Displaying and Describing Categorical Data

Al Nosedal
University of Toronto

Summer 2019

My momma always said: "Life was like a box of chocolates. You never know what you're gonna get."

Forrest Gump.

Definitions

A **variable** is some characteristic of a population or sample. We usually represent the name of a variable using uppercase letters such as X, Y, and Z.

The **values** of the variable are the possible observations of the variable.

Data are the observed values of a variable.

Types of Data

There are three types of data: interval, nominal, and ordinal.

- **Interval** data are real numbers, such as heights, weights, incomes, and distances. We also refer to this type of data as **quantitative** or **numerical**.
- The values of **nominal** data are categories. For example, responses to questions about marital status nominal data. Nominal data are also called **qualitative** or **categorical**.
- **Ordinal** data appear to be nominal, but the difference is that the order of their values has meaning. For example, at the completion of most university courses, students are asked to evaluate the course.

Calculations for Types of Data

- **Interval Data**. All calculations are permitted on interval data. We often describe a set of interval data by calculating the average.
- **Nominal Data**. Because the codes of nominal data are completely arbitrary, we **cannot** perform any calculations on these codes.
- **Ordinal Data**. The most important aspect of ordinal data is the order of the values. The only permissible calculations are those involving a ranking process.

Example. Babies

Medical researchers at a large city hospital investigating the impact of prenatal care on newborn health collected data from 882 births between 1998 and 2000. They kept track of the mother's age, the number of weeks the pregnancy lasted, the type of birth (Caesarean, induced, natural), the level of prenatal care the mother had had (none, minimal, adequate), the birth weight and sex of the baby, and whether the baby exhibited health problems (none, minor, major).

a) What are the individuals in this data set?

b) For each individual, what variables are given? Which of these variables are categorical and which are quantitative?

Fuel economy (solution)

a) The individuals are the babies.
b) There are three quantitative variables: mother's age, length of pregnancy, and birth weight of baby. There are four categorical variables: type of birth, level of prenatal care, gender of baby, and baby's health problems.

Distribution of a variable

The distribution of a variable tells us what values it takes and how often it takes these values.
The values of a categorical variable are labels for the categories. The distribution of a categorical variable lists the categories and gives either the count or the percent of individuals that fall in each category.

Summarizing Qualitative Data

Frequency distribution. A frequency distribution is a tabular summary of data showing the number (frequency) of items in each of several non-overlapping classes.

Relative frequency of a class $= \dfrac{\text{Frequency of the class}}{n}$

where n represents the total number of observations.

More about Qualitative Data

A **relative frequency distribution** gives a tabular summary of data showing the relative frequency for each class.

A **percent frequency distribution** summarizes the percent frequency of the data for each class.

Bar charts and pie charts

A **bar chart**, is a graphical device for depicting qualitative data summarized in a frequency, relative frequency, or percent frequency distribution. On one axis of the graph, we specify the labels that are used for the classes (categories). A frequency, relative frequency, or percent frequency scale can be used for the other axis of the graph.
The pie chart provides another graphical device for presenting relative frequency and percent frequency distributions for qualitative data.

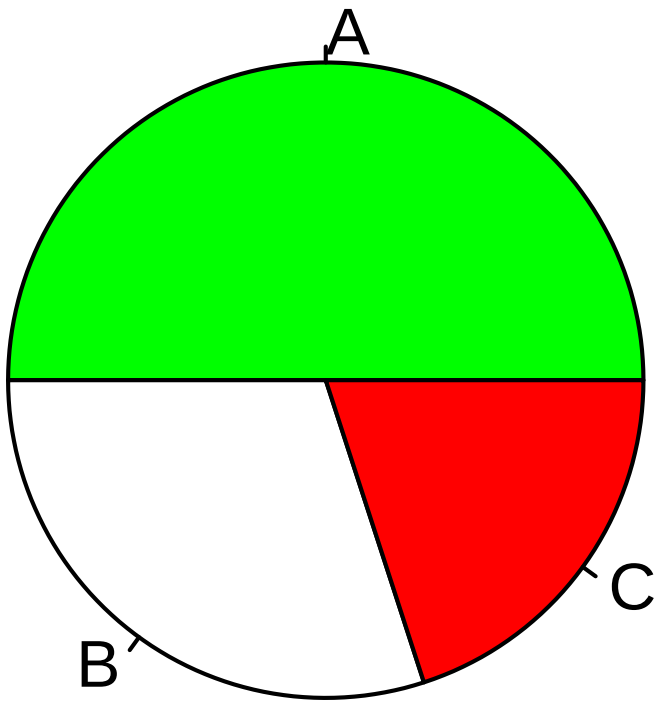
Toy Example

The response to a question has three alternatives: A, B, and C. A sample of 120 responses provides 60 A, 24 B, and 36 C.
a) Show the frequency, relative frequency and percent frequency distributions.
b) Construct a pie chart.
c) Construct a bar graph.

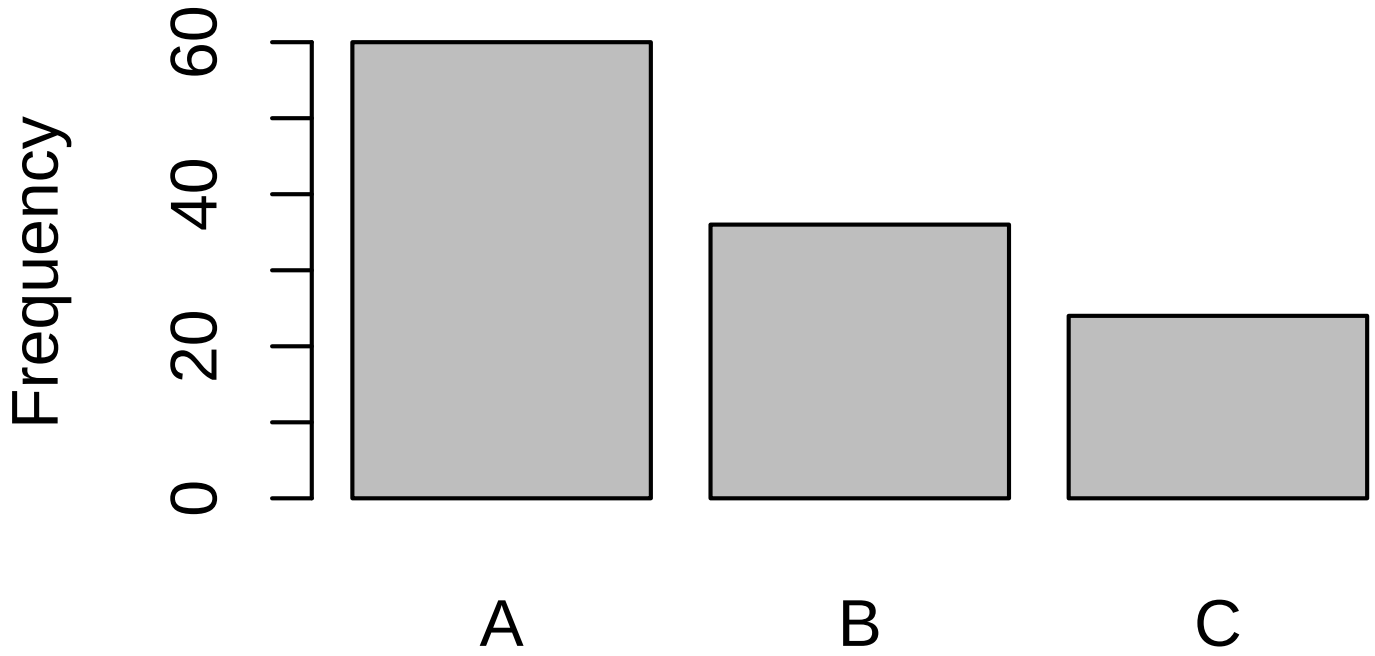
Solution

Class	Frequency	Relative Freq.	Percent Freq.
A	60	60/120	0.50
B	24	24/120	0.20
C	36	36/120	0.30

Solution (pie chart)



Solution (bar chart)



Infertility after Spontaneous and Induced Abortion

Description
The role of induced (and spontaneous) abortions in the aetiology of secondary sterility was investigated. Obstetric and gynaecologic histories were obtained from 100 women with secondary infertility admitted to the First Department of Obstetrics and Gynaecology of the University of Athens Medical School and to the Division of Fertility and Sterility of that Department. For every patient, an attempt was made to find two healthy control subjects from the same hospital with matching for age, parity, and level of education. Two control subjects each were found for 83 of the index patients.
Note. One case with two prior spontaneous abortions and two prior induced abortions is omitted.

R Code (Data set)

```
attach(infert);
names(infert);
head(infert);
all.cases=infert$case;
table.cases=table(all.cases);
table.cases;
```

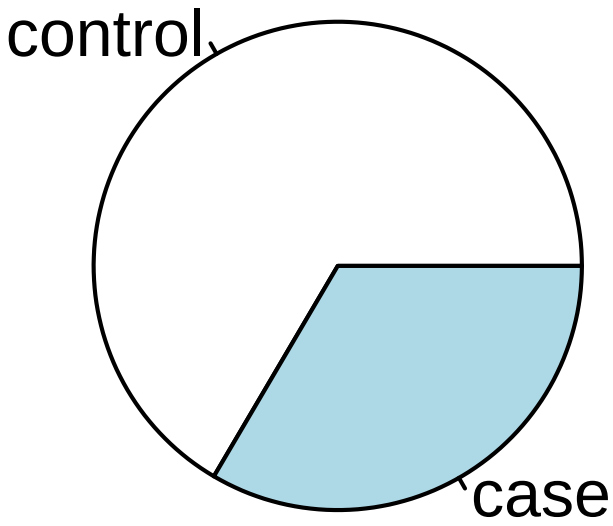
R Code (Data set)

```
## [1] "education"      "age"              "parity"          "inc
## [5] "case"           "spontaneous"      "stratum"         "poo
##    education age parity induced case spontaneous stratum poo
## 1     0-5yrs  26      6       1    1           2       1
## 2     0-5yrs  42      1       1    1           0       2
## 3     0-5yrs  39      6       2    1           0       3
## 4     0-5yrs  34      4       2    1           0       4
## 5    6-11yrs  35      3       1    1           1       5
## 6    6-11yrs  36      4       2    1           1       6
## all.cases
##   0   1
## 165  83
```

R Code (Pie chart)

```
attach(infert);
names(infert);
head(infert);
all.cases=infert$case;
table.cases=table(all.cases);
table.cases;
labels=c("control","case");
pie(table.cases,labels);
```

R Code (Pie chart)



R Code (subsetting)

```
attach(infert);
names(infert);
head(infert);
all.cases=infert$case;
all.cases[2];
all.cases[4];
```

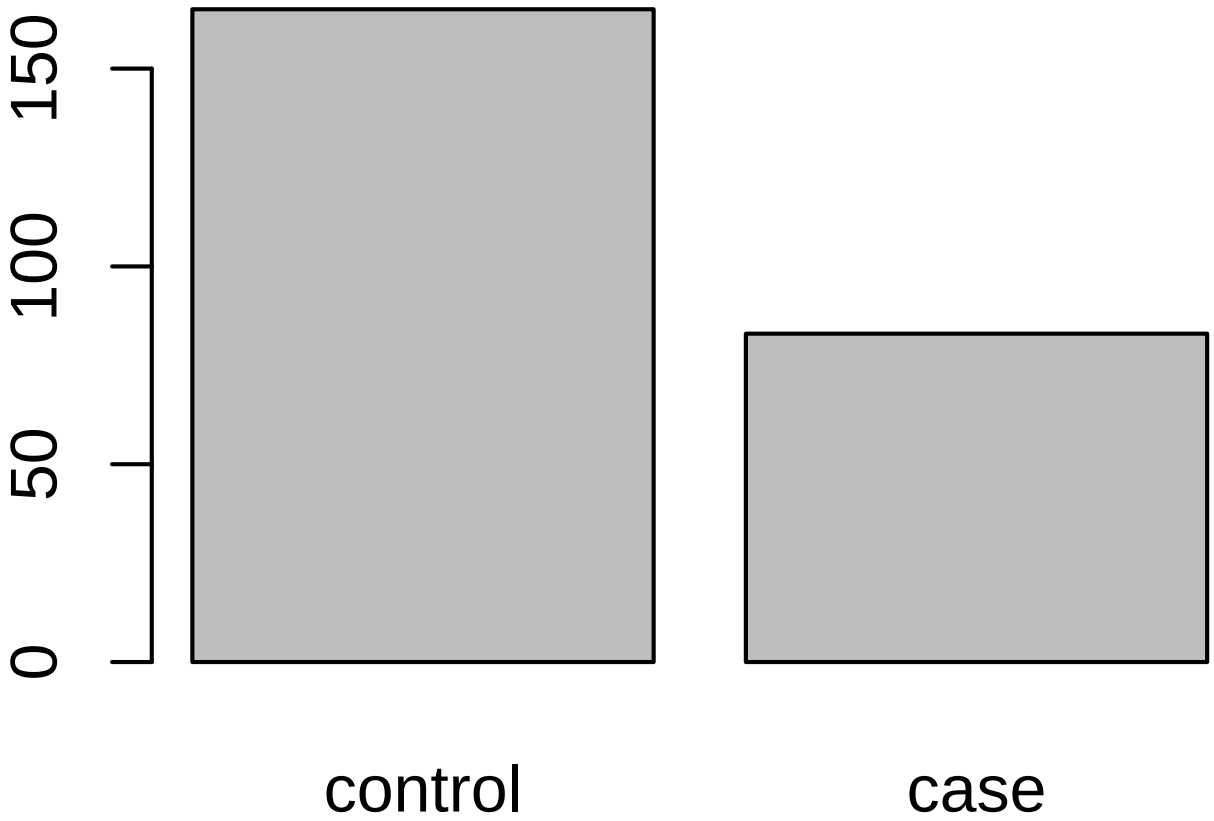
R Code (subsetting)

```
## [1] 1
## [1] 1
```

R Code (Bar chart)

```
attach(infert);
names(infert);
head(infert);
all.cases=infert$case;
table.cases=table(all.cases);
table.cases;
labels=c("control","case");
barplot(table.cases,names.arg=labels);
```

R Code (Bar chart)



Example. Never on Sunday?

Births are not, as you might think, evenly distributed across the days of the week. Here are the average numbers of babies born on each day of the week in 2008:

Day	Births
Sunday	7,534
Monday	12,371
Tuesday	13,415
Wednesday	13,171
Thursday	13,147
Friday	12,919
Saturday	8,617

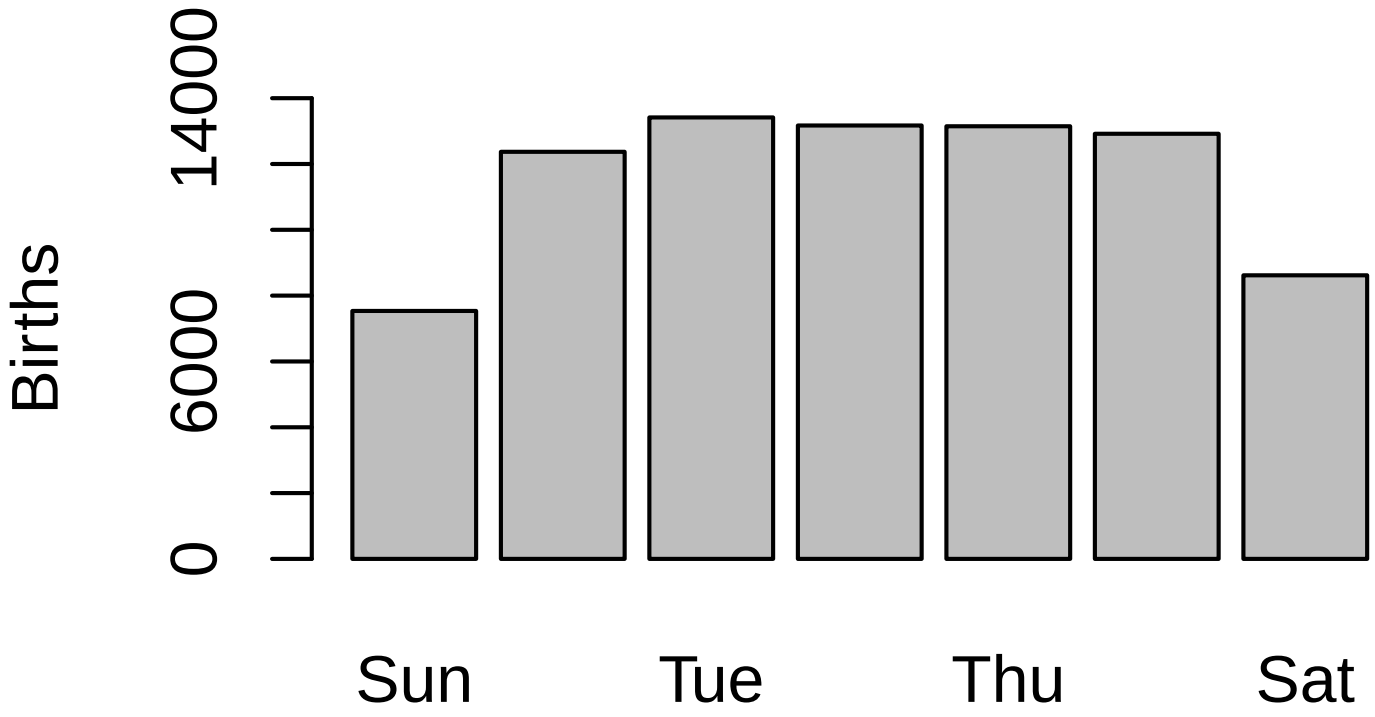
Example. Never on Sunday? (cont.)

Present these data in a well-labeled bar graph. Would it also be correct to make a pie chart? Suggest some possible reasons why there are fewer births on weekends.

Solution (bar chart)

```
## Step 1. Entering Data;
births=c(7534,12371,13415,13171,13147,12919,8617);
names=c("Sun","Mon","Tue","Wed","Thu","Fri","Sat");

## Step 2. Making bargraph;
barplot(births,names.arg=names,ylim=c(0,14000),ylab="Births");
```

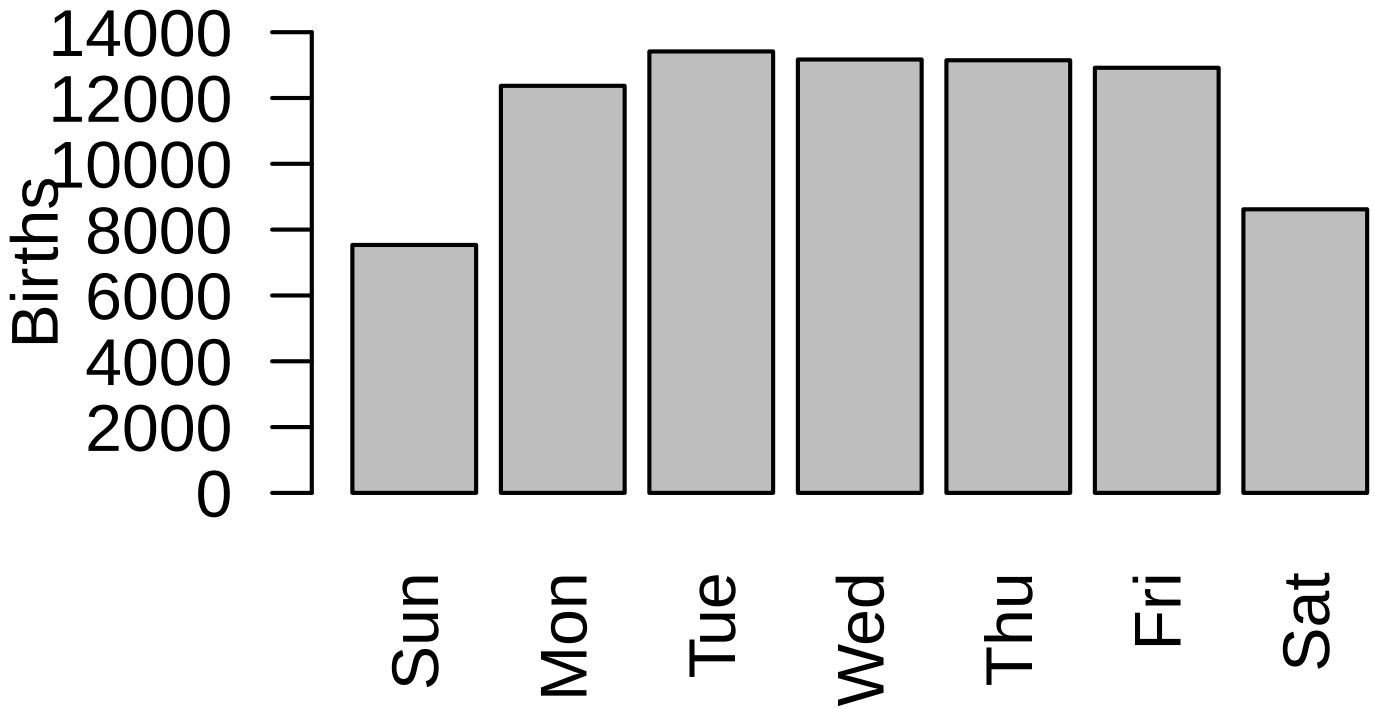
Solution (bar chart)



Solution (bar chart)

```
## Step 1. Entering Data;
births=c(7534,12371,13415,13171,13147,12919,8617);
names=c("Sun","Mon","Tue","Wed","Thu","Fri","Sat");

## Step 2. Making bargraph;
barplot(births,names.arg=names,ylim=c(0,14000),
ylab="Births",las=2);

# las=2 changes orientation of labels;
```

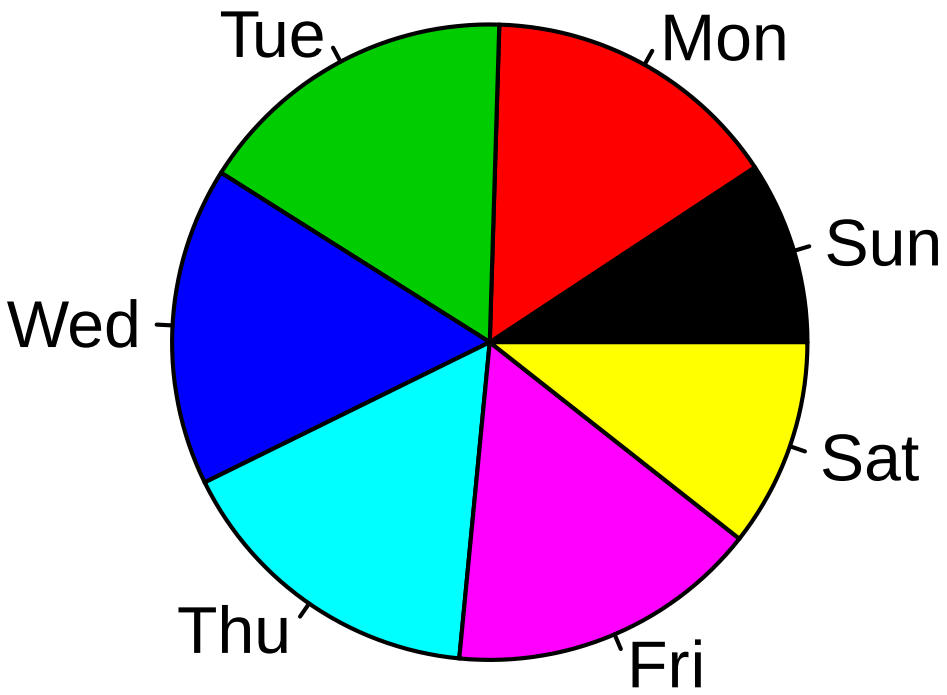
Solution (bar chart)



Solution (pie chart)

```
## Step 1. Entering Data.

births=c(7534,12371,13415,13171,13147,12919,8617);
names=c("Sun","Mon","Tue","Wed","Thu","Fri","Sat");

## Step 2. Making pie chart;

pie(births,names,col=c(1:7));
```

Solution (pie chart)



Example. Never on Sunday?

Solution.
It would be correct to make a pie chart but a pie chart would make it more difficult to distinguish between the weekend days and the weekdays. Some births are scheduled (e.g., induced labor), and probably most are scheduled for weekdays.

Example. What color is your car?

The most popular colors for cars and light trucks vary by region and over time. In North America white remains the top color choice, with black the top choice in Europe and silver the top choice in South America. Here is the distribution of the top colors for vehicles sold globally in 2010.

Color	Popularity (%)
Silver	26
Black	24
White	16
Gray	16
Red	6
Blue	5
Beige, brown	3
Other colors	

What color is your car? (cont.)

a) Fill in the percent of vehicles that are in other colors.
b) Make a graph to display the distribution of color popularity.

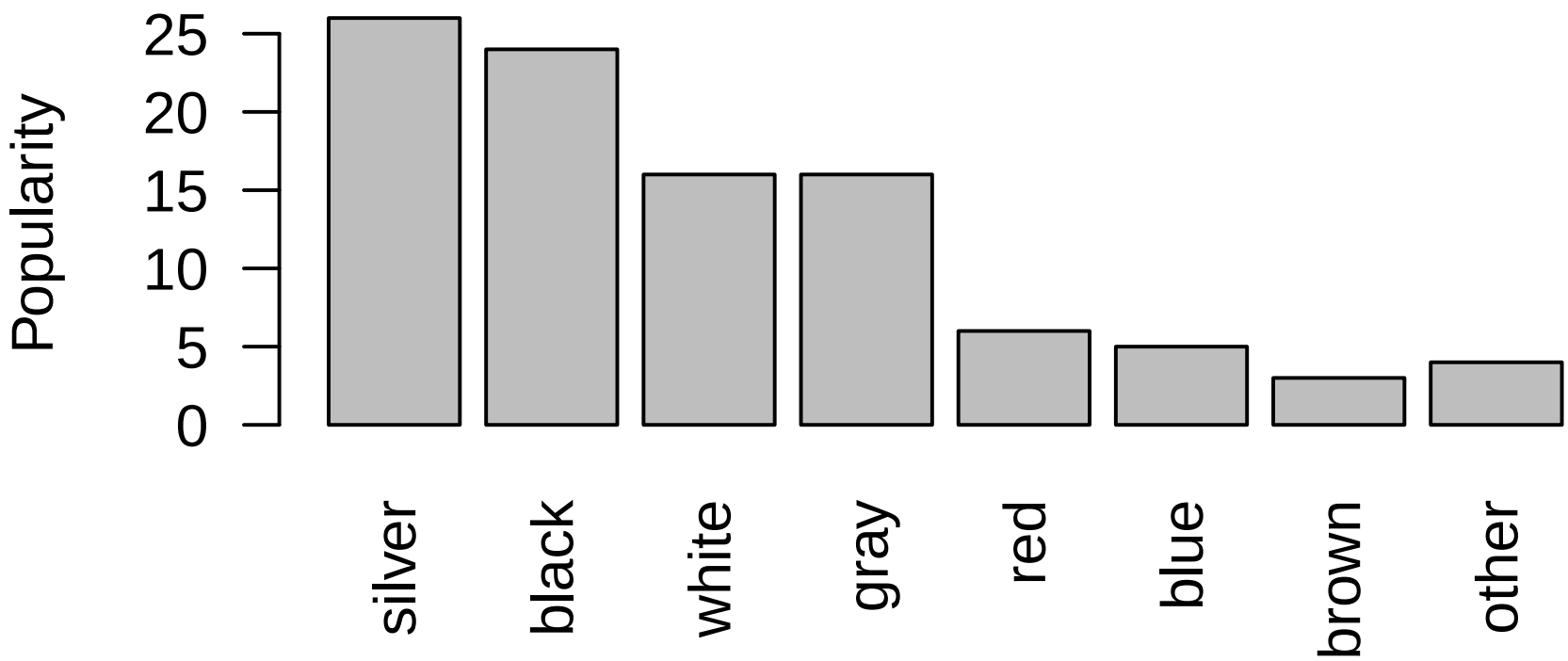
Solution

a) Other $= 100 - (26 + 24 + 16 + 16 + 6 + 5 + 3) = 4$.

Solution (bar chart)

```
# Step 1. Entering data;

popularity<-c(26,24,16,16,6,5,3,4);

color<-c("silver","black","white","gray",
"red","blue","brown","other");

# Step 2. Making bar graph;

barplot(popularity,names.arg=color,ylab="Popularity",las=2);
```

Solution (bar chart)



Another example

The following table lists the top 10 countries and amounts of oil (millions of barrels annually) they exported to the United States in 2010.

Country	Oil Imports (millions of barrels annually)
Algeria	119
Angola	139
Canada	720
Colombia	124
Iraq	151
Kuwait	71
Mexico	416
Nigeria	360
Saudi Arabia	394
Venezuela	333

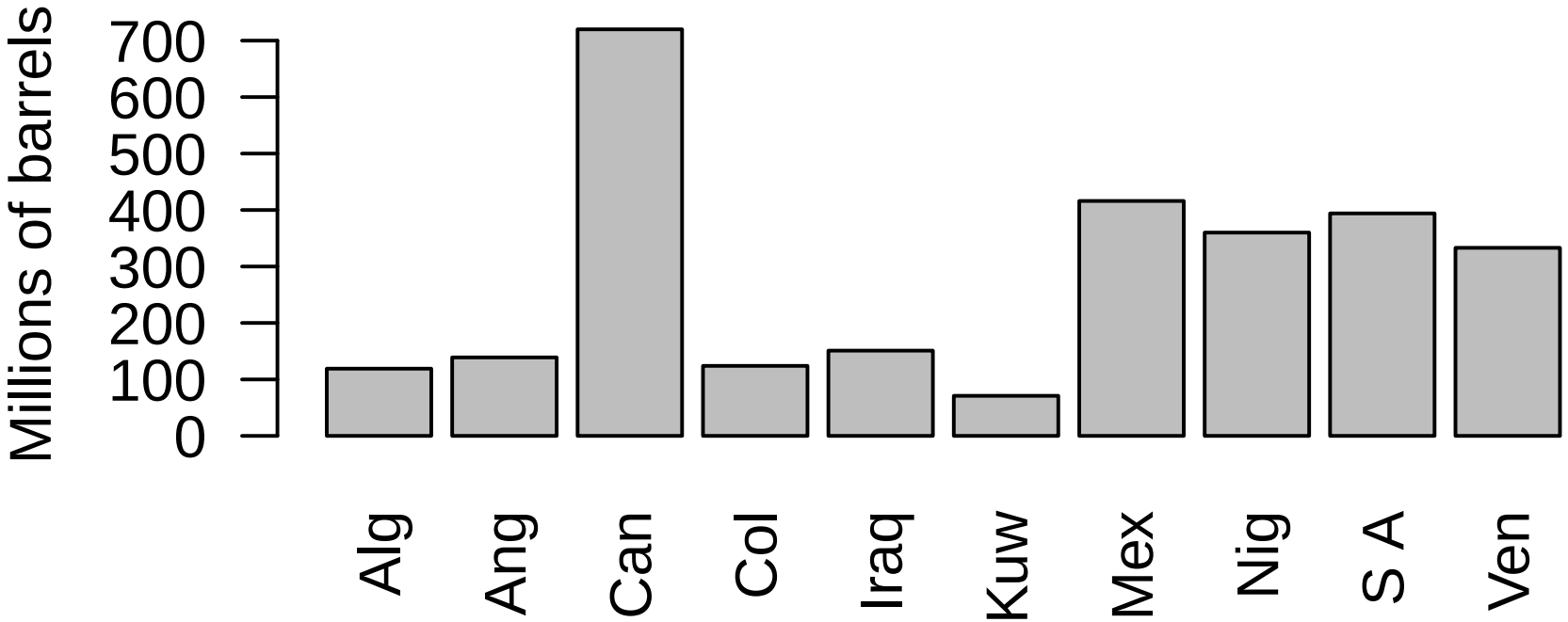
Another example (cont.)

a. Draw a bar chart.
b. Draw a pie chart.

R Code (Bar chart)

```r
# Step 1. Entering data;

barrels=c(119,139,720,124,151,71,416,360,394,333);
country=c("Alg","Ang","Can","Col","Iraq","Kuw","Mex",
"Nig","S A","Ven");


# Step 2. Making bar chart;

barplot(barrels,names.arg=country,ylab="Millions of barrels",
las=2);
```

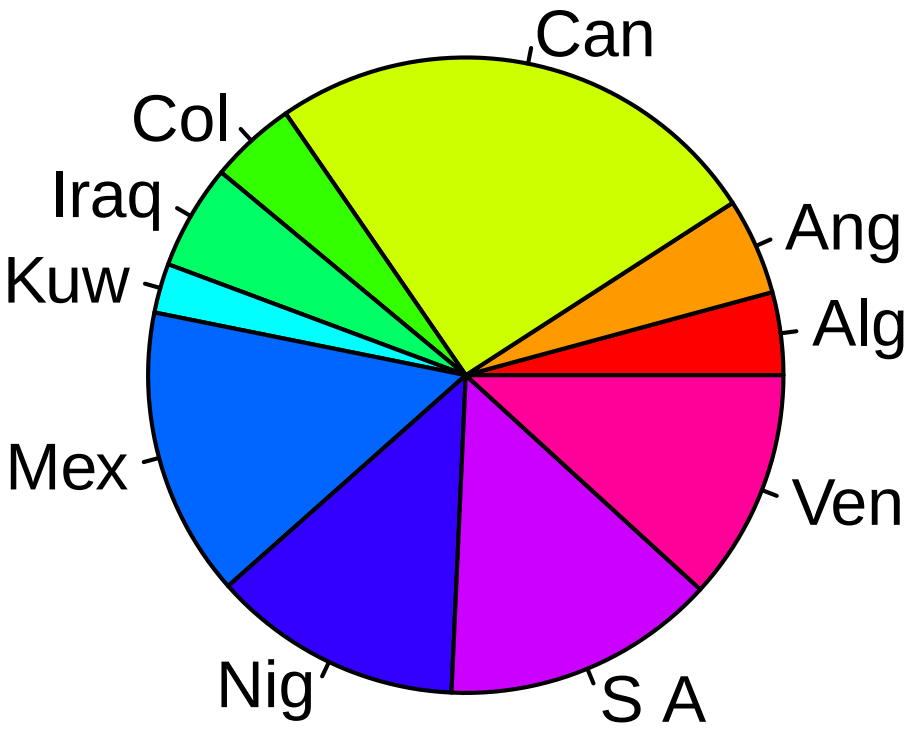
Bar chart


Millions of barrels
700
600
500
400
300
200
100
0
Alg
Ang
Can
Col
Iraq
Kuw
Mex
Nig
S A
Ven

R Code (Pie chart)

```
# Step 1. Entering data;

barrels=c(119,139,720,124,151,71,416,360,394,333);
country=c("Alg","Ang","Can","Col","Iraq","Kuw","Mex",
"Nig","S A","Ven");


# Step 2. Making pie chart;

pie(barrels,country,col=rainbow(10));
```

Pie chart



Age and Education

The table shown below presents Census Bureau data for the year 2000 on the level of education reached by Americans of different ages. Many people under 25 years of age have not completed their education, so they are left out of the table. Both variables, age and education, are grouped into categories. This is a **two-way table** (a.k.a. contingency table) because it describes two categorical variables. Education is the **row variable** because each row in the table describes people with one level of education. Age is the **column variable** because each column describes one age group. The entries in the table are the counts of persons in each age-by-education class. Although both age and education in this table are categorical variables, both have a natural order from least to most. The order of the rows and the columns in the table reflects the order of the categories.

Table

Years of school completed, by age (thousands of persons)

	Age group			
Education	25 to 34	35 to 54	55 and over	Total
Did not complete high school	4459	9174	14226	27859
Completed high school	11562	26455	20060	58077
College, 1 to 3 years	10693	22647	11125	44465
College, 4 or more years	11071	23160	10597	44828
Total	37786	81435	56008	175230

The distribution of a categorical variable says how often each outcome occurred. The distributions of education alone and age alone are called **marginal distributions** because they appear at the right and bottom margins of the two-way table.

Note. If you check the row and column totals in our Table, you will notice some discrepancies. For example, the sum of the entries in the "25 to 34" column is 37,785. The entry in the "Total" row for that column is 37,786. The explanation is **roundoff error**. The table entries are in thousands of persons, and each is rounded to the nearest thousand.

Calculating a marginal distribution

The percent of people 25 years of age and older who have at least 4 years of college is

$$\frac{\text{total with 4 years of college}}{\text{table total}} = \frac{44,828}{175,230} = 0.256 = 25.6\%$$

Calculating a marginal distribution

Do three more such calculations to obtain the marginal distributions of education level in percents. Here it is:

	Did not complete high school	Completed high school	1 to 3 years of college	4 or more years of college
Percent	15.9	33.1	25.4	25.6

The total is 100% because everyone is in one of the four education categories.
Each marginal distribution from a two-way table is a distribution for a single categorical variable.

Example. Calculating a conditional distribution

Information about the 25 to 34 age group occupies the first column in our Table. To find the complete distribution of education in this age group, look only at that column. Compute each count as a percent of the column total, which is 37,786. Here is the distribution:

	Did not complete high school	Completed high school	1 to 3 years of college	4 or more years of college
Percent	11.8	30.6	28.3	29.3

The four percents together are the **conditional distribution** of education, given that a person is 25 to 34 years of age. We use the term "conditional" because the distribution refers only to people who satisfy the condition that they are 25 to 34 years old.

Example. Calculating a conditional distribution (cont.)

Now focus in turn on the second column (people aged 35 to 54) and then the third column (people 55 and over) of our Table in order to find two more conditional distributions. Comparing the conditional distributions reveals the nature of the association between age and education. The distributions of education in the two younger groups are quite similar, but higher education is less common in the 55-and-over group (Homework?). Bar graphs can help make the association visible. We could make three side-by-side bar graphs to present the three conditional distributions (Homework?).

Do medical helicopters save lives?

Accident victims are sometimes taken by helicopter from the accident scene to a hospital. The helicopter may save time and also brings medical care to the accident scene. Does the use of helicopters save lives? We might compare the percents of accident victims who die with helicopter evacuation and with the usual transport to a hospital by road. Here are hypothetical data that illustrate a practical difficulty:

	Helicopter	Road
Victim died	64	260
Victim survived	136	840
Total	200	1100

Do medical helicopters save lives?

We see that 32% (64 out of 200) helicopter patients died, compared with only 23.64% (260 out 1100) of the others.

Do medical helicopters save lives?

That seems discouraging. The explanation is that the helicopter is sent mostly to serious accidents, so that the victims transported by helicopter are more often seriously injured that other victims. They are more likely to die with or without helicopter evacuation. Below we show the same data broken down by the seriousness of the accident.

Do medical helicopters save lives?

Serious Accidents

	Helicopter	Road
Victim died	48	60
Victim survived	52	40
Total	100	100

Do medical helicopters save lives?

Less Serious Accidents

	Helicopter	Road
Victim died	16	200
Victim survived	84	800
Total	100	1000

Do medical helicopters save lives?

Inspect these tables to convince yourself that they describe the same 1300 accidents as the original two-way table. For example, 200 were moved by helicopter, and 64 (48+16) of these died.
Among victims of serious accidents, the helicopter saves 52% compared with 40% for road transport. If we look only at less serious accidents, 84% of those transported by helicopter survive, versus 80% of those transported by road. Both groups of victims have a higher survival rate when evacuated by helicopter.

Do medical helicopters save lives?

At first, it seems paradoxical that the helicopter does better for both groups of victims but worse when all victims are lumped together. Examining the data makes the explanation clear. Half the helicopter transport patients are from serious accidents, compared with only 100 of the 1100 road transport patients. So the helicopter carries patients who are more likely to die. The seriousness of the accident was a lurking variable that made the relationship between survival and mode of transport to a hospital hard to interpret. Our example illustrates **Simpson's paradox**.

Simpson's paradox

An association or comparison that holds for all of several groups can reverse direction when the data are combined to form a single group. This reversal is called **Simpson's paradox**.

Risks of playing soccer

A study in Sweden looked at former elite soccer players, people who had played soccer but not at the elite level, and people of the same age who did not play soccer. Here is a two-way table that classifies these subjects by whether or not they had arthritis of the hip or knee by their mid-50s:

	Elite	Non-elite	Did not play
Arthritis	10	9	24
No arthritis	61	206	548

Risks of playing soccer

a) How many people do these data describe?
b) How many of these people have arthritis of the hip or knee?
c) Give the marginal distribution of participation in soccer, both as counts and as percents.

Solution

a) 858 people.
b) 43 had arthritis.
c) 71 (8.3%) played elite soccer,
215 (25.1%) played non-elite soccer,
and 572 (66.7%) did not play.

Risks of playing soccer (revisited)

Find the percent of each group in the soccer-risk data (see previous example) who have arthritis. What do these percents say about the association between playing soccer and later arthritis?

Solution

14.1% (10 out of 71) of elite players have arthritis, compared to 4.2% of the other two groups.
There is no difference between the non-elite group and those who did not play at all, but the percentage of elite players with arthritis is noticeably higher.

R Code

```r
# Step 1. Entering data;
table=matrix(c(10,61,9,206,24,528),nrow=2,ncol=3);
table;

# Note that R reads by column;
```

R Code

```
##       [,1] [,2] [,3]
## [1,]   10    9   24
## [2,]   61  206  528
```

R Code

```
# Giving names to columns and rows;
colnames(table)=c("Elite","Non-elite","Did not play");
rownames(table)=c("Arthritis",
"No arthritis");
table;
```

R Code

```
##                Elite Non-elite Did not play
## Arthritis         10         9           24
## No arthritis      61       206          528
```

R Code

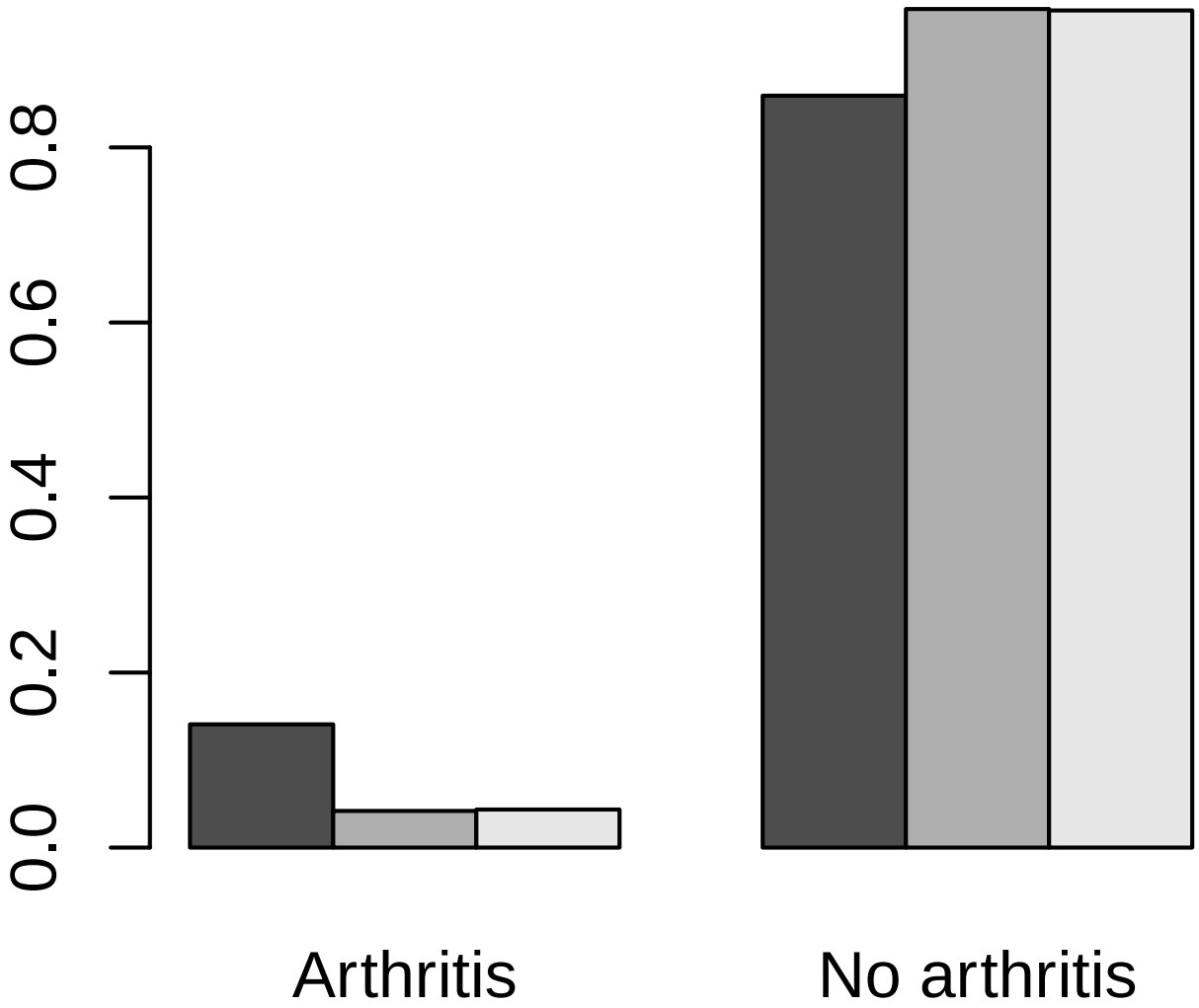
```
# Step 2. Making table of relative frequencies;

rel.freq.tab=prop.table(table,2);
rel.freq.tab;

# prop.table(table,2)
# that 2 is telling R to compute
# conditional distributions by column;
```

R Code

```
##                    Elite  Non-elite Did not play
## Arthritis     0.1408451 0.04186047   0.04347826
## No arthritis 0.8591549 0.95813953   0.95652174
```

R Code

```r
# Step 3. Graphing table of column relative
# frequencies;

barplot(t(rel.freq.tab),beside=T);
```


0.0
0.2
0.4
0.6
0.8
Arthritis
No arthritis

Majors for men and women in business

A study of the career plans of young women and men sent questionnaires to all 722 members of the senior class in the College of Business Administration at the University of Illinois. One question asked which major within the business program the student had chosen. Here are the data from the students who responded:

	Female	Male
Accounting	68	56
Administration	91	40
Economics	5	6
Finance	61	59

Majors for men and women in business

Find the two conditional distributions of major, one for women and one for men. Based on your calculations, describe the differences between women and men with a graph and in words.

Solution

For women: 30.2%, 40.4%, 2.2%, and 27.1%.
For men: 34.8%, 24.8%, 3.7%, and 36.6%.
The biggest difference between women and men is in administration: A higher percentage of women chose this major. Meanwhile, a greater proportion of men chose other fields, especially finance.