Topic 1
Displaying Data
Categorical Data

Outline

Types of Data

A data set provides information about a group of individuals.

These individuals are, typically, representatives chosen from a population under study. Data on the individuals are meant, either informally or formally, to allow us to make inferences about the population.

- Individuals are the objects described by the data.
- Variables are characteristics of an individual. In order to present data, we must first recognize the types of data under consideration.

Types of Data

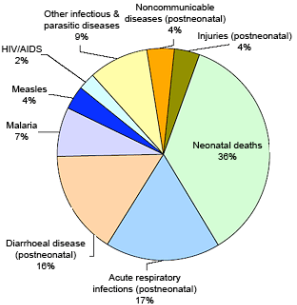
- Categorical variables partition the individuals into classes.
  - Other names for categorical variables are levels or factors.
- Quantitative variables are those for which arithmetic operations like addition and differences make sense.
  - Another name for a quantitative variable is feature.

Exercise. Give at least 8 variables for University students and classifying them as either categorical or quantitative.

Pie Charts

A pie chart is a circular chart divided into sectors, illustrating relative magnitudes in frequencies or percents. The area is proportional to the quantity it represents.

Example. From UNICEF, we read "The proportion of children who reach their fifth birthday is one of the most fundamental indicators of a country's concern for its people."



Major causes of death in neonates and children under five, 2004

Source: WHO The Global burden of disease:2004 update (2008)

Bar Charts

We will use R to make a bar chart of U.S. health expenditures in 2011.

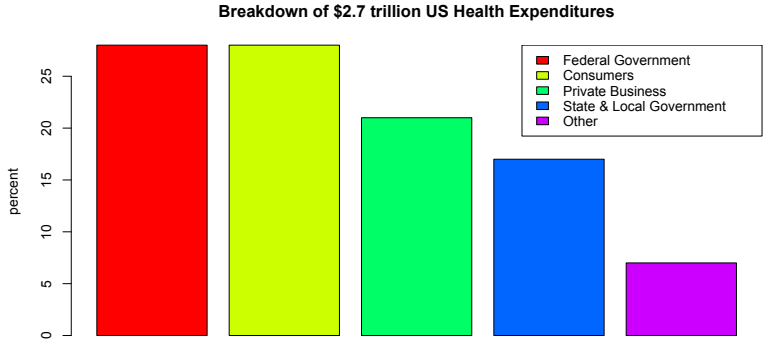
- First make a simple bar chart.
- Then, add colors and a legend on the *y*-axis.
- Finally, add a title and legend.

```
> expenditures<-c(28,28,21,17,7)
> barplot(expenditures)
```

```
> expenditures<-c(28,28,21,17,7)
> barplot(expenditures,ylab="percent",col=rainbow(5))
```

```
> expenditures<-c(28,28,21,17,7)
> barplot(expenditures,ylab="percent",col=rainbow(5),
  main="Breakdown of $2.7 trillion US Health Expenditures")
> legend("topright",c("Federal Government","Consumers","Private Business",
  "State & Local Government","Other"),fill=rainbow(5))
```

Bar Charts


Breakdown of $2.7 trillion US Health Expenditures
percent
0
5
10
15
20
25
Federal Government
Consumers
Private Business
State & Local Government
Other

Pie Chart versus Bar Chart

Because the human eye is good at judging linear measures and poor at judging relative areas, a bar chart is often preferable to pie charts as a way to display categorical data.

Exercise.

- Begin with four categories having 10, 25, 30, and 35 percent of the observations.
- Give successive pie charts and a box charts of the data.
  - Bring the final three categories values closer and closer to 30 percent, keeping the sum at 90 percent.
  - Save the pie chart and bar chart that is at the limit of your ability to discern the order of sizes of the categories.
    Choose for example, 10, 27, 30, 33 percent for the second pair of charts and continue to reduce the gap.

NB. Make pie charts using the `pie` command.

Two-way Table

Relationships between two categorical variables can be shown through a two-way table or contingency table (also known as a contingency table or cross tabulation).

Example. In 1964, Surgeon General Dr. Luther Leonidas Terry published a landmark report saying that smoking may be hazardous to health. This led to many influential reports on the topic, including the study of the smoking habits of 5375 high school children in Tucson in 1967. Here is a two-way table summarizing some of the results.

	student smokes	student does not smoke	total
2 parents smoke	400	1380	1780
1 parent smokes	416	1823	2239
0 parents smoke	188	1168	1356
total	1004	4371	5375

Two-way Table

	student smokes	student does not smoke	total
2 parents smoke	400	1380	1780
1 parent smokes	416	1823	2239
0 parents smoke	188	1168	1356
total	1004	4371	5375

	student smokes	student does not smoke	total
2 parents smoke	400	1380	1780
1 parent smokes	416	1823	2239
0 parents smoke	188	1168	1356
total	1004	4371	5375

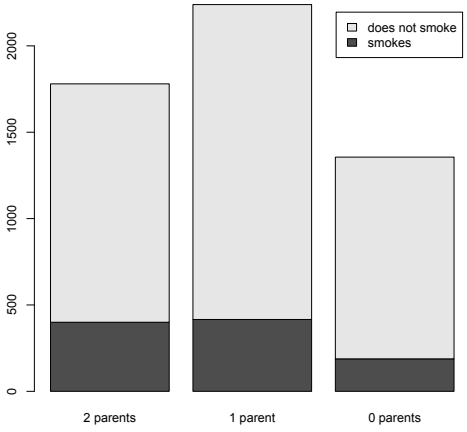
	student smokes	student does not smoke	total

Segmented Bar Chart

We can create a segmented bar chart as follows:

```
> smoking<-matrix(c(400,1380,416,1823,188,1168),ncol=3)
> colnames(smoking)<-c("2 parents","1 parent", "0 parents")
> rownames(smoking)<-c("smokes","does not smoke")
> smoking
               2 parents 1 parent 0 parents
smokes               400      416        188
does not smoke      1380     1823       1168
> barplot(smoking,legend=rownames(smoking))
```

Segmented Bar Chart



Segmented Bar Chart

Exercise. Hemoglobin E (HbE) is a variant of Hemoglobin A (HbA) with a mutation in the β globin. It has been suggested that HbE provides some protection against malaria virulence when heterozygous, but is causes anemia when homozygous.

The table below gives the counts of differing hemoglobin genotypes on two Indonesian islands.

genotype	AA	AE	EE
Flores	128	6	0
Sumba	119	78	4

Make a segmented bar chart of the data on hemoglobin genotypes. Have each bar display the distribution of genotypes on the two Indonesian islands.