BME 1132
Probability and Biostatistics

Instructor: Ali AJDER, *Ph.D.*

Week-3

Descriptive Statistics

- Table, Chart,
- Graphs: Line graphs, Bar graphs, Pie charts…

Measures of Locations

- Sample Mean,
- Median,
- Mode

Measures of Spread

- Sample Variance and Standard Deviation
- The Range
- Quantiles

Descriptive Statistics

- The first step in looking at data is to describe the data at hand in some concise way.

In smaller studies this step can be accomplished by **listing each data point**. In general, however, this procedure is tedious or impossible and, even if it were possible, would not give an overall picture of what the data look like.

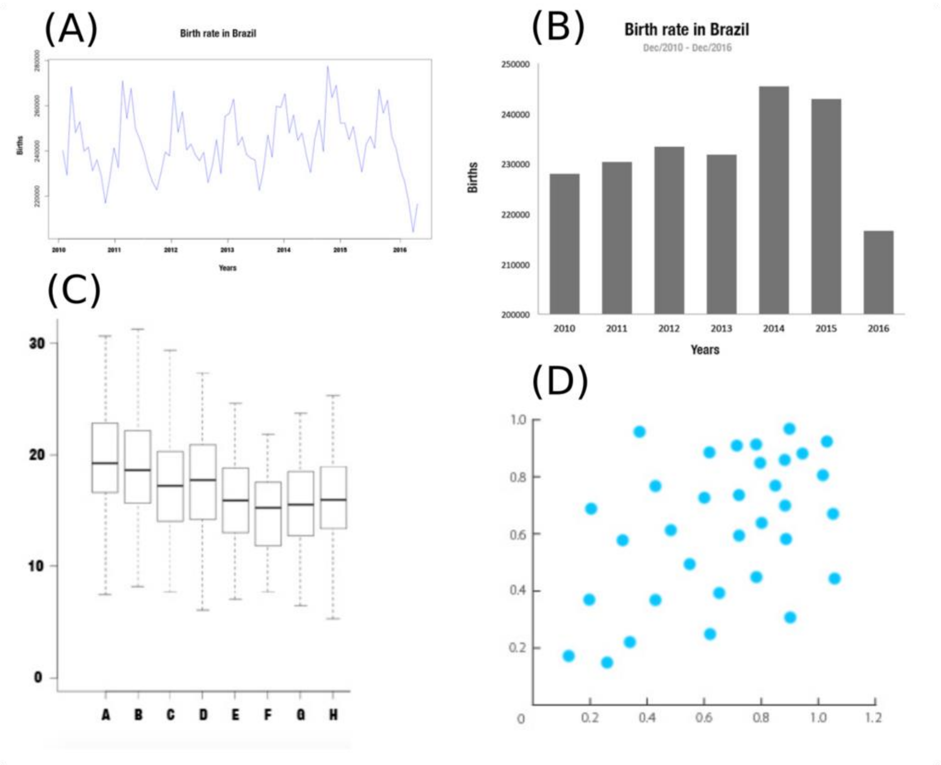
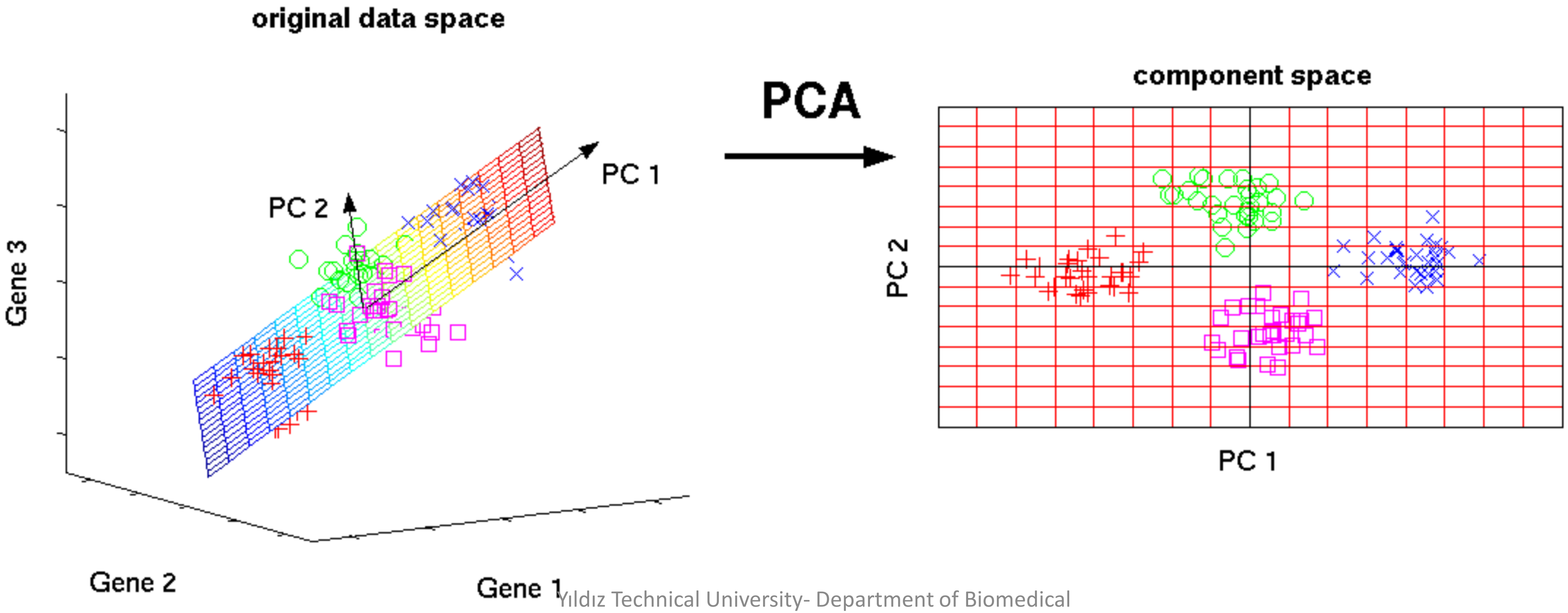


Figure A: **Line graph example**. The birth rate in Brazil (2010–2016);
Figure B: **Bar chart example.** The birth rate in Brazil for the December months from 2010 to 2016;
Figure C: **Example of Box Plot**: number of glycines in the proteome of eight different organisms (A-H);
Figure D: **Example of a scatter plot.**
https://www.wikiwand.com/en/Biostatistics

Describing Data with Charts, Tables and Graphs

Well-constructed data summaries and displays are essential to good **statistical thinking**, because they can focus the engineer on important features of the data or provide insight about the type of model that should be used in solving the problem.



What makes a good graphic or numeric display?

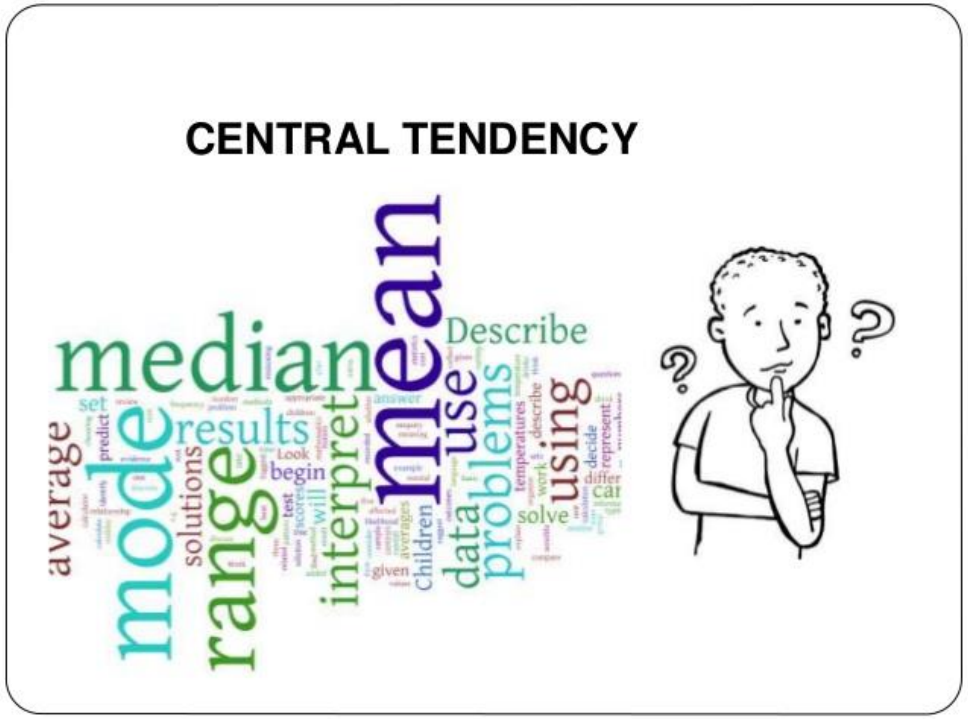
The main guideline is that the material should be as **self-contained** as possible and should be **understandable** without reading the text.

- ✓ The attributes require clear labeling.
- ✓ The captions, units, and axes on graphs should be clearly labeled, and
- ✓ The statistical terms used in tables and figures should be well defined.
- ✓ The quantity of material presented is equally important.

If bar graphs are constructed, then care must be taken to display neither too many nor too few groups. The same is true of tabular material.

Measures of Location- Central Tendency

The basic problem of statistics can be stated as follows:
Consider a **sample** of data $x_1, x_2, \ldots x_n$ where x_1 corresponds to the first sample point and x_n corresponds to the nth sample point. Presuming that the sample is drawn from some **population** $\boldsymbol{P}$, what inferences or conclusions can be made about P from the sample?



The Sample Mean

Sample Mean:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Sample Variance:

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

Sample **standard deviation** is the square root of the sample variance.

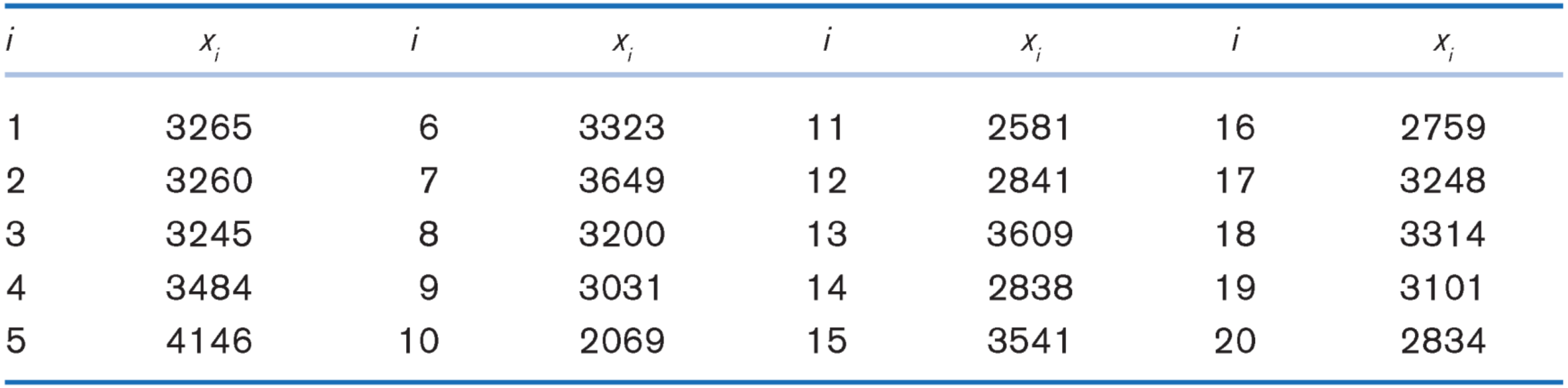
NOTE

$$\sum_{i=1}^{n} X_i = X_1 + X_2 + \cdots + X_n$$

$$\sum_{i=1}^{n} cX_i = c\left(\sum_{i=1}^{n} X_i\right)$$

The Sample Mean- Example

Suppose the sample consists of the birthweights of all live-born infants born at a private hospital, during an 1-week period.



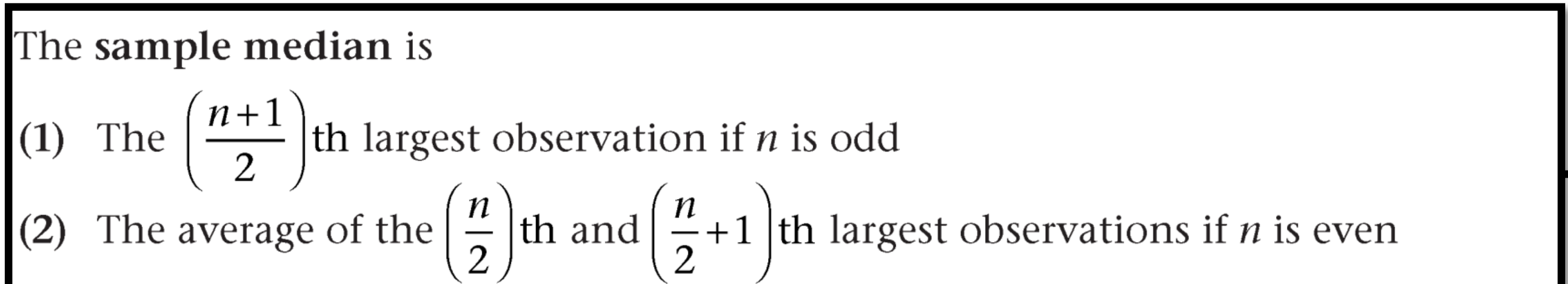
i	x_i	i	x_i	i	x_i	i	x_i
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

What is the mean and the standard deviation for the sample of birthweights in table?

The Median

An alternative measure of location, perhaps second in popularity to the arithmetic mean, is the median or, more precisely, the sample median.

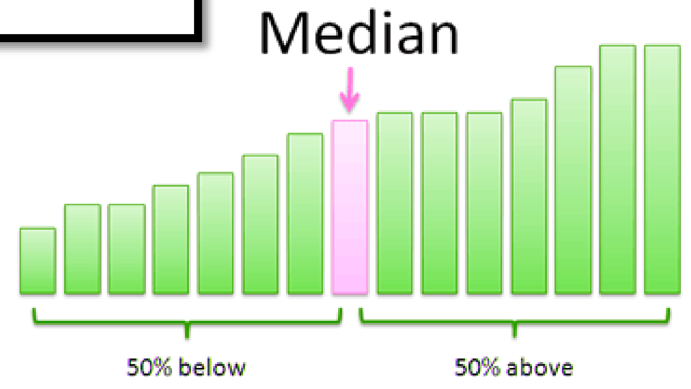
Suppose there are n observations in a sample. If these observations are ordered from smallest to largest, then the median is defined as follows:



The **sample median** is

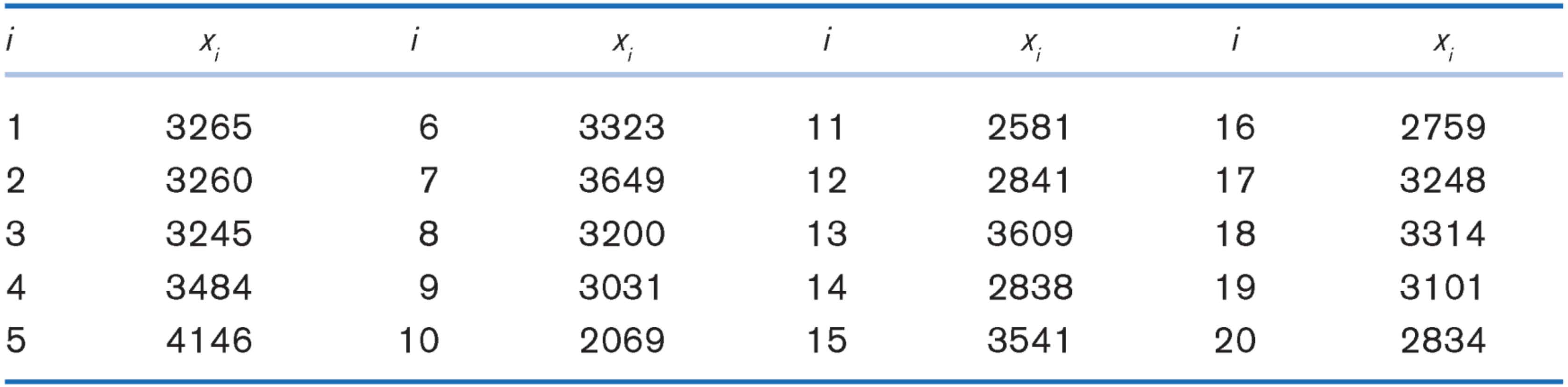
(1) The $\left(\frac{n+1}{2}\right)$th largest observation if n is odd

(2) The average of the $\left(\frac{n}{2}\right)$th and $\left(\frac{n}{2}+1\right)$th largest observations if n is even



The Median- Example

Suppose the sample consists of the birthweights of all live-born infants born at a private hospital, during a 1-week period.



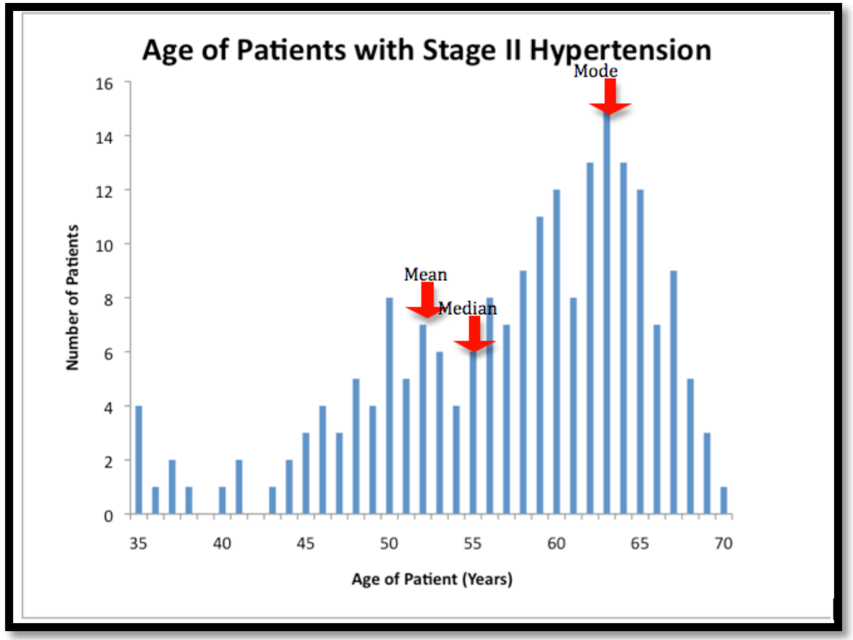
i	x_i	i	x_i	i	x_i	i	x_i
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

Compute the sample median for the sample in table?

The Mode

Another widely used measure of location is the mode.

The **mode** is the most frequently occurring value among all the observations in a sample.

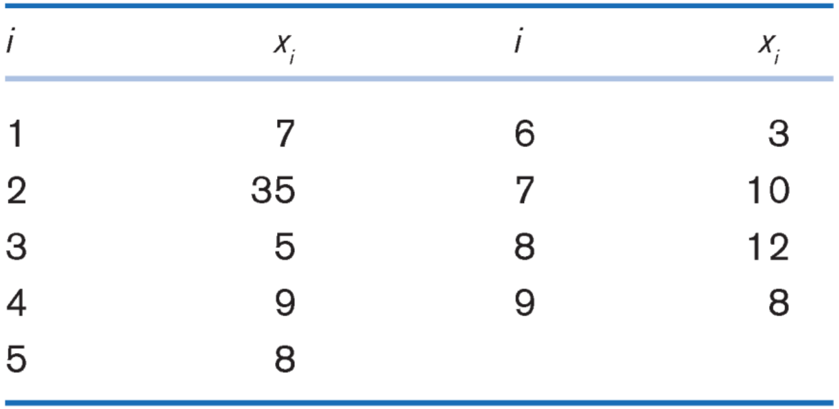


NOTE

A distribution with one mode is called **unimodal**; two modes, **bimodal**; three modes, **trimodal**; and so forth.

The Mode- Example

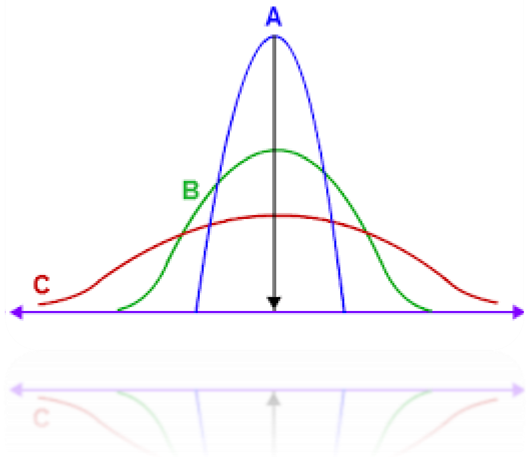
Consider the data set in Table, which consists of white-blood counts taken upon admission of all patients entering a small hospital on a given day.



i	x_i	i	x_i
1	7	6	3
2	35	7	10
3	5	8	12
4	9	9	8
5	8		

Compute the mode of the distribution in Table.

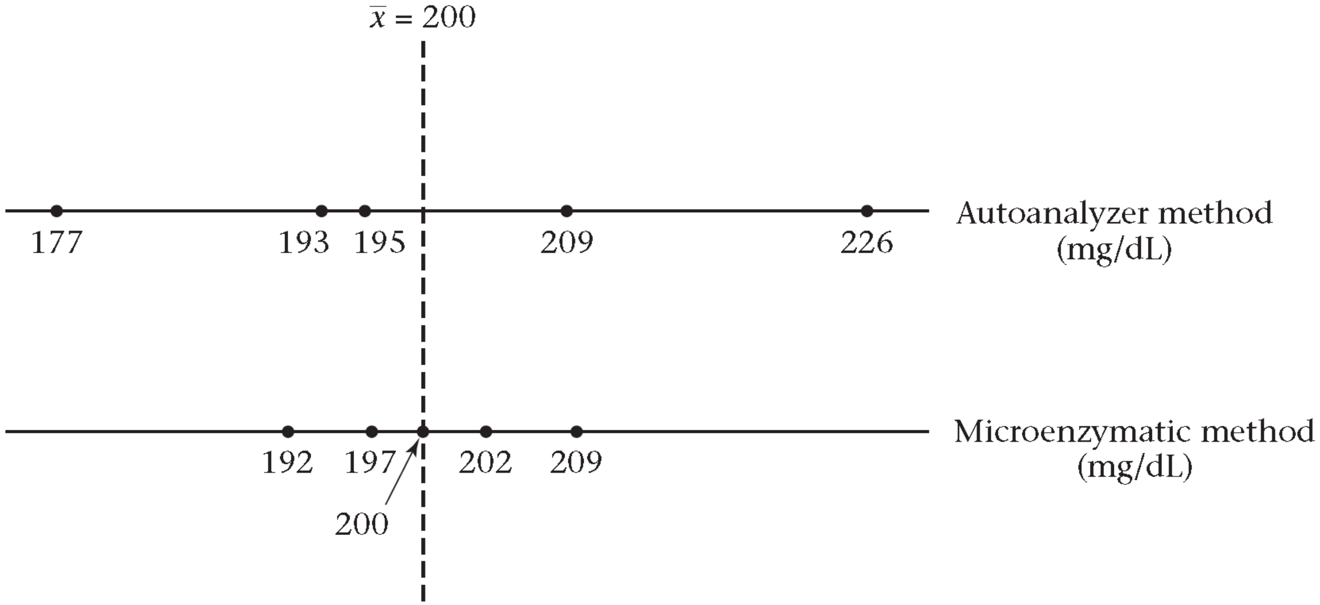
Measures of Spread



Following figure represents two samples of cholesterol measurements, each on the same person, but using different measurement techniques.

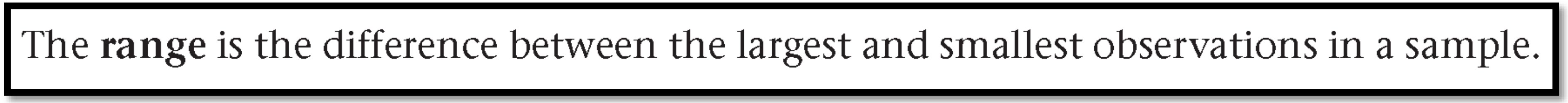
The samples appear to have about the **same center**, and whatever measure of central location is used is probably about the same in the two samples.

In fact, the arithmetic means are both 200 mg/dL. Visually, however, the two samples appear radically different. This difference lies in the greater **variability**, or **spread**, of the Autoanalyzer method relative to the Microenzymatic method.

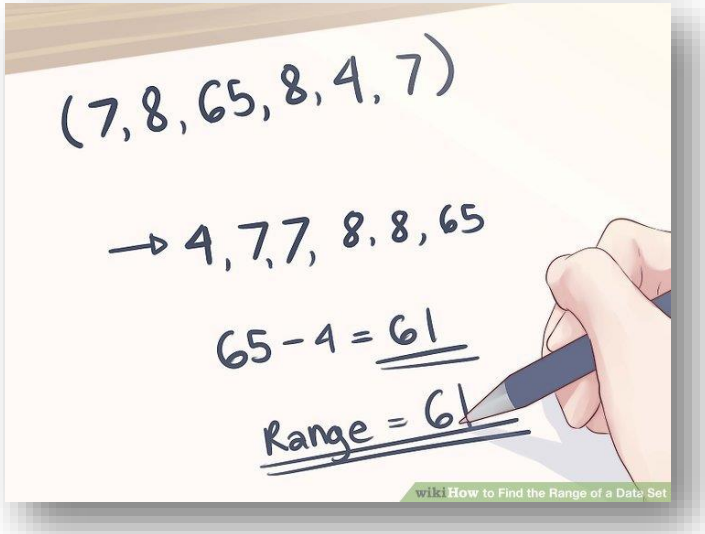


The Range

Several different measures can be used to describe the variability of a sample.
Perhaps the simplest measure is the range.



The **range** is the difference between the largest and smallest observations in a sample.

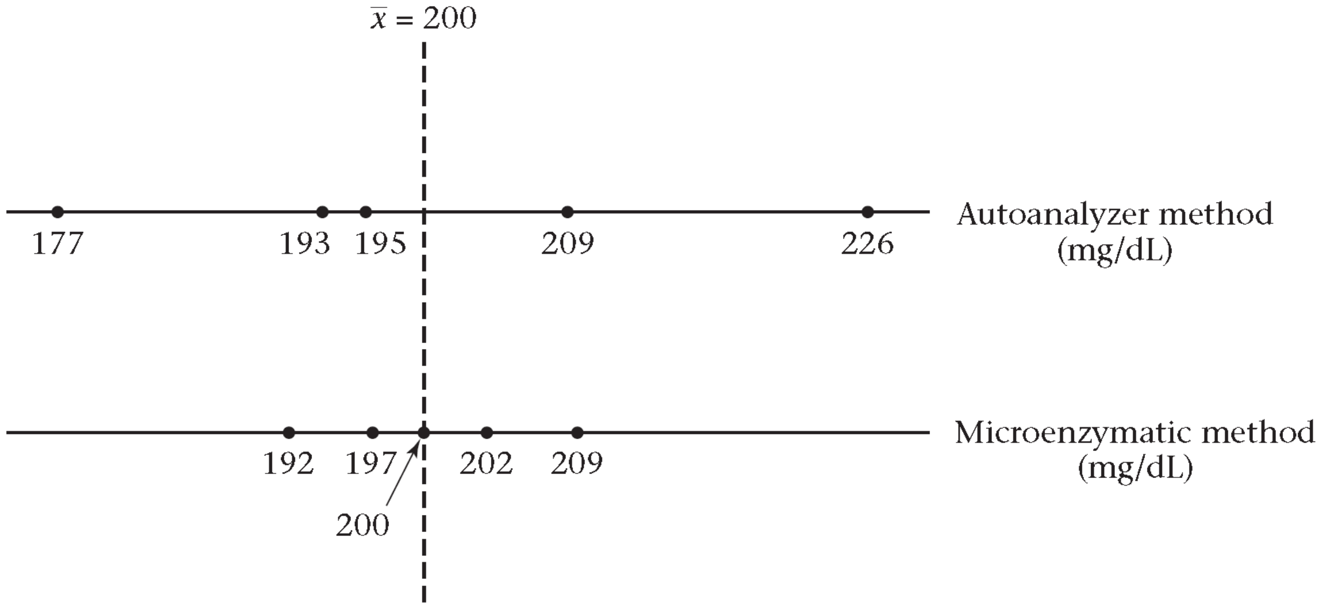


The Range- Examples

1. Compute the range in the sample of birthweights.

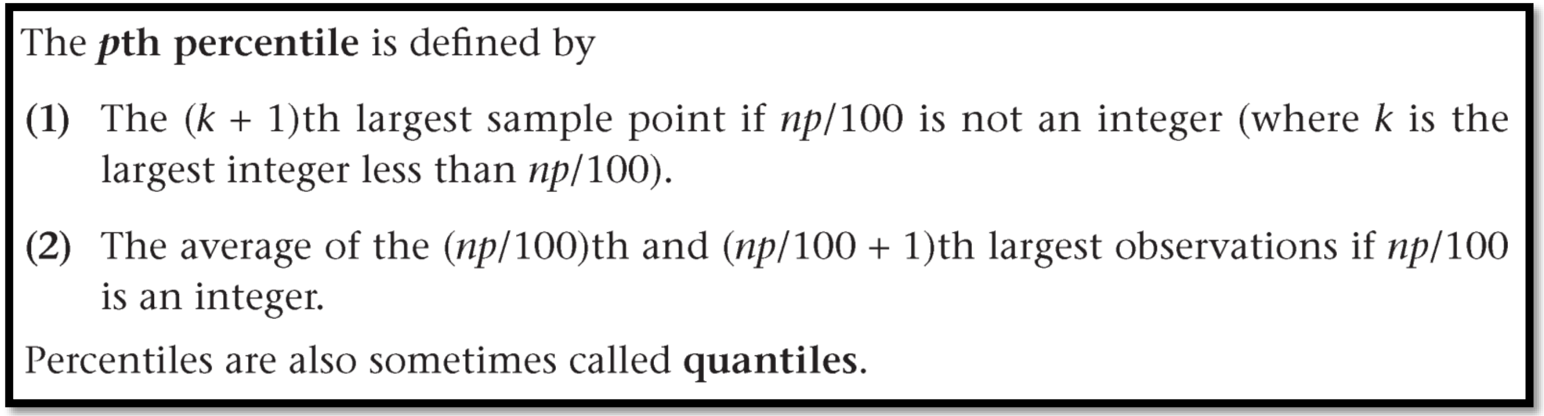
i	x_i	i	x_i	i	x_i	i	x_i
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

2. Compute the ranges for the Autoanalyzer-Microenzymatic-method data, and compare variability of the two methods.



Quantiles

Another approach that addresses some of the shortcomings of the range in quantifying the spread in a data set is the use of **quantiles** or **percentiles**.



The ***p*th percentile** is defined by

(1) The $(k + 1)$th largest sample point if $np/100$ is not an integer (where k is the largest integer less than $np/100$).

(2) The average of the $(np/100)$th and $(np/100 + 1)$th largest observations if $np/100$ is an integer.

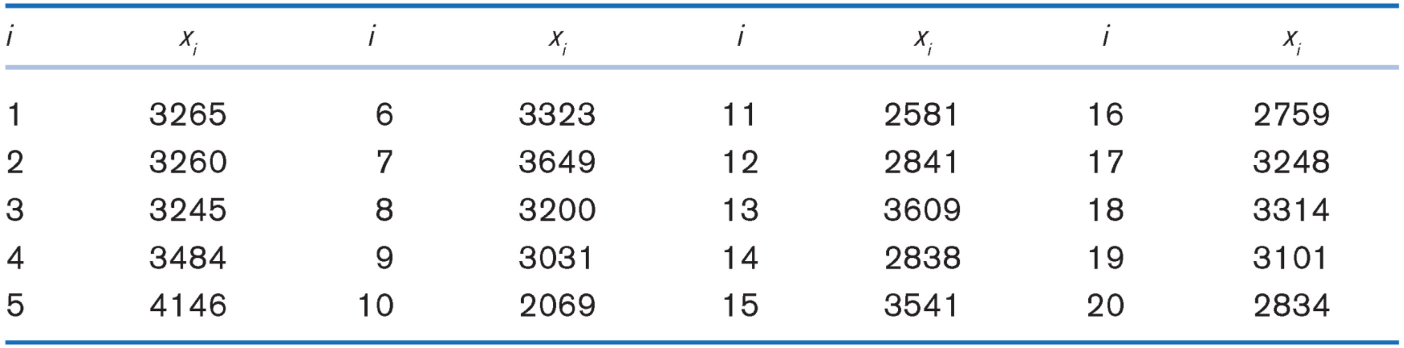
Percentiles are also sometimes called **quantiles**.

NOTE

There is no limit to the number of percentiles that can be computed.
The most useful percentiles are often determined by the sample size and by subject-matter considerations.
Frequently used percentiles are
tertiles (33rd and 67th percentiles),
quartiles (25th, 50th, and 75th percentiles),
quintiles (20th, 40th, 60th, and 80th percentiles), and
deciles (10th, 20th, . . . , 90th percentiles).
It is almost always instructive to look at some of the quantiles to get an overall impression of the spread and
the general shape of a distribution.

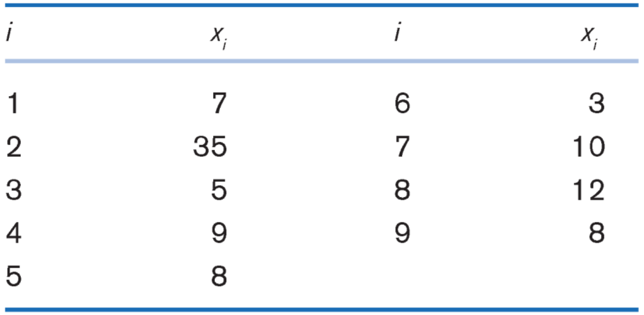
Quantiles Examples

1. Compute the 10th and 90th percentiles for the birthweight data in Table.



i	x_i	i	x_i	i	x_i	i	x_i
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

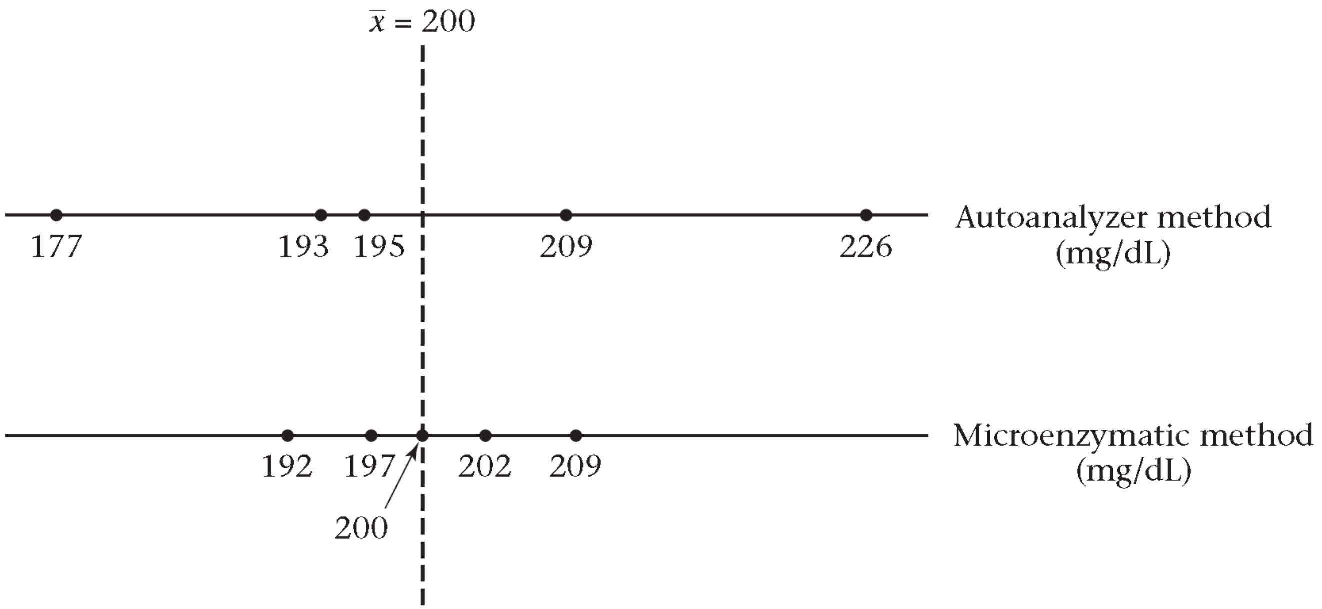
2. Compute the 20th percentile for the white-blood-count data in Table.



i	x_i	i	x_i
1	7	6	3
2	35	7	10
3	5	8	12
4	9	9	8
5	8		

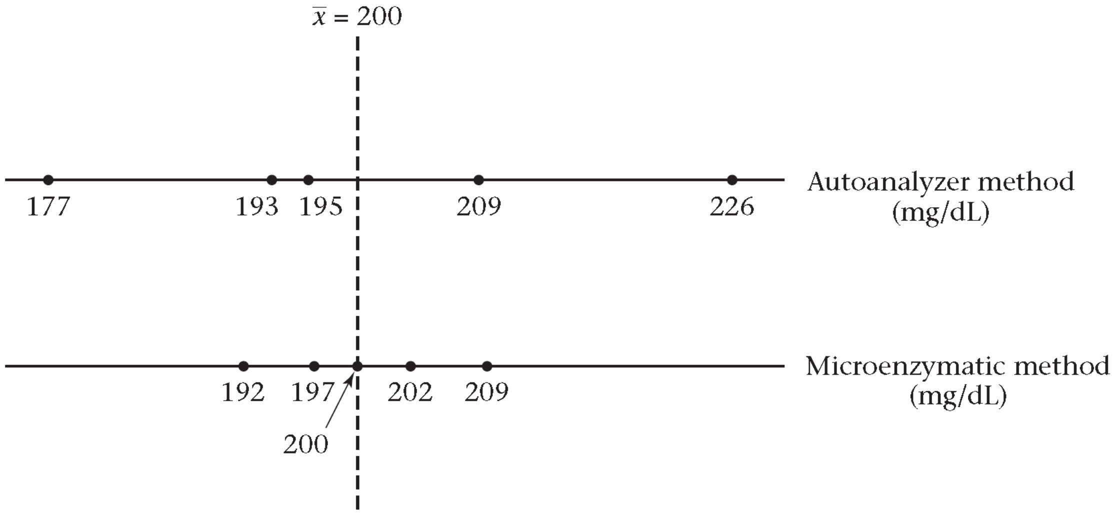
Examples- Variance and Standard Deviation

Compute the variance and standard deviation for the Autoanalyzer- and Microenzymatic-method data.



Examples- Variance and Standard Deviation

Compute the variance and standard deviation for the Autoanalyzer- and Microenzymatic-method data.



Solution: Autoanalyzer Method

$$s^2 = \left[(177-200)^2 + (193-200)^2 + (195-200)^2 + (209-200)^2 + (226-200)^2\right]/4$$

$$= (529+49+25+81+676)/4 = 1360/4 = 340$$

$$s = \sqrt{340} = 18.4$$

Microenzymatic Method

$$s^2 = \left[(192-200)^2 + (197-200)^2 + (200-200)^2 + (202-200)^2 + (209-200)^2\right]/4$$

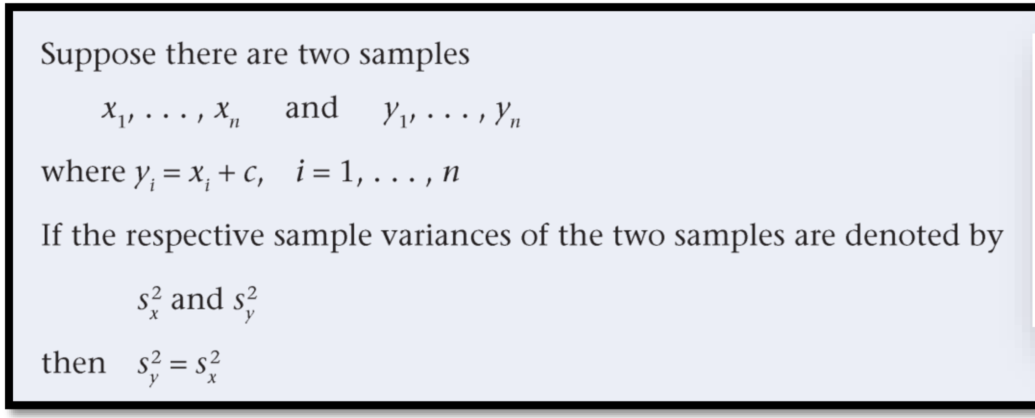
$$= (64+9+0+4+81)/4 = 158/4 = 39.5$$

$$s = \sqrt{39.5} = 6.3$$

Thus the Autoanalyzer method has a standard deviation roughly three times as large as that of the Microenzymatic method.

Some Properties of the Variance and Standard Deviation

How are they affected by a change in origin or a change in the units being worked with?



Suppose there are two samples

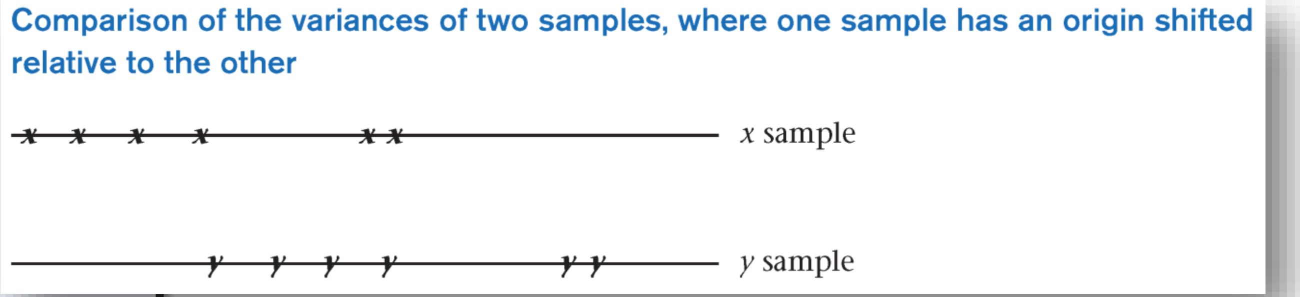
$$x_1, \ldots, x_n \quad \text{and} \quad y_1, \ldots, y_n$$

where $y_i = x_i + c, \quad i = 1, \ldots, n$

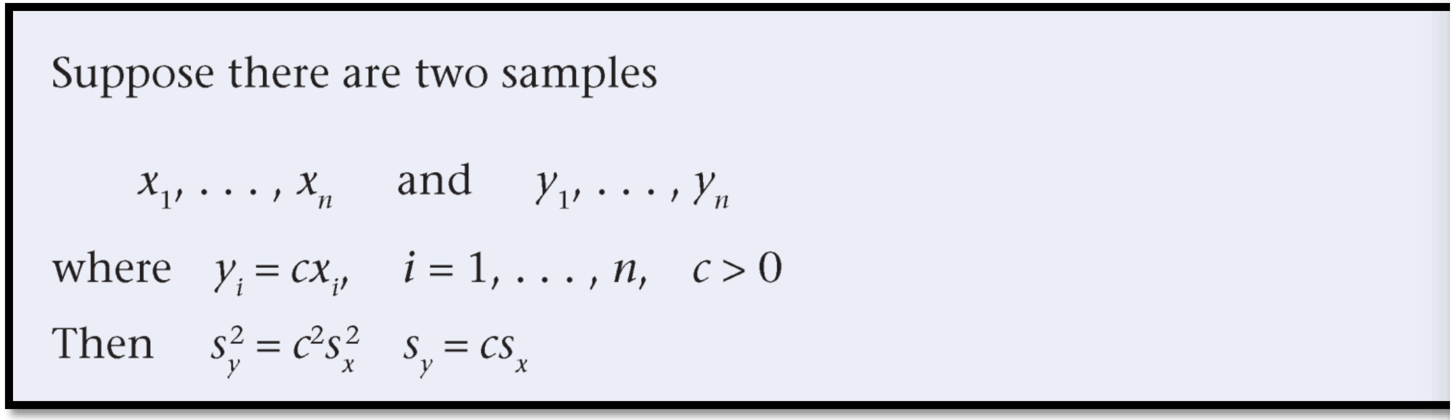
If the respective sample variances of the two samples are denoted by

$$s_x^2 \text{ and } s_y^2$$

then $s_y^2 = s_x^2$



Comparison of the variances of two samples, where one sample has an origin shifted relative to the other

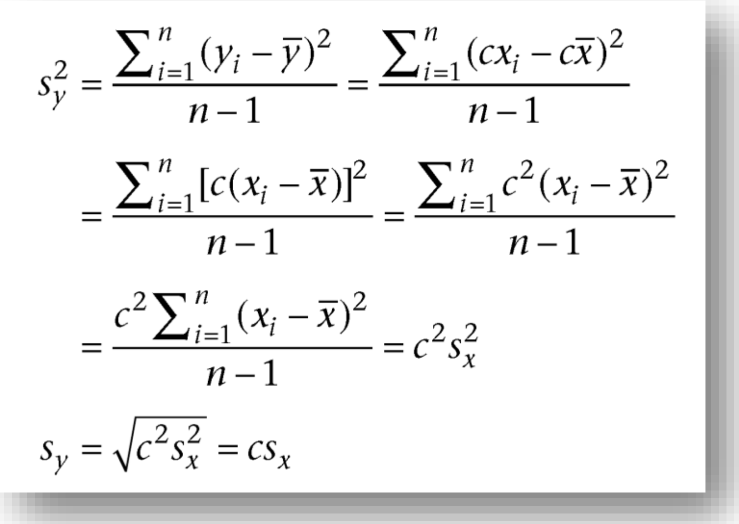


Suppose there are two samples

$$x_1, \ldots, x_n \quad \text{and} \quad y_1, \ldots, y_n$$

where $y_i = cx_i, \quad i = 1, \ldots, n, \quad c > 0$

Then $s_y^2 = c^2 s_x^2 \quad s_y = cs_x$



$$s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \frac{\sum_{i=1}^{n}(cx_i - c\bar{x})^2}{n-1}$$

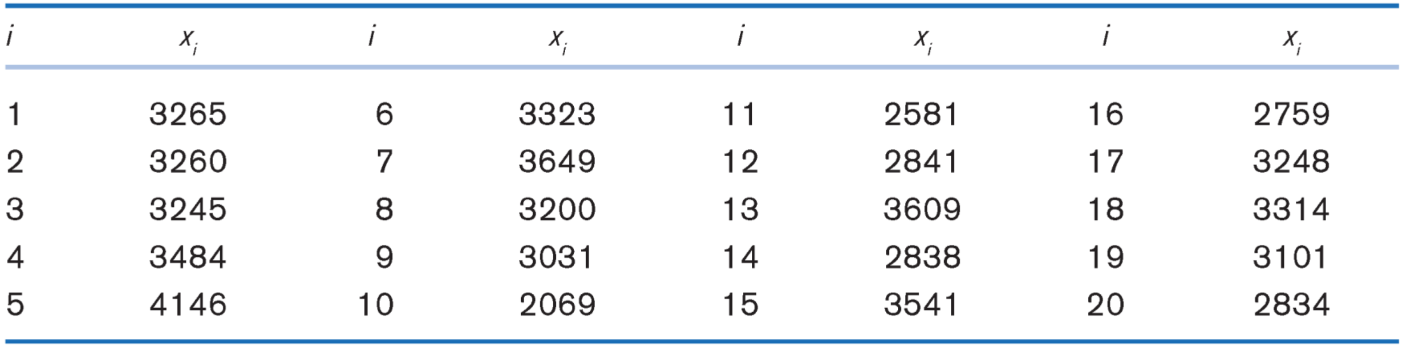
$$= \frac{\sum_{i=1}^{n}[c(x_i - \bar{x})]^2}{n-1} = \frac{\sum_{i=1}^{n}c^2(x_i - \bar{x})^2}{n-1}$$

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

$$s_y = \sqrt{c^2 s_x^2} = cs_x$$

Example- Variance and Standard Deviation

Compute the variance and standard deviation of the birthweight data in Table in both **grams** and **ounces**.



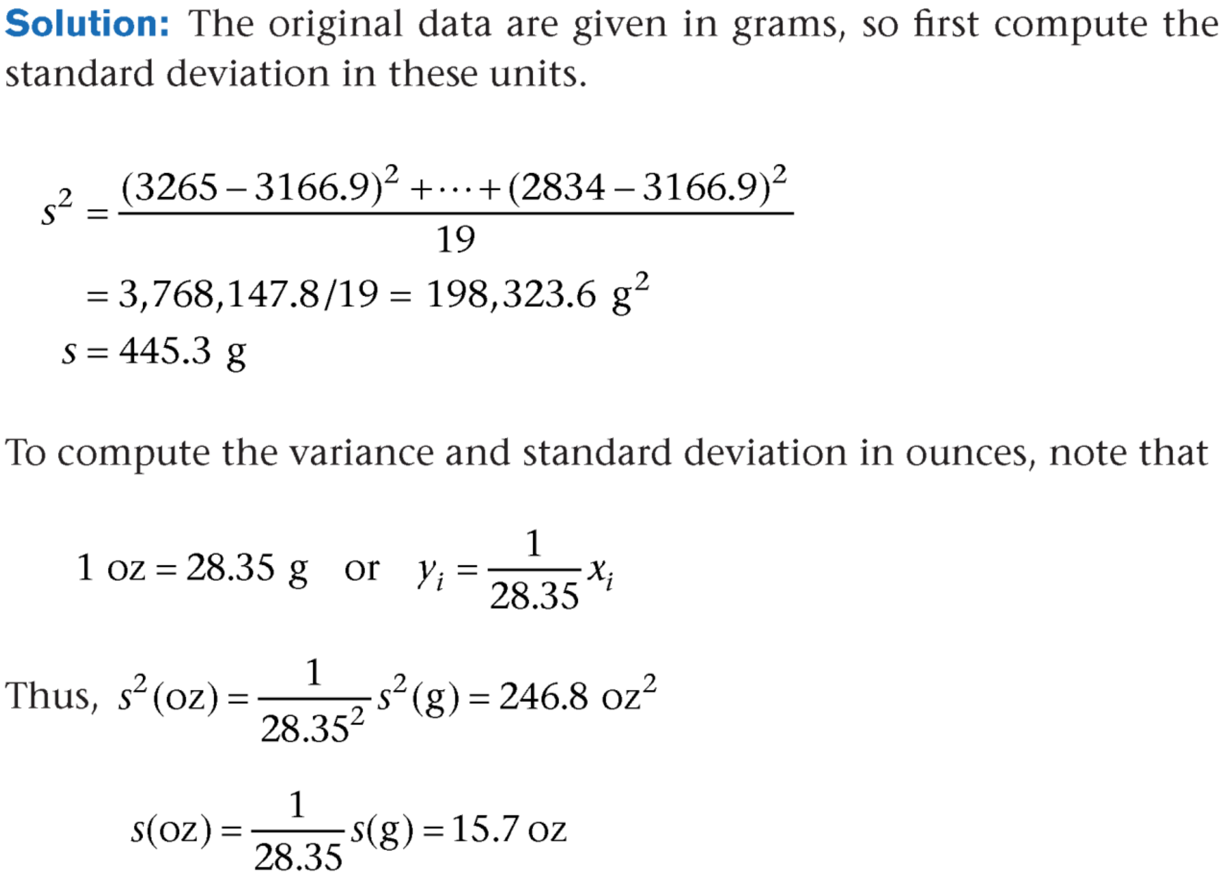
i	x_i	i	x_i	i	x_i	i	x_i
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

NOTE

$$\boxed{\begin{array}{c} 1\ oz = 28.35\ g\ or \\ y_i = \dfrac{1}{28.35} x_i \end{array}}$$

Example- Variance and Standard Deviation

Compute the variance and standard deviation of the birthweight data in Table in both **grams** and **ounces**.



Solution: The original data are given in grams, so first compute the variance and standard deviation in these units.

$$s^2 = \frac{(3265-3166.9)^2 + \cdots + (2834-3166.9)^2}{19}$$
$$= 3{,}768{,}147.8/19 = 198{,}323.6 \text{ g}^2$$
$$s = 445.3 \text{ g}$$

To compute the variance and standard deviation in ounces, note that

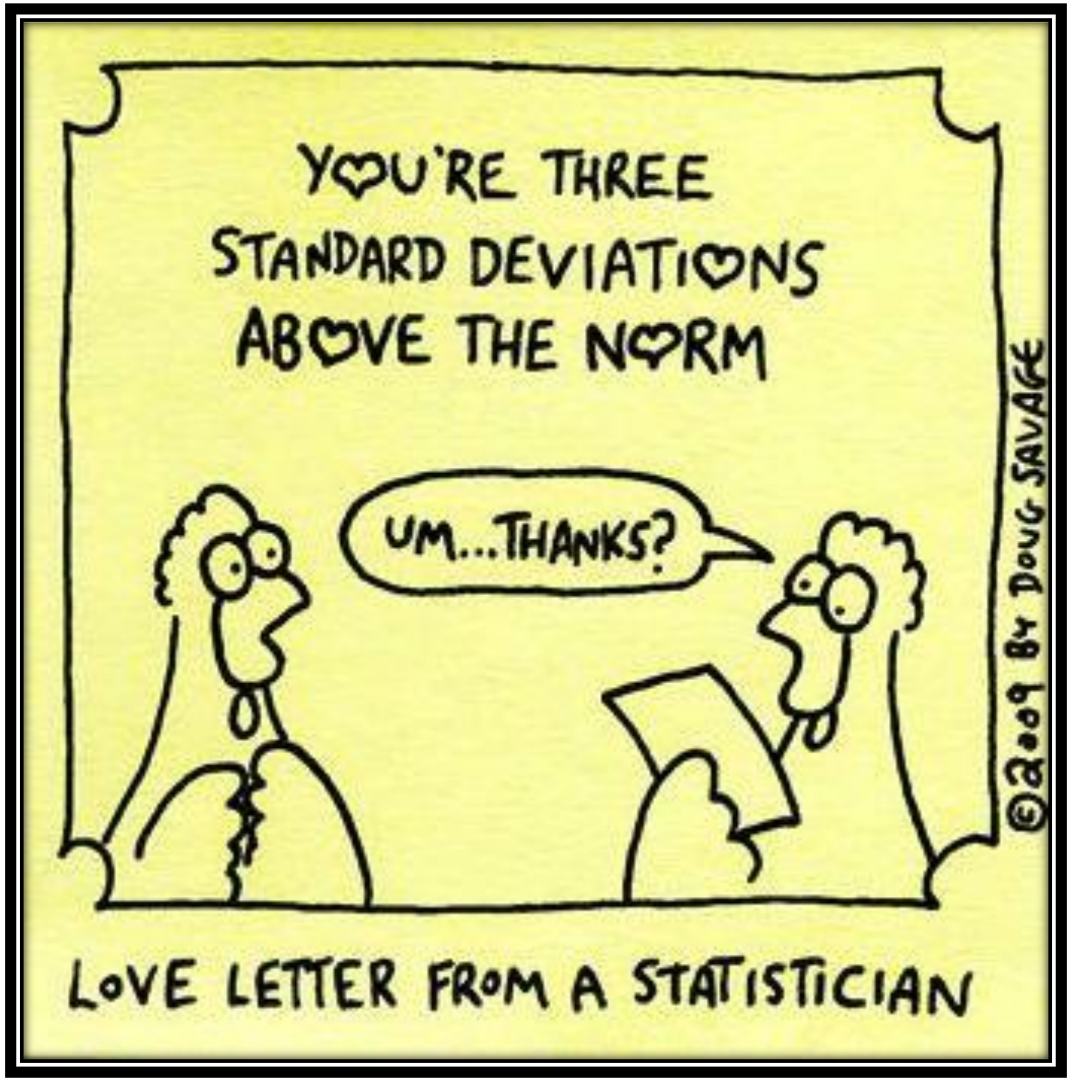
$$1 \text{ oz} = 28.35 \text{ g} \quad \text{or} \quad y_i = \frac{1}{28.35} x_i$$

Thus, $s^2(\text{oz}) = \frac{1}{28.35^2} s^2(\text{g}) = 246.8 \text{ oz}^2$

$$s(\text{oz}) = \frac{1}{28.35} s(\text{g}) = 15.7 \text{ oz}$$

Result:
Thus, if the sample points change in scale by a factor of c, the variance changes by a factor of c^2 and the standard deviation changes by a factor of c. This relationship is the main reason why the standard deviation is more often used than the variance as a measure of spread: the standard deviation and the arithmetic mean are in the same units, whereas the variance and the arithmetic mean are not.

Questions?


YOU'RE THREE STANDARD DEVIATIONS ABOVE THE NORM
UM...THANKS?
LOVE LETTER FROM A STATISTICIAN
©2009 BY DOUG SAVAGE