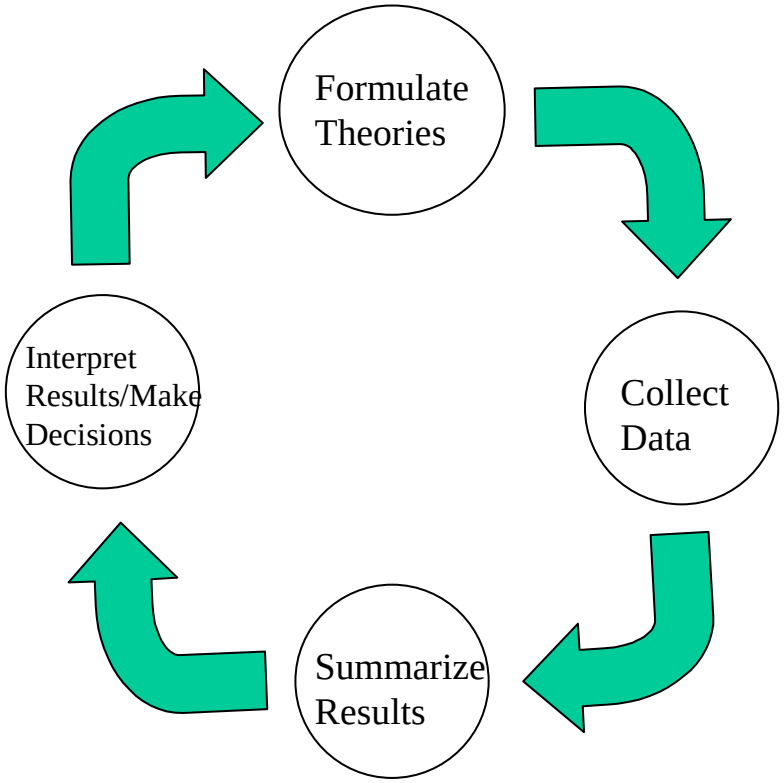
Chapter Goals

To establish the usefulness of summary measures of data.



Numerical Descriptive Measures

Measures of Position (Central Tendency)

1. Mean
2. Median
3. Mode

Measures of Variability

1. Range
2. Variance

The Arithmetic Mean

Definition *The mean of a set of quantitative data, $X_1, X_2, \ldots, X_n$, is equal to the sum of the measurements divided by the number of measurements.*

In notation,

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

$\bar{X}$ is usually called the sample mean. If the X_i's are values of a population, the mean is called the "population" mean, and is denoted by the symbol μ.

Example *Find the mean of the following* 5 *numbers:* 5, 3, 8, 5 *and* 6.

Limitations of the Mean

- The mean is influenced by extreme values.
- For non-symmetrical distributions, the mean is located away from the concentration of items.

Let's do it! 5.1

Kim's test scores are 7, 98, 25, 19, and 26. Calculate Kim's mean test score. Does the mean do a good job of capturing Kim's test scores?

Let's do it! 5.2

The mean score for 3 students is 54 and the mean score for 4 different students is 76.

What is the mean score for all 7 students?

Median

Definition *The median Md of a data set is the middle number when the measurements are arranged in ascending (or descending) order.*

Calculating the median:

1. Arrange the n measurements from the smallest to the largest.
2. If n is odd, the median is the middle number.
3. If n is even, the median is the mean (average) of the middle two numbers.

Example *Previous data set (number* 5, 3, 8, 5, 6*)*

Properties of Median

- From the above definition, the median *Md* of a set of items is in the middle of the values ordered by magnitude. What if the values in the data set are repeated?
- The median is not affected by any extreme items in a data set.

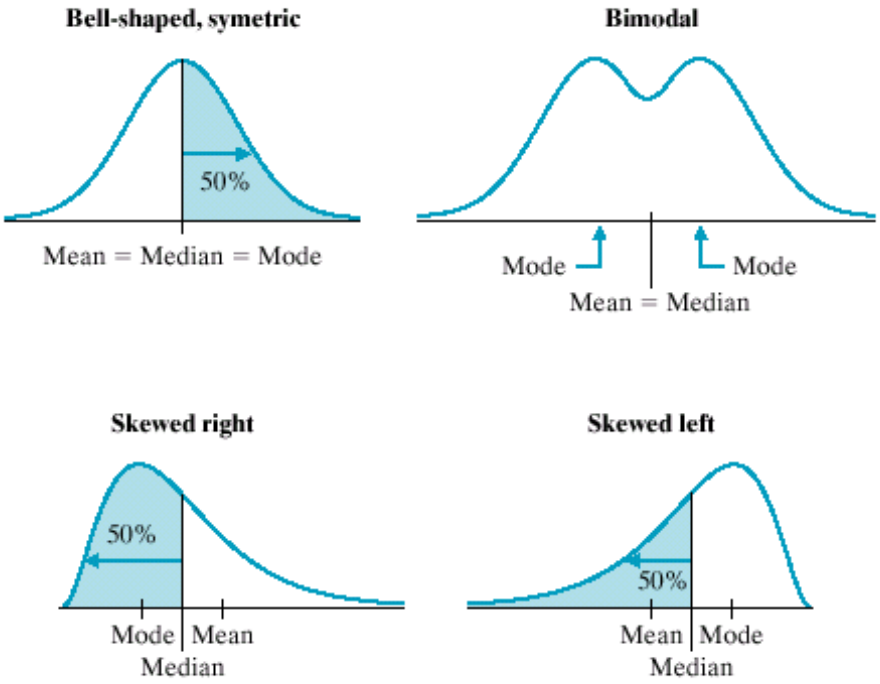
Mode

Definition *The mode is the measurement that occurs with the greatest frequency in the data set.*

Example: Previous data set (number 5, 3, 8, 5, 6)

The modal class in a frequency distribution with equal class intervals is the class with the largest frequency. If the frequency polygon has only a single peak, it is said to be unimodal. If the frequency polygon has two peaks, it is said to be bimodal.

Which Measure to Use?



Let's do it! 5.5

Consider a study to compare two antibiotics for treating strep throat in children, amoxicillin and cefadroxil. At one clinic for this study, 23 children (who met the study entrance criteria and for whom consent was given) were randomly assigned to one of the two treatment groups. One concern is that age of the child might influence the effectiveness of the antibiotics. The ages of the children in each treatment group are given. Calculate the mean, median, and mode of the ages for each of the two treatment groups.

Amoxicillin Group (n $=$ 11) AGE: 14, 17, 11, 10, 11, 14, 9, 12, 8, 10, 9

Mean

Median

Mode

Cefadroxil Group ($n = 12$**) AGE**: 9, 14, 8, 10, 13, 7, 9, 11, 16, 10, 12, 9

Mean

Median

Mode

Percentiles

Let $x_1, x_2, \cdots, x_n$ be a set of n measurements arranged in increasing (or decreasing) order. The pth percentileis a number **x** such that p% of the measurements fall below the pth percentile.

- In large data sets where values do not tend to repeat themselves extensively, the 25th percentile indicates that about 25 percent of the items are less than this value and about 75 percent are more. Other percentiles are interpreted correspondingly. When the values in a data set do tend to repeat themselves, this interpretation is no longer appropriate.
- A single value can correspond to more than one percentile.
- Many percentiles are known by other names. Percentiles that are multiples of 25 are called quartiles. Thus, Similarly, percentiles that are multiples of 10 are called deciles. For instance, the 70th percentile is also called the 7th decile.

Measures of Variability (Dispersion)

Range

Definition *The **range** is the difference between the largest and smallest values in a set of items.*

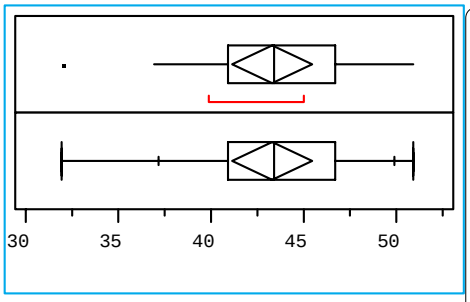
Definition *The **interquartile range** is the difference between the third and first quartiles of the data set.*

Example: $5, 3, 8, 5, 6$

- Considers only extreme items.
- With a frequency distribution, the range of original data cannot be determined exactly.

Box Plots

Example *The ages of the* 20 *subjects in a medical study are as follows:* 32, 37, 39, 40, 41, 41, 41, 42, 42, 43, 44, 45, 45, 45, 46, 47, 47, 49, 50, 51.



Quantiles

maximum	100.0%	51.000
	99.5%	51.000
	97.5%	51.000
	90.0%	49.900
quartile	75.0%	46.750
median	50.0%	43.500
quartile	25.0%	41.000
	10.0%	37.200
	2.5%	32.000
	0.5%	32.000
minimum	0.0%	32.000

Quantile Box Plot The quantile box plot shows selected quantiles on the response axis. The box shows the median as a line across the middle and the quartiles (25th and 75th percentiles) as its ends. The means diamond identifies the mean of the sample and the 95% confidence interval about the mean.

Outlier Box Plot The Outlier Box Plot is a schematic that lets you see the sample distribution and identify points with extreme values, or outliers. The ends of the box are the 25th and 75th quantiles, also called the quartiles. The difference between the quartiles is the interquartile range. The line across the middle identifies the median sample value. The ends of the whiskers are the outer-most data points from their respective quartiles that fall within the distance computed as 1.5 *(interquartile range). The bracket along the edge of the box identifies the shortest half, which is the most dense 50% of the observations.

Let's do it! 5.7

Here is the data from Let's do it! 5.5:

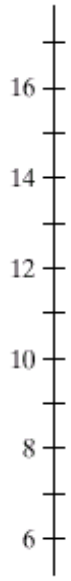
Amoxicillin Group ($n = 11$**)** 8, 9, 9, 10, 10, 11, 11, 12, 14, 14, 17

Five-number summary: Q1=9, Q3=14, Md=11, Min=8, Max=17

Cefadroxil Group ($n = 12$**)** 7, 8, 9, 9, 9, 10, 10, 11, 12, 13, 14, 16

Five-number summary: Q1 = 9, Q3 = , Md = 10, Min = 7, Max = 16

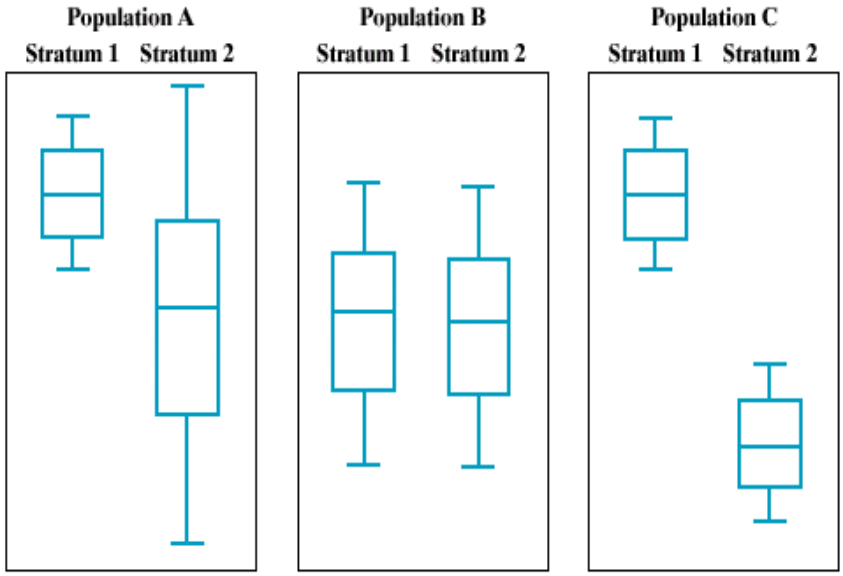
Exercise *Make side-by-side boxplots for the above data*



Exercise *Mark any outliers in your graph.*

Let's do it! 5.8

The following graphs are side-by-side box-plots of some variable for two strata in three hypothetical populations, A, B, and C. In each population the units are evenly divided between the two strata.



Consider three sampling designs to estimate the true population mean (the total sample size is the same for all three designs):

1. simple random sampling
2. stratified random sampling taking equal sample sizes from the two strata
3. stratified random sampling taking most units from one strata, but sampling a few units from the other strata

- For which population will design (1) and (2) be comparably effective?
- For which population will design (2) be the best?
- For which population will design (3) be the best?
- Which stratum in this population should have the higher sample size?

Variance

Definition *The **sample variance** s^2 of a set of data values $X_1, X_2, \ldots, X_n$ is defined as*

$$s^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1},$$

where $\overline{X}$ is the sample mean.

An alternative formula for calculating the variance s^2 is the following:

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

Example: $5, 3, 8, 5, 6$

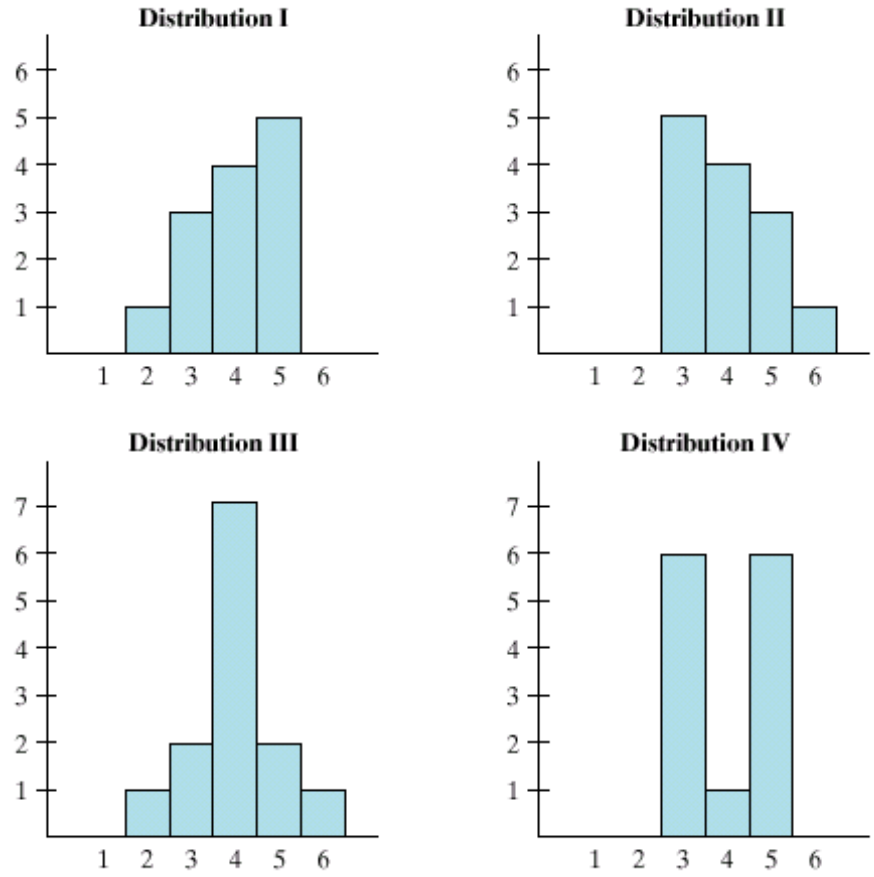
Comments

- The greater the variability of the values in a data set, the greater the variance is. If there is no variability of the values — that is, if all are equal and hence all are equal to the mean — then $s^2 = 0$.
- The variance s^2 is expressed in units that are the square of the units of measure of the characteristic under study. Often, it is desirable to return to the original units of measure which is provided by the standard deviation.
- The positive square root of the variance is called the sample **standard deviation** and is denoted by s:

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

Example *The table below gives summary measures for 4 datasets:*

Measure	*Data Set*			
	I	*II*	*III*	*IV*
Range	3	3	4	2
IQR	2	2	1	2
s	1	1	1	1



Which dataset has the greatest variability?

Linear Transformations

Why the need for transformation?

Example *The data on the number of children in a neighborhood of* 10 *households is as follows:* 2, 3, 0, 2, 1, 0, 3, 0, 1, 4. *The mean for this dataset is* $\overline{X} = 1.6$, *and the standard deviation is* $s = 1.43$.

Exercise *If there are two adults in each of the above households, what is the mean and standard deviation of the number of people (children + adults) living in each household?*

Exercise *If each child gets an allowance of* $3, *then the total amount of allowance given to children in each household is as follows:* 6, 9, 0, 6, 3, 0, 9, 0, 3, 12. *What is the mean and standard deviation of the amount of allowance in each household in this neighborhood?*

Definition *Let X be the variable representing a set of values, and s_x and $\overline{X}$ be the standard deviation and mean of X, respectively. Let $Y = aX + b$, where a and b are constants. Then, the mean and standard deviation of Y are given by $\overline{Y} = a\bar{X} + b$, and $s_Y = |a|s_X$.*

Definition *A variable X is said to be* ***standardized*** *if the variable has a mean of 0 and a standard deviation of 1. To standardize a variable, you first subtract its mean and then divide by its standard deviation. Note that the standardized variable $(X - \bar{X})/s_X$ is a linear transformation where $a = \frac{1}{s_X}$ and $b = -\frac{\bar{X}}{s_X}$.*

Example *During a recent week in Europe, the temperature was as follows:*

$X = temp$	M	T	W	H	F	Sa	Su
Celcius	40	41	39	41	41	40	38

Exercise *Based on this,* $\bar{X} = 40^o$ *Celsius, and* $s_X = 1.14^o$ *Celsius. Calculate the mean and standard deviation in Fahrenheit.*

Exercise *Calculate the standardized scores for the above data.*