Describing Data

Section 1.2 …part two

Measuring Spread

- The five-number summary isn't the most common description of a distribution of data.
- The mean (vs. median) is often used to measure center and the standard deviation (vs. IQR) is often used to measure spread.

Deviation

- A deviation is simply how far a value is from the mean…
- In other words, $x_1 - \bar{x}$
- Deviations can be positive or negative

Variance

The variance, s^2, of a set of observations is the average of the squares of the deviations of the observations from their mean. In symbols, the variance of n observations $x_1, x_2, \ldots, x_n$ is

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}$$

Or, $$s^2 = \frac{1}{n - 1} \sum (x_i - \bar{x})^2$$

1 2 3 4 5

$$s^2 = \frac{(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2}{5-1} = \frac{(-2)^2 + (-1)^2 + (0)^2 + (1)^2 + (2)^2}{4} = \frac{10}{4}$$

Standard Deviation

The standard deviation, s, is the square root of the variance, s^2:

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

1 2 3 4 5

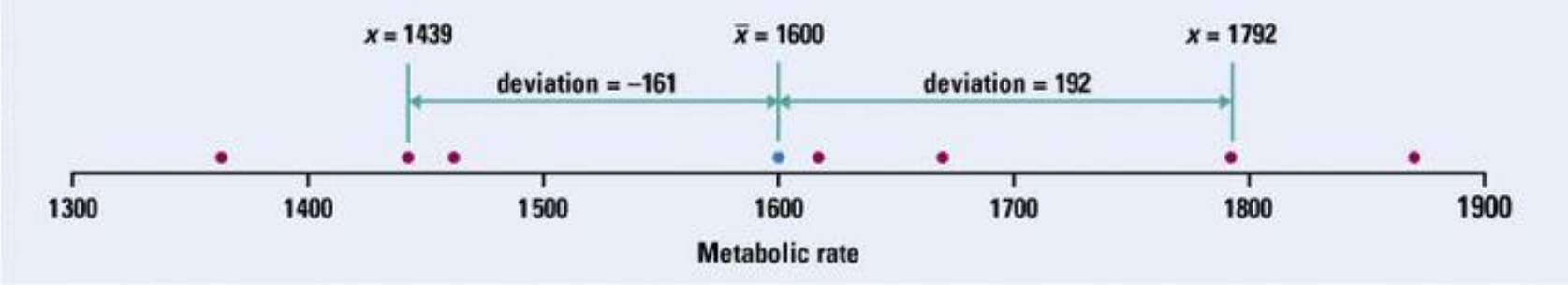
$$s = \sqrt{s^2} = \sqrt{\frac{10}{4}} = 1.58$$

1. Basically describes how much the values tend to vary from the average.
2. Has the same units as the values themselves.

Example 1.16 – Metabolic Rate

1792, 1666, 1362, 1614, 1460, 1867, 1439

$\overline{X}$ = 1600 calories and s = 189.24 calories



Properties of the Standard Deviation

1. s measures spread about the mean and should be used only when the mean is chosen as the measure of center.

2. $s = 0$ only when there is no spread/variability. This happens only when all observations have the same value. Otherwise, $s > 0$. As the observations become more spread out about their mean, s gets larger.

2 2 2 2 2

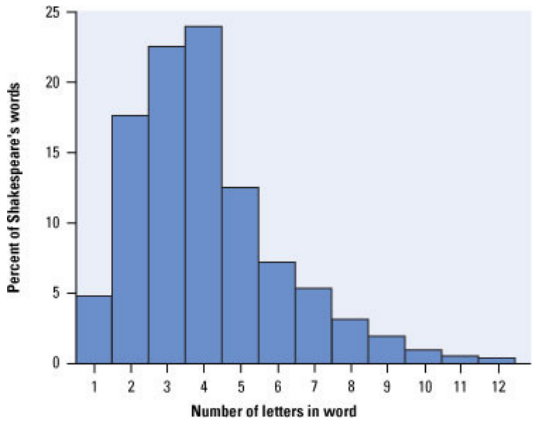
3. s, like the mean, $\bar{x}$, is not resistant. A few outliers can change s a lot.

For example, dropping the Honda Insight from our list of two-seater cars reduces the mean highway mileage from 24.7 to 22.6 mpg. It cuts the standard deviation by more than half, from 10.8 mpg with the Insight to 5.3 mpg without it.

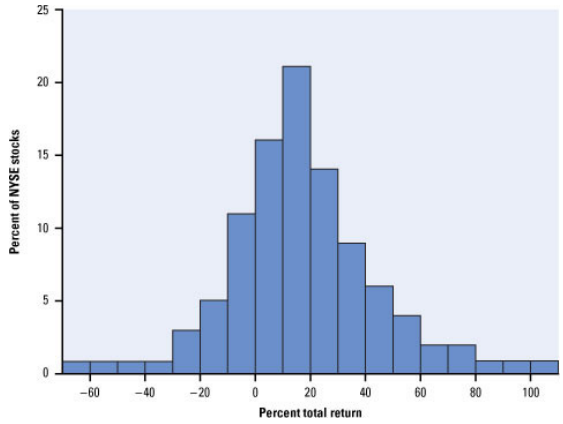
http://en.wikipedia.org/wiki/Honda_Insight

Choosing Measures of Center and Spread

- The five-number summary is usually better than the mean and standard deviation for describing a skewed distribution or a distribution with strong outliers. Use $\bar{x}$ and s only for reasonably symmetric distributions that are free of outliers.



Five-Number Summary



Mean and Standard Deviation

Example 1.17 - Investments

Investment	Mean return	Standard deviation
Common stocks	13.2%	17.6%
Treasury bills	5.0%	2.9%

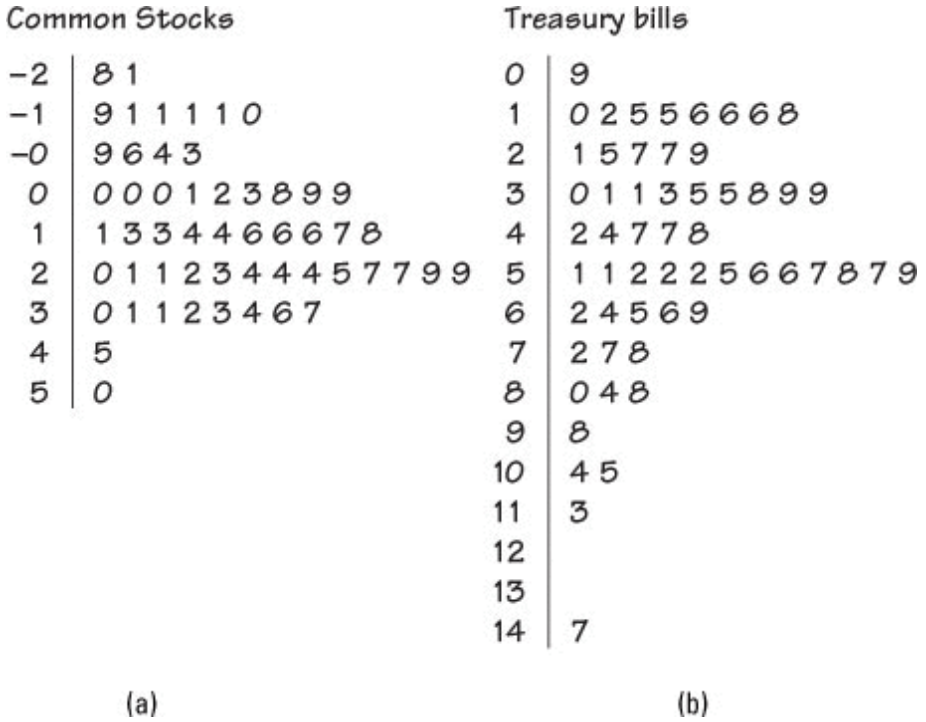


Figure 1.22 Stemplots of annual returns for stocks and Treasury bills, 1950 to 2003. (a) Stock returns, in whole percents. (b) Treasury bill returns, in percents and tenths of a percent.

Graphs vs. Numerical Summaries

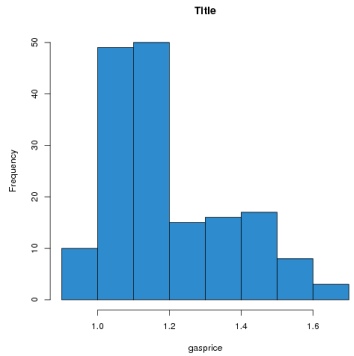
Remember that a graph gives the best overall picture of a distribution. Numerical measures of center and spread report specific facts about a distribution, but they do not describe its entire shape. Numerical summaries do not disclose the presence of multiple modes or gaps, for example. **So, it's always best to plot your data.**

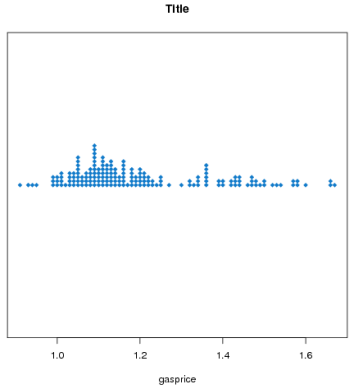
Gas Prices…Quick Example

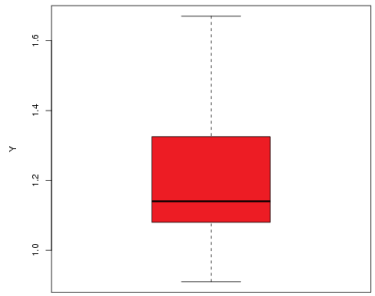
N = 168	
Sum	201.7100
Mean	1.2007
Variance	0.0296
Standard Deviation	0.1720
Standard Error	0.0133
Median	1.1400
Range	0.7600
Minimum	0.9100
Maximum	1.6700
First Quartile	1.0800
Third Quartile	1.3300
Mode	1.0900

leaf unit: 0.01
n: 168

```
 9 | 133
 9 | 4888
10 | 0001111222224444
10 | 55555555666777788888999999999999
11 | 00000111111112222222222222233333
11 | 4445555555777777888
12 | 111122222399999
12 | 5557
13 | 0223444
13 | 66666699
14 | 2233344499
14 | 6777889
15 | 00234
15 | 7788
16 | 0
16 | 557
```







Linear Transformations

- Sometimes it becomes necessary to adjust all of our data by some multiple or some fixed amount.
- A **linear transformation** changes the original variable x into the new variable x_{new} given by an equation of the form

$$x_{new} = a + bx$$

- Adding the constant a shifts all values of x upward or downward by the same amount.
- Multiplying by the positive constant b changes all values of x by a factor of b.

$$x_{new} = a + bx$$

Example 1.18 – Turning Up the Heat

Table 1.7 *Year 2005 salaries for Miami Heat players (in millions of dollars)*

Player	Salary	Player	Salary
Shaquille O'Neal	27.70	Christian Laettner	1.10
Eddie Jones	13.46	Steve Smith	1.10
Dwyane Wade	2.83	Shandon Anderson	0.87
Damon Jones	2.50	Keyon Dooling	0.75
Michael Doleac	2.40	Zhizhi Wang	0.75
Rasual Butler	1.20	Udonis Haslem	0.62
Dorell Wright	1.15	Alonzo Mourning	0.33
Qyntel Woods	1.13		

Source: www.hoopshype.com/salaries/miami.htm.

1 | 2: represents 1.2
leaf unit: 0.1 million
n: 15

0 | 3
0 | 6778
1 | 11112
1 |
2 | 4
2 | 58

HI: 13.46 27.7

a

$100,000 bonus…

What happens to the shape?
What happens to the center?
What happens to the spread?

Shape does not change.
Center increases by a.
Spread does not change.

b

10% Salary Increase

What happens to the shape?
What happens to the center?
What happens to the spread?

Shape does not change.
Center increases by a factor of b.
Spread changes by a factor of b.

Comparing Distributions

- **Data** - Organize and examine the data. Answer the ***key questions***:
  - **Who** are the individuals described by the data?
  - **What** are the variables? In what units is each variable recorded?
  - **Why** were the data gathered?
  - **When, where, how, and by whom** were the data produced?
- **Graphs** - Construct appropriate graphical displays.
- **Numerical summaries** - Calculate relevant summary statistics (5# Summary or Mean and *s*).
- **Interpretation** - Discuss what the data, graphs, and numerical summaries tell you in the context of the problem. Answer the question!

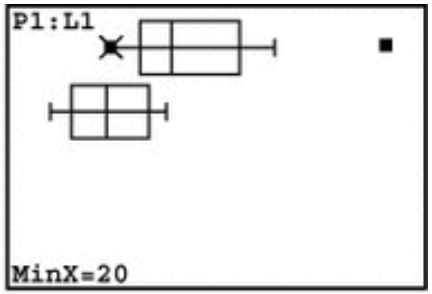
Example 1.20 – Swiss Doctors and C-Sections

Male Doctors… 20 25 25 27 28 31 33 34 36 37 44 50 59 85 86

Female Doctors… 5 7 10 14 18 19 25 29 31 33

Male		Female
	0	5 7
	1	0 4 8 9
8 7 5 5 0	2	5 9
7 6 4 3 1	3	1 3
4	4	
9 0	5	
	6	
	7	
6 5	8	

(a)



(b)

Who?
What?
Why?
When?
Where?
How?
By Whom?

Graphs

Numerical Summaries

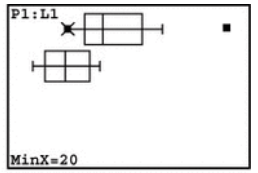
Interpretation
(on next slide)

	$\bar{x}$	s	Min.	Q_1	M	Q_3	Max.	IQR
Male doctors	41.333	20.607	20	27	34	50	86	23
Female doctors	19.1	10.126	5	10	18.5	29	33	19

Interpretation – SOCS Summary

Male		Female
	0	5 7
	1	0 4 8 9
8 7 5 5 0	2	5 9
7 6 4 3 1	3	1 3
4	4	
9 0	5	
	6	
	7	
6 5	8	

(a)



(b)

	$\bar{x}$	s	Min.	Q_1	M	Q_3	Max.	IQR
Male doctors	41.333	20.607	20	27	34	50	86	23
Female doctors	19.1	10.126	5	10	18.5	29	33	19

- **Shape** – Right-skewed for male doctors, roughly symmetric for female doctors. Both unimodal.
- **Outliers** – 85 and 86 C-sections per year are outliers.
- **Center** – More than half of the female doctors in the study performed fewer than 20 cesarean sections in a year; 20 was the minimum number of cesarean sections performed by male doctors. The mean and median numbers of cesarean sections performed are higher for the male doctors.
- **Spread** – Both *s* and the *IQR* for the male doctors are much larger than *s* and *IQR* for female doctors, so there is much greater variability in the number of cesarean sections performed by male doctors.

Finally…Answer the Research Question

Due to the outliers in the male doctor data and the lack of symmetry of their distribution of cesareans, we should use the resistant medians and IQRs in our numerical comparisons.

In Switzerland, it does seem that male doctors generally perform more cesarean sections each year (median = 34) than do female doctors (median = 18.5).

In addition, male Swiss doctors are more variable in the number of cesarean sections performed each year (IQR = 23) than female Swiss doctors (IQR = 19).

We may want to do more research on why this apparent discrepancy exists.