Chapter 1 - Graphical and Numerical Summaries

Read sections 1.6 - 1.7

Graphical Methods

How to display data:

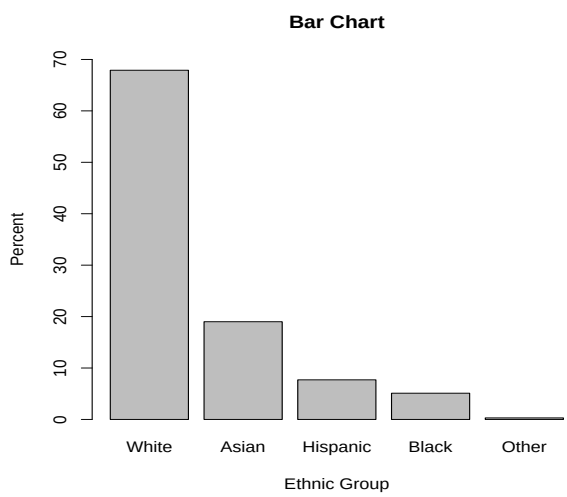
1. Give the observed values of a categorical or numerical variable taken from a sample, denoted as x_1, x_2, x_3, ..., x_n. The **sample size** is n
2. Indicate how often the variable takes on these values.

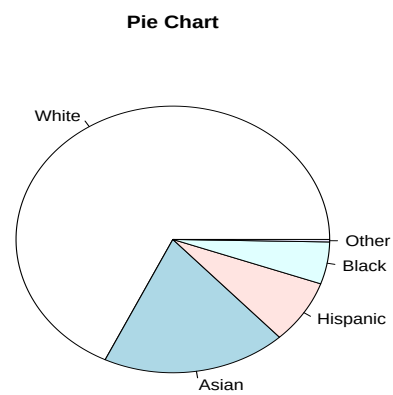
Displaying Categorical Data

To display a categorical variable measured from a sample:

- Encode the values of the variable with respect to each category into binary numbers: a 0 or a 1.
- **Frequency** is the count of 1's there are in each category, or the number of times a category appears
- **Relative Frequency** is the proportion $\frac{\text{Frequency}}{\text{Total Number of Observations}}$
- **Percentage** = (Relative Frequency) $\times$ 100

1. **Bar Chart:** A bar chart is a graph of frequencies, relative frequencies, or percentages on the vertical axis versus the categories of a categorical variable on the horizontal axis.



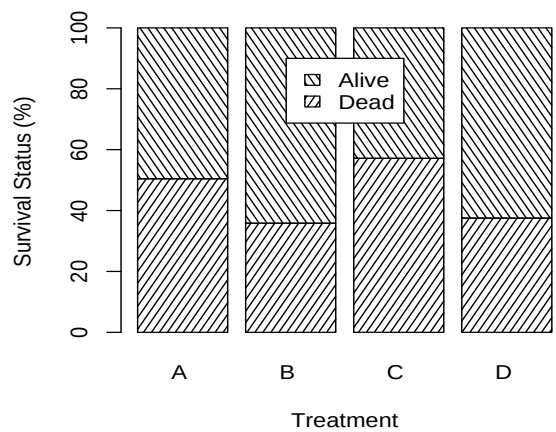



```
> freq = c(679,190,77,51,3)     # the frequencies on each category
> rel.freq=prop.table(freq)     # relative frequencies
> rel.freq
[1] 0.679 0.190 0.077 0.051 0.003
> percent=rel.freq*100
> percent
[1] 67.9 19.0  7.7  5.1  0.3
> Race = c("White","Asian","Hispanic","Black","Other")
> rbind(Race,percent)
        [,1]    [,2]    [,3]       [,4]    [,5]
Race    "White" "Asian" "Hispanic" "Black" "Other"
percent "67.9"  "19"    "7.7"      "5.1"   "0.3"

> barplot(percent,main="Bar Chart",names=Race,xlab="Ethnic Group",ylab="Percent",ylim=c(0,70))
> pie(percent,main="Pie Chart",labels=Race)
```

2. **Pie Chart:** A graph of the categories of a categorical variable as pieces of a pie, where the size of each piece is proportional to the frequency, relative frequency, or percentage of the category. Pie charts are inferior to bar charts because humans have a more difficult time judging the difference between angles than the difference between heights or lengths of bars.

3. **Segmented Bar Chart:** A segmented bar chart compares two categorical variables. The categories of one of the categorical variables are on the horizontal axis, and percentage is on the vertical axis. Each bar is partitioned into pieces, where each piece represents the categories of the second categorical variable. The textbook describes a **comparative** or **side-by-side** bar chart which serves the same purpose as a segmented bar chart.




```
> freq = matrix(c(58,43,56,45,57,77,42,75),ncol=4,byrow=TRUE)
> freq
     [,1] [,2] [,3] [,4]
[1,]   58   43   56   45
[2,]   57   77   42   75

> rownames(freq) = c("Dead","Alive")
> colnames(freq) = c("A ","B ","C ","D ")
> freq
      A  B  C  D
Dead  58 43 56 45
Alive 57 77 42 75

> prop.table(freq,2)   # Takes proportions along each column
             A         B         C     D
Dead  0.5043478 0.3583333 0.5714286 0.375
Alive 0.4956522 0.6416667 0.4285714 0.625


> percent = prop.table(freq,2)*100
> ang=c(60,120)
> index=c(2,1)
> barplot(percent,beside=FALSE,angle=ang,density=20,col="black",
    ylab="Survival Status (%)",xlab="Treatment")
> legend(1.85,90,fill=TRUE,legend=rownames(freq)[index],angle=ang[index],
    density=20,merge=TRUE,bg="white")
```

Displaying Numerical Data:

To display a numerical variable measured from a sample:

- Partition the **range** of the data into equal sized **bins**.
- **Frequency** is the number of data points in each bin
- **Relative Frequency** is the proportion $\frac{\text{Frequency}}{\text{Total Number of Observations}}$
- **Percentage** = (Relative Frequency) $\times$ 100

1. **Stem-and-leaf Plot:** A graph of the numerical data where the bins are defined by "stems."

   - split each data value into a stem and a leaf
   - 5 to 15 stems is best
   - good display for small data sets
   - If there are too many leaves per stem, you can split the stems, where the upper stem takes the leaves 0 through 4 and the lower stem takes the leaves of 5 through 9.

```
> Crunchy = c(34,34,36,40,42,42,47,47,50,52,53,56,62,62,62,75,75,80)
> stem(Crunchy)
  The decimal point is 1 digit(s) to the right of the |

  3 | 446
  4 | 02277
  5 | 0236
  6 | 222
  7 | 55
  8 | 0
```

   - can compare two distributions.

   An article on peanut butter in *Consumer Reports* reported the following scores (quality ratings on a scale from 0 to 100) for various brands.

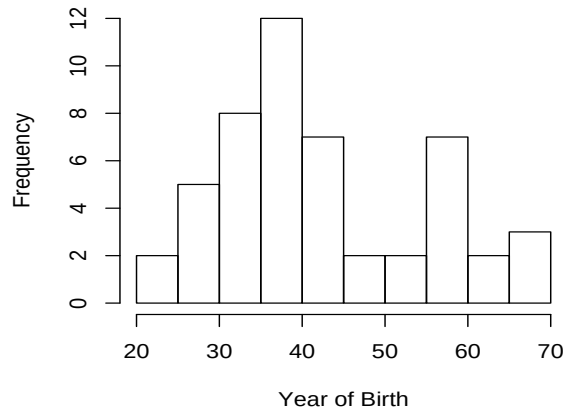
   QUESTION: Compare the distributions.

```
Creamy        Crunchy
     2 | 2 |
  9600 | 3 | 446
 54100 | 4 | 02277
666300 | 5 | 0236
   852 | 6 | 222
       | 7 | 55
       | 8 | 0
```

2. **Histogram:** A graph of frequencies, relative frequencies or percentages on the vertical axis versus the values of the numerical variable on the horizontal axis.

```
> hist(x=BirthYear,xlab="Year of Birth",main=" ")
```

Bins	Frequency
$[20, 25)$	2
$[25, 30)$	5
$[30, 35)$	8
$[35, 40)$	12
$[40, 45)$	7
$[45, 50)$	2
$[50, 55)$	2
$[55, 60)$	7
$[60, 65)$	2
$[65, 70)$	3



<u>What to Look For in Displays of Numerical Data</u>:

- Center
- Spread (narrow or wide?)
- Shape (modes and symmetry)
- Are there outliers (data values that do not follow the overall pattern)?

<u>Modes</u>:	**<u>Shapes</u>:**
Unimodal - one major peak	Symmetric
Bimodal - two majors peaks	Right-skewed (Positively-skewed)
Multimodal - more than two major peaks	Left-skewed (Negatively-skewed)

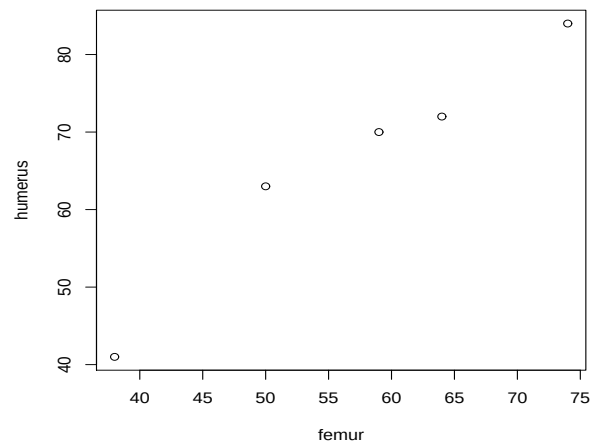
Do not expect perfection in the histogram of sample data! Due to sampling variability, there will be small peaks, valleys, and gaps. Do not focus on slight irregularities! Do not put too much weight on features caused by one or a few data values.

Graphs For Paired Numerical Data:

Paired or **bivariate** means that there are two variables to be studied. One variable, called the *explanatory variable* X, is used to describe the other variable, the *response* Y. So a sample of paired data consists of ordered pairs (x_1, y_1), (x_2, y_2), ..., (x_n, y_n).

1. **Scatterplot** A graphical display of the relationship between two numerical variables. The explanatory variable is along the horizontal axis. The response variable is along the vertical axis.

```
# Bone lengths (in) from n=5 dinosaurs
> femur=c(38,50,59,64,74)
> humerus=c(41,63,70,72,84)
> plot(femur,humerus)
```



Different colors or symbols can be used to distinguish between groups.

```
> levels(Status)
[1] "Kept"    "LaidOff"
> n=length(Status)
> Status.num = rep(1,n)
> Status.num[Status=="Kept"]=16
> plot(x=HireYear,y=BirthYear,
    pch=Status.num,xlab="Year of Hire",
    ylab="Year of Birth")
> Status.leg=levels(Status)
> Status.leg[2] = "Laid Off"
> Status.leg
[1] "Kept"    "Laid Off"
> legend(x=43,y=69,legend=Status.leg,pch=c(16,1))
```

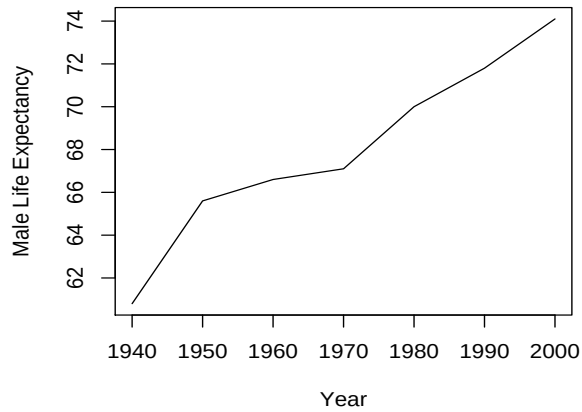
How to Describe the Relationship between two variables:

(a) **Form** - linear, non-linear (curved), clustered, etc.

(b) **Association** - positive or negative. A positive association indicates that increasing values of one variable are associated with increasing values of the other variable. A negative association indicates that increasing values of one variable are associated with decreasing values of the other variable.

(c) **Strength** - strong, moderate, or weak

2. **Time-series Plot** - A plot of time, the explanatory variable x, on the horizontal axis versus a response y on the vertical axis. The points are connected to show a trend.




```
> plot(time,male.life.expect,type="l",ylab="Male Life Expectancy",xlab="Year")
```

Numerical Summaries

Data: The observed values of a variable taken from a sample, denoted as x_1, x_2, x_3, ..., x_n. The **sample size** is n.

Statistic: A numerical value calculated from a sample of individuals. In other words, a statistic is a function of the data x_1, x_2, x_3, ..., x_n.

<u>Measures Of Center</u>:

1. **Sample Mean:** $\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum x}{n}$

   - The statistic $\bar{x}$ is the "balance point" (of center of gravity or fulcrum) of the of sample data.
   - The mean value of a discrete, numerical variable need not be a possible value. Example: The average number of children per household is 2.3.
   - **Sample Proportion:** a special type of sample mean. After encoding a categorical variable as 0's and 1's, then $\bar{x}$ is equal to

$$p = \text{sample proportion of successes} = \frac{0 + 1 + 1 + \ldots + 0}{n} = \frac{\#\text{ of } 1'\text{s}}{n}.$$

2. **Sample Median:** $\tilde{x}$ is the 50th percentile of the data (50% of the data below, 50% above)

   <u>How to Find the Median</u>:

   - Order the data values from smallest to largest.
   - $\tilde{x}$ = the middle value $[\left(\frac{n+1}{2}\right)^{\text{th}}$ ordered value] if n is odd and
     $\tilde{x}$ = the average of the 2 middle values $[\left(\frac{n}{2}\right)^{\text{th}}$ and $\left(\frac{n}{2}+1\right)^{\text{th}}$ ordered values] if n is even.

QUESTION: House prices in \$1000's: 143.5 132.0 154.5 169.3 134.7 2500

(a) Find the sample mean.

(b) Find the sample median.

(c) Why are the mean and median so different?

IMPORTANT! The mean is strongly affected by (not resistant to) outliers and skewness, whereas the median is not affected by (resistant to) outliers and skewness.

Outliers - the mean is pulled toward the outlier(s)

Skewness - the mean is pulled toward the longer tail

- Symmetric: Mean = Median
- Left-skewed (Negatively-skewed): Mean < Median
- Right-skewed (Positively-skewed): Mean > Median

NOTE: The mean is sensitive to outliers because it uses all the data values. The median is insensitive to outliers because it uses only 1 or 2 of the middle values in the ordered list.

Measures Of Variability (Spread):

1. **Sample Variance:** $s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}$
2. **Sample Standard Deviation:** s = $+\sqrt{s^2}$
   - A deviation is the distance from a data value to the sample mean ($\bar{x}$).
   - Standard deviation should be thought of as the "average (or typical) deviation".
   - Deviations sum to zero, $\sum(x_i - \bar{x}) = 0$.
   - The mean and standard deviation have the same units as the data values (e.g. inches, pounds). The variance has units^2 (e.g. inches^2, pounds^2).
3. **Interquartile Range (IQR):** IQR = Q_3 - Q_1

   where Q_1 is the first quartile (25% below, 75% above) and

   Q_3 is the third quartile (75% below, 25% above).

Note, Q_1 is the median of the lower half of the ordered list and Q_3 is the median of the upper half of the ordered list.

QUESTION:
Data: 1 1 2 4 5 7 7 7 8 9 10
Find IQR.

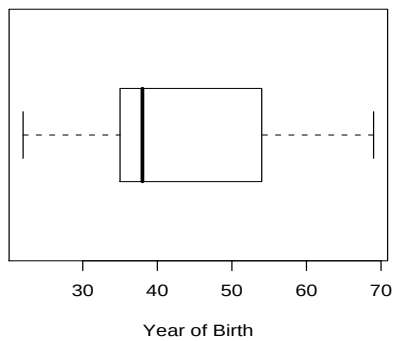
IMPORTANT! Standard deviation and variance are both strongly affected by (not resistant to) outliers and skewness, whereas IQR is not affected by (resistant to) outliers and skewness.

- Use the mean and standard deviation (or variance) as the measures of center and spread (respectively) when neither outliers nor skewness are present.
- Use the median and IQR as the measures of center and spread (respectively) when either outliers or skewness are present.

Five-number Summary: Minimum, Q_1, Median, Q_3, Maximum

- The five-number summary provides measures of center (median) and spread (IQR and range).

Boxplot: plot of the five-number summary

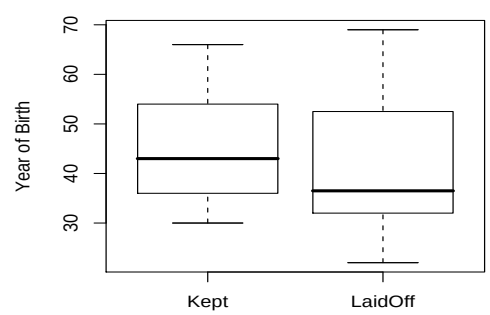


Outlier Guidelines for a Boxplot:

1. A "mild" outlier falls between 1.5(IQR) and 3(IQR) away from the nearest quartile. Use a solid circle to denote a mild outlier.
2. An "extreme" outlier falls more than 3(IQR) away from the nearest quartile. Use an open circle to denote an extreme outlier.
3. Plot the whiskers to the most extreme *non-outlier* data value.

Comparative (Side-by-Side) Boxplots:

- Great for comparing two or more distributions.
- Compare centers and spread.



STATISTICS vs. PARAMETERS:

Statistic: A numerical value calculated from a sample of individuals.

Parameter: A numerical value calculated from all individuals in a population.

Statistics	Parameters
$\bar{x}$	μ
$\tilde{x}$	$\tilde{\mu}$
s	σ
s^2	σ^2
p	π

R code

```
> # Measures of Center
> wt=c(1,1,2,4,5,7,7,7,8,9,10)
> mean(wt)
[1] 5.545455
> median(wt)
[1] 7
> mean(wt,trim=.25)
[1] 5.714286


> # Sample Proportion
> d=c("a","a","b","b","b")
> as.numeric(d=="a")    % encode the "a" category as a 1
[1] 1 1 0 0 0
> sum(as.numeric(d=="a"))/length(d)
[1] 0.4

> # Measures of Spread
> var(wt)
[1] 10.07273
> sd(wt)
[1] 3.173756
> iqr(wt)
Error: could not find function "iqr"
> IQR(wt)
[1] 4.5
> range(wt)
[1]  1 10
> summary(wt)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   3.000   7.000   5.545   7.500  10.000
> boxplot(wt)
> boxplot(wt1,wt2)
> boxplot(Response ~ CatVar)
```

Exercises

Graphs for paired data, p. 65:1.39 and 1.41
Numerical summaries, p. 66: 1.43 and 1.45
Graphs for numerical variables, p. 66: 1.47 (instead of a dot plot, consider a stem and leaf plot), 1.49 - 1.61 odd
Graphs for categorical variables, p. 71: 1.65, 1.67, 1.69abc