Measures of central tendency (center, location): measures the middle point of a distribution or data; these include mean and median.

Measures of dispersion (variability, spread): measures the extent to which the observations are scattered; these include standard deviation and range.

Parameter: a numerical descriptive measure computed from the population measurements (census data).

Statistic: a numerical descriptive measure computed from sample measurements.

Throughout we denote
N: the number of observations in the population, that is the population size,
n: the number of observations in a sample, that is, the sample size, and
x_i : the *i*-th observation.

Measures of Central Tendency (Location)

Population mean

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

Sample mean

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

Formulas> Insert Function (f_x) >Statistical>AVERAGE

Note that the population mean and sample mean are **arithmetic means** or **average**, and computed in the same manner.

Example 1. Find the sample mean of the following data.
5, 7, 2, 0, 4

Deviation from the mean: observation minus the mean, i.e. $x_i - \mu$ or $x_i - \bar{x}$.

The mean is a measure of the center in the sense that it "balances" the deviations from the mean.

Median: The middle value of data when ordered from smallest to largest
Formulas> Insert Function (f_x) >Statistical>MEDIAN

Example 2. Find the sample median of the data in Example 1.

The median is a measure of the center in the sense that it "balances" the number of observations on both sides of the median.

Example 3. Find the median of the following data.

5, 7, 2, 4, 2, 1

The mode is useful as a measure of central tendency for qualitative data.

Graphical comparison of the mean, median, and mode

Example 4. Find the mean and median of the following data and compare the results with those of Example 1.

5, 70, 2, 0, 4

HC Homework: 3.1-a & d only, 3.2-a & d only on p.117.

Geometric mean: the *n*-th root of the product of *n* values

$G = \sqrt[n]{x_1 \times x_2 \times \cdots \times x_n}$.

Formulas>Insert Function (f_x) >Statistical>GEOMEAN

Geometric mean rate of return (change): an average rate of change; a good way to sort out effects that are multiplicative

$$\sqrt[n]{(1+r_1) \times (1+r_2) \times \cdots \times (1+r_n)} - 1$$

where r_i the rate of change in decimals of *i*-th period.

Example 5. Suppose the interest rates on a savings account during a five-year period are given below. Find the average interest rate per year during the five year period.

Year	1	2	3	4	5
Interest rate	3.7%	2.7%	2.4%	2%	1.3%

For the above example the numbers the you enter for **GEOMEAN** in Excel are 1.037, 1.027, 1.024, 1.02, and 1.013.

MSL Homework: 3.21
HC Homework: 3.22

Measures of Dispersion

Range=largest observation - smallest observation

Mean absolute deviation (**MAD**):

$$\frac{1}{N}\sum_{i=1}^{N} |x_i - \mu|$$

for census data and

$$\frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|$$

for sample data

Formulas >Insert Function (f_x)>Statistical>AVEDEV

Example 6. Find the mean absolute deviation of the data: 40, 55, 75, 95, 95

Population variance:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \frac{1}{N}(\sum_{i=1}^{N} x_i^2 - N\mu^2)$$

The (population) variance is the average of the squared deviations from the population mean.

Formulas >Insert Function (f_x)>Statistical>VAR.P

Population standard deviation:

$$\sigma = \sqrt{\sigma^2}$$

Formulas >Insert Function (f_x)>Statistical>STDEV.P

Sample variance:

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

The (sample) variance is the "average" of the squared deviations from the (sample) mean.

Formulas >Insert Function (f_x)>Statistical>VAR.S

Sample standard deviation:

$$s = \sqrt{s^2}$$

Formulas >Insert Function (f_x)>Statistical>STDEV.S

Example 7. Find the variance and standard deviation of the sample data in Example 6.

Example 8. Find the standard deviation of the data in Example 7 assuming they are census data.

Coefficient of variation:

$\frac{\sigma}{\mu} \times 100$ % for population and $\frac{s}{\bar{x}} \times 100$ % % for sample

a relative measure of the amount of variation with respect to the mean.

Note the standard deviation is an absolute measure of variation.

Example 9. Find the coefficient of variation of the data in Example 6.

HC Homework: 3.1-b only, 3.2-b only on p.117

Example 10. Suppose based on the history of two stocks A and B we have the following information. Find the C.V.'s of the stocks.

	Stock A	Stock B
Mean Price	\$100/share	\$10/share
St. Dev.	\$5	\$5

You may get summary measures using MS Excel: **Data >Data Analysis….> Descriptive Statistics**

Note that the standard deviation (as well as variance) given by the above is always based on the "sample" formula. Therefore for census data we need to make the following correction for the correct standard deviation.

$\sqrt{\frac{N-1}{N}}$ (St. Dev. from Excel)

or use the function wizard as follows: **Formulas >Insert Function (f_x)>Statistical>STDEV.P**.

Standardized score:

$z = \frac{x-\mu}{\sigma}$ for population data and $z = \frac{x-\bar{x}}{s}$ for sample data,

sometimes called a *z*-score represents the signed distance from the mean measured in units of standard deviation. The standardized variable has mean zero and standard deviation one and does not have a measurement unit.

Formulas >Insert Function (f_x)>Statistical>STANDARDIZE.

Example 11. Suppose the test score of a student is 75 on the accounting test with mean 70 and standard deviation 10 and the test score of the student is 65 on the statistics test with mean 60 and standard deviation 5. Compute the standardized test scores of the accounting and statistics test scores of this student, and interpret the standardized scores.

HC Homework: 3.1-c only, 3.2-c only on p.117.
MSL Homework: 3.13, 3.14, 3.17

Chebychev's Theorem: For any sample or population data the proportion of observations that lie within k standard deviations from the mean is at least $1 - 1/k^2$.

This theorem describes the distribution of data using the mean and the standard deviation together.

Example 12. When the sample mean and standard deviation are 77 and 9.04, respectively, find an interval around the mean that covers at least 75% of the data. Also find an interval that covers at least 95% of the data.

Empirical rule: When the distribution of (sample or population) data is approximately bell shaped (or mound-shaped), then
approximately 68% of the data are within 1 standard deviation from the mean,
approximately 95% of the data are within 2 standard deviation from the mean, and
approximately 99% of the data are within 3 standard deviation from then mean.

Example 13. When the sample mean and standard deviation are 77 and 9.04, respectively, find an interval around the mean that covers approximately 95 % of the data. Assume the distribution of the data is bell shaped.

Outliers: Extreme observations not conforming to the rest of the observation. As a rule of thumb observations that are three standard deviations above or below the mean are considered as outliers.

MSL Homework: 3.37
HC Homework: 3.40

Suppose the observations $x_1, x_2, \cdots, x_n$ have been arranged in ascending order. The ***p*-th percentile** is the value denoted by $x_{(p)}$ such that at least *p* percent of the observations are less than or equal to $x_{(p)}$ and at least (*100-p*) percent of the observations are greater than or equal to $x_{(p)}$.

Formulas >Insert Function (f_x)>Statistical>PERCENTILE.INC

Procedure for calculating percentiles:

1. Arrange *n* observations in ascending order.
2. Calculate ***i=np/100.***
3. If ***i*** is an integer, the *p*-th percentile is the arithmetic average of x_i and x_{i+1}.
4. If ***i*** is not an integer, the *p*-th percentile is x_j, where *j* is the smallest integer greater than ***i.***

Example 14: Suppose we have data

40, 37, 50, 26, 30, 59, 8, 50

a. Find a 25-th percentile.
b. Find a 90-th percentile.
c. Find a 75th percentile.

The **first quartile**, also called lower quartile and denoted by Q_1, is the 25th percentile.

The **third quartile**, also called upper quartile and denoted by Q_3, is the 75th percentile.

Formulas >Insert Function (f_x)>Statistical>QUARTILE.INC

The **inter-quartile range** (IQR) is $Q_3 - Q_1$, and is a measure of variability.

Midhinge: $\frac{Q_1+Q_3}{2}$, a measure of location.

Example 15: In Example 14, find the inter-quartile range and midhinge.

A **box-(and whisker) plot** is a graphical method of displaying the quartiles, ranges, and extreme values of data. In a box plot

the left "I"(also called the left hinge) is at Q_1, the right "I" (right hinge) at Q_3, and "I" inside of the box is at the median,

the right (left) whisker extends out from the right (left) hinge to the largest (smallest) observation with 1.5 IQR above (below) the right (left) hinge,

observations beyond the reach of the whiskers are considered outlying observation.

The Inner fences are located 1.5× IQR below Q_1 and above Q_3.

The Outer fences are located 3× IQR below Q_1 and above Q_3.

HC Homework: 3.27 and Calculate 20th and 90th percentile of the data for the problem.
MSL Homework: 3.32

Example of MS Excel output for the 227 observations of the 1-year return % of the Growth Funds in the data file "Retirement Funds."

Growth	
Mean	14.27797357
Standard Error	0.332135457
Median	14.18
Mode	16.95
Standard Deviation	5.004125231
Sample Variance	25.04126933
Kurtosis	5.147918131
Skewness	0.203923123
Range	45.26
Minimum	-11.28
Maximum	33.98
Sum	3241.1
Count	227

f_n >Statistical>AVEDEV 3.45945

AVERAGE 14.27797

MEDIAN 14.18

QUARTILE.INC 11.790 (with quart 1)

16.635 (with quart 3)

STDEV.P 4.993 (Is this appropriate for the data?)

PERCENTILE.INC 18.86 (90^{th} percentile with k=0.9)

(Do you get the same value when you apply the procedure for percentiles discussed in class?)