Vancouver R User Group

17 May, 2011

Data Summarization in R

[Describing Features of Data Frames 2](#_Toc285474350)

[Summarizing Data Frames 3](#_Toc285474352)

Summarizing Quantitative Variables…………………………..… …………....5

[Summarizing Qualitative Variables 8](#_Toc285474397)

[Summarizing Conditional Distributions of Quantitative Variables 9](#_Toc285474405)

**Isabella R. Ghement, Ph.D.**   
Ghement Statistical Consulting Company Ltd.  
301-7031 Blundell Road, Richmond, B.C., Canada, V6Y 1J5   
Tel: 604-767-1250   
Fax: 604-270-3922   
E-mail: isabella@ghement.ca   
Web: www.ghement.ca

# Describing Features of Data Frames

Throughout this section, we will work with the data set *fish,* which contains measurements on 159 fish caught in the lake Laengelmavesi in Finland. Specifically, this data set contains 159 observations on the following 7 variables:

**Weight**

Weight of the fish (in grams)

**Length1**

Length from the nose to the beginning of the tail (in cm)

**Length2**

Length from the nose to the notch of the tail (in cm)

**Length3**

Length from the nose to the end of the tail (in cm)

**Height**

Maximal height as % of Length3

**Width**

Maximal width as % of Length3

**Species**

Species

This data set can be imported into R as a data frame called **fish** by using the command below and browsing for the **fish.csv** file in your R Workshop folder (i.e., R Workshop 🡪 Data Sets 🡪 fish 🡪 fish.csv):

fish <- read.csv(file.choose())

When working with a data frame such as **fish** for the first time, it is important that you get familiar with that data frame’s dimensions, structure, first and last few records, and so on. This will help you understand the various features of data frame as you prepare for summarizing and visualizing the distribution of the variables stored in this data frame.

|  |  |
| --- | --- |
| Functions to be used for gaining insight into the structure of a data frame | |
| **R Function** | **Description** |
| dim()  nrow()  ncol() | Check the dimensions of a data frame (i.e., number of rows and number of columns).  e.g.: dim(fish); nrow(fish); ncol(fish) |
| str() | Check the structure of a data frame.  e.g.: str(fish) |
| head() | Access the first six rows of a data frame.  e.g: head(fish) |
| tail() | Access the last six rows of a data frame.  e.g.: tail(fish) |

# Of the above R commands, the str() command is particularly important. This command provides insights into the nature of each variable and will suggest whether each variable is treated appropriately by R.

# 

# Summarizing Data Frames

R offers several functions for summarizing all of the variables in a data frame simultaneously. Perhaps the most important of these functions is the summary() function, which is able to recognize whether a variable is treated by R as quantitative or qualitative and provide appropriate summary measures for each of these two variable types.

|  |  |
| --- | --- |
| Functions to be used for computing summary statistics for all variables in a data set | |
| **R Function** | **Description** |
| summary() | Summarize each of the variables in a data frame.   * For numerical variables, the summary() function computes the following descriptive statistics: minimum, 1st quartile, median, mean, 3rd quartile and maximum. * For categorical variables, the summary() function computes the frequency of data values falling in each category.   e.g.: summary(fish) |
| describe()  [Hmisc] | Describe each of the variables in a data frame.   * For numerical variables, report the total number of observations (*n*), the number of missing observations (*nmiss*), the number of unique values (*unique*), the mean value (*mean*), selected percentiles (*5th, 10th, 25th, 50th, 75th, 90th, 95th percentiles*), five lowest values and five highest values. * For categorical variables, compute the frequency of data values falling in each category, as well as the percentage of data values falling in each category (possibly rounded off).   e.g:  install.packages("Hmisc");  library(Hmisc);  describe(fish);  detach(package:Hmisc); |

Some of the functions provided by R for the summarization of all of the variables in a data frame only work if these data frames contain quantitative variables only. Two such functions – stat.desc() and describe() – are presented below.

|  |  |
| --- | --- |
| Functions to be used for computing summary statistics for numerical variables in a data set | |
| **R Function** | **Description** |

|  |  |
| --- | --- |
| stat.desc()  [pastecs] | Compute various summaries for the numerical variables in a data set (e.g., median, mean, standard error on the mean, 95% confidence interval for the true mean, variance, standard deviation and variation coefficient).  e.g.:  install.packages("pastecs");  library(pastecs);  stat.desc(fish[ ,1:6]);  detach(package:pastecs); |
| describe()  [psych] | Compute various summaries for the numerical variables in a data set (including mean, standard deviation, median, median absolute deviation, minimum, maximum, skewness, kurtosis and standard error).  e.g:  install.packages("psych");  library(psych);  describe(fish[ ,1:6]);  detach(package:psych); |

**Summarizing Quantitative Variables**

|  |  |  |
| --- | --- | --- |
| Measures of Location | Description | R Syntax |
| Arithmetic Mean | The sum of the values of x divided by the number n of values of x | mean(x) |
| Trimmed Mean | The arithmetic mean calculated  after a fraction (typically 0.05  or 5%) of the lower and upper  values of x have been discarded | mean(x, trim=0.05) |
| Winsorized Mean | The arithmetic mean of x is calculated after the trimmed values are replaced by the upper and lower trimmed quantiles | library(psych) winsor(x, trim=0.05)  detach(package:psych) |
| Median | The middle value in the list of ordered values of x | median(x) |
| Quantiles | The values having a certain rank among the ordered values of x | quantile(x)quantile(x, probs=c(0.25,0.75)) |
| Minimum | Smallest value of x | min(x) |
| Maximum | Largest value of x | max(x) |

|  |  |  |
| --- | --- | --- |
| Measures of Spread | Description | R Syntax |
| Standard Deviation | Square-root of the variance of x | sd(x) |
| Variance | Average deviation of values of x  from their mean value | var(x) |
| Median Absolute Deviation | The median difference of the values of x from the median  Value | mad(x) |
| Interquartile Range | Difference between the 75% and 25% ranked values of x | IQR(x) |
| Coefficient of Variation | Ratio of the standard deviation of x to the mean of x | co.var<-function(x)(sd(x)/mean(x))co.var(x) |

# Note: x is a quantitative variable having n values.

|  |  |  |
| --- | --- | --- |
| Measure of Skewness | Description | R Syntax |
| Coefficient of Skewness | Compute the skewness coefficient of the distribution of a numerical variable. If the skewness coefficient is zero, the distribution is symmetric.  A negative skewness coefficient indicates a negative skew (i.e., mean is smaller than median), and a positive one indicates a positive skew (i.e., mean is larger than median). | library(e1071);  skewness(x);  detach(package:e1071); |

# Note: x is a quantitative variable having n values.

|  |  |  |
| --- | --- | --- |
| Measure of Peakedness | Description | R Syntax |
| Coefficient of Kurtosis | Compute the standardized kurtosis coefficient of the distribution of a numerical variable. Recall that evaluation of a distribution’s kurtosis is especially useful after it has been determined that the distribution is not unduly skewed. It is not very useful for asymmetric or skewed distributions. A normal distribution has a standardized kurtosis coefficient equal to zero. A positive value for the standardized kurtosis coefficient implies that the distribution is more peaked than the normal distribution. A negative value implies that the distribution is flatter than the normal distribution. | library(e1071);  kurtosis(x);  detach(package:e1071); |

# Note: x is a quantitative variable having n values.

|  |  |  |
| --- | --- | --- |
| Precision and Confidence | Description | R Syntax |
| Standard Error of Sample Mean | Precision of the sample mean | sd(x)/sqrt(length(x)) |
| 95% Confidence Interval for Population Mean | 95% confidence interval for the population mean | library(gmodels) ci(x) detach(package:gmodels) |

# Note: x is a quantitative variable having n values.

R Exercise

# Compute various summary statistics for the variable Height in the fish data frame.

# *# measures of location*

# mean(fish$Height)

# mean(fish$Height, trim=0.05)

# median(fish$Height)

# quantile(fish$Height)

# min(fish$Height)

# max(fish$Height)

# *# measures of spread*

# sd(fish$Height)

# var(fish$Height)

mad(fish$Height)

# IQR(fish$Height)

# co.var<-function(x)(

# sd(x)/mean(x)

# )

# co.var(fish$Height)

# *# measures of skewness and kurtosis*

library(e1071)

skewness(fish$Height)

kurtosis(fish$Height)

detach(package:e1071)

**Summarizing Qualitative Variables**

Usually, the distribution of the values of a qualitative variable is summarized by reporting how often each category of this variable appears in the data set.

The joint distribution of the values of two qualitative variables is summarized by reporting how often each combination of values of these two variables appears in the data set.

|  |  |  |
| --- | --- | --- |
| Frequency Distribution | Description | R Syntax |
| Frequency Table Describing the Distribution of a Single Qualitative Variable | Lists the categories of the qualitative variable and gives an indication of how often each of these categories is represented in the data | table(f) |
| Contingency Table Describing the Joint Distribution of Two Qualitative Variables | Lists the categories of one variable across rows and the categories of the other variable across columns and gives an indication of how often each combination of categories is represented in the data | table(f1,f2) library(gmodels);  CrossTable(f1,f2);  detach(packages:gmodel); |

R Exercise

# Compute various summary statistics for the variable Height in the fish data frame, separately for each species.

table(fish$Species)

**Note:** Other useful functions for summarizing information on two qualitative variables include:

* margin.table()
* addmargins()
* prop.table()

The margin.table() function adds marginal totals to a 2 x 2 contingency table, while the function addmargins() calculates and returns the marginal totals. The function prop.table() computes the conditional distribution of a qualitative variable for each level of another qualitative variable. All three functions can be used with the option margin=1 (for rows) or margin=2 (for columns). E.g.: margin.table(table(f1,f2), margin=2)

addmargins(table(f1,f2), margin=2)

prop.table(table(f1,f2), margin=2)

# Summarizing Conditional Distributions of Quanitative Variables

# In many statistical problems, interest lies in describing the distribution of the values of a quantitative variable separately for each level of a qualitative variable. R has a variety of functions for describing the distribution of a quantitative variable conditional on the values of a qualitative variable. Some of these functions are displayed in the table below.

|  |  |
| --- | --- |
| Functions to be used for computing summary statistics by grouping variables | |
| **R Function** | **Description** |
| by() | Produce summaries of the data by relevant categories of a categorical variable.  e.g.: by(fish, fish$Species, summary);  by(fish, fish$Species, mean);  by(fish, fish$Species, sd); |
| describe.by()  [psych] | Generate summary statistics by a single grouping variable.  e.g.  install.packages("psych");  library(psych);  describe.by(x=fish[,1:6], group=fish$Species) ;  detach(package:psych); |
| summaryBy()  [doBy] | Generate summary statistics by grouping variables.  e.g.:  *# no missing values in Height*  install.packages("doBy");  library(doBy);  summaryBy(Height ~ Species, data=fish,  FUN= function(x){c(m = mean(x), s = sd(x))});  detach(package:doBy);  *# missing values in Weight*  library(doBy);  summaryBy(Weight ~ Species, data=fish,  FUN= function(x){c(m = mean(x, na.rm=T),  s = sd(x, na.rm=T))  }  );  detach(package:doBy); |

Other R functions that you may find useful when computing summary statistics for your data are given below.

|  |  |
| --- | --- |
| **R Function** | **Description** |
| apply, lapply, sapply, tapply | Calculations on rows and columns of matrices and arrays (apply), on components of lists (lapply, sapply), and data subsets (tapply). |
| aggregate | Splits the data into subsets, computes summary statistics for each, and returns the result in a convenient form. |
| CrossTable  [gmodels] | Creates a contingency table from categorical (factor) data. |
| split | Splits up a data set and gets a list with one component per group value. |

Example: Using the function apply()

* Why use the function apply()? The function *apply()* is a great tool for avoiding memory-consuming looping in R. This function allows you to simultaneously perform the same computation(s) on all the columns (or rows) of a matrix or data frame. For instance, you can use *apply()* to compute the means of all variables in a data frame containing numeric variables. [Note: More generally, *apply()* can be used to operate on various dimensions of an array.]
* Syntax:

apply(*matrix\_na*me, **1**, *function\_name*)

apply a function to all rows of a matrix or data frame

apply(*matrix\_na*me, **2**, *function\_name*)

apply a function to all columns of a matrix or data frame

* E.g.:

apply(fish[,c("Length1","Length2","Length3")], 2, mean)

Example: Using the function aggregate()

Often, you may wish to split a data frame into subsets and then compute summary statistics for each subset. You can accomplish this using the function *aggregate().* For instance, if you have daily data on ozone, solar radiation, wind speed and maximum daily temperature, you can first subset these data by month and then compute summary statistics such as sample means and medians for each subset.

* The first argument of *aggregate()* consists of the data frame to be subseted.
* The second argument of *aggregate()*, called *by*, specifies how subsetting should be done. The argument *by* consists of a list of categorical variables (even if there is only one categorical variable). Each subset is defined as a combination of levels of these categorical variables. For instance, the option *by=list(Month)* instructs R to subset the data frame by month.
* The third argument of *aggregate(),* called *FUN,* isa built-in or user-defined function that determines what type of summary statistics will be computed for each data subset. For example, using the option *FUN=mean* amounts to computing the sample mean for each subset.

aggdata <- aggregate(fish, by=list(fish$Species), FUN=mean, na.rm=T)

print(aggdata)

Output:

**Species**

Group.1 Weight Length1 Length2 Length3 Height Width Species

1 1 626.00000 30.30571 33.10857 38.35429 39.52571 14.13143 1

2 2 531.00000 28.80000 31.31667 34.31667 29.20000 15.90000 2

3 3 152.05000 20.64500 22.27500 24.97000 26.73500 14.60500 3

4 4 154.81818 18.72727 20.34545 22.79091 39.30909 14.08182 4

5 5 11.17857 11.25714 11.92143 13.03571 16.88571 10.22143 5

6 6 718.70588 42.47647 45.48235 48.71765 15.84118 10.43529 6

7 7 382.23929 25.73571 27.89286 29.57143 26.25714 15.83929 7