

# Elementary Statistics with R

## Qualitative Data

The tutorials in this section are based on an R built-in data frame named `painters`. It is a compilation of technical information of a few eighteenth century classical painters. The data set belongs to the MASS package, and has to be pre-loaded into the R workspace prior to its use.

```
> library(MASS)           # load the MASS package
> painters
```

|               | Composition | Drawing | Colour | Expression | School |
|---------------|-------------|---------|--------|------------|--------|
| Da Udine      | 10          | 8       | 16     | 3          | A      |
| Da Vinci      | 15          | 16      | 4      | 14         | A      |
| Del Piombo    | 8           | 13      | 16     | 7          | A      |
| Del Sarto     | 12          | 16      | 9      | 8          | A      |
| Fr. Penni     | 0           | 15      | 8      | 0          | A      |
| Guilio Romano | 15          | 16      | 4      | 14         | A      |
| .....         |             |         |        |            |        |

The last School column contains the information of school classification of the painters. The schools are named as A, B, ..., etc, and the School variable is qualitative.

```
> painters$School
[1] A A A A A A A A A A B B B B B B C C C C C C D D D D
[27] D D D D D D E E E E E E E F F F F G G G G G G H H
[53] H H
Levels: A B C D E F G H
```

For further details of the painters data set, please consult the R documentation.

```
> help(painters)
```

## Frequency Distribution of Qualitative 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 `painters`, the frequency distribution of the `School` variable is a summary of the number of painters in each school.

### Problem

Find the frequency distribution of the painter schools in the data set `painters`.

### Solution

We apply the `table` function to compute the frequency distribution of the `School` variable.

```
> library(MASS)           # load the MASS package
> school = painters$School # the painter schools
> school.freq = table(school) # apply the table function
```

### Answer

The frequency distribution of the schools is:

```
> school.freq
school
 A B C D E F G H
10 6 6 10 7 4 7 4
```

### Enhanced Solution

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

```
> cbind(school.freq)
school.freq
 A      10
 B       6
 C       6
 D      10
 E       7
 F       4
 G       7
 H       4
```

### Exercise

1. Find the frequency distribution of the composition scores in `painters`.

## Relative Frequency Distribution of Qualitative 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 `painters`, the relative frequency distribution of the `School` variable is a summary of the proportion of painters in each school.

### Problem

Find the relative frequency distribution of the painter schools in the data set `painters`.

### Solution

We first apply the `table` function to compute the frequency distribution of the `School` variable.

```
> library(MASS)           # load the MASS package
> school = painters$School # the painter schools
> school.freq = table(school) # apply the table function
```

Then we find the sample size of painters with the `nrow` function, and divide the frequency distribution with it. Therefore the relative frequency distribution is:

```
> school.relfreq = school.freq / nrow(painters)
```

### Answer

The relative frequency distribution of the schools is:

```
> school.relfreq
school
  A    B    C    D    E    F
0.185185 0.111111 0.111111 0.185185 0.129630 0.074074
  G    H
0.129630 0.074074
```

### Enhanced Solution

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

```
> old = options(digits=1)
> school.relfreq
school
  A  B  C  D  E  F  G  H
0.19 0.11 0.11 0.19 0.13 0.07 0.13 0.07
```

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

In addition, we can apply the cbind function to print the result in column format.

```
> old = options(digits=1)
```

```
> cbind(school.relfreq)
```

```
school.relfreq
```

```
A      0.19
```

```
B      0.11
```

```
C      0.11
```

```
D      0.19
```

```
E      0.13
```

```
F      0.07
```

```
G      0.13
```

```
H      0.07
```

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

### **Exercise**

Find the relative frequency distribution of the composition scores in painters.

## Bar Graph

A bar graph of a qualitative data sample consists of vertical parallel bars that shows the frequency distribution graphically.

### Example

In the data set painters, the bar graph of the School variable is a collection of vertical bars showing the number of painters in each school.

### Problem

Find the bar graph of the painter schools in the data set painters.

### Solution

We first apply the table function to compute the frequency distribution of the School variable.

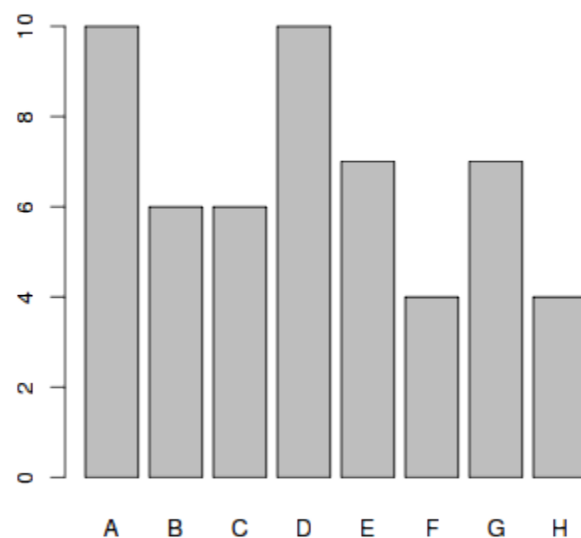
```
> library(MASS)           # load the MASS package  
> school = painters$School # the painter schools  
> school.freq = table(school) # apply the table function
```

Then we apply the barplot function to produce its bar graph.

```
> barplot(school.freq)      # apply the barplot function
```

### Answer

The bar graph of the school variable is:

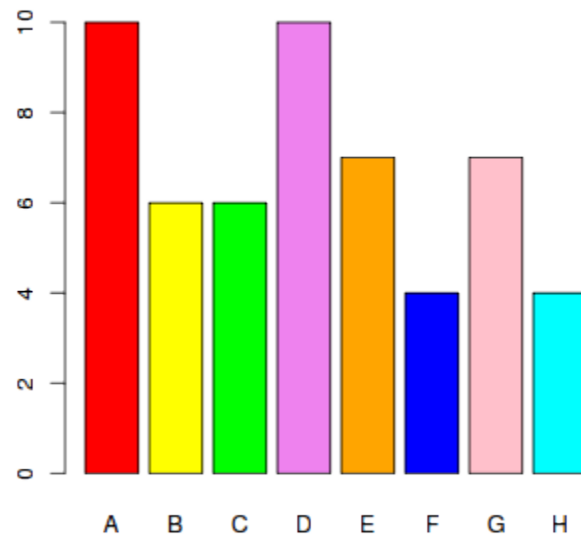


### Enhanced Solution

To colorize the bar graph, we select a color palette and set it in the col argument of barplot.

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

```
> barplot(school.freq, # apply the barplot function  
+ col=colors)          # set the color palette
```



### Exercise

Find the bar graph of the composition scores in painters.

## Pie Chart

A pie chart of a qualitative data sample consists of pizza wedges that shows the frequency distribution graphically.

### Example

In the data set painters, the pie chart of the School variable is a collection of pizza wedges showing the proportion of painters in each school.

### Problem

Find the pie chart of the painter schools in the data set painters.

### Solution

We first apply the table function to produce the frequency distribution of School.

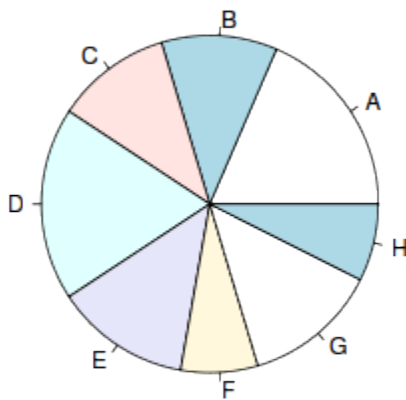
```
> library(MASS)           # load the MASS package  
> school = painters$School # the painter schools  
> school.freq = table(school) # apply the table function
```

Then we apply the pie function to produce its pie chart.

```
> pie(school.freq)          # apply the pie function
```

### Answer

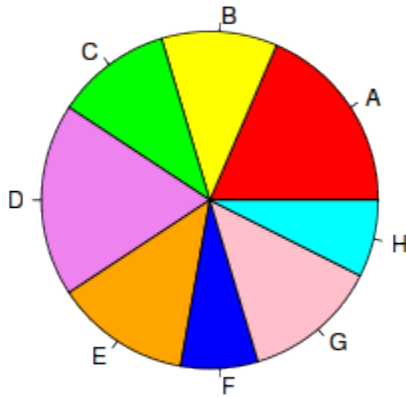
The pie chart of the school variable is:



### Enhanced Solution

To colorize the pie chart, we select a color palette and set it in the col argument of pie.

```
> colors = c("red", "yellow", "green", "violet",  
+ "orange", "blue", "pink", "cyan")  
> pie(school.freq,      # apply the pie function  
+ col=colors)           # set the color palette
```



### Exercise

Find the pie chart of the composition scores in painters.



## Category Statistics

In the built-in data set `painters`, the painters are classified according to the schools they belong. Each school can be characterized by its various statistics, such as mean composition, drawing, coloring and expression scores.

Suppose we would like to know which school has the highest mean composition score. We would have to first find out the mean composition score of each school. The following shows how to find the mean composition score of an arbitrarily chosen school.

### Problem

Find out the mean composition score of school C in the data set `painters`.

### Solution

The solution consists of a few steps:

1. Create a logical index vector for school C.  

```
> library(MASS)           # load the MASS package  
> school = painters$School # the painter schools  
> c_school = school == "C" # the logical index vector
```
2. Find the child data set of painters for school C. For explanation, please consult the tutorial of Data Frame Row Slice.  

```
> c_painters = painters[c_school, ] # child data set
```
3. Find the mean composition score of school C.  

```
> mean(c_painters$Composition)  
[1] 13.167
```

### Answer

The mean composition score of school C is 13.167.

### Alternative Solution

Instead of computing the mean composition score manually for each school, use the `tapply` function to compute them all at once.

```
> tapply(painters$Composition, painters$School, mean)  
A      B      C      D      E      F      G      H  
10.400 12.167 13.167 9.100 13.571 7.250 13.857 14.000
```

### Exercise

1. Find programmatically the school with the highest composition scores.
2. Find the percentage of painters whose color score is equal to or above 14.