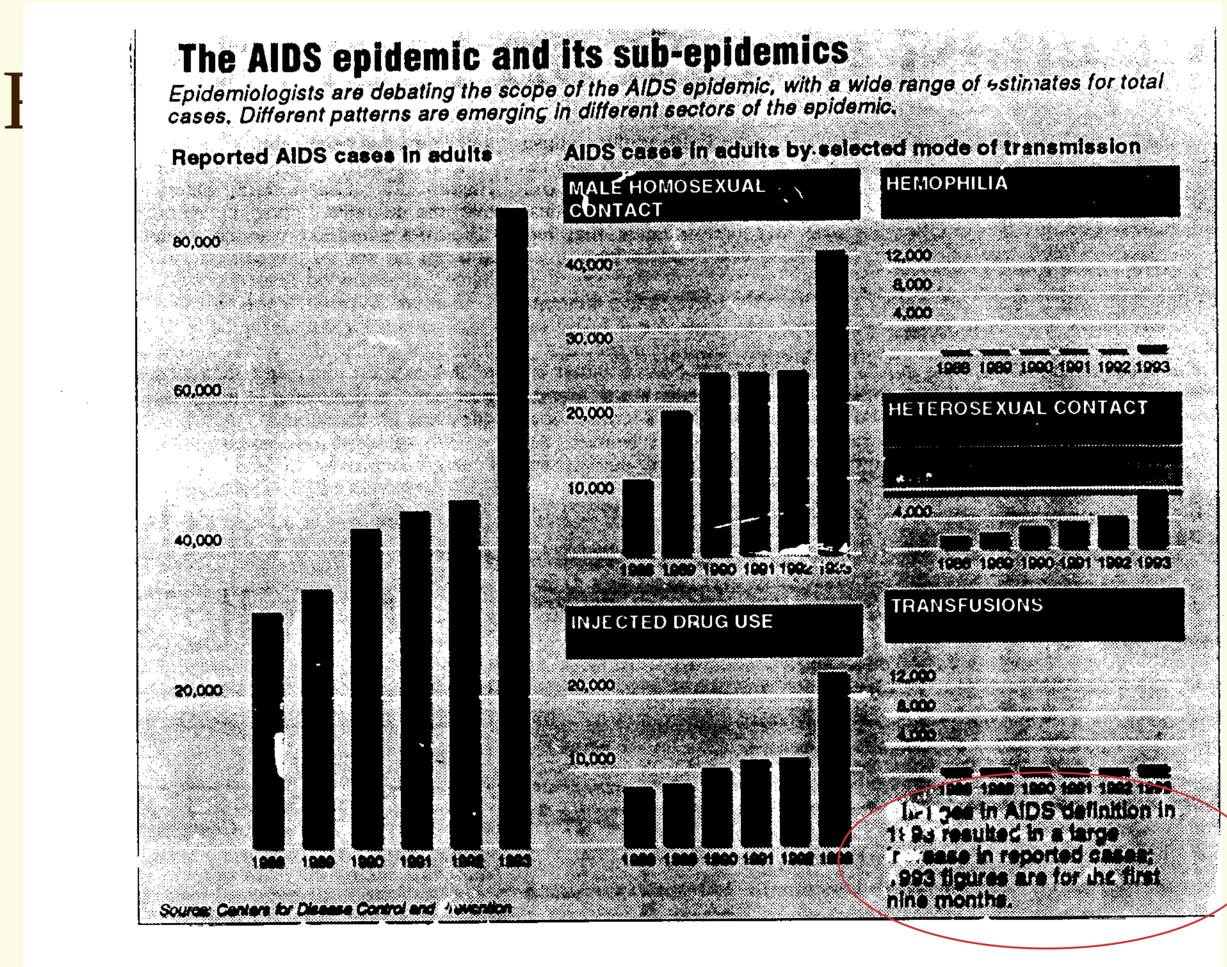
Misleading Graphs

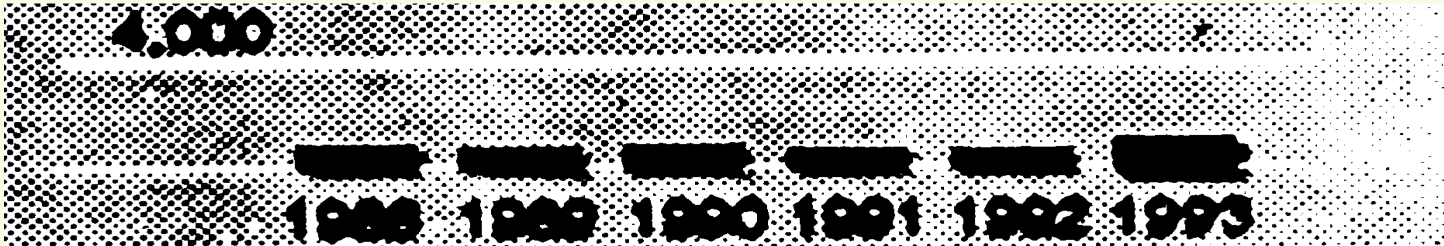
- Examples
- Compare unlike quantities
- Truncate the *y*-axis
- Improper scaling
- “Chart Junk”
- Impossible to interpret

Pretty Bleak Picture

- F



But Wait…..!



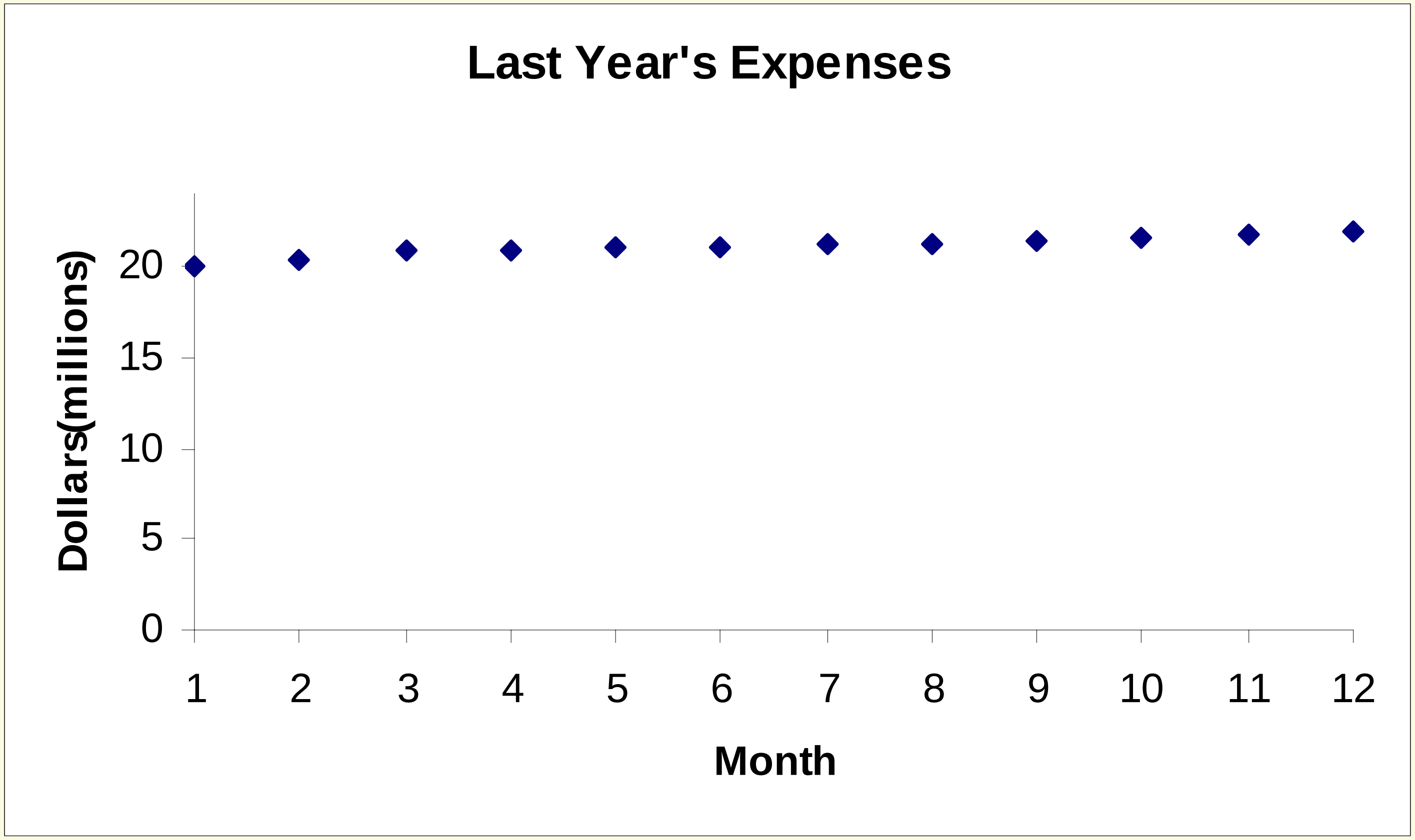
Changes in AIDS definition in 1993 resulted in a large increase in reported cases; 1993 figures are for the first nine months.

Turk Incorporated

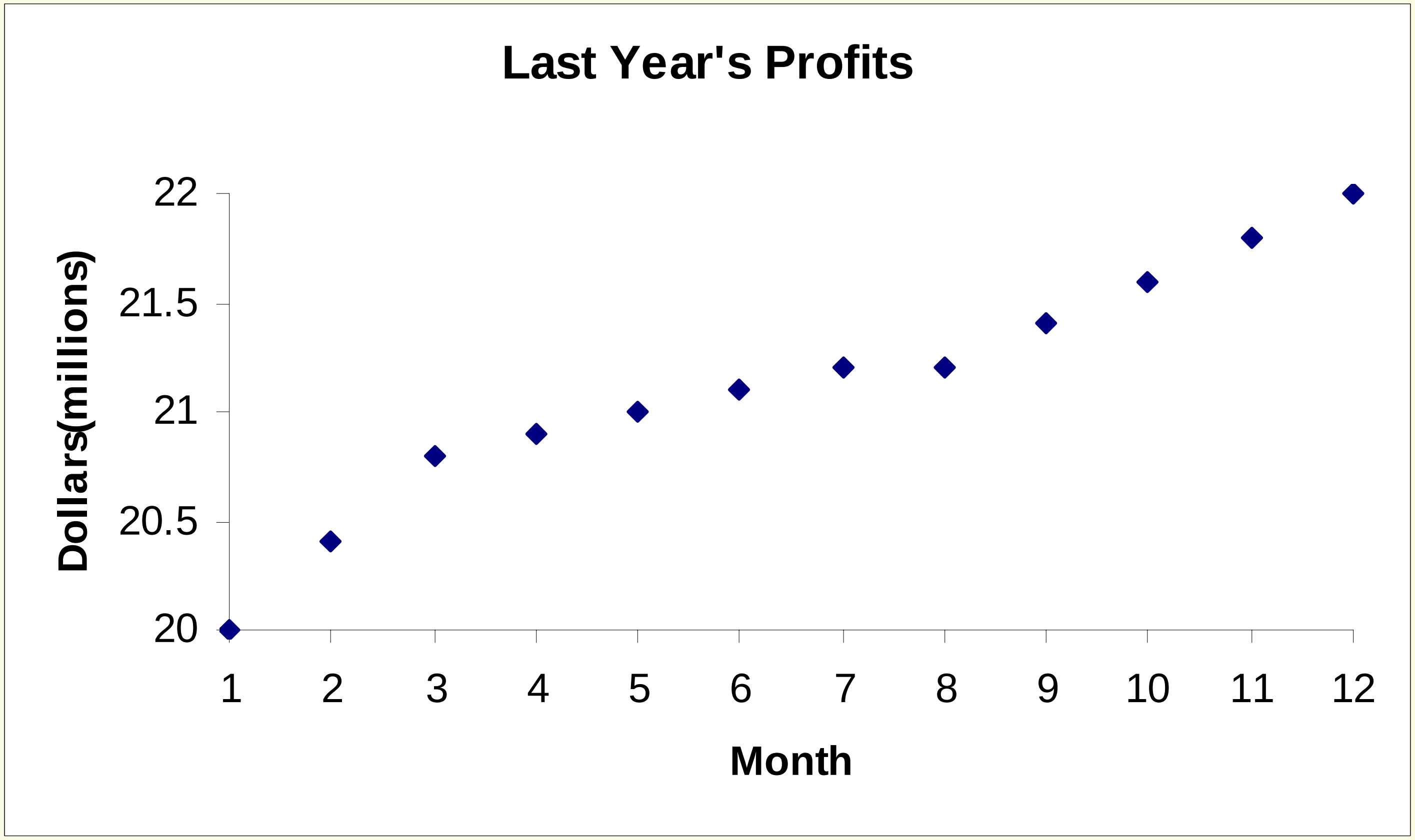
- Company report

	Jan	Feb	Mar	Apr	May	Jun
$ mill	20	20.4	20.8	20.9	21	21.1
	July	Aug	Sept	Oct	Nov	Dec
$ mill	21.2	21.2	21.4	21.6	21.8	22

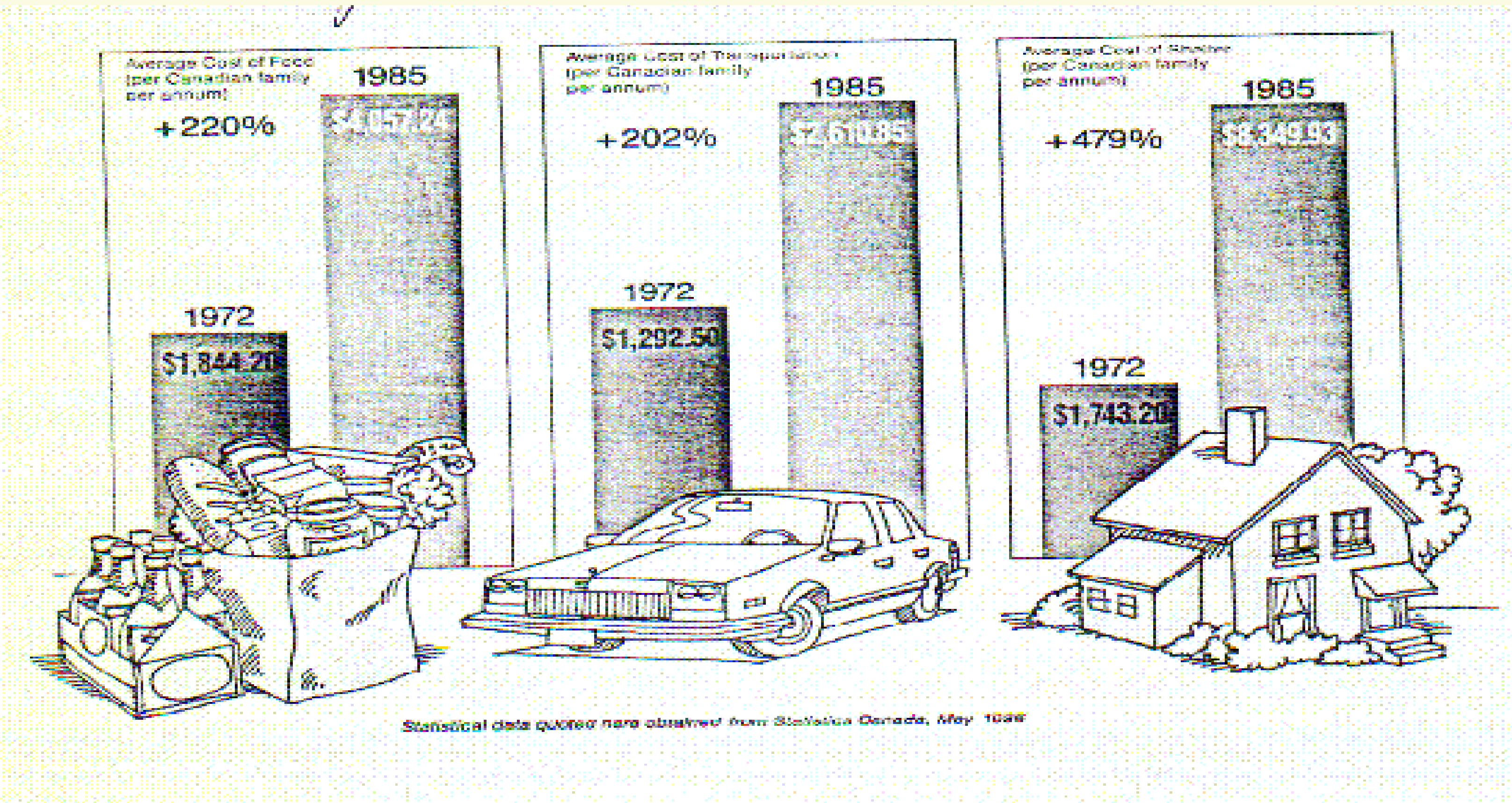
Last Year’s Expenses



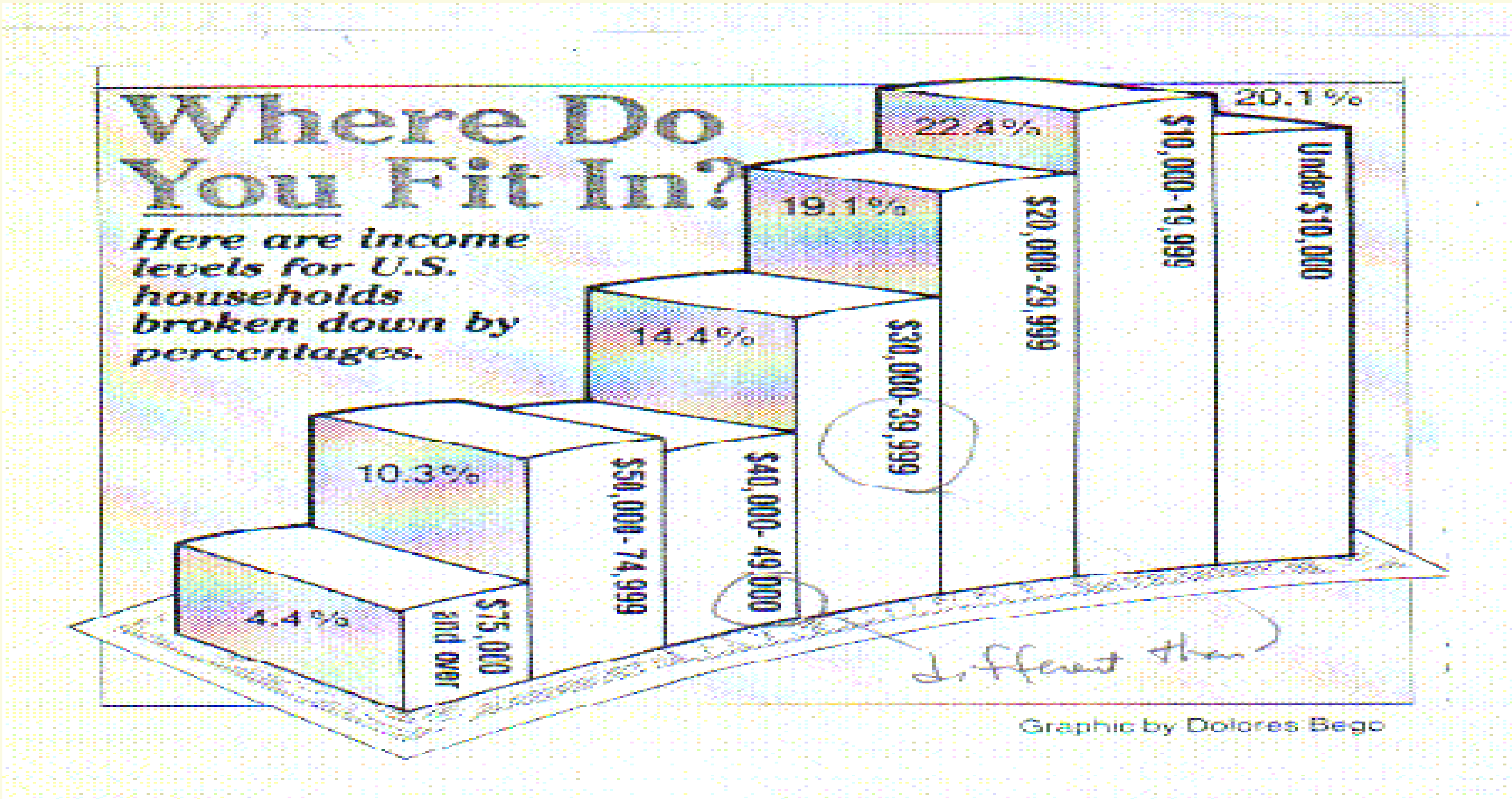
Last Year’s Profits



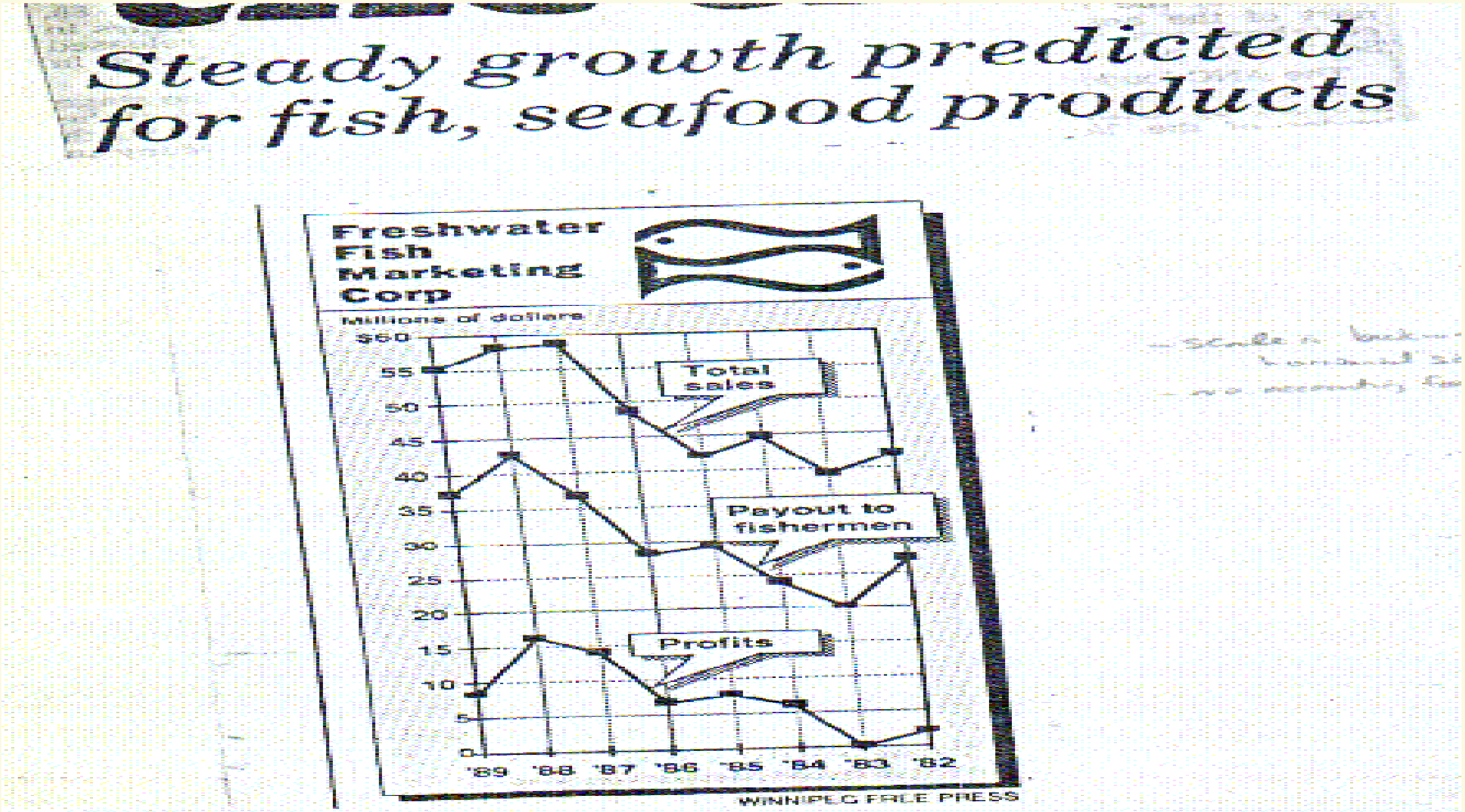
The Cost of Living



Income Levels



Am I missing something here?



Wanted Dead or Alive

- Bad Graphs
- All media are fair game
- Reward? Coffee, extra credit, enhanced self worth,…

Review of Summation Notation

- Letters such as *x*, *y*,and *z* denote variables
- We use the subscript *i* to represent the *i*th observation of the variable
- *n* is the sample size

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \cdots + x_n$$

index

Example of Using Summation Notation

- The total number of cars I saw turning right onto Babcock (out of the Molly parking lot) during a week a few years back.
- I saw 2 on Monday, none on Wednesday, and 4 on Friday

$$x_1 = 2; x_2 = 0; x_3 = 4$$

$$\sum_{i=1}^{n} x_i = 2 + 0 + 4 = 6$$

Other Important Sums

$$\sum_{i=1}^{n} x_i^2$$

$$\sum_{i=1}^{n} \left(x_i - \overline{x} \right)$$

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

Measures of Central Tendency

- Descriptive measures that indicate where the center or most typical value of a data set lies, a.k.a. “averages”

1. Mean
2. Median
3. Mode

Mean – arithmetic average

Mean of a Data Set

The *mean* of a data set is the sum of the observations divided by the number of observations.

Notation for the Mean

- The mean is simply the average value of the observations.
- For a variable, the mean of the observations is denoted:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{x_i + x_i + \ldots + x}{n}$$

Median – Middle Value

Median of a Data Set

Arrange the data in increasing order.

- If the number of observations is odd, then the ***median*** is the observation exactly in the middle of the ordered list.
- If the number of observations is even, then the ***median*** is the mean of the two middle observations in the ordered list.

In both cases, if we let n denote the number of observations, then the median is at position $(n+1)/2$ in the ordered list.

Mode – most common value

(s)

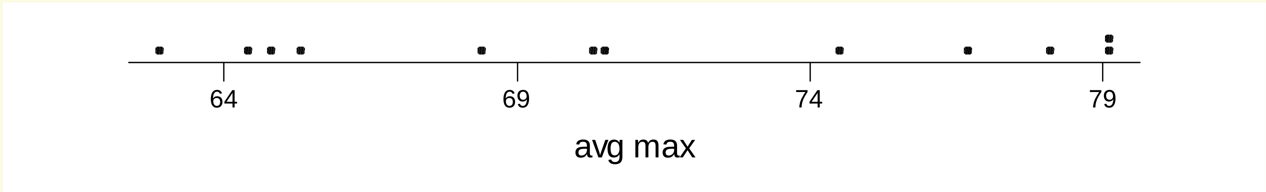
Mode of a Data Set

Obtain the frequency of occurrence of each value and note the greatest frequency.

- If the greatest frequency is 1 (i.e., no value occurs more than once), then the data set has no mode.
- If the greatest frequency is 2 or greater, then any value that occurs with that greatest frequency is called a ***mode*** of the data set.

Example – Average Daily Maximum Temperatures in San Luis Obispo, CA

Jan	62.9
Feb	64.8
Mar	65.3
Apr	68.4
May	70.3
Jun	74.5
Jul	78.1
Aug	79.1
Sep	79.1
Oct	76.7
Nov	70.5
Dec	64.4



$$\text{Mean} = \frac{62.9+64.8+\ldots+64.4}{12} = 71.175$$

Median = ?

Mode = 79.1

What about the median average?

62.9 64.4 64.8 65.3 68.4 70.3 | 70.5 74.5 76.7 78.1 79.1 79.1

Avg = 70.4

- Location = between 6th and 7th values
- Value = 70.4

Example – SRS of n = 15 Swiss doctors

- Mean

  41.3 hysterectomies done per year
- Median

  34 hysterectomies done per year
- Why are these measures of center so different?

Example continued

2	05578
3	13467
4	4
5	09
6	
7	
8	56

median=34

mean=41.3

The median uses the *location*, not the value, and will be more *resistant* to extreme observations

The mean will be pulled up by the two high values, i.e. in the direction of the skewness

Resistant = value is insensitive to outliers; median - yes; mean - no

A fix? - trimmed mean = 36.7

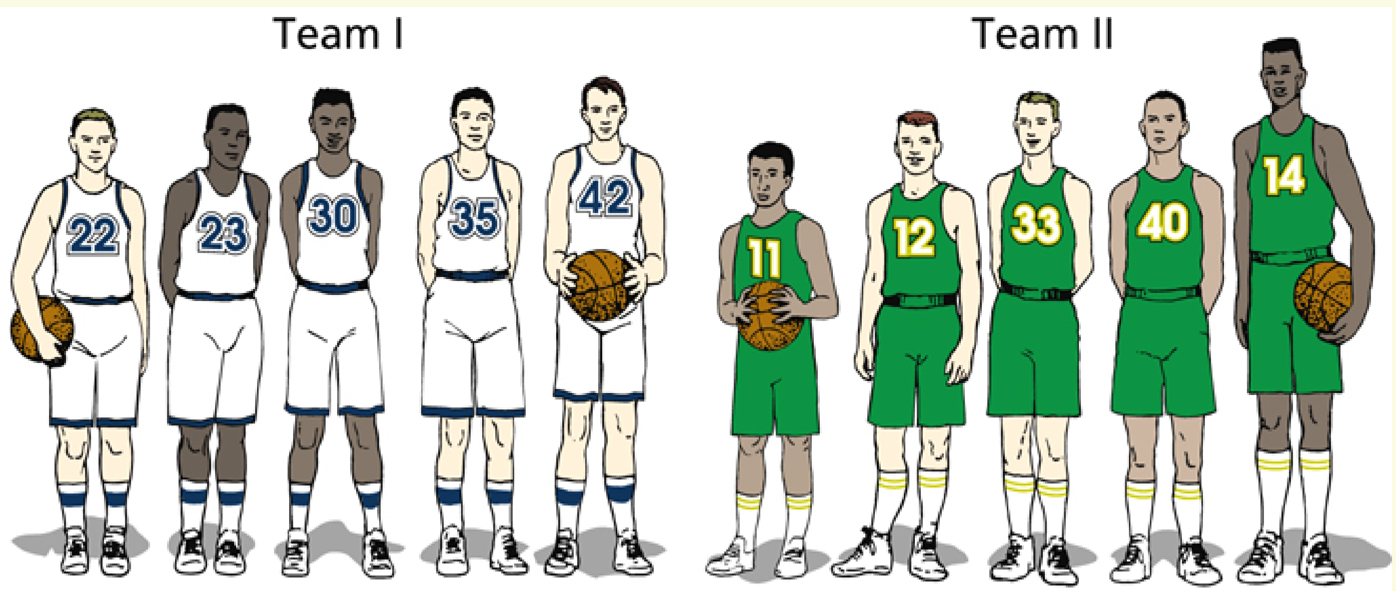
Which is the right answer?

- Depends!
  - Mean is generally preferred when histogram is bell shaped and symmetric
  - Median is often preferred for skewed data
  - Median is used to represent a *typical* value
  - Mean is used to represent average of *all* values
  - Mode may not be near the center
- Must look at graph and question asked to decide which is appropriate

Measures of Variation (Spread)

- Range
- Sample Standard Deviation
- Interquartile Range

Example

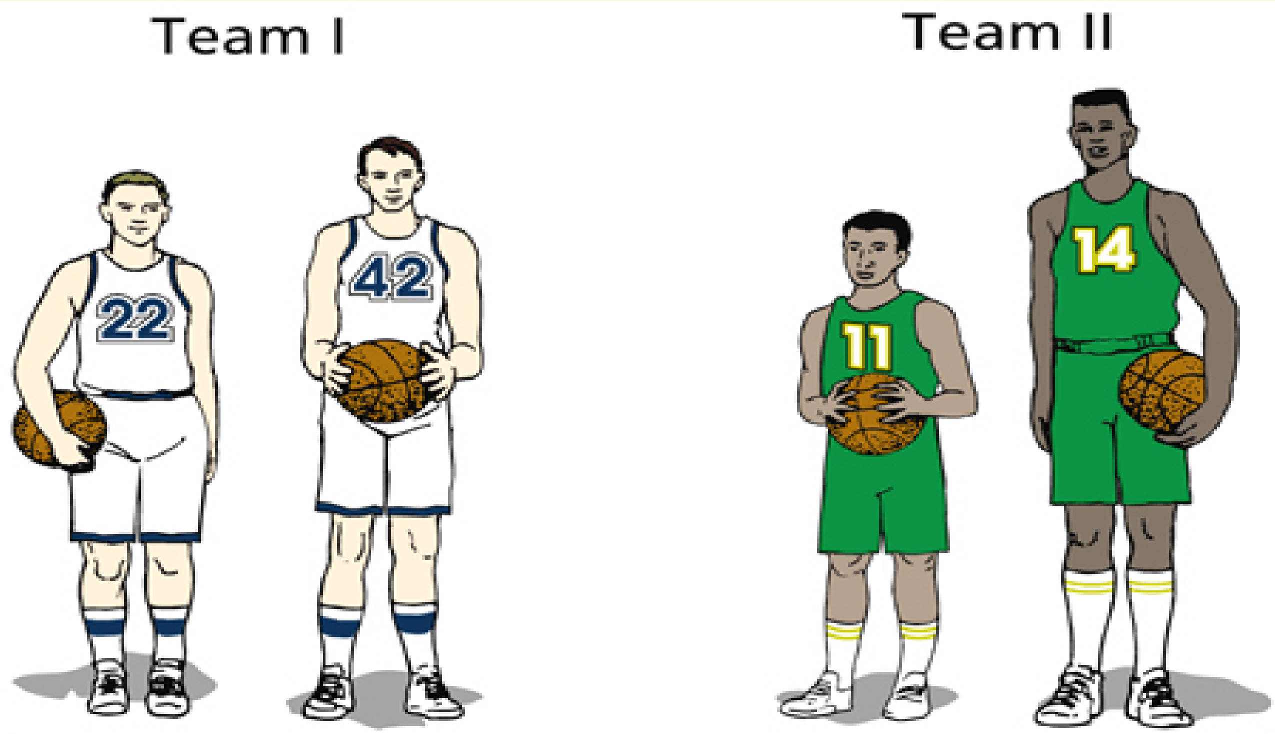


Feet and inches	6'	6'1"	6'4"	6'4"	6'6"	5'7"	6'	6'4"	6'4"	7'
Inches	72	73	76	76	78	67	72	76	76	84

Range of a Data Set

- The range of a data set is equal to the maximum observed value minus the minimum observed value
- Disadvantage? Information from other observations is ignored!

Example: What are the ranges?



	Team I		Team II	
Feet and inches	6'	6'6"	5'7 "	7 '
Inches	72	78	67	84

The Sample Standard Deviation

Sample Standard Deviation

For a variable x, the standard deviation of the observations for a sample is called a ***sample standard deviation.*** It is denoted s_x or, when no confusion will arise, simply s. We have

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

where n is the sample size.

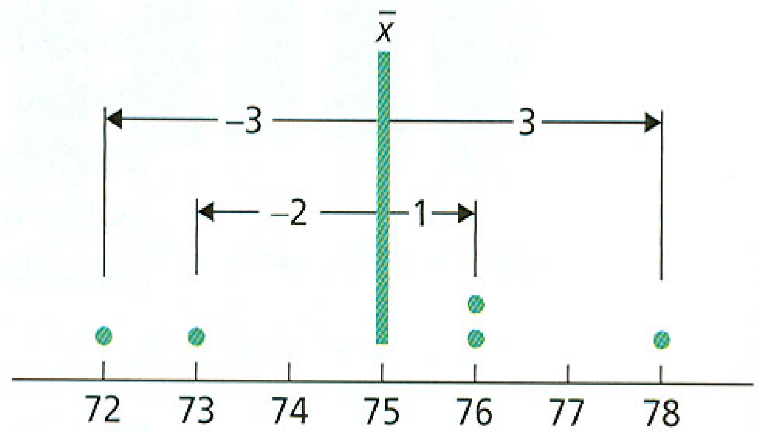
- A measure of variation by indicating how far, on average, the observations are from the mean
- Do not confuse with the population standard deviation which we will discuss later on

Deviations from the Mean

- The first step in calculating the sample standard deviation is to find how far each observation is from the mean.

Height x	Deviation from mean $x - \bar{x}$
72	−3
73	−2
76	1
76	1
78	3

Graphical display of the deviations from the mean (dots represent observations)



Deviations from the Mean

- Problem: Taking an average deviation won’t work. Do you know why?
- Solution: We will square the deviations first, and then take the average. Thus, we now have a measure of average deviation from the mean for all the observations.

Squared Deviations from the Mean

Height x	Deviation from mean $x - \bar{x}$	Squared deviation $(x - \bar{x})^2$
72	−3	9
73	−2	4
76	1	1
76	1	1
78	3	9
		24

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

a.k.a. "sum of squares"

The Sample Variance

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

- Can be thought of as an average squared deviation.
- So what's up with the *n* – 1?
- Two reasons – neither are obvious!

The Sample Variance - Example

$$s^2 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1} = \frac{24}{5-1} = 6 \text{ inches}^2$$

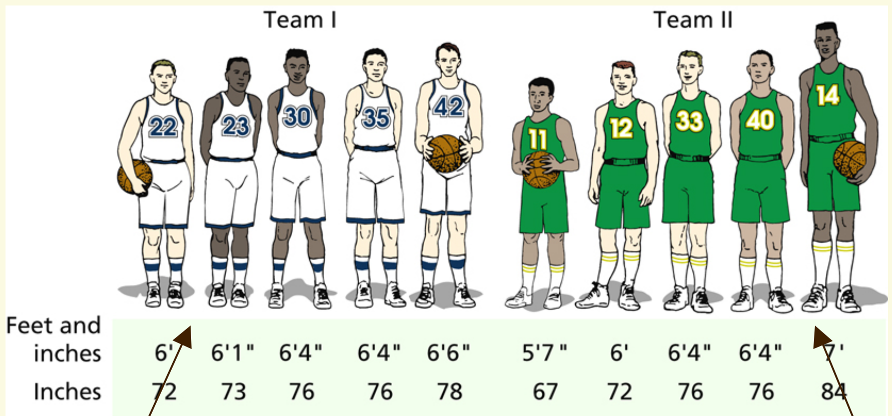
The Sample Standard Deviation - Example

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}} = \sqrt{6} = 2.4 \text{ inches}$$

- On average, the heights of the players on Team I vary from the mean height of 75 inches by 2.4 inches (notice we ditched the “squared”!)
- Get to know your calculator!

So What Does *s* Tell Us?

- The more variation there is in a data set, the larger is its standard deviation



s = 2.4 inches

s = 6.2 inches

The Downside

- *s* is not resistant: its value can be strongly affected by a few extreme observations
- Can anyone tell me why? Hint: inspect the formula for *s*

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

Alternative Computing Formula for *s*

Computing Formula for a Sample Standard Deviation

A sample standard deviation can be computed using the formula

$$s = \sqrt{\frac{\Sigma x^2 - (\Sigma x)^2/n}{n-1}},$$

where n is the sample size.

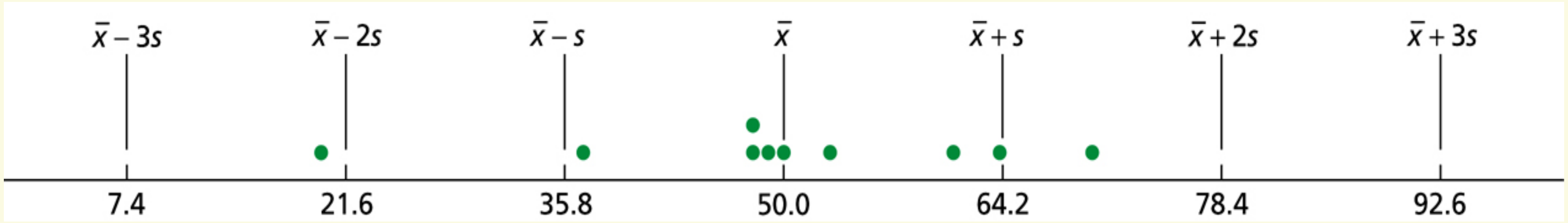
- We won't emphasize this formula

Rounding

- Do not perform any rounding until the computation is complete; otherwise, substantial roundoff error can result.
- Book: round final answers to one more decimal place than the raw data
- Me: round intermediate steps to four decimal places and the final answer to two decimal places

Further Interpretation of the Sample Standard Deviation – An Example

- Data -> 20, 37, 48, 48, 49, 50, 53, 61, 64, 70
- Sample Mean = 50.0
- Sample Standard Deviation = 14.2



Three-Standard-Deviation Rule

- Almost all the observations in any data set lie within three standard deviations to either side of the mean
- What does "almost all" mean?
  - For all data sets, at least 89%
  - For bell-shaped data sets, about 99.7%

Properties of Standard Deviation

- s measures spread about the mean and should be used only when the mean is chosen as a measure of center
- s=0 only when there is NO spread. (all observations have the same value)
- As the observations become more spread out about their mean, s gets larger.
- s, like the mean, is NOT resistant. A few ouliers can make s large.