Chapter 7

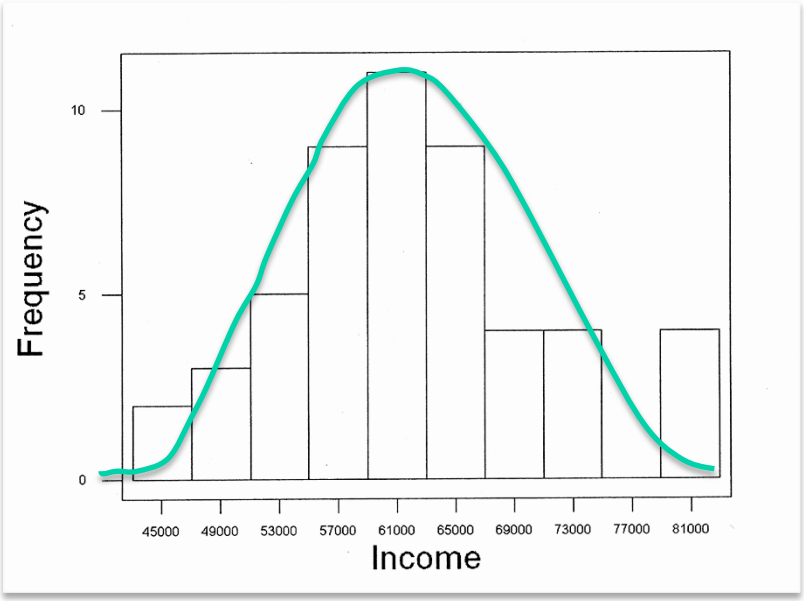
Summarizing and Displaying Measurement Data

You will need your calculators for this

Interpreting Data

Four kinds of useful information about a set of data:

- **Center (median)**
- **Unusual values (outliers)**
- **Variability**
- **Shape**



Mean, Variance, and Standard Deviation

- **Mean:** represents center
- **Standard Deviation:** represents spread or variability in the values;
- **Variance = (standard deviation)^2**

Mean and standard deviation are most useful for *symmetric* sets of data with *no outliers*.

The Standard Deviation and Variance

Consider two sets of numbers, both with mean of 100.

Numbers	Mean	Standard Deviation
100, 100, 100, 100, 100	100	0
90, 90, 100, 110, 110	100	10

- **First** set of numbers has **no spread** or variability at all.
- **Second** set has some spread to it; **on average, the numbers are about 10 points away from the mean**.

The standard deviation is roughly the average distance of the observed values from their mean.

Computing the Standard Deviation

1. Find the mean.
2. Find the deviation of each value from the mean.
   Deviation = value – mean.
3. Square the deviations.
4. Sum the squared deviations.
5. Divide the sum by (the number of values) – 1,
   resulting in the variance.
6. Take the square root of the variance.
   The result is the standard deviation.

Computing the Standard Deviation

Try it for the set of values: 90, 90, 100, 110, 110.

1. The mean is 100.
2. The deviations are -10, -10, 0, 10, 10.
3. The squared deviations are 100, 100, 0, 100, 100.
4. The sum of the squared deviations is 400.
5. The variance = 400/(5 – 1) = 400/4 = 100.
6. The standard deviation is the square root of 100, or 10.

Computing the Standard Deviation

The accepted formula for Standard Deviation is this:

$$\text{Variance} = \frac{\sum_{k=1}^{n}\left(x_k - \bar{x}\right)^2}{n-1}$$

x_k= each value in the data table

$\bar{x}$ = each value in the data table

$\sum_{k=1}^{n}$ → add each of the n terms

$$\text{Standard Deviation} = \sqrt{\text{Variance}}$$

The Mean, Median, and Mode

Ordered Listing of 28 Exam Scores

32, 55, 60, 61, 62, 64, 64, 68, 73, 75, 75, 76, 78, 78, 79, 79, 80, 80, 82, 83, 84, 85, 88, 90, 92, 93, 95, 98

- **Mean (numerical average): 76.04**
- **Median: 78.5 (halfway between 78 and 79)**
- **Mode (most common value): no single mode exists, many occur twice.**

Ordered Listing of 28 Exam Scores

32, 55, 60, 61, 62, 64, 64, 68, 73, 75, 75, 76, 78, 78,
79, 79, 80, 80, 82, 83, 84, 85, 88, 90, 92, 93, 95, 98

Outliers:

Outliers = values far removed from rest of data.
Median of 78.5 higher than mean of 76.04 because one very low score (32) pulled down mean.

Variability:

How spread out are the values? A score of 80 compared to mean of 76 has different meaning if scores ranged from 72 to 80 versus 32 to 98.

Ordered Listing of 28 Exam Scores

32, 55, 60, 61, 62, 64, 64, 68, 73, 75, 75, 76, 78, 78,
79, 79, 80, 80, 82, 83, 84, 85, 88, 90, 92, 93, 95, 98

Minimum, Maximum and Range:

Range = max – min = 98 – 32 = 66 points.
Other variability measures include interquartile range and standard deviation.

Shape:

Are most values clumped in middle with values tailing off at each end? Are there two distinct groupings? Pictures of data will provide this info.

Stemplots and Histograms

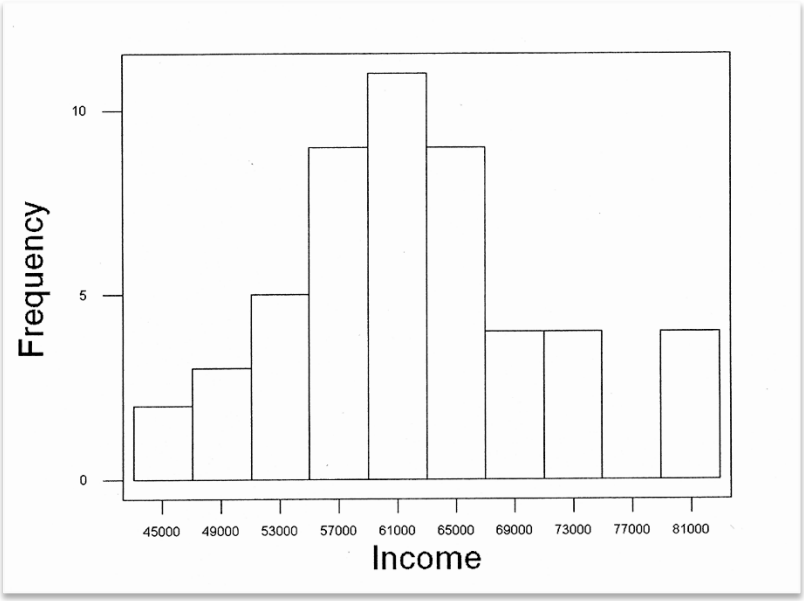
Stemplot: Break numbers down into places. Note how it gives a rough picture of shape.

Histogram: gives a good picture of where the data is centered if at all. Also provides picture of shape.

Stemplot for Exam Scores

3|2
4|
5|5
6|024418
7|56598398
8|5430820
9|53208

Example: 3|2 = 32



Splitting Stems:

Reusing digits two or five times.

Stemplot A:

5|4
5|789
6|023344
6|55567789
7|00124
7|58

Stemplot B:

5|4
5|7
5|89
6|0
6|233
6|44555
6|677
6|89
7|001
7|2
7|45
7|
7|8

Two times:

1st stem = leaves 0 to 4

2nd stem = leaves 5 to 9

Five times:

1st stem = leaves 0 and 1

2nd stem =leaves 2 and 3, etc.

Using a Stemplot to determine shape, identify outliers, and locate the center.

Pulse Rates:

5|4
5|789
6|023344
6|55567789
7|00124
7|58

Bell-shape
Centered mid 60's
no outliers

Exam Scores

3|2
4|
5|5
6|024418
7|56598398
8|5430820
9|53208

Outlier of 32.
Apart from 55,
rest uniform from
the 60's to 90's.

Median Incomes:

4|66789
5|11344
5|56666688899999
6|011112334
6|556666789
7|01223
7|
8|0022

Wide range with 4
unusually high values.
Rest bell-shape around
high $50,000s.

The Shape of a graph is…

- **Symmetric** if you could draw line through center and the picture on one side would be mirror image of picture on other side.
  Example: bell-shaped data set.
- **Unimodal** there is a single prominent peak
- **Bimodal** if there are two prominent peaks
- **Skewed to the Right** if the *higher* values are more spread out than lower values
- **Skewed to the Left:** if the *lower* values are more spread out and higher ones tend to be clumped

The five-number summary display

	Median	
Lower Quartile		**Upper Quartile**
Lowest		**Highest**

- **Lowest** = Minimum
- **Highest** = Maximum
- **Median** = number such that half of the values are at or above it and half are at or below it (middle value or average of two middle numbers in ordered list).
- **Quartiles** = medians of the two halves.

Five-Number Summary for Income

***n* = 51 observations**

- **Lowest:** \$46,xxx => \$46,596
- **Highest:** \$82,xxx => \$82,879
- **Median:** (51+1)/2 => 26th value \$61,xxx => \$61,036
- **Quartiles:** Lower quartile = median of lower 25 values => 13th value, \$56,xxx => \$56,067; Upper quartile = median of upper 25 values => 13th value, \$66,xxx => \$66,507

Median Incomes:

```
4|66789
5|11344
5|56666688899999
6|011112334
6|556666789
7|01223
7|
8|0022
```

Five-number summary for family income

	\$61,036	
\$56,067		**\$66,507**
\$46,596		**\$82,879**

Provides center and spread. Can compare gaps between extremes and quartiles, gaps between quartiles and median.

Creating a Boxplot for Sleep Hours

190 Students at a large university were asked to answer a series of questions one of which was how many hours they had slept the night before. A five number summary for the reported number of hours of sleep is 3, 6, 7, 8, 16 with 16 being reported by only two individuals. Aside from those two the maximum hours reported was 12.

A five number summary consists of:

- Minimum
- First Quartile (or the 25^{th} percentile)
- Median (half-way mark of the ascending list)
- Third Quartile (or the 75^{th} percentile)
- Maximum

The IQR or InterQuartile Range is simply Q3 – Q1

We need this value to determine is 16 hours qualifies as an *outlier*

Creating a Boxplot for Sleep Hours

190 Students at a large university were asked to answer a series of questions one of which was how many hours they had slept the night before. A five number summary for the reported number of hours of sleep is 3, 6, 7, 8, 16 with 16 being reported by only two individuals. Aside from those two the maximum hours reported was 12.

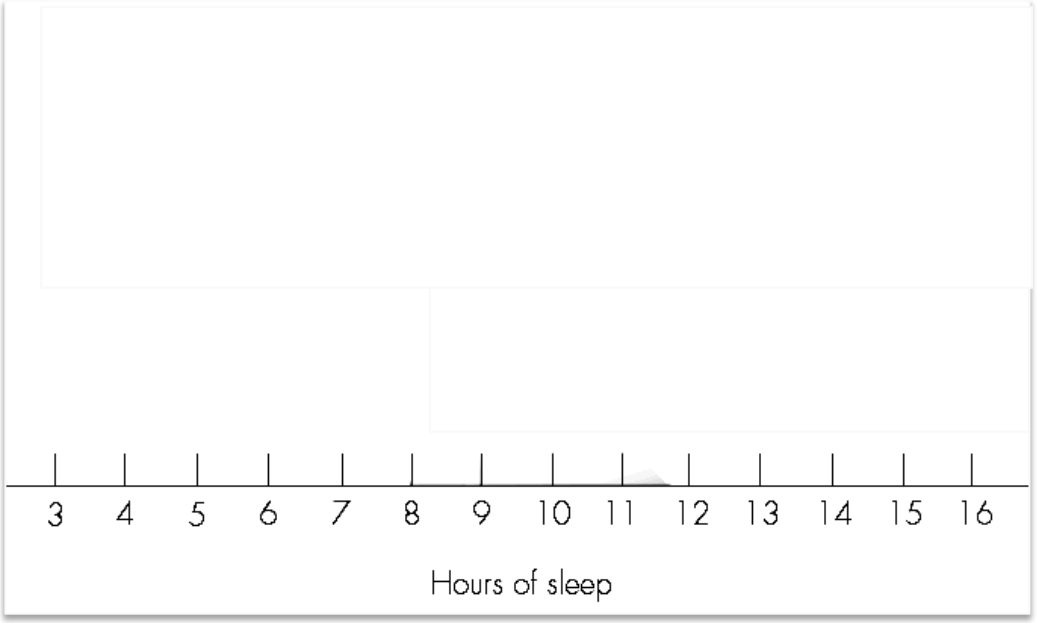
A five number summary consists of:

- Minimum
- First Quartile (or the 25^{th} percentile)
- Median (half-way mark of the ascending list)
- Third Quartile (or the 75^{th} percentile)
- Maximum

An *outlier* is a value that is more than 1.5 × IQR away from the closest quartile

Creating a Boxplot for Sleep Hours

1. Draw horizontal line and label it from 3 to 16.
2. Draw rectangle (box) with ends at 6 and 8.
3. Draw line in box at median of 7.
4. Compute IQR = 8 – 6 = 2.
5. Compute 1.5(IQR) = 1.5(2) = 3; outlier is any value below 6 – 3 = 3, or above 8 + 3 = 11.
6. Draw line from each end of box extending down to 3 but up to 11.
7. Draw asterisks at outliers of 12 and 16 hours.



Generate a Five-Number Summary from your student height data

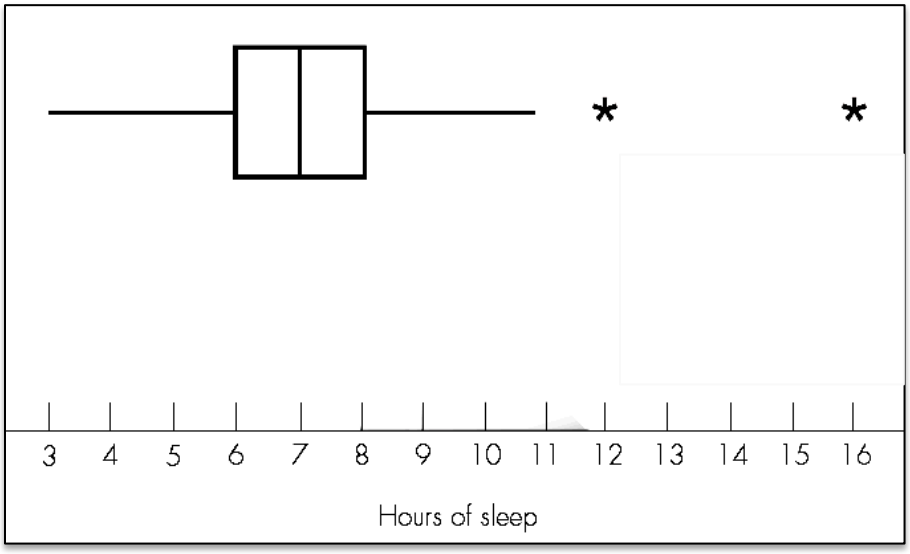
1. Draw horizontal line and label it.
2. Draw rectangle (box) with ends at the two Quartiles?
3. Draw line in box at median.
4. Compute IQR.
5. Compute 1.5(IQR); outlier is any value below…

6. Draw line from each end of box extending down .
7. Draw asterisks at outliers if any.

Interpreting Boxplots

- Divide the data into fourths.
- Easily identify outliers.
- Useful for comparing two or more groups.

Outlier: any value more than 1.5(IQR) beyond closest quartile.



¼ of students slept between 3 and 6 hours, ¼ slept between 6 and 7 hours, ¼ slept between 7 and 8 hours, and final ¼ slept between 8 and 16 hours

For Those Who Like Formulas

The Data

n = number of observations

x_i = the *ith* observation, $i = 1,2, \ldots, n$

The Mean

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The Variance

$$s^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

The Computational Formula for the Variance

$$s^2 = \frac{1}{(n-1)}\left(\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}\right)$$

The Standard Deviation

Use either formula to find s^2; then simply take the square root to get the standard deviation s.