

Quantitative Data

Quantitative data, also known as continuous data, consists of numeric data that support arithmetic operations.

This is in contrast with qualitative data, whose values belong to pre-defined classes with no arithmetic operation allowed.

We will explain how to apply some of the R tools for quantitative data analysis with examples.

The tutorials in this section are based on a built-in data frame named `faithful`.

It consists of a collection of observations of the Old Faithful geyser in the USA Yellowstone National Park.

The following is a preview via the `head` function.

```
> head(faithful)
  eruptions waiting
1    3.600     79
2    1.800     54
3    3.333     74
4    2.283     62
5    4.533     85
6    2.883     55
```

There are two observation variables in the data set.

The first one, called `eruptions`, is the duration of the geyser eruptions.

The second one, called `waiting`, is the length of waiting period until the next eruption.

It turns out there is a correlation between the two variables, as shown in the Scatter Plot tutorial.

Frequency Distribution of Quantitative Data

The frequency distribution of a data variable is a summary of the data occurrence in a collection of non-overlapping categories.

Example

In the data set `faithful`, the frequency distribution of the eruptions variable is the summary of eruptions according to some classification of the eruption durations.

Problem

Find the frequency distribution of the eruption durations in `faithful`.

Solution

The solution consists of the following steps:

1. We first find the range of eruption durations with the `range` function. It shows that the observed eruptions are between 1.6 and 5.1 minutes in duration.

```
> duration = faithful$eruptions  
> range(duration)  
[1] 1.6 5.1
```

2. Break the range into non-overlapping sub-intervals by defining a sequence of equal distance break points. If we round the endpoints of the interval `[1.6, 5.1]` to the closest half-integers, we come up with the interval `[1.5, 5.5]`. Hence we set the break points to be the half-integer sequence `{ 1.5, 2.0, 2.5, ... }`.

```
> breaks = seq(1.5, 5.5, by=0.5) # half-integer sequence  
> breaks  
[1] 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5
```

3. Classify the eruption durations according to the half-unit-length sub-intervals with `cut`. As the intervals are to be closed on the left, and open on the right, we set the `right` argument as `FALSE`.

```
> duration.cut = cut(duration, breaks, right=FALSE)
```

4. Compute the frequency of eruptions in each sub-interval with the `table` function.

```
> duration.freq = table(duration.cut)
```

Answer

The frequency distribution of the eruption duration is:

```
> duration.freq
duration.cut
[1.5,2) [2,2.5) [2.5,3) [3,3.5) [3.5,4) [4,4.5) [4.5,5) [5,5.5)
      51      41       5       7      30      73      61       4
```

Enhanced Solution

We apply the *cbind* function to print the result in column format.

```
> cbind(duration.freq)
      duration.freq
[1.5,2)         51
[2,2.5)         41
[2.5,3)          5
[3,3.5)          7
[3.5,4)         30
[4,4.5)         73
[4.5,5)         61
[5,5.5)          4
```

Note

Per R documentation, you are advised to use the *hist* function to find the frequency distribution for performance reasons.

Exercise

1. Find the frequency distribution of the eruption waiting periods in *faithful*.
2. Find programmatically the duration sub-interval that has the most eruptions.

Histogram

A histogram consists of parallel vertical bars that graphically shows the frequency distribution of a quantitative variable. The area of each bar is equal to the frequency of items found in each class.

Example

In the data set `faithful`, the histogram of the `eruptions` variable is a collection of parallel vertical bars showing the number of eruptions classified according to their durations.

Problem

Find the histogram of the eruption durations in `faithful`.

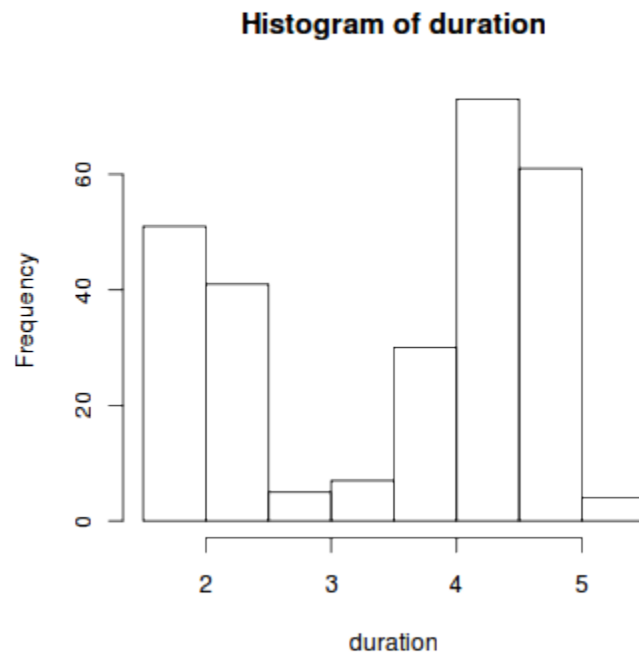
Solution

We apply the `hist` function to produce the histogram of the `eruptions` variable.

```
> duration = faithful$eruptions  
> hist(duration,      # apply the hist function  
+ right=FALSE)      # intervals closed on the left
```

Answer

The histogram of the eruption durations is:



Enhanced Solution

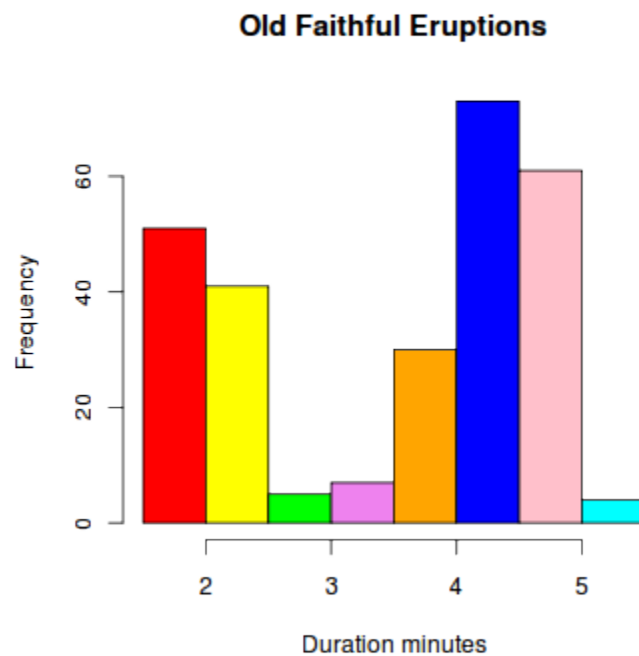
To colorize the histogram, we select a color palette and set it in the col argument of hist. In addition, we update the titles for readability.

```
> colors = c("red", "yellow", "green", "violet", "orange", "blue", "pink", "cyan")
```

```
> hist(duration,                # apply the hist function
+ right=FALSE,                 # intervals closed on the left
+ col=colors,                  # set the color palette
+ main="Old Faithful Eruptions", # the main title
+ xlab="Duration minutes")      # x-axis label
```

Exercise

Find the histogram of the eruption waiting period in faithful.



Relative Frequency Distribution of Quantitative Data

The relative frequency distribution of a data variable is a summary of the frequency proportion in a collection of non-overlapping categories.

The relationship of frequency and relative frequency is:

Example

In the data set `faithful`, the relative frequency distribution of the eruptions variable shows the frequency proportion of the eruptions according to a duration classification.

Problem

Find the relative frequency distribution of the eruption durations in `faithful`.

Solution

We first find the frequency distribution of the eruption durations as follows. Further details can be found in the Frequency Distribution tutorial.

```
> duration = faithful$eruptions
> breaks = seq(1.5, 5.5, by=0.5)
> duration.cut = cut(duration, breaks, right=FALSE)
> duration.freq = table(duration.cut)
```

Then we find the sample size of `faithful` with the `nrow` function, and divide the frequency distribution with it. As a result, the relative frequency distribution is:

```
> duration.relfreq = duration.freq / nrow(faithful)
```

Answer

The frequency distribution of the eruption variable is:

```
> duration.relfreq
duration.cut
[1.5,2) [2,2.5) [2.5,3) [3,3.5) [3.5,4) [4,4.5) [4.5,5) [5,5.5)
0.187500 0.150735 0.018382 0.025735 0.110294 0.268382 0.224265 0.014706
```

Enhanced Solution

We can print with fewer digits and make it more readable by setting the digits option.

```
> old = options(digits=1)
> duration.relfreq
duration.cut
[1.5,2) [2,2.5) [2.5,3) [3,3.5) [3.5,4) [4,4.5) [4.5,5) [5,5.5)
  0.19   0.15   0.02   0.03   0.11   0.27   0.22   0.01

> options(old)    # restore the old option
```

We then apply the *cbind* function to print both the frequency distribution and relative frequency distribution in parallel columns.

```
> old = options(digits=1)
> cbind(duration.freq, duration.relfreq)
      duration.freq duration.relfreq
[1.5,2)         51          0.19
[2,2.5)         41          0.15
[2.5,3)          5          0.02
[3,3.5)          7          0.03
[3.5,4)         30          0.11
[4,4.5)         73          0.27
[4.5,5)         61          0.22
[5,5.5)          4          0.01

> options(old)    # restore the old option
```

Exercise

Find the relative frequency distribution of the eruption waiting periods in faithful.

Cumulative Frequency Distribution

The cumulative frequency distribution of a quantitative variable is a summary of data frequency below a given level.

Example

In the data set `faithful`, the cumulative frequency distribution of the `eruptions` variable shows the total number of eruptions whose durations are less than or equal to a set of chosen levels.

Problem

Find the cumulative frequency distribution of the eruption durations in `faithful`.

Solution

We first find the frequency distribution of the eruption durations as follows. Further details can be found in the Frequency Distribution tutorial.

```
> duration = faithful$eruptions
> breaks = seq(1.5, 5.5, by=0.5)
> duration.cut = cut(duration, breaks, right=FALSE)
> duration.freq = table(duration.cut)
```

We then apply the `cumsum` function to compute the cumulative frequency distribution.

```
> duration.cumfreq = cumsum(duration.freq)
```

Answer

The cumulative distribution of the eruption duration is:

```
> duration.cumfreq
[1.5,2) [2,2.5) [2.5,3) [3,3.5) [3.5,4) [4,4.5) [4.5,5) [5,5.5)
   51    92    97   104   134   207   268   272
```


Enhanced Solution

We apply the *cbind* function to print the result in column format.

```
> cbind(duration.cumfreq)
```

	duration.cumfreq
[1.5,2)	51
[2,2.5)	92
[2.5,3)	97
[3,3.5)	104
[3.5,4)	134
[4,4.5)	207
[4.5,5)	268
[5,5.5)	272

Exercise

Find the cumulative frequency distribution of the eruption waiting periods in faithful.

Cumulative Frequency Graph

A cumulative frequency graph or give of a quantitative variable is a curve graphically showing the cumulative frequency distribution.

Example

In the data set `faithful`, a point in the cumulative frequency graph of the eruptions variable shows the total number of eruptions whose durations are less than or equal to a given level.

Problem

Find the cumulative frequency graph of the eruption durations in `faithful`.

Solution

We first find the frequency distribution of the eruption durations as follows. Further details can be found in the Frequency Distribution tutorial.

```
> duration = faithful$eruptions
> breaks = seq(1.5, 5.5, by=0.5)
> duration.cut = cut(duration, breaks, right=FALSE)
> duration.freq = table(duration.cut)
```

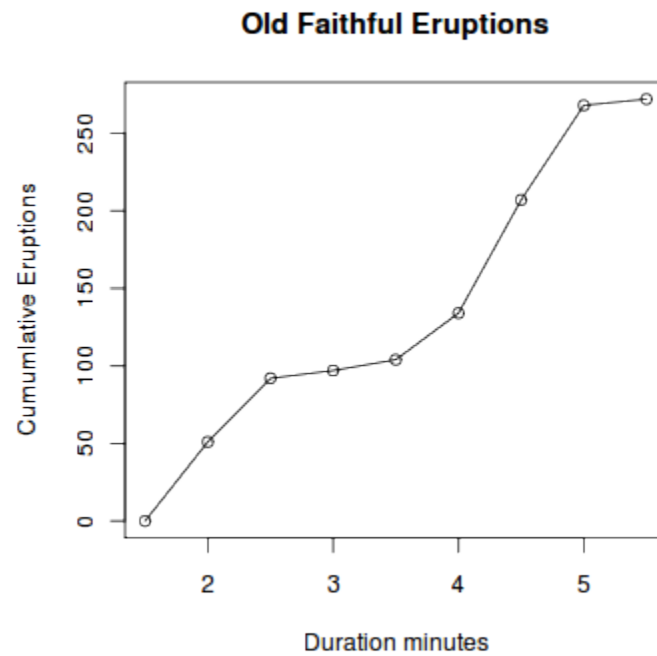
We then compute its cumulative frequency with `cumsum`, and plot it along with the starting zero element.

```
> cumfreq0 = c(0, cumsum(duration.freq))
> plot(breaks, cumfreq0,          # plot the data
+   main="Old Faithful Eruptions", # main title
+   xlab="Duration minutes",      # x-axis label
+   ylab="Cumumulative Eruptions") # y-axis label

> lines(breaks, cumfreq0)          # join the points
```

Answer

The cumulative frequency graph of the eruption durations is:

**Exercise**

Find the cumulative frequency graph of the eruption waiting periods in faithful.

Cumulative Relative Frequency Distribution

The cumulative relative frequency distribution of a quantitative variable is a summary of frequency proportion below a given level.

The relationship between cumulative frequency and relative cumulative frequency is:

$$\text{Cumulative Relative Frequency} = \frac{\text{Cumulative Frequency}}{\text{Sample Size}}$$

Example

In the data set `faithful`, the cumulative relative frequency distribution of the eruptions variable shows the frequency proportion of eruptions whose durations are less than or equal to a set of chosen levels.

Problem

Find the cumulative relative frequency distribution of the eruption durations in `faithful`.

Solution

We first find the frequency distribution of the eruption durations as follows. Further details can be found in the Frequency Distribution tutorial.

```
> duration = faithful$eruptions
> breaks = seq(1.5, 5.5, by=0.5)
> duration.cut = cut(duration, breaks, right=FALSE)
> duration.freq = table(duration.cut)
```

We then apply the `cumsum` function to compute the cumulative frequency distribution.

```
> duration.cumfreq = cumsum(duration.freq)
```

Then we find the sample size of `faithful` with the `nrow` function, and divide the cumulative frequency distribution with it. As a result, the cumulative relative frequency distribution is:

```
> duration.cumrelfreq = duration.cumfreq / nrow(faithful)
```

Answer

The cumulative relative frequency distribution of the eruption variable is:

```
> duration.cumrelfreq
[1.5,2) [2,2.5) [2.5,3) [3,3.5) [3.5,4) [4,4.5) [4.5,5) [5,5.5)
0.18750 0.33824 0.35662 0.38235 0.49265 0.76103 0.98529 1.00000
```

Enhanced Solution

We can print with fewer digits and make it more readable by setting the digits option.

```
> old = options(digits=2)
> duration.cumrelfreq
[1.5,2) [2,2.5) [2.5,3) [3,3.5) [3.5,4) [4,4.5) [4.5,5) [5,5.5)
  0.19   0.34   0.36   0.38   0.49   0.76   0.99   1.00

> options(old)    # restore the old option
```

We then apply the *cbind* function to print both the cumulative frequency distribution and relative cumulative frequency distribution in parallel columns.

```
> old = options(digits=2)
> cbind(duration.cumfreq, duration.cumrelfreq)
      duration.cumfreq duration.cumrelfreq
[1.5,2)           51           0.19
[2,2.5)           92           0.34
[2.5,3)           97           0.36
[3,3.5)          104           0.38
[3.5,4)          134           0.49
[4,4.5)          207           0.76
[4.5,5)          268           0.99
[5,5.5)          272           1.00

> options(old)
```

Exercise

Find the cumulative frequency distribution of the eruption waiting periods in faithful.

Cumulative Relative Frequency Graph

A cumulative relative frequency graph of a quantitative variable is a curve graphically showing the cumulative relative frequency distribution.

Example

In the data set `faithful`, a point in the cumulative relative frequency graph of the eruptions variable shows the frequency proportion of eruptions whose durations are less than or equal to a given level.

Problem

Find the cumulative relative frequency graph of the eruption durations in `faithful`.

Solution

We first find the frequency distribution of the eruption durations as follows. Further details can be found in the Frequency Distribution tutorial.

```
> duration = faithful$eruptions
> breaks = seq(1.5, 5.5, by=0.5)
> duration.cut = cut(duration, breaks, right=FALSE)
> duration.freq = table(duration.cut)
```

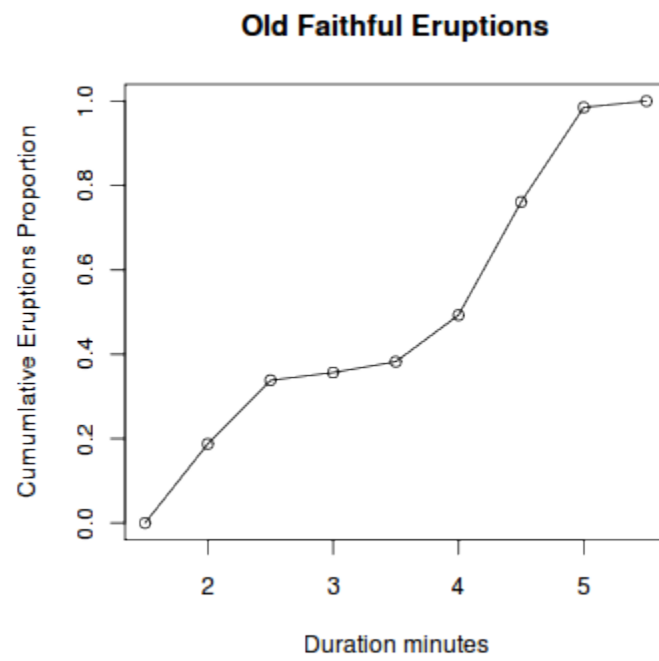
We then compute its cumulative frequency with `cumsum`, divide it by `nrow(faithful)` for the cumulative relative frequency, and plot it along with the starting zero element.

```
> cumfreq0 = c(0, cumsum(duration.freq))
> cumrelfreq0 = cumfreq0 / nrow(faithful)
> plot(breaks, cumrelfreq0,                # plot the data
+      main="Old Faithful Eruptions",      # main title
+      xlab="Duration minutes",
+      ylab="Cumumulative Eruptions Proportion")

> lines(breaks, cumrelfreq0)              # join the points
```

Answer

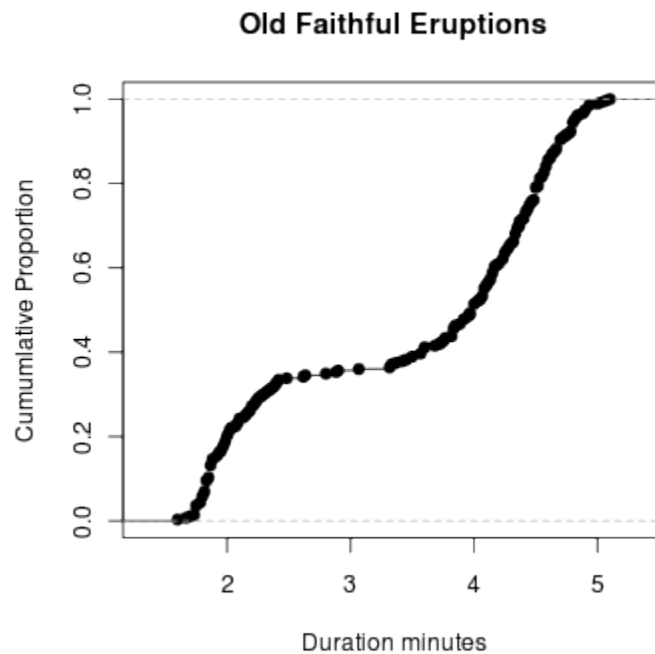
The cumulative relative frequency graph of the eruption duration is:



Alternative Solution

We create an interpolate function F_n with the built-in *ecdf* method. Then we produce a plot of F_n right away. There is no need to compute the cumulative frequency distribution a priori.

```
> Fn = ecdf(duration)           # compute the interpolate
> plot(Fn,                      # plot Fn
+   main="Old Faithful Eruptions", # main title
+   xlab="Duration minutes",      # x-axis label
+   ylab="Cumumulative Proportion") # y-axis label
```



Exercise

Find the cumulative relative frequency graph of the eruption waiting periods in faithful.

Stem-and-Leaf Plot

A stem-and-leaf plot of a quantitative variable is a textual graph that classifies data items according to their most significant numeric digits. In addition, we often merge each alternating row with its next row in order to simplify the graph for readability.

Example

In the data set `faithful`, a stem-and-leaf plot of the `eruptions` variable identifies durations with the same two most significant digits, and queue them up in rows.

Problem

Find the stem-and-leaf plot of the eruption durations in `faithful`.

Solution

We apply the `stem` function to compute the stem-and-leaf plot of `eruptions`.

Answer

The stem-and-leaf plot of the eruption durations is

```
> duration = faithful$eruptions  
> stem(duration)
```

The decimal point is 1 digit(s) to the left of the |

```
16 | 070355555588  
18 | 000022233333335577777777888822335777888  
20 | 00002223378800035778  
22 | 0002335578023578  
24 | 00228  
26 | 23  
28 | 080  
30 | 7  
32 | 2337  
34 | 250077  
36 | 0000823577  
38 | 2333335582225577  
40 | 0000003357788888002233555577778  
42 | 03335555778800233333555577778  
44 | 02222335557780000000023333357778888  
46 | 0000233357700000023578  
48 | 00000022335800333  
50 | 0370
```

Exercise

Find the stem-and-leaf plot of the eruption waiting periods in `faithful`.

Scatter Plot

A scatter plot pairs up values of two quantitative variables in a data set and display them as geometric points inside a Cartesian diagram.

Example

In the data set `faithful`, we pair up the eruptions and waiting values in the same observation as (x,y) coordinates. Then we plot the points in the Cartesian plane. Here is a preview of the eruption data value pairs with the help of the `cbind` function.

```
> duration = faithful$eruptions    # the eruption durations
> waiting = faithful$waiting       # the waiting interval
> head(cbind(duration, waiting))
  duration waiting
[1,]   3.600     79
[2,]   1.800     54
[3,]   3.333     74
[4,]   2.283     62
[5,]   4.533     85
[6,]   2.883     55
```

Problem

Find the scatter plot of the eruption durations and waiting intervals in `faithful`. Does it reveal any relationship between the variables?

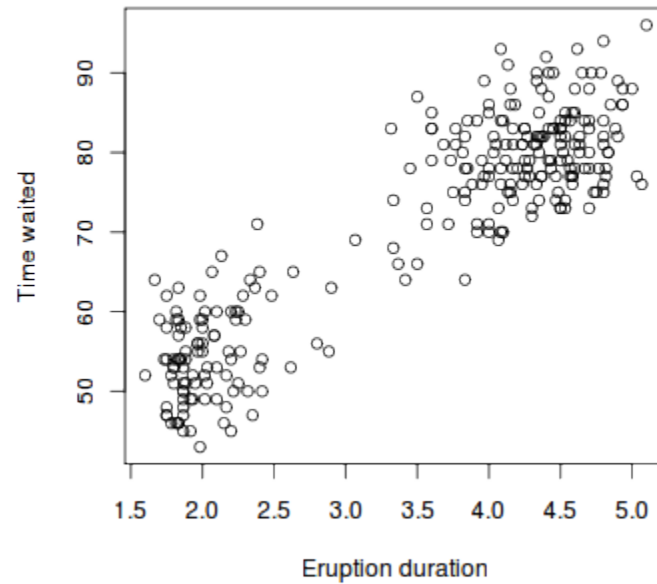
Solution

We apply the `plot` function to compute the scatter plot of eruptions and waiting.

```
> duration = faithful$eruptions    # the eruption durations
> waiting = faithful$waiting       # the waiting interval
> plot(duration, waiting,         # plot the variables
+       xlab="Eruption duration", # x-axis label
+       ylab="Time waited")      # y-axis label
```

Answer

The scatter plot of the eruption durations and waiting intervals is as follows. It reveals a positive linear relationship between them.



Enhanced Solution

We can generate a linear regression model of the two variables with the *lm* function, and then draw a trend line with *abline*.

```
> abline(lm(waiting ~ duration))
```

