Topic 1

Displaying Data

There are two goals when presenting data: convey your story and establish credibility. - Edward Tufte

Statistics is a mathematical science that is concerned with the collection, analysis, interpretation or explanation, and presentation of data.

The first encounters one has to data are through graphical displays and numerical summaries. The goal is to find an elegant method for this presentation that is at the same time both objective and informative - making clear with a few lines or a few numbers the salient features of the data. In this sense, data presentation is at the same time an art, a science, and an obligation to impartiality.

In the section, we will describe some of the standard presentations of data and at the same time, taking the opportunity to introduce some of the commands that the software package R provides to draw figures and compute summaries.

1.1 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. We shall later discuss how to define a population, how choose individuals in the population and how to collect data on these individuals.

- **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.
  - **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**.

Example 1.1 (individuals and variables)**.** *We consider two populations - the first is the nations of the world and the second is the people who live in those countries. Below is a collection of variables that might be used to study these populations.*

nations	people
population size	age
time zones	height
average rainfall	gender
life expectancy	ethnicities
mean income	annual income
literacy rate	literacy
capital city	mother's maiden name
largest river	marital status

Exercise 1.2. *Classify the variables as quantitative or categorical in the example above.*

The naming of variables and their classification as categorical or quantitative may seem like a simple, even trite, exercise. However, the first steps in designing an experiment and deciding on which individuals to include and which information to collect are vital to the success of the experiment. For example, if your goal is to measure the time for an animal (insect, bird, mammal) to complete some task under different (genetic, environmental, learning) conditions, then, you may decide to have a single quantitative variable - the time to complete the task. However, an animal in your study may not attempt the task, may not complete the task, or may perform the task. As a consequence, your data analysis will run into difficulties if you do not add a categorical variable to include these possible outcomes of an experiment.

Exercise 1.3. *Give examples of variables for the population of vertebrates, of proteins.*

1.2 Categorical Data

1.2.1 Pie Chart

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

Example 1.4. *As the nation debates strategies for delivering health insurance, let's look at the sources of funds and the types of expenditures.*

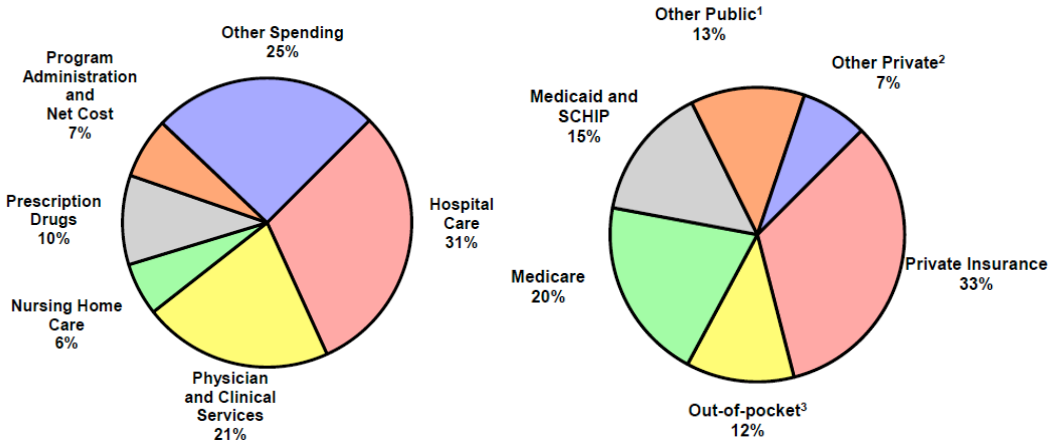
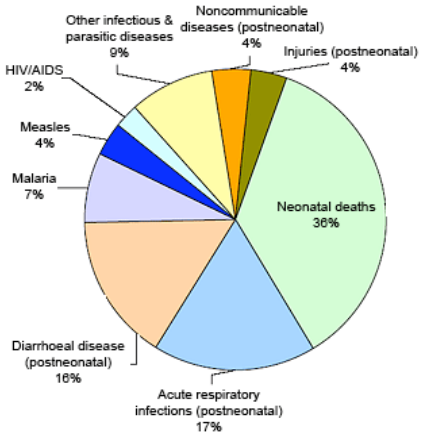


Figure 1.1: 2008 United States health care (a) expenditures (b) income sources, Source: Centers for Medicare and Medicaid Services, Office of the Actuary, National Health Statistics Group

Exercise 1.5. *How do you anticipate that this pie chart will evolve over the next decade? Which pie slices are likely to become larger? smaller? On what do you base your predictions?*

Example 1.6. *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. Child survival statistics are a poignant indicator of the priority given to the services that help a child to flourish: adequate supplies of nutritious food, the availability of high-quality health care and easy access to safe water and sanitation facilities, as well as the family's overall economic condition and the health and status of women in the community. "*



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

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

Example 1.7. Gene Ontology (GO) *project is a bioinformatics initiative whose goal is to provide unified terminology of genes and their products. The project began in 1998 as a collaboration between three model organism databases,* Drosophila*, yeast, and mouse. The GO Consortium presently includes many databases, spanning repositories for plant, animal and microbial genomes. This project is supported by National Human Genome Research Institute. See*

`http://www.geneontology.org/`

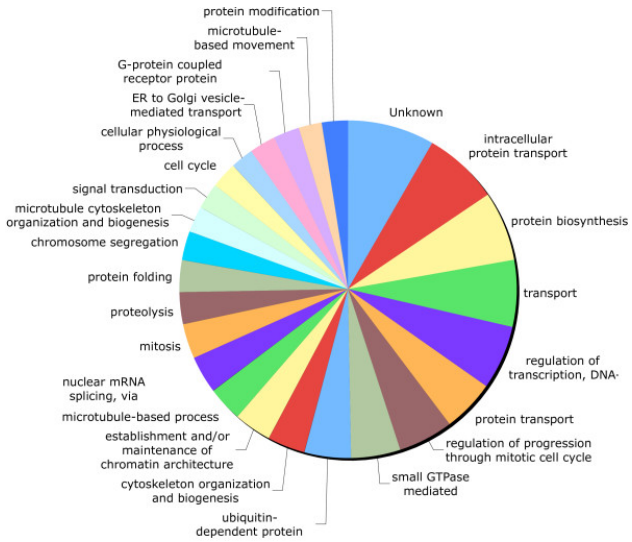


Figure 1.2: The 25 most frequent Biological Process Gene Ontology (GO) terms.

To make a simple **pie chart** in R for the proportion of AIDS cases among US males by transmission category.

```
> males<- c(58,18,16,7,1)
> pie(males)
```

This many be sufficient for your own personal use. However, if we want to use a pie chart in a presentation, we will have to provide some essential details. For a more descriptive pie chart, one has to become accustomed to learning to interact with the software to settle on a graph that is satisfactory to the situation.

- Define some colors ideal for black and white print.

  ```
  > colors <- c("white","grey70","grey90","grey50","black")
  ```

- Calculate the percentage for each category

  ```
  > male_labels <- round(males/sum(males)*100, 1)
  ```

  The number 1 indicates rounded to one decimal place.

  ```
  > male_labels <- paste(male_labels, "\%", sep=" ")
  ```

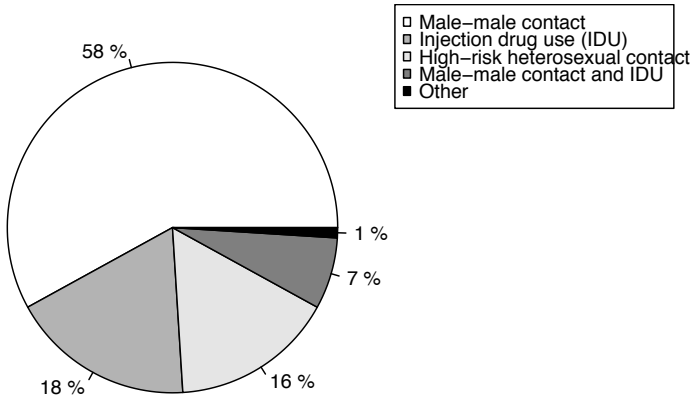
  This adds a space and a percent sign.

- Create a pie chart with defined heading and custom colors and labels and create a legend.

  ```
  > pie(males, main="Proportion of AIDS Cases among Males by  Transmission Category
  + Diagnosed - USA, 2005", col=colors, labels=male_labels, cex=0.8)
  > legend("topright", c("Male-male contact","Injection drug use (IDU)",
  + "High-risk heterosexual contact","Male-male contact and IDU","Other"),
  + cex=0.8,fill=colors)
  ```

  The entry `cex=0.8` indicates that the legend has a type set that is 80% of the font size of the main title.

Proportion of AIDS Cases among Males by Transmission Category Diagnosed – USA, 2005



1.2.2 Bar Charts

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

To make a simple bar graph in R,

```
> barplot(males)
```

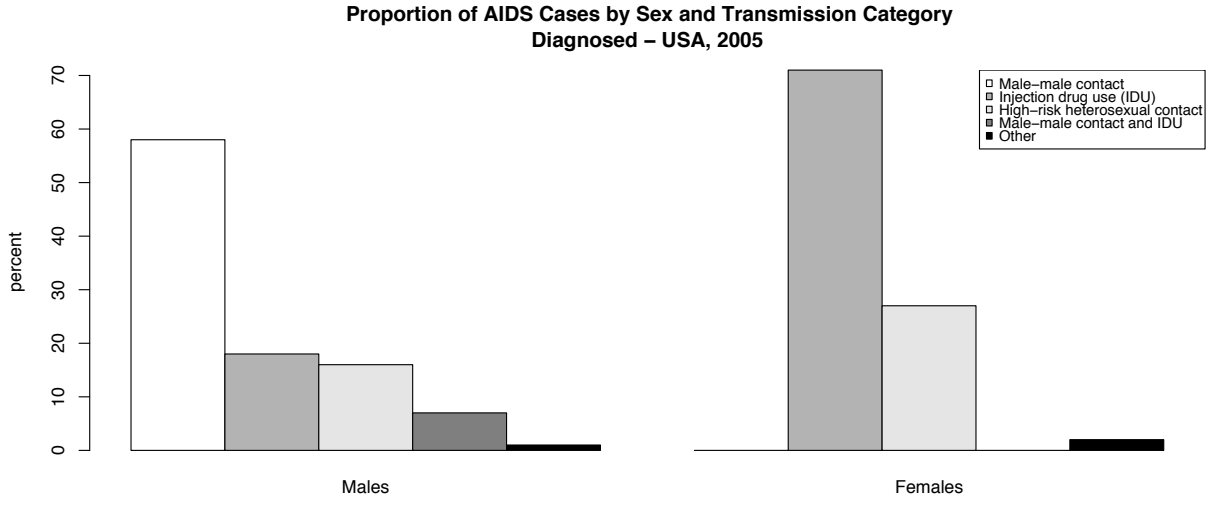
For a more descriptive bar chart with information on females:

- Enter the data for females and create a 5×2 array.

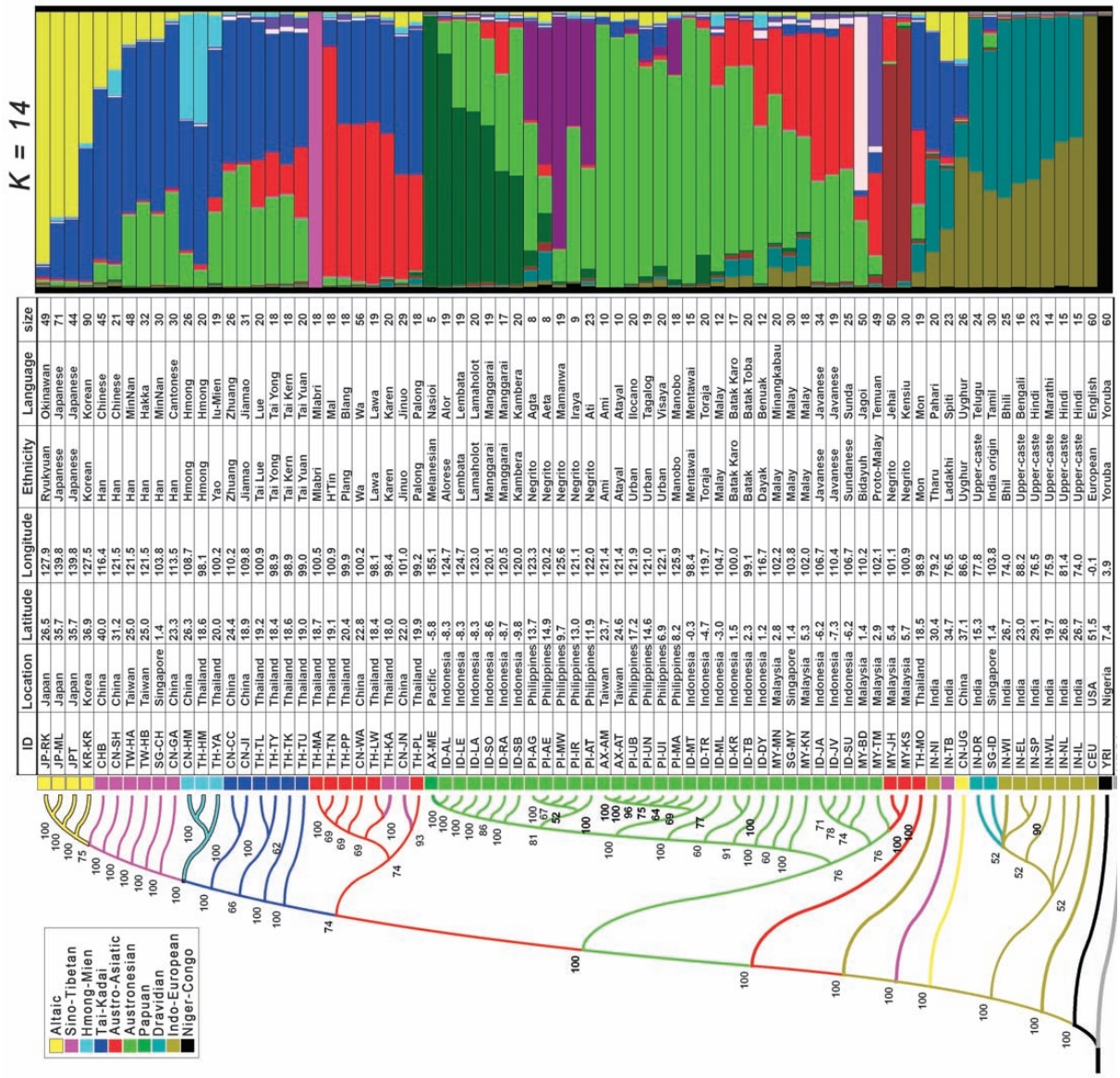
```
> females <- c(0,71,27,0,2)
> hiv<-array(c(males,females), dim=c(5,2))
```

- Generate side-by-side bar graphs and create a legend,

```
> barplot(hiv, main="Proportion of AIDS Cases by Sex and Transmission Category
+ Diagnosed - USA, 2005", ylab= "percent",  beside=TRUE,
+ names.arg = c("Males", "Females"),col=colors)
> legend("topright", c("Male-male contact","Injection drug use (IDU)",
+ "High-risk heterosexual contact","Male-male contact and IDU","Other"),
+ cex=0.8,fill=colors)
```

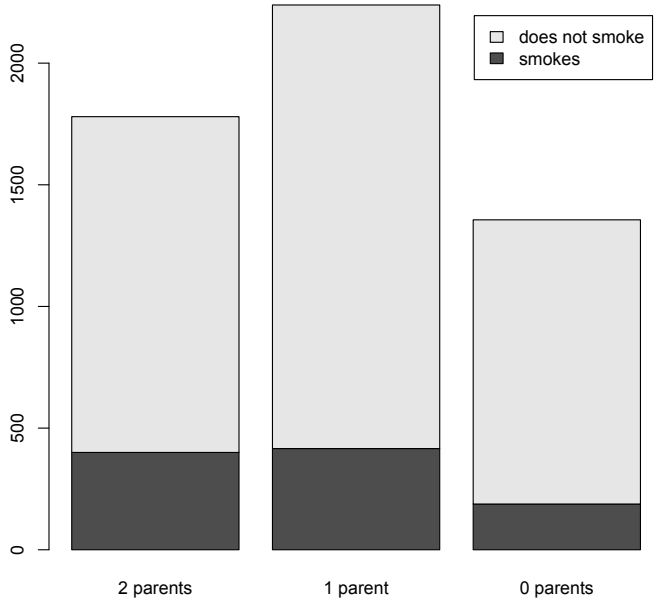


Example 1.8. *Next we examine a segmented bar plot. This shows the ancestral sources of genes for 75 populations throughout Asia. the data are based on information gathered from 50,000 genetic markers. The designations for the groups were decided by the software package STRUCTURE.*



ID	Location	Latitude	Longitude	Ethnicity	Language	size
JP-RK	Japan	26.5	127.9	Ryukyuan	Okinawan	49
JP-ML	Japan	35.7	139.8	Japanese	Japanese	71
JPT	Japan	35.7	139.8	Japanese	Japanese	44
KR-KR	Korea	36.9	127.5	Korean	Korean	90
CHB	China	40.0	116.4	Han	Chinese	45
CN-SH	China	31.2	121.5	Han	Chinese	21
TW-HA	Taiwan	25.0	121.5	Han	MinNan	48
TW-HB	Taiwan	25.0	121.5	Han	Hakka	32
SG-CH	Singapore	1.4	103.8	Han	MinNan	30
CN-GA	China	23.3	113.5	Han	Cantonese	30
CN-HM	China	26.3	108.7	Hmong	Hmong	26
TH-HM	Thailand	18.6	98.1	Hmong	Hmong	20
TH-YA	Thailand	20.0	100.2	Yao	Iu-Mien	19
CN-CC	China	24.4	110.2	Zhuang	Zhuang	26
CN-JI	China	18.9	109.8	Jiamao	Jiamao	31
TH-TL	Thailand	19.2	100.9	Tai Lue	Lue	20
TH-TY	Thailand	18.4	98.9	Tai Yong	Tai Yong	18
TH-TK	Thailand	18.6	98.9	Tai Kern	Tai Kern	18
TH-TU	Thailand	19.0	99.0	Tai Yuan	Tai Yuan	20
TH-MA	Thailand	18.7	100.5	Mlabri	Mlabri	18
TH-TN	Thailand	19.1	100.9	H'Tin	Mal	18
TH-PP	Thailand	20.4	99.9	Pliang	Blang	18
CN-WA	China	22.8	100.2	Wa	Wa	56
TH-LW	Thailand	18.4	98.1	Lawa	Lawa	19
TH-KA	Thailand	18.0	98.4	Karen	Karen	20
CN-JN	China	22.0	101.0	Jinuo	Jinuo	29
TH-PL	Thailand	19.9	99.2	Palong	Palong	18
AX-ME	Pacific	-5.8	155.1	Melanesian	Nasioi	5
ID-AL	Indonesia	-8.3	124.7	Alorese	Alor	19
ID-LE	Indonesia	-8.3	124.7	Lembata	Lembata	19
ID-LA	Indonesia	-8.3	123.0	Lamaholot	Lamaholot	20
ID-SO	Indonesia	-8.6	120.1	Manggarai	Manggarai	19
ID-RA	Indonesia	-8.7	120.5	Manggarai	Manggarai	17
ID-SB	Indonesia	-9.8	120.0	Kambera	Kambera	20
PI-AG	Philippines	13.7	123.3	Negrito	Agta	8
PI-AE	Philippines	14.9	120.2	Negrito	Aeta	8
PI-MW	Philippines	9.7	125.6	Negrito	Mamanwa	19
PI-IR	Philippines	13.0	121.1	Negrito	Iraya	9
PI-AT	Philippines	11.9	122.0	Negrito	Ati	23
AX-AM	Taiwan	23.7	121.4	Ami	Ami	10
AX-AT	Taiwan	24.6	121.4	Atayal	Atayal	10
PI-UB	Philippines	17.2	121.9	Urban	Ilocano	20
PI-UN	Philippines	14.6	121.0	Urban	Tagalog	19
PI-UI	Philippines	6.9	122.1	Urban	Visaya	20
PI-MA	Philippines	8.2	125.9	Manobo	Manobo	18
ID-MT	Indonesia	-0.3	98.4	Mentawai	Mentawai	15
ID-TR	Indonesia	-4.7	119.7	Toraja	Toraja	20
ID-ML	Indonesia	-3.0	104.7	Malay	Malay	12
ID-KR	Indonesia	1.5	100.0	Batak Karo	Batak Karo	17
ID-TB	Indonesia	2.3	99.1	Batak	Batak Toba	20
ID-DY	Indonesia	1.2	116.7	Dayak	Benuak	12
MY-MN	Malaysia	2.8	102.2	Malay	Minangkabau	20
SG-MY	Singapore	1.4	103.8	Malay	Malay	30
MY-KN	Malaysia	5.3	102.0	Malay	Malay	18
ID-JA	Indonesia	-6.2	106.7	Javanese	Javanese	34
ID-JV	Indonesia	-7.3	110.4	Javanese	Javanese	19
ID-SU	Indonesia	-6.2	106.7	Sundanese	Sunda	25
MY-BD	Malaysia	1.4	110.2	Bidayuh	Jagoi	50
MY-TM	Malaysia	2.9	102.1	Proto-Malay	Temuan	49
MY-JH	Malaysia	5.4	101.1	Negrito	Jehai	50
MY-KS	Malaysia	5.7	100.9	Negrito	Kensiu	30
TH-MO	Thailand	18.5	98.9	Mon	Mon	19
IN-NI	India	30.4	79.2	Tharu	Pahari	20
IN-TB	India	34.7	76.5	Ladakhi	Spiti	23
CN-UG	China	37.1	86.6	Uyghur	Uyghur	26
IN-DR	India	15.3	77.8	Upper-caste	Telugu	24
SG-ID	Singapore	1.4	103.8	India origin	Tamil	30
IN-WI	India	26.7	74.0	Bhil	Bhili	25
IN-EL	India	23.0	88.2	Upper-caste	Bengali	16
IN-SP	India	29.1	76.5	Upper-caste	Hindi	23
IN-WL	India	19.7	75.9	Upper-caste	Marathi	14
IN-NL	India	26.8	81.4	Upper-caste	Hindi	15
IN-IL	India	26.7	74.0	Upper-caste	Hindi	15
CEU	USA	51.5	-0.1	European	English	60
YRI	Nigeria	7.4	3.9	Yoruba	Yoruba	60

Figure 1.3: Dispaying human genetic diversity for 75 populations in Asia. The software program STRUCTURE here infers 14 source populations, 10 of them major. The length of each segment in the bar is the estimate by STRUCTURE of the fraction of the genome in the sample that has ancestors among the given source population.



1.3 Two-way Tables

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

Example 1.9. *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

- The **column variable** is the student smoking habits.
- The **row variable** is the parents smoking habits.

The totals along each of the rows and columns give the **marginal distributions**.
We can create a **segmented bar graph** 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))
```

Example 1.10. *Hemoglobin E is a variant of hemoglobin with a mutation in the β globin gene causing substitution of glutamic acid for lysine at position 26 of the β globin chain. HbE (E is the one letter abbreviation for glutamic acid.) is the second most common abnormal hemoglobin after sickle cell hemoglobin (HbS). HbE is common from India to Southeast Asia. The β chain of HbE is synthesized at a reduced rate compare to normal hemoglobin (HbA) as the HbE produces an alternate splicing site within an exon.*

It has been suggested that Hemoglobin E provides some protection against malaria virulence when heterozygous, but is causes anemia when homozygous. The circumstance in which the heterozygotes for the alleles under consideration have a higher adaptive value than the homozygote is called **balancing selection**.

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

Because the heterozygotes are rare on Flores, it appears malaria is less prevalent there since the heterozygote does not provide an adaptive advantage.

Exercise 1.11. *Make a segmented barchart of the data on hemoglobin genotypes. Have each bar display the distribution of genotypes on the two Indonesian islands.*

1.4 Histograms

Histograms are a common visual representation of a quantitative variable. Histograms visual the data using rectangles to display frequencies and proportions as normalized frequencies. In making a histogram, we

- Divide the range of data into bins of equal width (usually, but not always)
- Count the number of observations in each class.
- Draw the histogram rectangles representing frequencies or percents by *area*.

Interpret the histogram by giving

- the overall pattern
  - the center
  - the spread
  - the shape (symmetry, skewness, peaks)
- and deviations from the pattern
  - outliers
  - gaps

The direction of the skewness is the direction of the longer of the two tails (left or right) of the distribution.

No one choice for the number of bins is considered best. One possible choice for larger data sets is Sturges' formula to choose $\lfloor 1 + \log_2 n \rfloor$ bins. ($\lfloor \cdot \rfloor$, the **floor function**, is obtained by rounding down to the next integer.)

Exercise 1.12. *The histograms in Figure 4 shows the distribution of lengths of a normal strain and mutant strain of* Bacillus subtilis*. Describe the distributions.*

Example 1.13. *Taking the age of the presidents of the United States at the time of their inauguration and creating its histogram in* R *is accomplished as follows.*

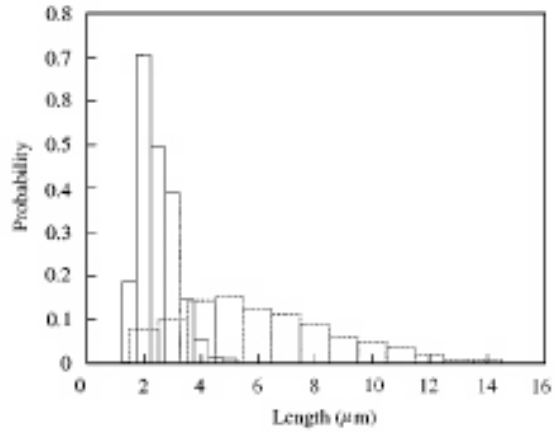
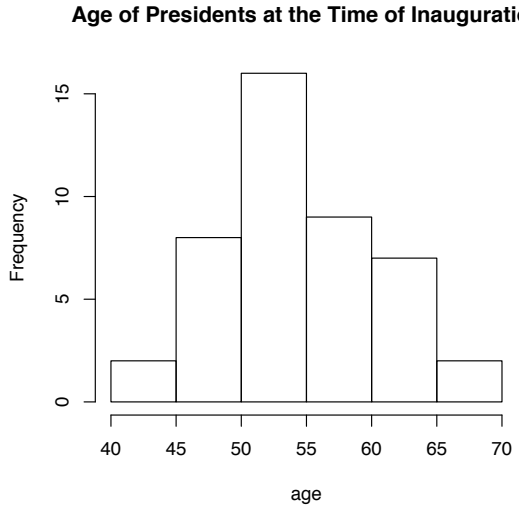


Figure 1.4: Histogram of lengths of *Bacillus subtilis*. Solid lines indicate wild type and dashed line mutant strain.

```
> age<- c(57,61,57,57,58,57,61,54,68,51,49,64,50,48,65,52,56,46,54,49,51,47,55,55,
54,42,51,56,55,51,54,51,60,61,43,55,56,61,52,69,64,46,54,47)
> hist(age, main = c("Age of Presidents at the Time of Inauguration"))
```



So the age of presidents at the time of inauguration range from the early forties to the late sixties with the frequency starting their tenure peaking in the early fifties. The histogram in generally symmetric about 55 years with spread from around 40 to 70 years.

The **empirical cumulative distribution function** $F_n(x)$ *gives, for each value* x*, the fraction of the data* **less** *than or equal to* x*. If the number of observations is* n*, then*

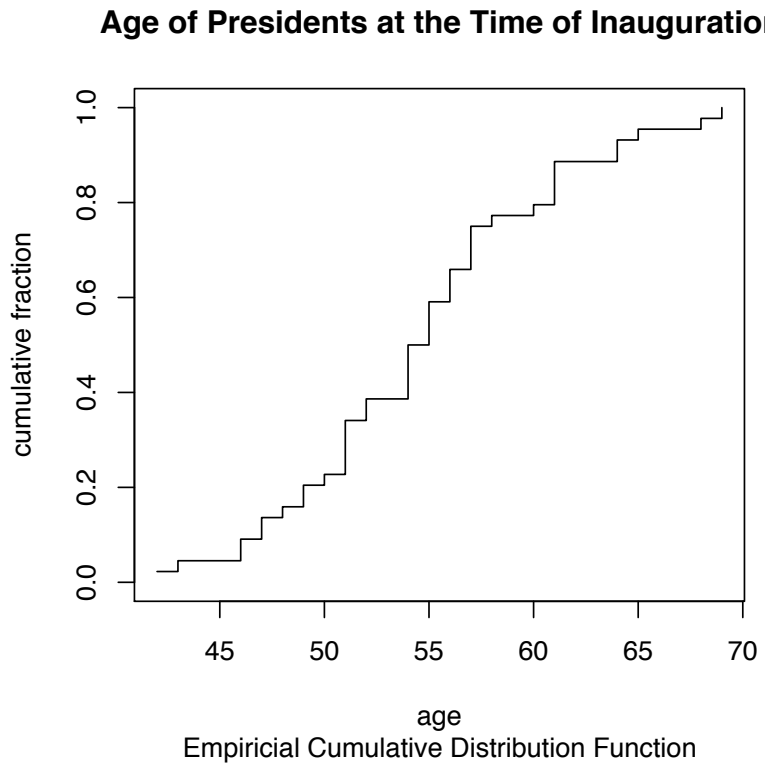
$$F_n(x) = \frac{1}{n}\#(\text{observations less than or equal to } x).$$

Thus, $F_n(x) = 0$ *for any value of* x *less than all of the observed values and* $F_n(x) = 1$ *for any* x *greater than all of the observed values. In between, we will see jumps that are multiples of the* $1/n$*. For example, in the empirical*

cumulative distribution function for the age of the presidents, we will see a jump of size $4/44 = 1/11$ *at* $x = 57$ *to indicate the fact that 4 of the 44 presidents were 57 at the time of their inauguration.*

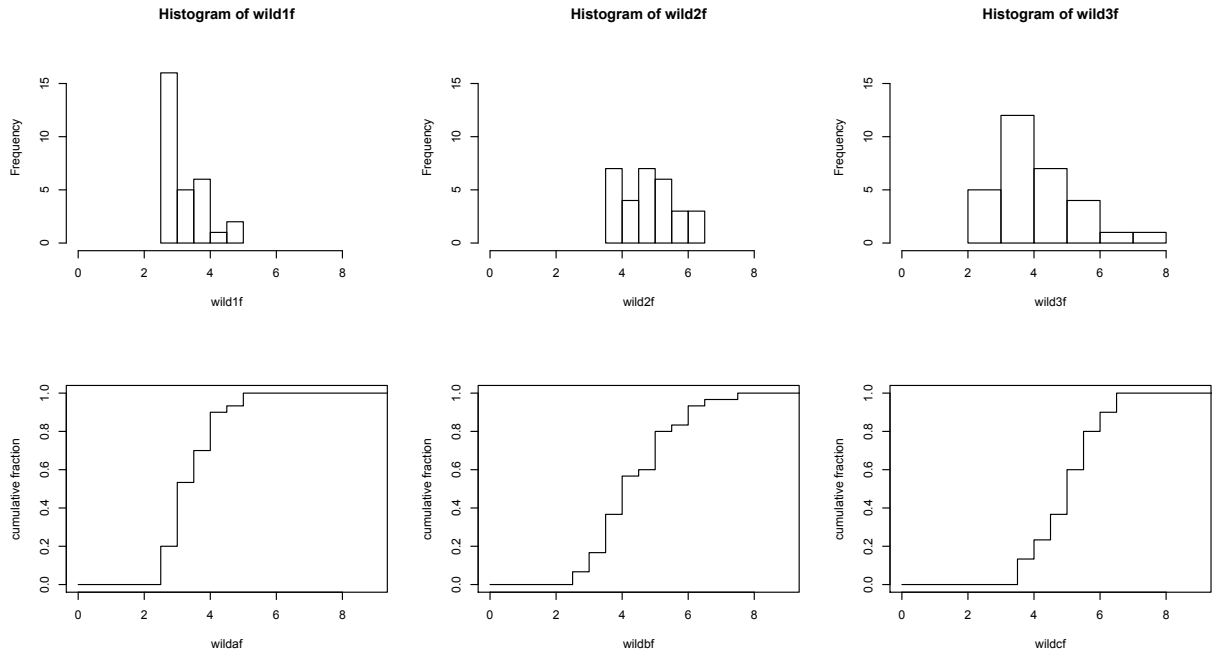
In order to create a graph of the empirical cumulative distribution function, first place the observations in order from smallest to largest. For the age of presidents data, we can accomplish this in R *by writing* `sort(age)`. *Next match these up with the integral multiples of the 1 over the number of observations. In* R*, we enter* `1:length(age)/length(age)`. *Finally,* `type="s"` *to give us the steps described above.*

```
> plot(sort(age),1:length(age)/length(age),type="s",ylim=c(0,1),
main = c("Age of Presidents at the Time of Inauguration"),
sub=("Empiricial Cumulative Distribution Function"),
xlab=c("age"),ylab=c("cumulative fraction"))
```



Exercise 1.14. *Give the fraction of presidents whose age at inauguration was under 60. What is the range for the age at inauguration of the youngest fifth of the presidents?*

Exercise 1.15. *The histogram for data on the length of three bacterial strains is shown below. Lengths are given in microns. Below the histograms (but not necessarily directly below) are empirical cumulative distribution functions corresponding to these three histograms.*



Match the histograms to their respective empirical cumulative distribution functions.

In looking at life span data, the natural question is "What fraction of the individuals have survived a given length of time?" The **survival function** $S_n(x)$ gives, for each value x, the fraction of the data **greater** than or equal to x. If the number of observations is n, then

$$\begin{aligned} S_n(x) &= \frac{1}{n}\#(\text{observations greater than } x) = \frac{1}{n}(n - \#(\text{observations less than or equal to } x)) \\ &= 1 - \frac{1}{n}\#(\text{observations less than or equal to } x) = 1 - F_n(x) \end{aligned}$$

1.5 Scatterplots

We now consider two dimensional data. The values of the first variable $x_1, x_2, \ldots, x_n$ are assumed known and in an experiment and are often set by the experimenter. This variable is called the **explanatory, predictor, or discriptor variables** and in a two dimensional **scatterplot** of the data display its values on the horizontal axis. The values $y_1, y_2 \ldots, y_n$, taken from observations with input $x_1, x_2, \ldots, x_n$ are called the **response variable** and its values are displayed on the vertical axis. In describing a scatterplot, take into consideration

- the form, for example,
  - linear
  - curved relationships
  - clusters
- the direction,
  - a positive or negative association
- and the strength of the aspects of the scatterplot.

Example 1.16 (Fossils of the *Archeopteryx*)**.** *The name Archeopteryx derives from the ancient Greek meaning "ancient feather" or "ancient wing". Archeopteryx is generally accepted by palaeontologists as being the oldest known bird. Archaeopteryx lived in the Late Jurassic Period around 150 million years ago, in what is now southern Germany during a time when Europe was an archipelago of islands in a shallow warm tropical sea. The first complete specimen of Archaeopteryx was announced in 1861, only two years after Charles Darwin published* On the Origin of Species*, and thus became a key piece of evidence in the debate over evolution. Below are the lengths in centimeters of the femur and humerus for the five specimens of* Archeopteryx *that have preserved both bones.*

femur	38	56	59	64	74
humerus	41	63	70	72	84

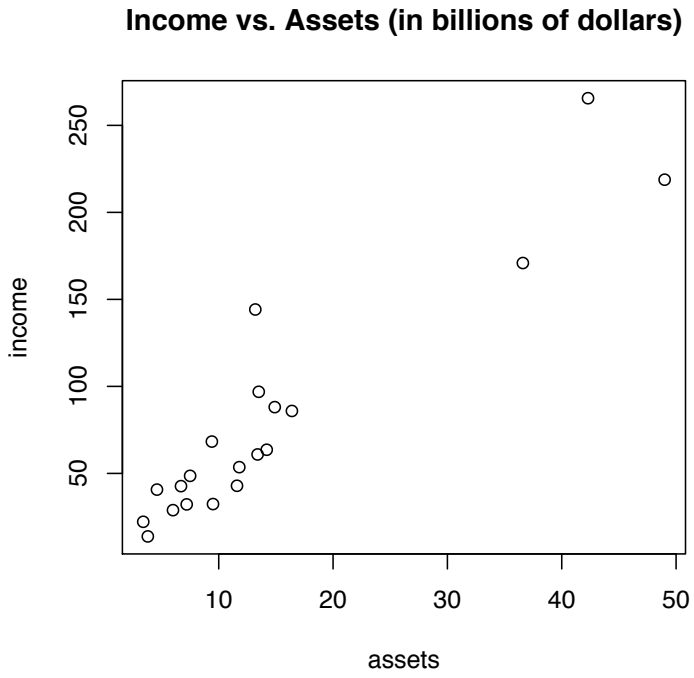
```
> femur<-c(38,56,59,64,74)
> humerus<-c(41,63,70,72,84)
> plot(femur, humerus,main=c("Bone Lengths for Archeopteryx"))
```

Unless we have a specific scientific question, we have no real reason for a choice of the explanatory variable.

Describe the scatterplot.

Example 1.17. *This historical data show the 20 largest banks in 1974. Values given in billions of dollars.*

Bank	1	2	3	4	5	6	7	8	9	10
Assets	49.0	42.3	36.6	16.4	14.9	14.2	13.5	13.4	13.2	11.8
Income	218.8	265.6	170.9	85.9	88.1	63.6	96.9	60.9	144.2	53.6
Bank	11	12	13	14	15	16	17	18	19	20
Assets	11.6	9.5	9.4	7.5	7.2	6.7	6.0	4.6	3.8	3.4
Income	42.9	32.4	68.3	48.6	32.2	42.7	28.9	40.7	13.8	22.2



Describe the scatterplot.

In 1972, Michele Sindona, a banker with close ties to the Mafia, along with a purportedly bogus Freemasonic lodge, and the Nixon administration purchased controlling interest in Bank 19, Long Island's Franklin National Bank. As a result of his acquisition of a controlling stake in Franklin, Sindona had a money laundering operation to aid his alleged ties to Vatican Bank and the Sicilian drug cartel. Sindona used the bank's ability to transfer funds, produce letters of credit, and trade in foreign currencies to begin building a banking empire in the United States. In mid-1974, management revealed huge losses and depositors started taking out large withdrawals, causing the bank to have to borrow over $1 billion from the Federal Reserve Bank. On 8 October 1974, the bank was declared insolvent due to mismanagement and fraud, involving losses in foreign currency speculation and poor loan policies.

What would you expect to be a feature on this scatterplot of a failing bank? Does the Franklin Bank have this feature?

1.6 Time Plots

Some data sets come with an order of events, say ordered by time.

Example 1.18. *The modern history of petroleum began in the 19th century with the refining of kerosene from crude oil. The world's first commercial oil wells were drilled in the 1850s in Poland and in Romania.The first oil well in North America was in Oil Springs, Ontario, Canada in 1858. The US petroleum industry began with Edwin Drake's drilling of a 69-foot deep oil well in 1859 on Oil Creek near Titusville, Pennsylvania for the Seneca Oil Company. The industry grew through the 1800s, driven by the demand for kerosene and oil lamps. The introduction of the internal combustion engine in the early part of the 20th century provided a demand that has largely sustained the industry to this day. Today, about 90% of vehicular fuel needs are met by oil. Petroleum also makes up 40% of total energy consumption in the United States, but is responsible for only 2% of electricity generation. Oil use increased exponentially until the world oil crises of the 1970s.*

Worldwide Oil Production

Year	Million Barrels	Year	Million Barrels	Year	Million Barrels
1880	30	1940	2150	1972	18584
1890	77	1945	2595	1974	20389
1900	149	1950	3803	1976	20188
1905	215	1955	5626	1978	21922
1910	328	1960	7674	1980	21722
1915	432	1962	8882	1982	19411
1920	689	1964	10310	1984	19837
1925	1069	1966	12016	1986	20246
1930	1412	1968	14014	1988	21338
1935	1655	1970	16690		

With the data given in two columns `oil` *and* `year`*, the time plot* `plot(year,oil,type="b")` *is given on the left side of the figure below. This uses* `type="b"` *that puts* **both** *lines and circles on the plot.*

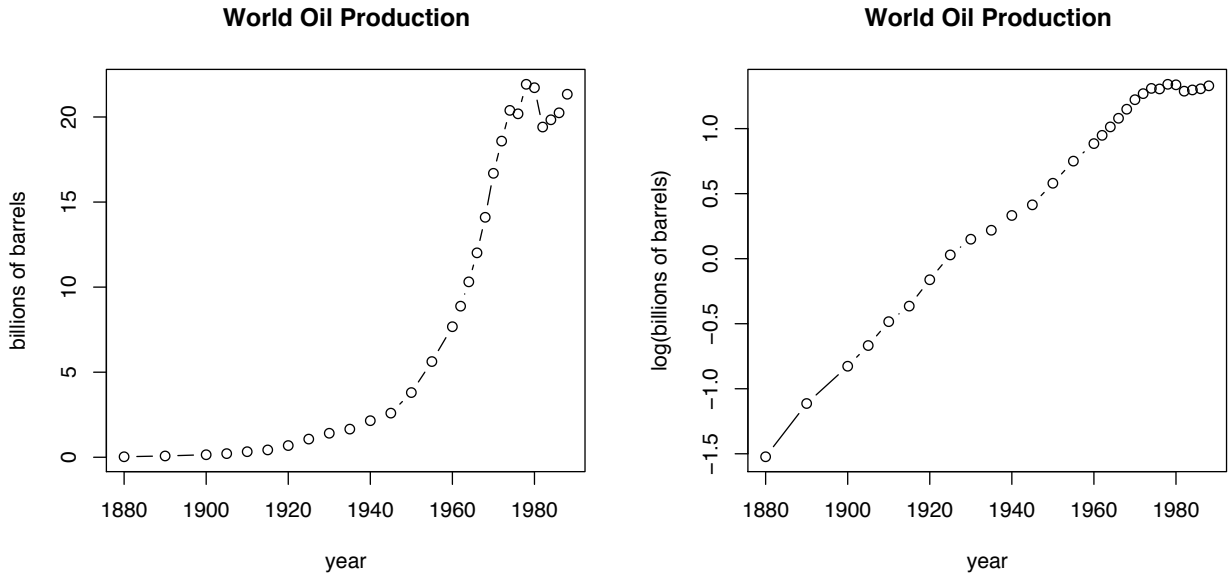


Figure 1.5: Oil production (left) and the logarithm of oil production (right) from 1880 to 1988.

Sometimes a **transformation of the data** *can reveal the structure of the time series. For example, if we wish to examine an exponential increase displayed in the oil production plot, then we can take the base 10 logarithm of the production and give its time series plot. This is shown in the plot on the right above. (In* R*, we write* `log(x)` *for the natural logarithm and* `log(x,10)` *for the base 10 logarithm.)*

Exercise 1.19. *What happened in the mid 1970s that resulted in the long term departure from exponential growth in the use of oil?*

Example 1.20. *The Intergovernmental Panel on Climate Change (IPCC) is a scientific intergovernmental body tasked with evaluating the risk of climate change caused by human activity. The panel was established in 1988 by the World Meteorological Organization and the United Nations Environment Programme, two organizations of the United Nations. The IPCC does not perform original research but rather uses three working groups who synthesize research and prepare a report. In addition, the IPCC prepares a summary report. The Fourth Assessment Report (AR4) was completed in early 2007. The fifth is scheduled for release in 2014.*

Below is the first graph from the 2007 Climate Change Synthesis Report: Summary for Policymakers.

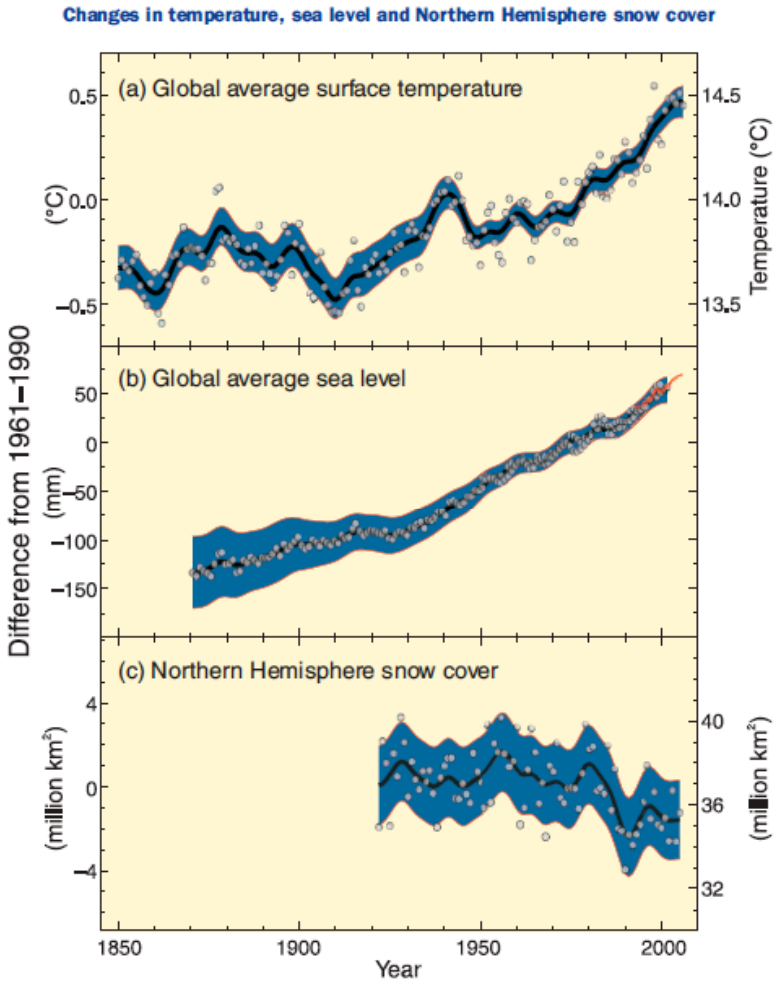
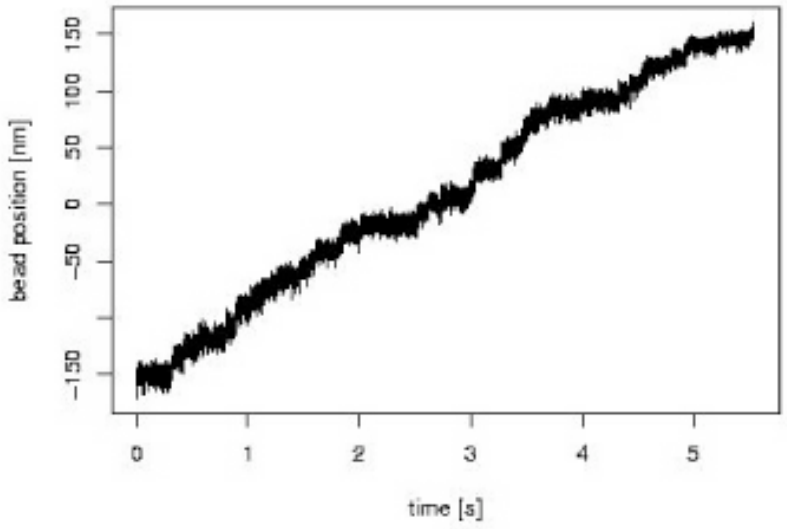
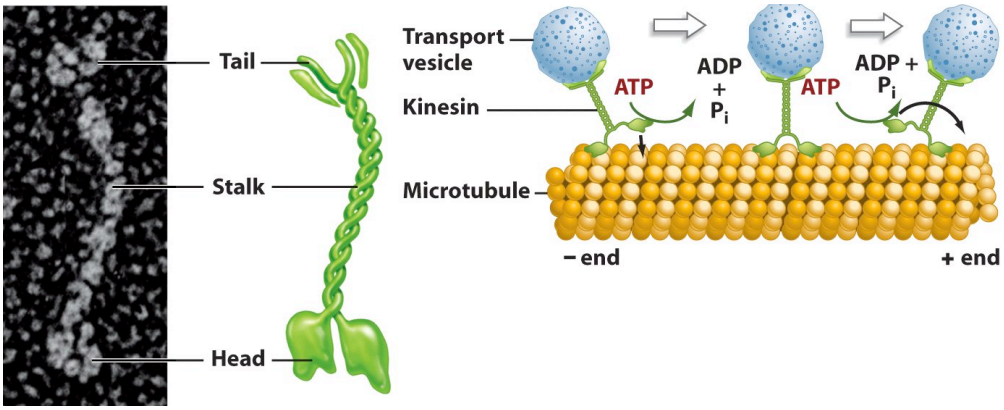


Figure SPM.1. Observed changes in (a) global average surface temperature; (b) global average sea level from tide gauge (blue) and satellite (red) data and (c) Northern Hemisphere snow cover for March-April. All differences are relative to corresponding averages for the period 1961-1990. Smoothed curves represent decadal averaged values while circles show yearly values. The shaded areas are the uncertainty intervals estimated from a comprehensive analysis of known uncertainties (a and b) and from the time series (c). (Figure 1.1)

The technique used to draw the curves on the graphs is called **local regression**. At the risk of discussing concepts that have not yet been introduced, let's describe the technique behind local regression. Typically, at each point in the data set, the goal is to draw a linear or quadratic function. The function is determined using weighted least squares, giving most weight to nearby points and less weight to points further away. The graphs above show the approximating curves. The blue regions show areas within two standard deviations of the estimate (called a confidence interval). The goal of local regression is to provide a smooth approximation to the data and a sense of the uncertainty of the data. In practice, local regression requires a large data set to work well.

Example 1.21. *The next figure give a time series plot of a single molecule experiment showing the movement of kinesin along a microtubule. In this case the kinesin has at its foot a glass bead and its heads are attached to a microtubule. The position of the glass bead is determined by using a laser beam and the optical properties of the bead to locate the bead and provide a force on the kinesin molecule. In this time plot, the load on the microtubule has a force of 3.5 pN and the concentration of ATP is 100μM. What is the source of fluctuations in this time series plot of bead position? How would you expect this time plot to change with changes in ATP concentration and with changes in force?*

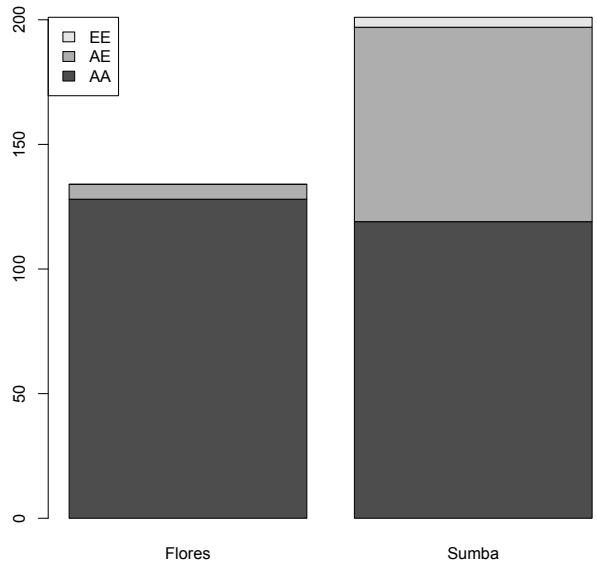




1.7 Answers to Selected Exercises

1.11. Here are the R commands:

```
> genotypes<-matrix(c(128,6,0,119,78,4),ncol=2)
> colnames(genotypes)<-c("Flores","Sumba")
> rownames(genotypes)<-c("AA","AE","EE")
> genotypes
   Flores Sumba
AA    128   119
AE      6    78
EE      0     4
> barplot(genotypes,legend=rownames(genotypes),args.legend=list(x="topleft"))
```



The legend was moved to the left side to avoid crowding with the taller bar for the data on Sumba.

1.12. The lengths of the normal strain has its center at 2.5 microns and range from 1.5 to 5 microns. It is somewhat skewed right with no outliers. The mutant strain has its center at 5 or 6 microns. Its range is from 2 to 14 microns and it is slightly skewed right. It has not outliers.

1.14. Look at the graph to the point above the value 60 years. Look left from this point to note that it corresponds to a value of 0.80.

Look at the graph to the point right from the value 0.20. Look down to note that it corresponds to 49 years. .

1.15. Match histogram *wild1f* to *wilddaf*. Note that both show the range is from 2 to 5 microns and that about half of the data lies between 2 and 3 microns. Match histogram *wild2f* with *wildcf*. The data is relatively uniform from 3.5 to 6.5 microns. Finally, match histogram *wild3f* with *wildbf*. The range is from 2 to 8 microns with most of the data between 3 and 6 microns. .

1.21. The fluctuation are due to the many bombardments with other molecules in the cell, most frequently, water molecules.

As force increases, we expect the velocity to increase - to a point. If the force is too large, then the kinesin is ripped away from the microtubule. As ATP concentration increases, we expect the velocity to increase - again, to a point. If ATP concentration is sufficiently large, then the biochemical processes are saturated.