcmdr point-and-click:**

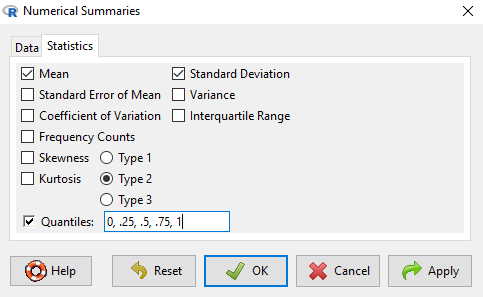
Statistics -> Summaries -> Numerical summaries



* In the dialogue window that opens up, select all variables you want to summarize.



* Click on the statistics tab and ensure that Mean, Standard Deviation, and Quantiles are checked.



* Calculate summary statistics of numeric variables “age”, “new\_datediff” in RStudio.

**Using R code:**

> numSummary(leptodbfinal[,c("age", "new\_datediff"), drop=FALSE], statistics=c("mean", "sd", "quantiles"), quantiles=c(0,.25,.5,.75,1))

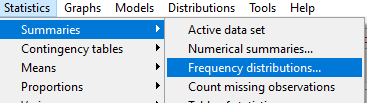
mean sd 0% 25% 50% 75% 100% n NA  
age 32.73333 18.987338 3 18.0 28 47.75 84 150 0  
new\_datediff 10.66667 3.137858 6 8.5 10 12.50 18 27 123

### 4.1.2 Univariable summaries of categorical variables

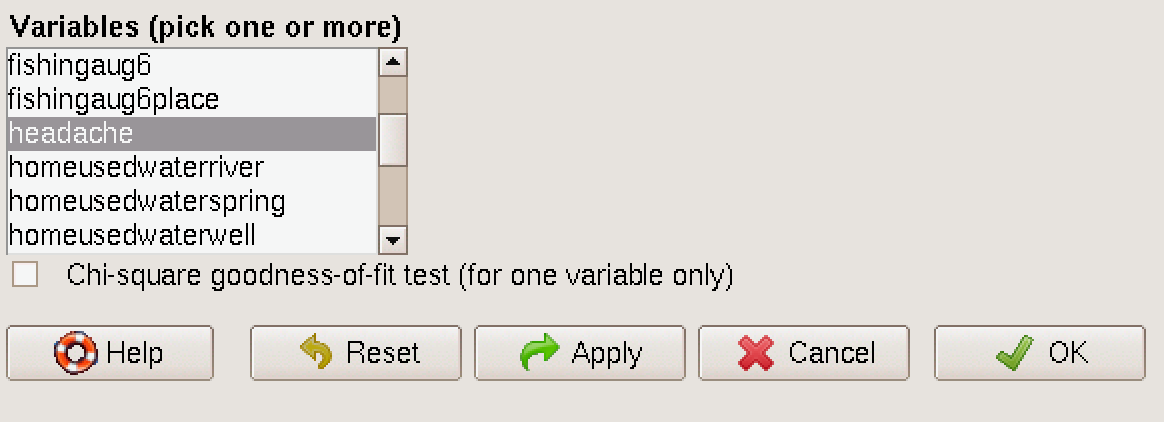
* Create summaries of the categorical variables for symptom variable: headdache, myalgia

**Using Rcmdr point-and-click:**

Statistics -> Summaries -> Frequency distributions



* Select variables for frequency distribution.



**Using R code:**

> ## count categories headache  
> table(leptodbfinal$headache)

1 2 3 4   
17 9 1 2

> ## count proportions of headache  
> round(prop.table(table(leptodbfinal$headache)),2)

1 2 3 4   
0.59 0.31 0.03 0.07

> ## count categories myalgia  
> table(leptodbfinal$myalgia)

0 1   
 7 22

> ## count proportions of myalgia  
> round(prop.table(table(leptodbfinal$myalgia)),2)

0 1   
0.24 0.76

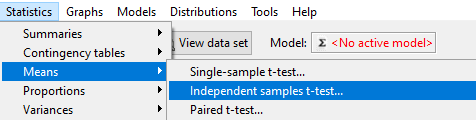
* Repeat same for all variables you need to analize.

## 4.2 Bivariable analysis - t-test

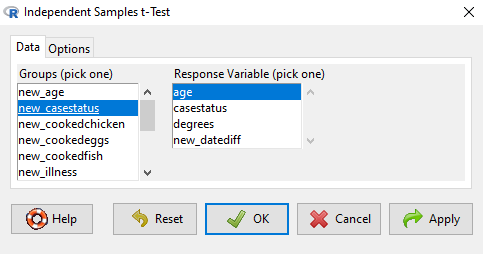
* Create summaries of the numerical variables by category, for example, age by case\_status

**Using Rcmdr point-and-click:**

Statistics -> Means -> Independent samples t-test



* In the dialogue window that opens up, select the groups variable: status
* Select in the response variable: age

 - Calculate age difference usig ttest

**Using R code:**

> t.test(age ~ status, data = leptodbfinal)

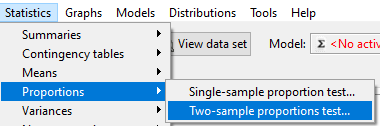
Welch Two Sample t-test  
  
data: age by status  
t = -0.25411, df = 40.786, p-value = 0.8007  
alternative hypothesis: true difference in means between group Case and group Control is not equal to 0  
95 percent confidence interval:  
 -9.282865 7.208200  
sample estimates:  
 mean in group Case mean in group Control   
 31.89655 32.93388

## 4.3 Bivariable analysis - chi-square

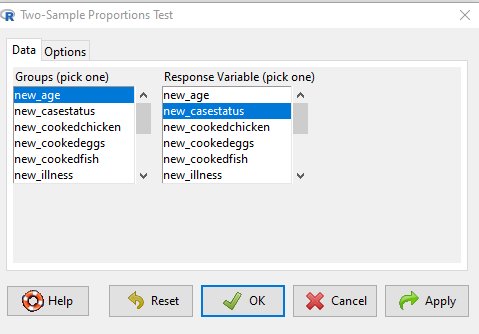
* Create summaries of the categorical variables by category, for example, age\_group by status

**Using Rcmdr point-and-click:**

Statistics -> Proportions -> Two-sample proportions test



* In the dialogue window that opens up, select the groups variable: age\_group
* Select in the response variable: status



* Use RStudio to calculate data.

**Using R code:**

> chiqtbl <- xtabs(~ age\_group + status, data = leptodbfinal)  
> chiqtbl

status  
age\_group Case Control  
 <15 5 18  
 10-39 16 62  
 40+ 8 41

> chisq.test(chiqtbl)

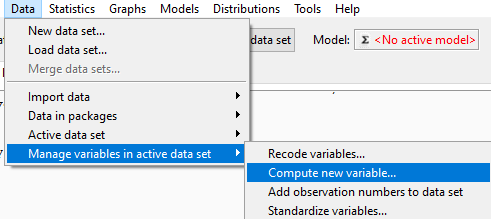
Warning in chisq.test(chiqtbl): Chi-squared approximation may be incorrect

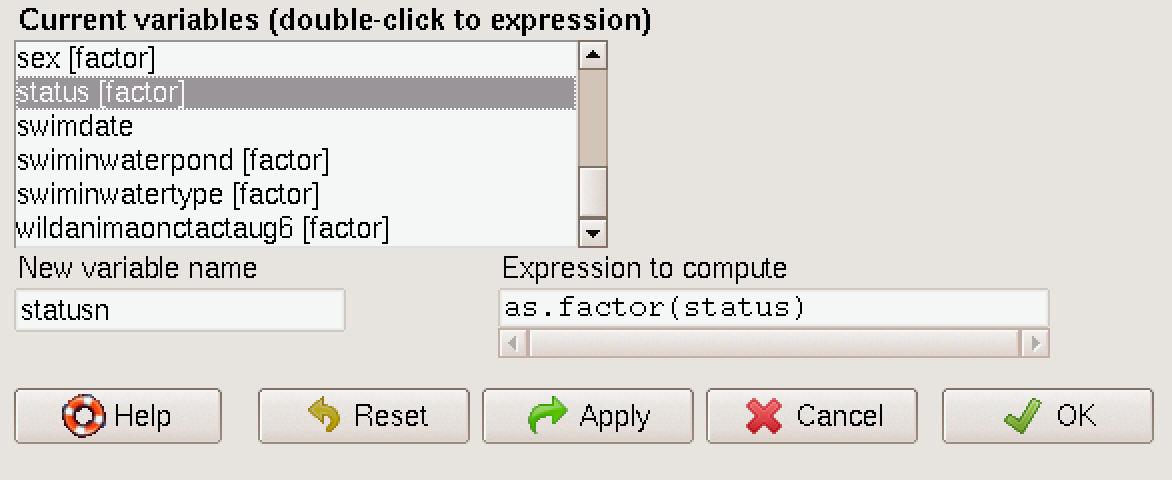
Pearson's Chi-squared test  
  
data: chiqtbl  
X-squared = 0.43899, df = 2, p-value = 0.8029

## 4.4 Bivariable analysis - odds ratio

* FIrst create new variable **statusn** from **status** and make it as factor variable, this is necesary to fit models and calcualate **Odds Ratio**.

**Using Rcmdr point-and-click:**



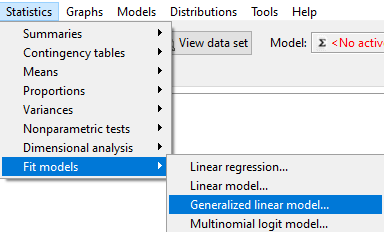


**Using R code:**

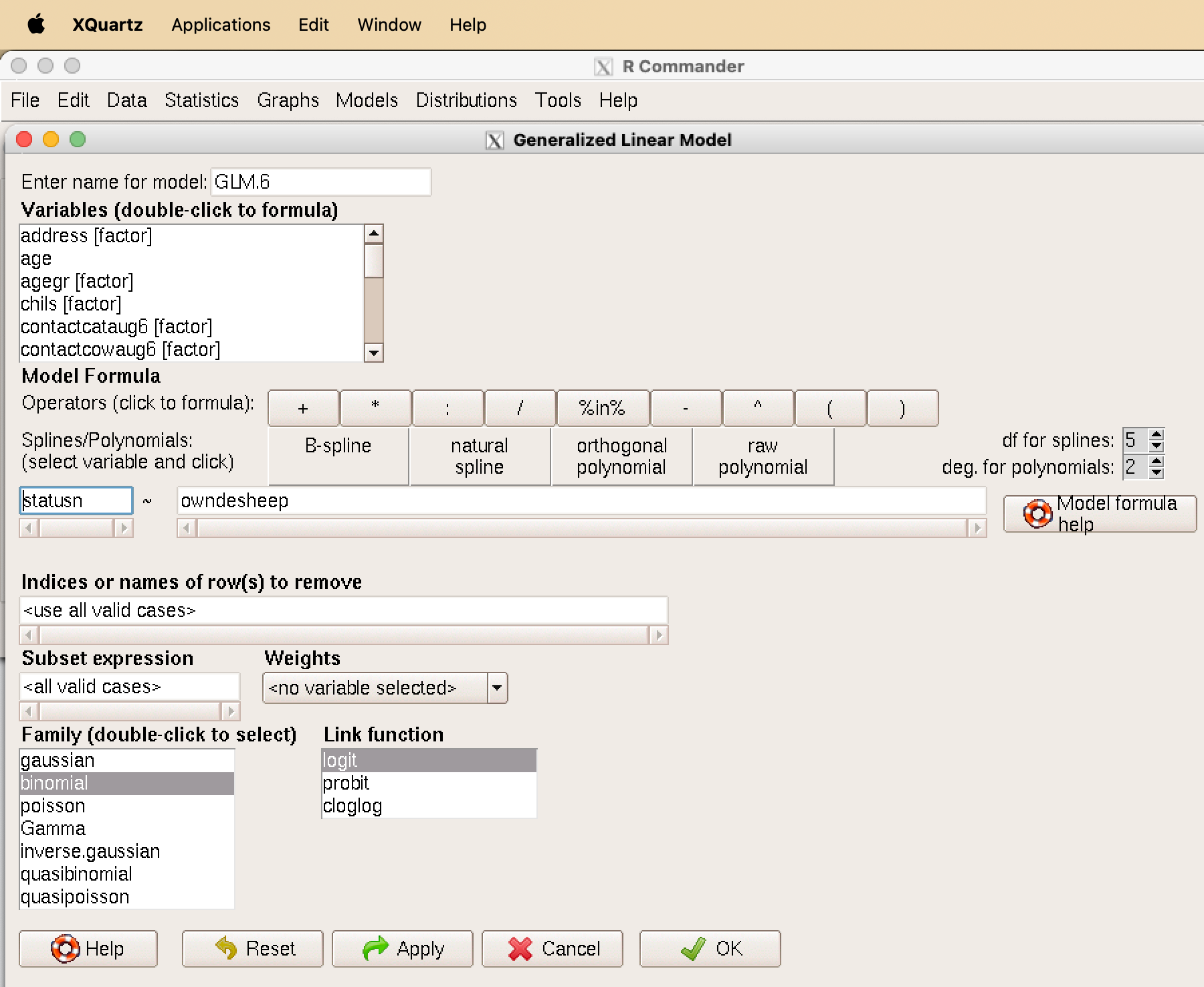
> leptodbfinal <- leptodbfinal %>% mutate(statusn = factor(status))

**Using Rcmdr point-and-click:**

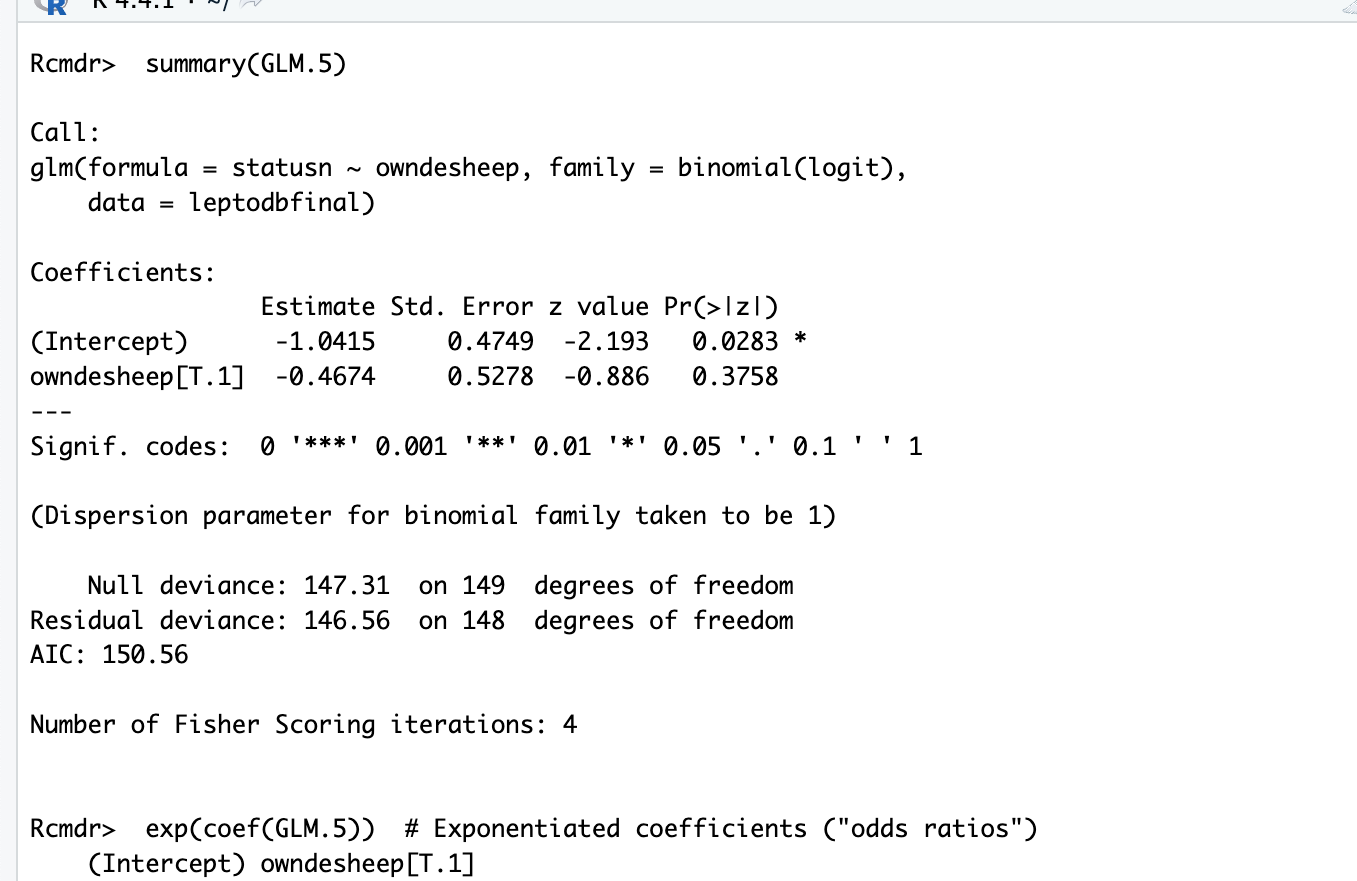
Statistics -> Fit models -> Generalized linear models



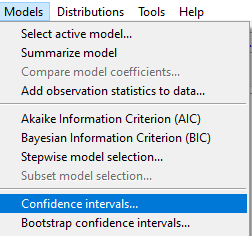
* In the left side of the model insert: statusn
* Select in the response variable: owndsheep
* Ensure that family = binomial and link function = logit are selected



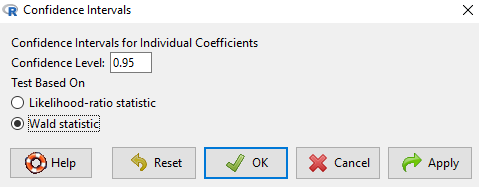
* Resutls



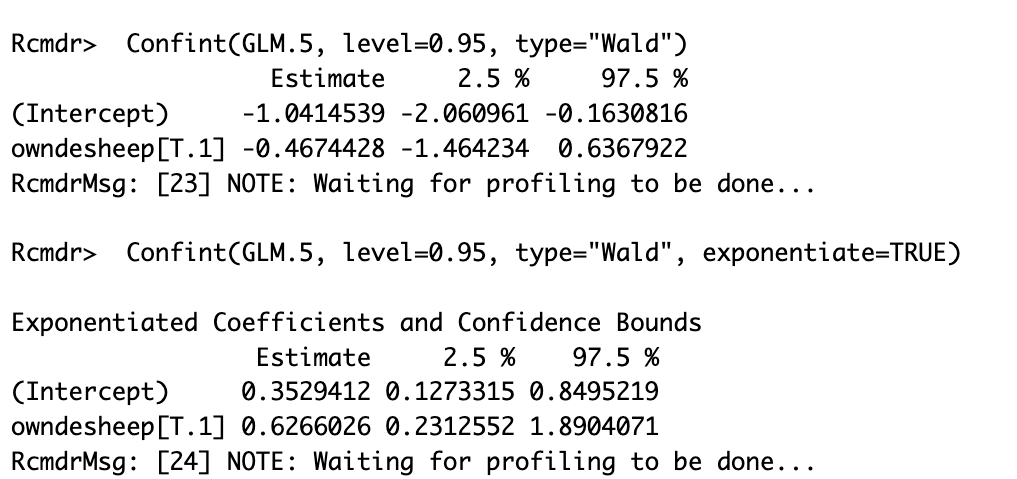
* Next get confidence intervals for our outputs



* In the window you should select



* You get results:



**Using R code:**

> ## Calculate OR using, create model  
> GLMOR <- glm(statusn ~ owndesheep, family=binomial(logit), data=leptodbfinal)  
> ## Calculate summary of OR  
> summary(GLMOR)

Call:  
glm(formula = statusn ~ owndesheep, family = binomial(logit),   
 data = leptodbfinal)  
  
Coefficients:  
 Estimate Std. Error z value Pr(>|z|)   
(Intercept) 1.7346 0.6262 2.770 0.00561 \*\*  
owndesheep1 -0.2257 0.6673 -0.338 0.73517   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
(Dispersion parameter for binomial family taken to be 1)  
  
 Null deviance: 137.18 on 146 degrees of freedom  
Residual deviance: 137.07 on 145 degrees of freedom  
 (3 observations deleted due to missingness)  
AIC: 141.07  
  
Number of Fisher Scoring iterations: 4

> ## Exponentiated coefficients  
> exp(coef(GLMOR))

(Intercept) owndesheep1   
 5.666667 0.797954

> ## Calcualte confidence intervals  
> Confint(GLMOR, level=0.95, type="LR", exponentiate=TRUE)

Exponentiated Coefficients and Confidence Bounds

Estimate 2.5 % 97.5 %  
(Intercept) 5.666667 1.9040507 24.27290  
owndesheep1 0.797954 0.1760366 2.62891

# 5 Descriptive analysis and distribution of symptoms of Leptospirosis in the outbreak data.

* Codes for R to create complex nice looking tables using **GTsummary** package.

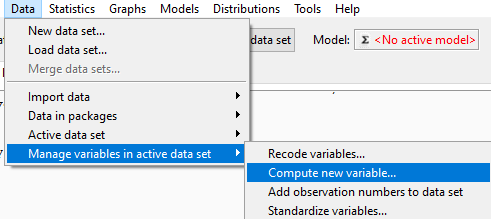
## 5.1 First calculate N of cases (case patients) of Leptospirosis.

* Calculate N of cases of leptospirosis.
* Create new variable nlpcs count N of cases by counting id using formula:
* nlpcs <- NROW(id)

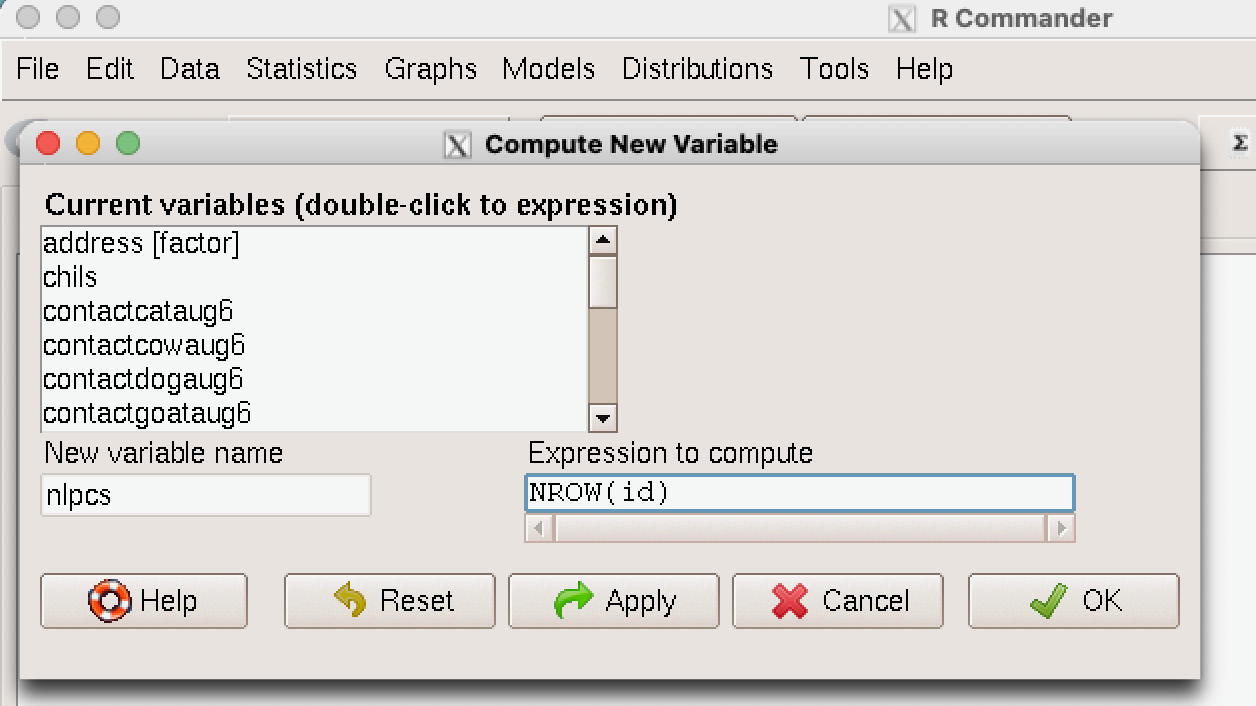
Here we use NROW with UPPERCASE which works with vectors different from nrow.

**Using Rcmdr point-and-click:**

Data -> Manage variables in active data set -> Compute new variable



* In the window **Expression to compute** write code **NROW(id)**



**Using R code:**

> ## calcualte N of cases   
> nlpcs <- NROW(leptocase$id)

> #table for gender by age years  
> leptb1 <- leptocase %>%   
+ select(c(sex, age)) %>%   
+ tbl\_summary(by = sex,  
+ percent = "column",  
+ sort = all\_categorical() ~ "frequency",  
+ statistic = list(  
+ all\_continuous() ~ "{mean} ({sd})",  
+ all\_categorical() ~ "{n} ({p})")  
+ )  
> leptb1 <- as\_flex\_table(leptb1)  
>   
> #table for sex and by age\_groups  
> leptb2 <- leptocase %>%   
+ select(c(sex, age\_group)) %>%   
+ tbl\_summary(by = sex)  
> leptb2 <- as\_flex\_table(leptb2)  
>   
> # table for symptoms  
> leptb3 <- leptocase %>%   
+ select(c(headache, fever, chils, myalgia)) %>%   
+ mutate(headache = fct\_recode(headache, "SEVERE" = "1", "INTERMED" = "2", "MILD" = "3", "NO PAIN" = "4", "NA" = "0"), fever = as.numeric(as.character(fever))) %>%   
+ tbl\_summary()

Warning: There was 1 warning in `mutate()`.  
ℹ In argument: `headache = fct\_recode(...)`.  
Caused by warning:  
! Unknown levels in `f`: 0

> leptb3 <- as\_flex\_table(leptb3)

**Task 1.2** Demographic characteristics of cases by gender and age. Are there more males or females affected? Create frequency distribution of N of cases by gender and detect mean values and standard deviation of age? Calculate mean age and standard deviation by gender?.

**Answer 1.2** Frequency distribution of cases by gender and age are provided in the Table 2.

**Table 2.** Cases of Leptospirosis by age and by gender.

| **Characteristic** | **F**  N = 121 | **M**  N = 171 |
| --- | --- | --- |
| age | 34 (20) | 30 (20) |
| 1Mean (SD) | | |

**Table 3.** Cases of Leptospirosis by age groups and by gender.

| **Characteristic** | **F**  N = 121 | **M**  N = 171 |
| --- | --- | --- |
| age\_group |  |  |
| <15 | 2 (17%) | 3 (18%) |
| 10-39 | 6 (50%) | 10 (59%) |
| 40+ | 4 (33%) | 4 (24%) |
| 1n (%) | | |

**Task 3.1** Create a table that summarizes the clinical features of the 29 cases. What does the distribution suggest to you?

**Answer 3.2** Frequency distribution of symptoms are provided in the Table 4.

**Table 4.** Distribution of leptospurosis symptoms among case patients. Clinical Features of case Patients with Leptospirosis, Kenkolat, East Kazakhstan, August 2004.

| **Characteristic** | **N = 29**1 |
| --- | --- |
| headache |  |
| SEVERE | 17 (59%) |
| INTERMED | 9 (31%) |
| MILD | 1 (3.4%) |
| NO PAIN | 2 (6.9%) |
| fever | 38.70 (38.00, 39.20) |
| chils |  |
| 1 | 29 (100%) |
| myalgia |  |
| 0 | 7 (24%) |
| 1 | 22 (76%) |
| 1n (%); Median (Q1, Q3) | |

# 6 Create plots/Histogram **Epicurve** in R commaneder and R using **ggplot** packages

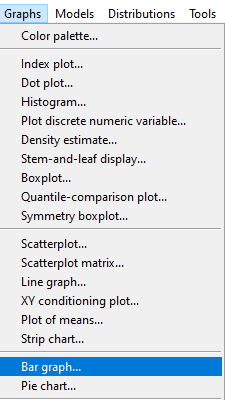
**Task 3.1** Create distribution of cases by date of onset of symptoms. Create Epi courve in R using histogram?

**Instructor’s Note:** Interval on X-axis should maximize the clarity of the pattern while preserving important detail. One general guideline is to use ¼ of the average incubation period for the interval. The average incubation period of leptospirosis is 10 days. One fourth of 10 days is 2.5. So, ask the class to use 2 graphs: one with 1-day intervals (for detail), and the other with 3-day intervals (for a general pattern).

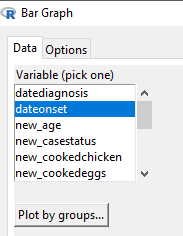
**Answer 3.1** Leptospirosis by date of onset of symptoms, Kenkolat, Kazakhstan, July-August 2004.

**Table 5.** Epi curve of cases of Leptospirosis by date of onset of symptoms during outbreak of Kenkolat, Kazakhstan, July-August 2004. (Interval 1 day).

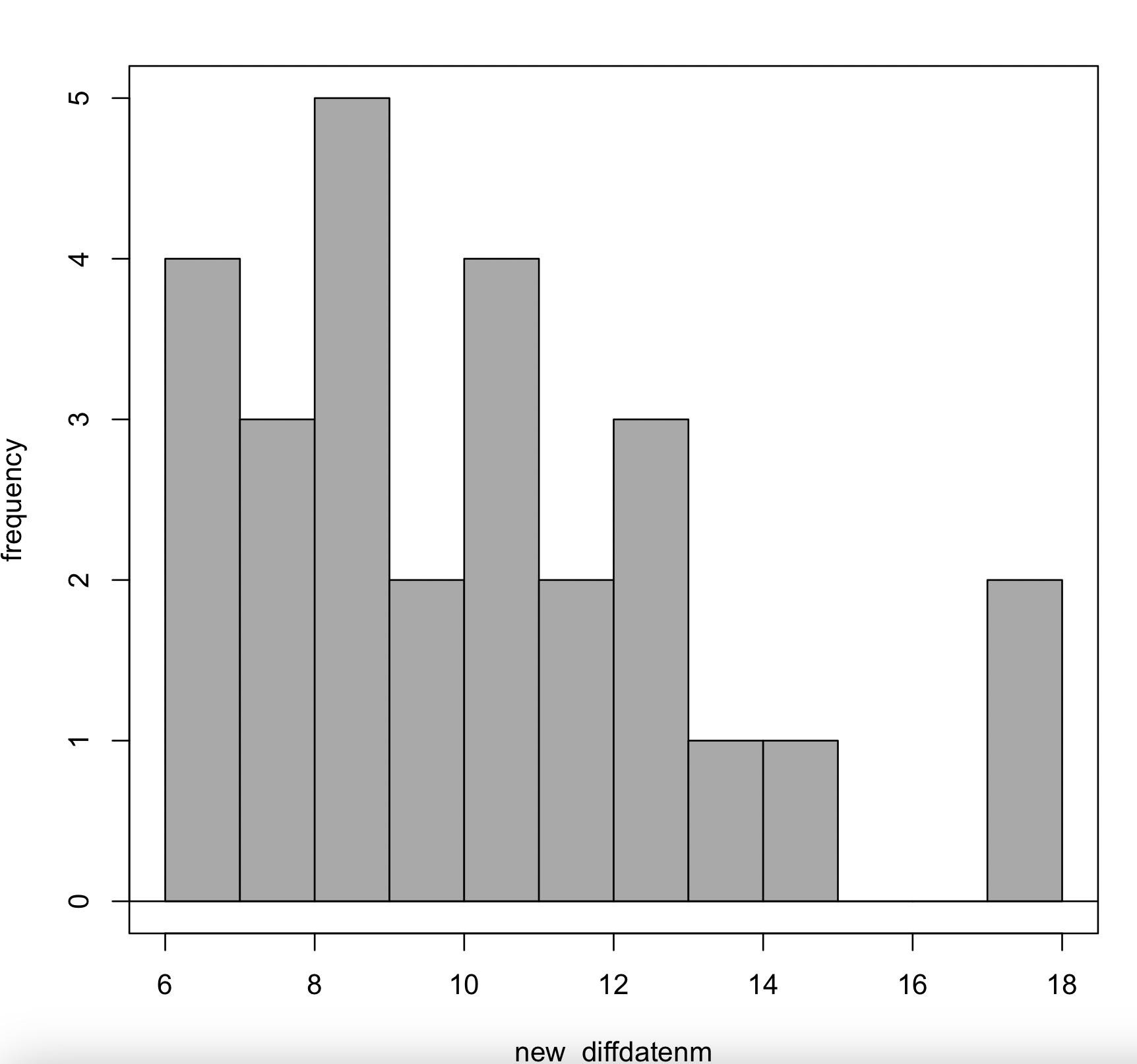
* Create histogram using R commander



* Select **dateofsymptoms** variable to create histogram

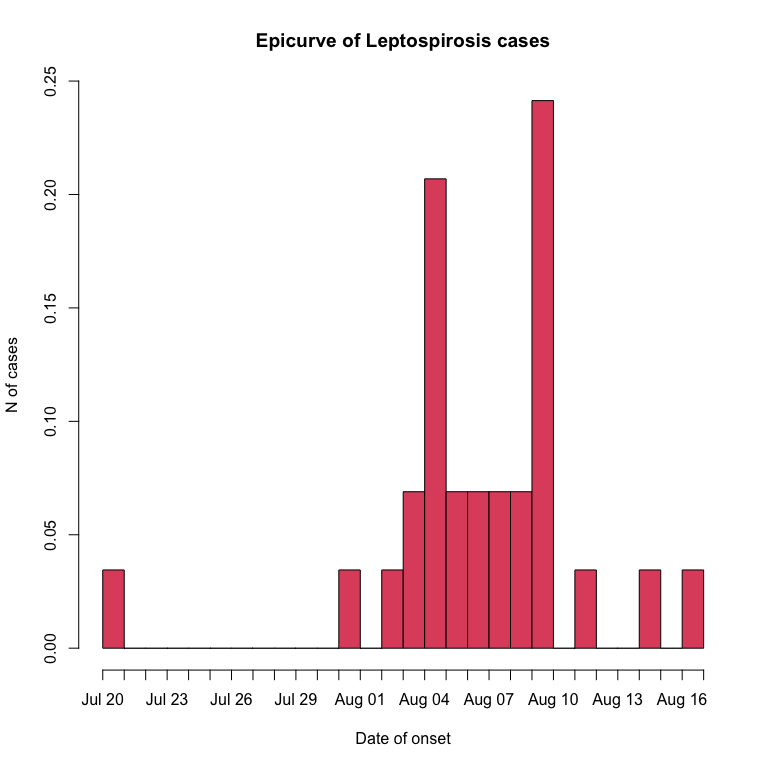


* Create histogram in R Commander.



* Create histogram using R code.

> plot2 <- hist(leptocase$new\_dateonset, breaks = 20, col = 2,   
+ main = "Epicurve of Leptospirosis cases",  
+ xlab = "Date of onset",  
+ ylab = "N of cases")



> #plot2

**Track 4.1** Calculate risk ratios to identify possible factors for the outbreak of Leptospirosis.

**Answer 4.1** Risk Ratios should be calculated for variabels considered as possible risk factors.

**Results of twobytwo table for variable regarding the use of river water**

* You can use same calcualtions to calculate other risk factors or add variables direcltly into the multivariate model.

> leptodbfinal <- leptodbfinal %>%   
+ mutate(rivwtr = factor(homeusedwaterriver))  
>   
> ## Calculate OR using tbl\_uvregression  
>   
> leptodbfinal %>%  
+ select(statusn, rivwtr, ratinyard) %>%  
+ tbl\_uvregression(  
+ method = glm,  
+ y = statusn,  
+ method.args = list(family=binomial(logit)),  
+ exponentiate = TRUE  
+ ) %>%  
+ add\_global\_p()

| **Characteristic** | **N** | **OR***1* | **95% CI***1* | **p-value** |
| --- | --- | --- | --- | --- |
| rivwtr | 147 |  |  | 0.005 |
| NO |  | — | — |  |
| YES |  | 0.27 | 0.09, 0.68 |  |
| ratinyard | 147 |  |  | 0.001 |
| 0 |  | — | — |  |
| 1 |  | 0.24 | 0.09, 0.57 |  |
| *1*OR = Odds Ratio, CI = Confidence Interval | | | | |

**Instructor’s Note:** Additional judgement and information for instructors to take into account for training.

**Answer 5a.** The outbreak begins with a single case on August 1 and rises rapidly to a broad peak from August 4-10. Three trailing cases a seen from August 11 to 17.

**Question 5b.** Is the epidemic curve consistent with a point source epidemic?

**Answer 5b.** Yes, this outbreak is consistent with a point source. Since the bulk of leptospirosis cases should have onset from 5 to 14 days after exposure, the bulk of the cases in a point source should occur over 9 days. In this outbreak 10 days pass from the first case onset to August 10 after which cases drop to a very low level. The additional 3 cases from August 12 to 17 could simply represent the tendency of leptospirosis to have a small proportion of cases with incubation periods of up to 4 weeks. Alternatively, these outlying cases could represent some additional or prolonged exposure secondary to the main event.

**Question 5c.** Assume that this is a point source. Using the known facts about leptospirosis, what type of exposure would you expect to explain the point source pattern.

**Answer 5c.** Assuming a point source one can simply determine the midpoint of the epidemic (median) and subtract the median incubation period (9-10 days) to find the most probably date of exposure. The median is August 6-7 (14th case occurs on August 6 and the 15th on August 7). Thus, a good estimate of the date of exposure is July 27-28. One may also subtract the minimum incubation (5 days) from the first case (August 1) yielding July 27. This strengthens the July-27-28 estimate. Counting back only two days (the minimum) from the first case or 28 days (the maximum) from the last case is not reliable because these very short and very long incubations are rare and have a low probability of appearing among the relatively small numbers of cases (29) in this outbreak.

The exposure would need to be brief and intense. Possibilities include transient contamination of water used for drinking, swimming, playing, bathing, or household purposes. Other possibilities could include brief exposure of many people around the same time to an ongoing source, for example gathering temporary workers to harvest rice or holding an outdoor swimming competition in a contaminated lake. Transmission directly from infected mammals or from ongoing rodent infestations would be very unlikely to appear as a point source.

# 7 Hypothesis generating interviews.

The epidemiologists questioned villagers about activities or events that occurred during the last week of July. The villagers reported that several continuous days of rain had caused excessive runoff to come down the normally dry gully, alongside the cattle corral complex, and into the permanent stream. The flow in the permanent stream also increased from the watershed upstream from the village. This stream was normally too small and too shallow for swimming or bathing. It was used to water livestock and gather non-potable water for household use. However, the flood had filled normally shallow pools in the streambed. The villagers had availed themselves of this welcome opportunity for swimming and other water related recreation in the heat of the summer.

Wrap-up **Question 6.** Summarize your findings.

**Answer 6. Clinical:** All cases had an illness with characteristics of leptospirosis. Frequent findings were moderate to high fever, chills, severe headache, and myalgias. MAT results confirmed leptospirosis in 85% of suspected cases.

**Person:** All age groups except the elderly (70 to 99 years old) were affected. Data were too sparse to make strong inferences about other details in the age and sex distribution. However, an excess risk was seen in males from 10 to 29 years old and in both sexes from 60 to 69 years old. The excess risk in 10 to 29 year old males accounted for all of the differences in case count and attack rates between males and females.

**Place:** Cases clustered in the downstream segment of the stream that ran along the northern edge of the village. This area of case clustering was across the river from a dry gully running down a hillside past a cattle corral complex. A spring from which villagers collected drinking water was also close to this gully the cattle corral complex.

**Time:** Of the 29 cases 25 fell from August 1 to August 10 which is consistent with the breadth of the more common incubation of 5 to 14 days. Adding the additional 3 cases (August 12 to 17) still placed all 29 cases within the more extreme range of incubation (2 to 28 days). This is very strong evidence of a point source in time within the village. Based on a median incubation period of 9 to 10 days for leptospirosis, the exposure centered on July 26 to 27. These dates followed a period of excess rainfall that flooded the gully that ran into the permanent stream across from a concentration of case houses.

**Question 7.** The investigation team decided to perform a case control study to gather more specific data about the possible causes of this outbreak. Based on the descriptive epidemiology, develop a hypothesis that can be tested in a case control study. Explain your reasoning.

**Answer 7:** The strongest information is the time distribution which can only be interpreted as exposure during a very narrow time window (July 26 and 27). The continuous rainfall during the last week of July would have set up favorable conditions for contamination of existing water sources for drinking, general household use, and swimming. Supporting this hypothesis is the proximity in space these water sources to a complex of cattle corrals and to a large cluster of cases. Moreover, the corrals were upslope from the water sources and would have drained the excess rainfall into the water sources. The age and sex distribution suggests widespread exposure that would be consistent with, but not indicative of a waterborne outbreak.

Accordingly, the hypothesis would be: “A point source outbreak of leptospirosis among residents of Kenkolat from August 1 to 17 resulted from exposure to water for drinking, general household use, watering livestock or swimming all after an unusual period of continuous rainfall during the last week of July. Water sources suspected of being contaminated were a spring used for drinking water and a permanent stream.”

At the end of the case study the participants may be prompted to open a Calibrated Peer Review assignment to develop the questionnaire to test this hypothesis for the case control study.

**Question 8:** How did this investigation address the basics of [One Health](https://www.cdc.gov/onehealth/basics/index.html)? What additional technical inputs could have strengthened the One Health approach?

**Answer 8:** This investigation involved One Health insomuch that leptospirosis in an important zoonotic disease. The outbreak occurred among a high-risk group for leptospirosis, residents of a rural village where the primary livelihood was raising livestock. A key environmental factor was heavy runoff or flooding from unusual, continuous rain. However, this investigation was limited to epidemiologists. A full One Health investigation could have also used input from microbiologists, veterinarians, environmentalists, and hydrologists. Microbiologist inputs - Isolate, identify, and serotype leptospires from livestock and rodents in the cattle corral complex - Test water from the spring, the well, and the pools in the permanent stream for microbial indicators of water quality - Isolate and serotype pathogenic leptospires from rodents, livestock, and suspected water sources Veterinarian inputs: - Assess the size and composition of the livestock population - Assess the livestock population for evidence of leptospirosis disease and infection - Assess livestock husbandry practices. - Recommend practices to reduce transmission of leptospirosis to and among the livestock Environmentalist inputs - Assess the rodent population with a focus on pest species (rats and mice) - Identify practices that increase rodent harborage and access to food. - Recommend and initiate measures to control pest rodents Hydrologist inputs: - Evaluate rainfall patterns and flood events from records at the nearest weather station. - Demonstrate a pathway for surface water to contaminate the spring. - Recommend solutions to provide sufficient potable water for the village. - Recommend solutions to provide water to livestock to minimize contamination of the stream. - Develop a system to divert runoff from the livestock corrals (commercial and household) away from the permanent stream.

**Question 9:** Before the case control study is started what control measures could be taken.

**Answer 9:** Although waterborne transmission of leptospirosis is only a hypothesis, several problems with the water supply for the village were obvious. These could also cause other waterborne diseases besides leptospirosis. Interim measures could include:

* Recommend that all villagers boil their drinking water until a safe potable water supply is developed.
* Advise the villagers to use rubber boots when engaged in activities at the permanent stream. These activities include but are not limited to collecting household water, watering livestock, washing clothes, wading, swimming, and fishing.
* Discourage villagers from swimming or bathing in the permanent stream.
* Provide livestock with a watering place separate from the stream, preferably downstream from the village.
* Reroute any water runoff channels from the animal corral complex to a point downstream from the village

Investigators and Supervisors from Central Asia Field Epidemiology Training Program:

* G. Aumoldaeva
* A. Turcunbaeva
* S. M. Ajeilat
* M. O. Favorov
* D. Nabirova