In this course:

**descriptive statistics** simply quantitatively and visually analyze data

**inferential statistics** try to reach conclusions (make judgements) by modeling relationships within data
Outline

1 terminology

2 exploring data in R

3 summary statistics
terminology

- **population**: all items of interest
- **sample**: portion of population that you have data for
- **parameter**: summary measure about population
- **statistic**: summary measure about the sample
describing the distribution of one variable

- Central tendency
- Dispersion
- Skew
Outline

1. terminology

2. exploring data in R

3. summary statistics
example: Great Lake water levels

Declining Water Levels In The Great Lakes May Signal Global Warming

ScienceDaily (Jan. 1, 2008) — Researchers in Michigan report new evidence that water levels in the Great Lakes, which are near record low levels, may be shrinking due to global warming. Their study, which examines water level data for Lakes Michigan and Huron over more than a century.

In the new study, Craig Stow and colleagues point out that water levels in the Great Lakes, which supply drinking water to more than 40 million U.S. and Canadian residents, have fluctuated over thousands of years. But recent declines in water levels have raised concern because the declines are consistent with many climate change projections, they say.

To evaluate the factors behind this decline, the scientists examined water level data for Lakes Michigan and Huron from 1860 to 2006, including precipitation, evaporation and runoff data. The results reveal an underlying gradual decline in water levels since 1973. This underlying drop may be due to an increase in evaporation levels, they say.

Lake Huron shoreline near Georgian Bay, Canada. Researchers have found a gradual decline in water level in Lake Huron and Lake Michigan, two of the Great Lakes of North America, since 1973. (Credit: iStockphoto)
data sets in R

> data() shows all available data sets
> ?LakeHuron is a built-in data set with data on Lake Huron water levels from 1875 to 1972. The ? in front prompts R to tell you what it knows about a dataset or function
> plot(LakeHuron) will plot a time series
Looks like water levels were up again right around 1972. So what happened from 1973 to 2007? Is global warming having an effect? Or was 1973 just a convenient place to "start" to observe a drop?
another data set

lakedata = read.table("somedir/lakehuron1918_2006.txt", head=T)

head = T instructs R to treat the first line of text as the column names
Make sure you include the correct path to the file.
You can also have R prompt you to browse to the file location:

lakedata=read.table(file=file.choose(),...)
Click on individual gauges to access water level data.

- Thessalon, Ontario
- Little Current, Ontario
- Parry Sound, Ontario
- Georgian Bay
- Collingwood, Ontario
- Tobermory, Ontario
- Goderich, Ontario
- Lake Huron
- Harbor Beach, MI
- Essexville, MI
- Lakeport, MI

**NOAA Gauges**

**DFO Gauges**

- Mackinaw City, MI
- De Tour Village, MI
Accessing columns by index or column name

```r
> colnames(lakedata)
[1] "Year"     "Level_meters"

> lakedata$Year
[27] ...

> lakedata[,1]
[27] ....
```
Adding and manipulating columns

> lakedata$LevelInFeet = lakedata$Level_meters * 3.2808

> lakedata$LevelInFeet
[22] ...
## Accessing subsets of the data

```r
> lakedata[10:15,]
   Year Level_meters LevelInFeet
  10  1927       175.92    577.1935
  11  1928       176.10    577.7841
  12  1929       176.65    579.5887
  13  1930       176.64    579.5558
  14  1931       176.21    578.1450
  15  1932       175.94    577.2591
```

```r
> lakedata[lakedata$Year > 2002,]
   Year Level_meters LevelInFeet
  86  2003       175.82    576.8654
  87  2004       175.87    577.0295
  88  2005       176.08    577.7185
  89  2006       175.88    577.0623
```
Outline

1. terminology
2. exploring data in R
3. summary statistics
The mean, or average, is the sum of the values \((X_1, X_2, X_3, \ldots X_N)\) divided by their count \((N)\):

\[
\bar{X} = \frac{X_1 + X_2 + X_3 + \ldots + X_N}{N}
\]

```r
> mean(lakedata$LevelInFeet)
[1] 578.5177
```
The sample median is the “middle value”. Half of the data points have equal or lower value, and the other half have equal or higher value.

If there is an even number of sampled data points, then the median is the average of the middle two.

```r
> median(lakedata$LevelInFeet)
[1] 578.5715
```
The variance, measures the amount of dispersion in the sample.

\[
var(X) = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \ldots + (X_N - \bar{X})^2)}{N - 1}
\]

Why \( N - 1 \) instead of \( N \)? If we have only a sample of the whole population, we don’t know the true mean of the population, only the sample mean. By dividing by \( N - 1 \) we have a more generous estimate of the variance.

\[
> \text{var(lakedata$LevelInFeet)}
\]

[1] 1.557329
The standard deviation is nothing more than the square root of the variance

\[ \sigma(X) = \sqrt{\frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \ldots + (X_N - \bar{X})^2}{N - 1}} \]

Now we can say that the water levels

\[ > \text{sd(lakedata$LevelInFeet)} \]
\[ [1] \ 1.247930 \]

So we can say that the lake water level (wherever the measuring gage may actually be located) is 578.52 ± 1.25 feet.
We can also observe that the median is very close to the mean, there is little skew.

```
hist(LevelInFeet,20)
plot(Year,LevelInFeet)
```
If you are using a data set, it can be tiresome to always have to type yourdataset$columnname. Using the `attach()` function you can start address it simply by the column name

```r
attach(lakedata)
hist(LevelInFeet, 20)
plot(Year, LevelInFeet)
detach(lakedata)
```
where there is skew: income

```r
> salaries = read.table("umichsalariesall.txt", head=T)
> hist(salaries$salary, 50)
> median(salaries$salary) # 48000
> mean(salaries$salary) # 57788.35
```
the *tails* of the distribution are the very large and very small values.

- a *long tailed distribution* has values that are far from the mean
- a *left skewed* distribution has a longer left tail
- a *right skewed* distribution has a longer right tail

Are umich salaries left or right skewed?
In class you took a handedness survey. What do you think the histogram looks like?

```r
> hand = read.table("handedness2008.txt",head=T)
> hand

<table>
<thead>
<tr>
<th>left</th>
<th>right</th>
<th>specialization</th>
<th>second_specialization</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>19</td>
<td>11</td>
<td>SC</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>19</td>
<td>HCI</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>21</td>
<td>HCI</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>15</td>
<td>ARM</td>
</tr>
</tbody>
</table>

...
Now we compute the handedness ratio: \((r - l)/(r + l)\), and plot a histogram.

\[
\text{hand.handedness} = (\text{hand}\$right - \text{hand}\$left)/(\text{hand}\$right + \text{hand}\$left)
\]

> hist(hand.handedness)
> hist(hand.handedness,breaks=20)

The second command tells the histogram function that we would like 20 bins. The rest awaits in the problem set...
Things you should feel comfortable with in R

- Entering in data
- Bringing up help pages
- Plotting and binning
- Loading data
- Selecting subsets of rows and columns in data
- Attaching and detaching data sets
we want to describe data
we want to quantify the central tendency, dispersion, and skew
we want to visualize the data
we can use R to get summary stats (mean, median, variance...)
we can use R to bin (histogram) and plot the data

Next time: probability