Introduction to Statistics

Objectives

To review basic concepts in used making quantitative or numerical judgments.

Eg :Frequency measures used in epidemiology

- Measures of central location and dispersion
- Group comparisons
- Hypothesis testing versus estimation

Proportions and Percentages

Proportion

A ratio with a unique characteristic: those in the numerator are also included in the denominator

"The proportion of adults in Ebonia who have hypertension is 0.2"

i.e. 2 million affected persons/10 million people

Percentage

Mathematically, the proportion expressed per 100 or *percent.*

"20% of adults in Ebonia have hypertension"

Rates

Rate

a ratio in which there is a distinct relationship between the numerator and denominator and a measure of time is an intrinsic part of the denominator

Example:

“Number of cases of stroke per 100,000 adults during 2004”

Note: "rate" often used [wrongly] for measures that have no time element.

Using count data

To measure disease or risk factor frequency, we need:

1. A count of affected population
2. The size of the source population
3. The time period at or over which the data were collected

In epidemiology and public health, we use these quantities to calculate prevalence (and index of how frequent a disease is) and incidence (an index of the rate at which a disease occurs)

Prevalence

$$\text{Prevalence} = \frac{\text{no. of existing cases of a disease}}{\text{total population at a given time}}$$

Example: 2 million adults in Ebonia have hypertension, out of a population of 10 million. The prevalence of hypertension is 20% (i.e. 2 million/10 million)

Incidence

Incidence quantifies the number of new cases or events that develop in a population of individuals at risk during a specified time period or interval

Two types of incidence measures: cumulative incidence and incidence rate.

The incident rate is also known as the incident density

Cumulative Incidence

$$\text{Cumulative incidence} = \frac{\text{number of new cases during a given time period}}{\text{total population at risk}}$$

Provides an estimate of the probability or risk that an individual will develop a disease during a specified time period.

Example: The cumulative incidence of hypertension among adults in Ebonia is 2.5 percent over a one year period.

Incidence rate

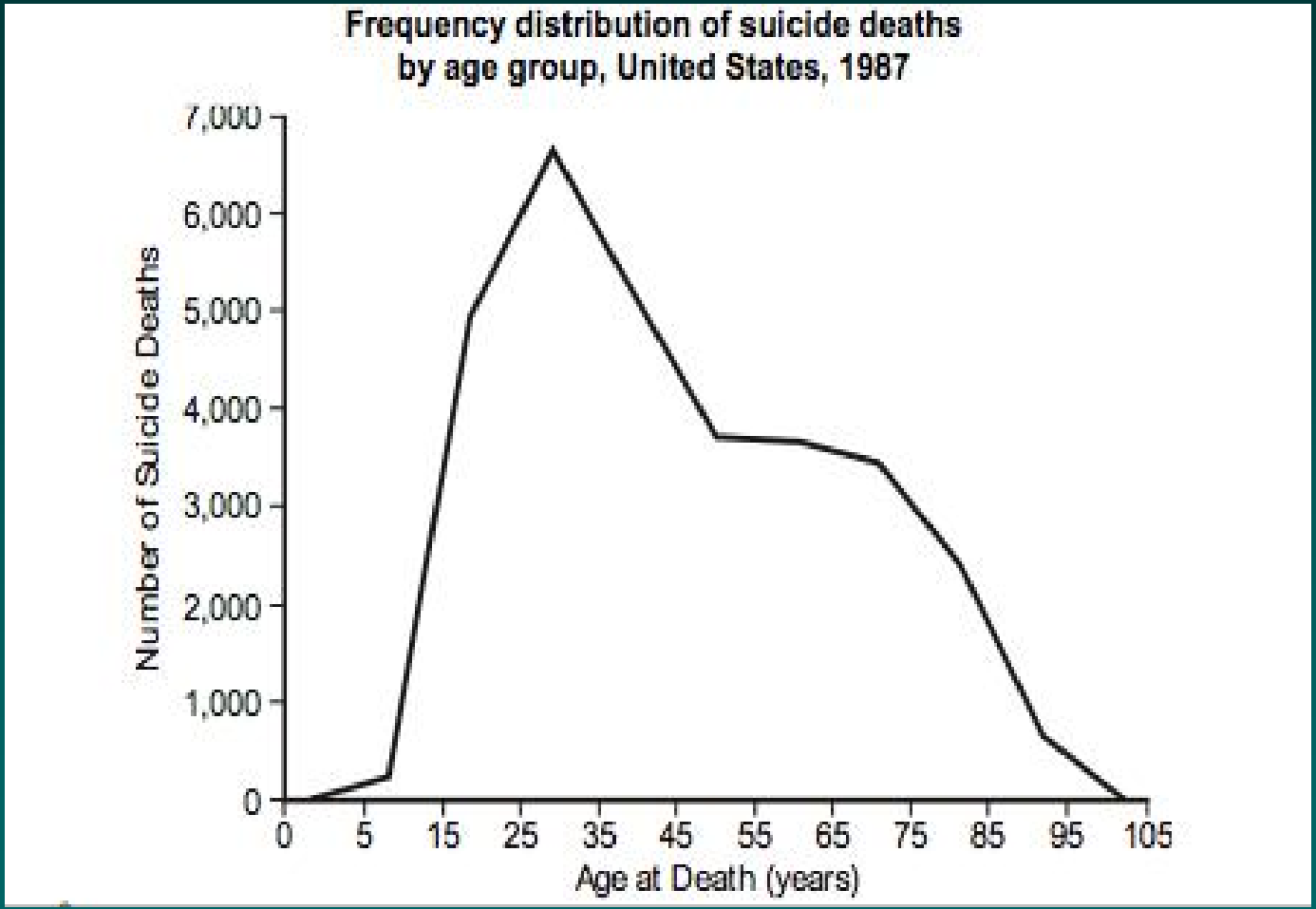
$$\text{Incidence rate} = \frac{\text{number of new cases during a given time period}}{\text{total person-time of observation}}$$

People enter study at various times and may default, move away, develop the disease or die

The incidence rate accounts for varying times of follow up for each individual

Important to specify time units: number of cases or events per person-day, person-month or person-year

Frequency distributions can also be shown as graphs



Summary

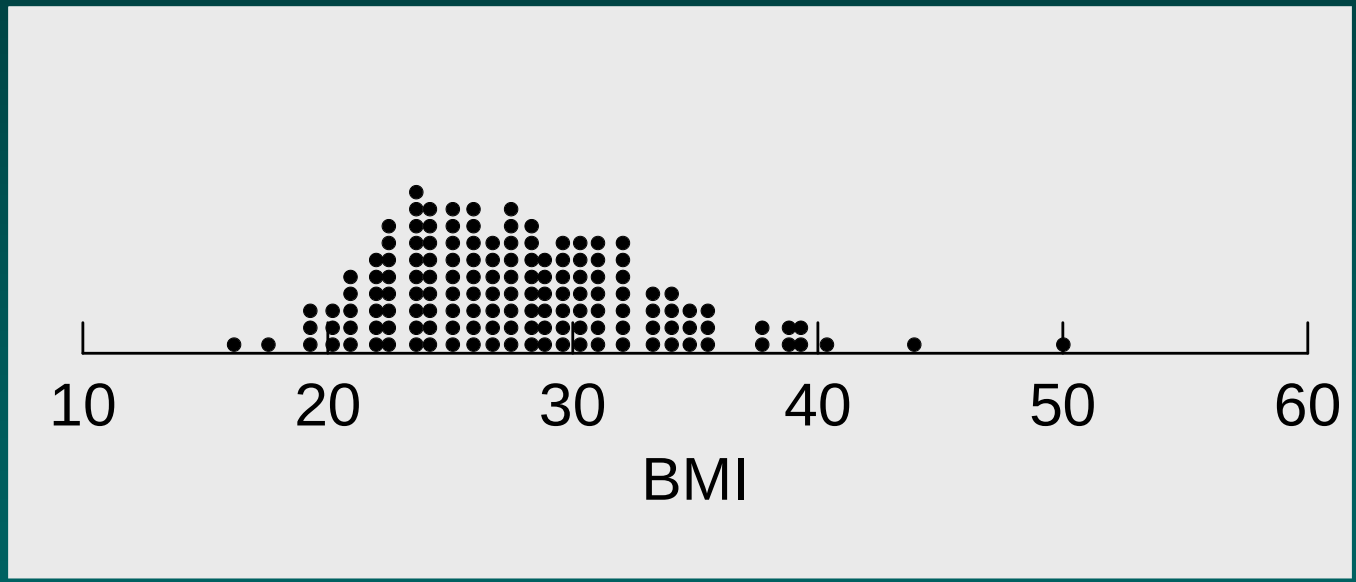
- To make useful statements about disease or risk factors, we need to count
- Basic information we need include: a count of affected individuals, the size of the source population and a time element
- We can calculate several rates and ratios using these pieces of information
- Where there are several categories under consideration, a frequency distribution is a useful way of summarizing information

Measures of central tendency and Measures of dispersion

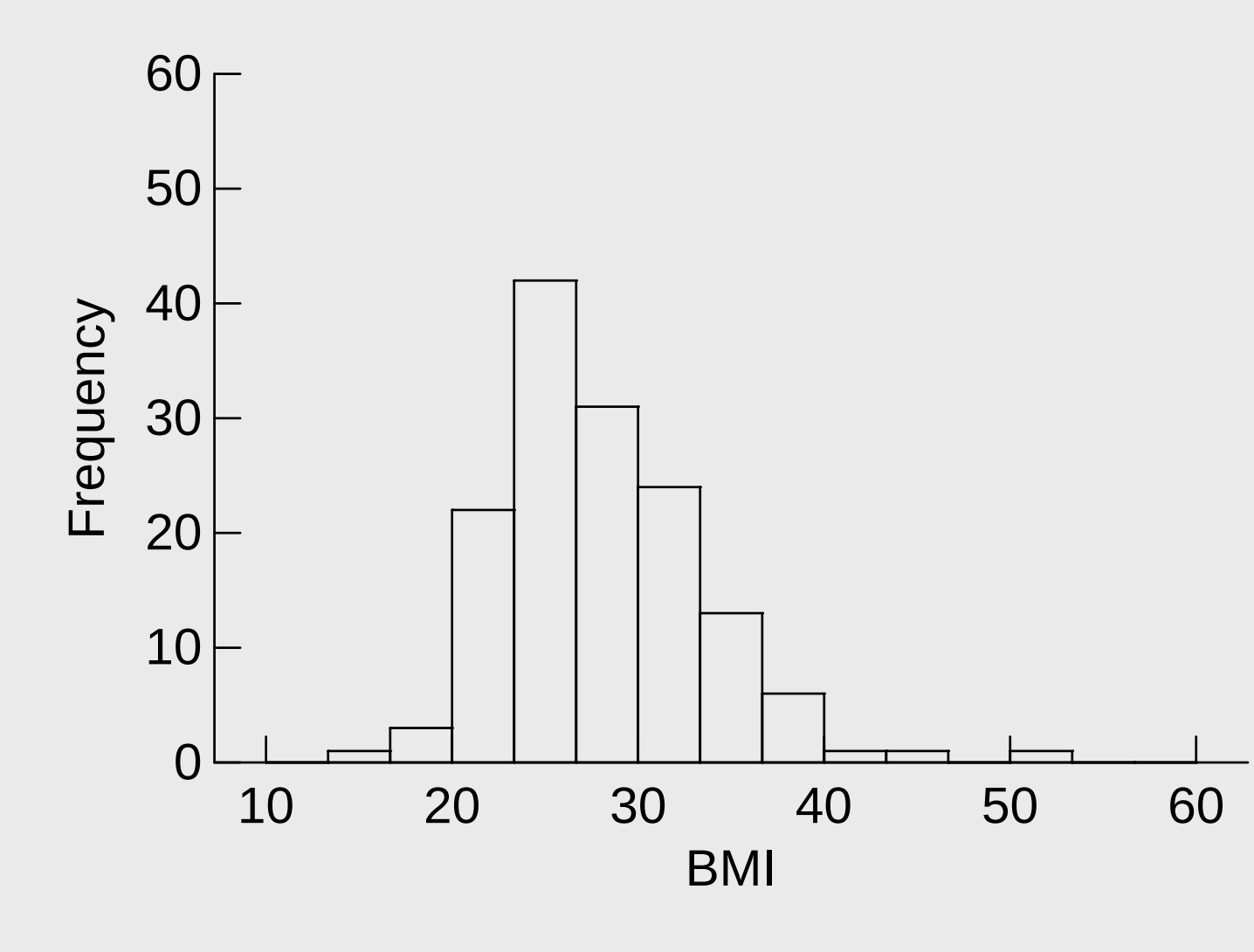
Introduction

- We often measure things on a continuous scale
- Examples: weight, height, blood pressure, cholesterol
- When we measure these on a group of people, we need to summarize them in a meaningful way

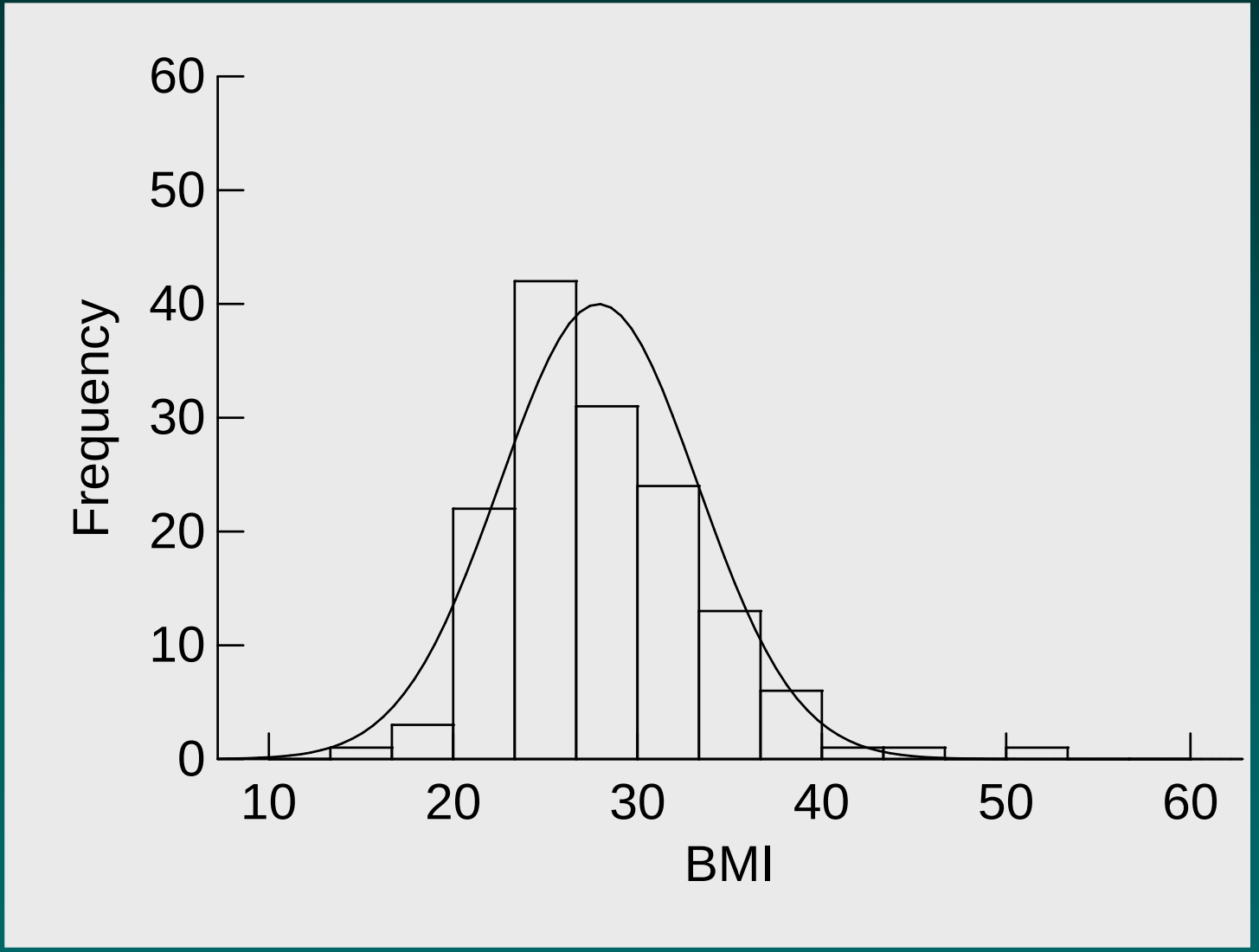
Plot of BMI values



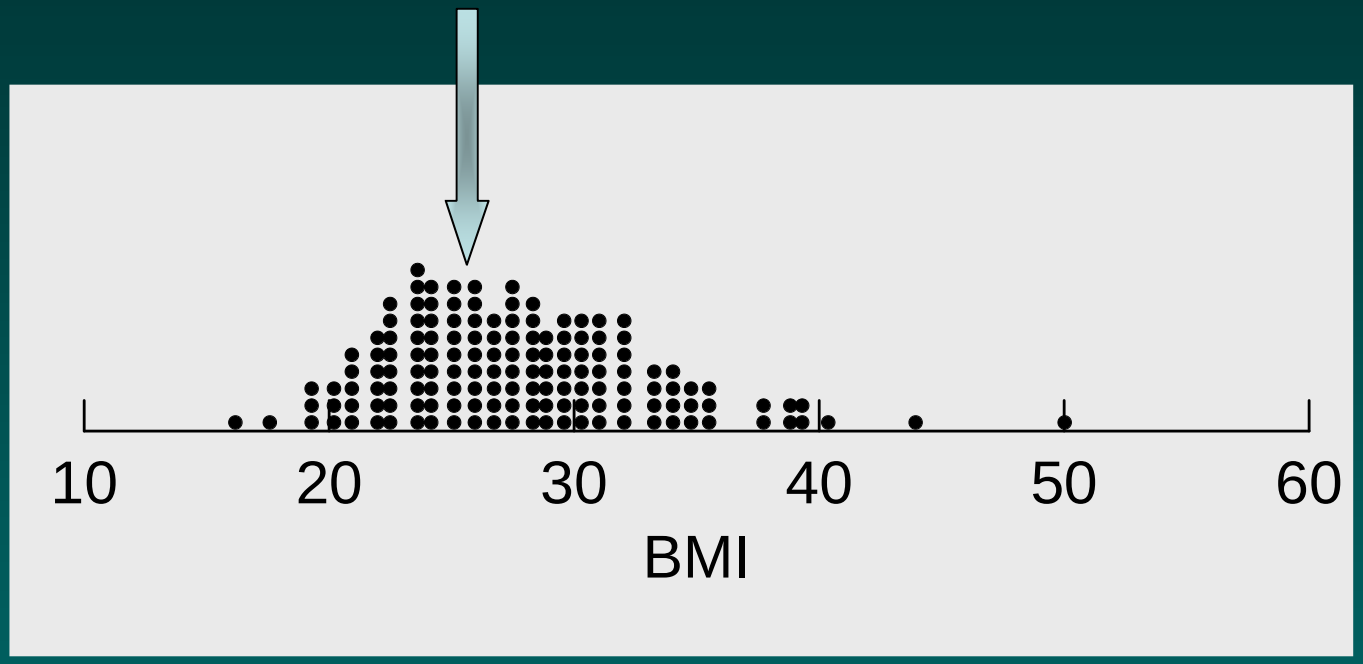
Histogram of BMI values



Frequency distribution curve BMI values



Plot of BMI values



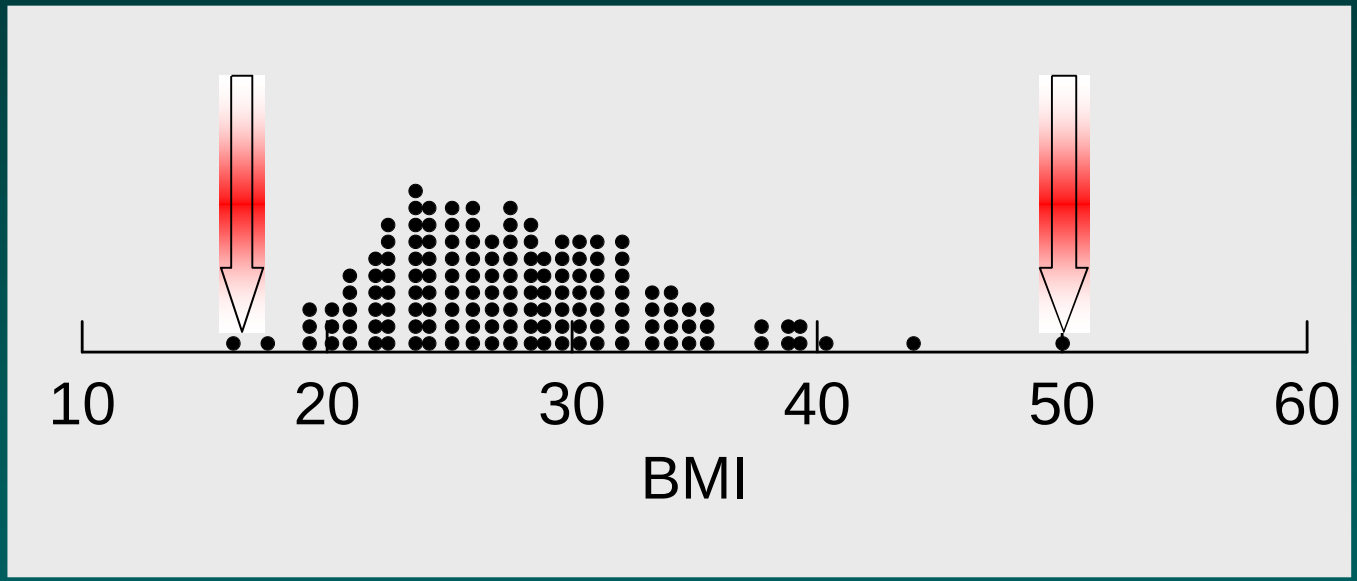
A large part of the observations clustered around a central value

This clustering is known as the central location or central tendency of a frequency distribution

Measures of central location

- The value that a distribution centers around is an important characteristic of the distribution. Once it is known, it can be used to characterize all of the data in the distribution.
- We can calculate a central value by several methods, and each method produces a somewhat different value.
- The central values that result from the various methods are known collectively as measures of central location.
- Of the possible measures of central location, we commonly use three in epidemiologic investigations: the arithmetic mean, the median, and the mode.

Plot of BMI values



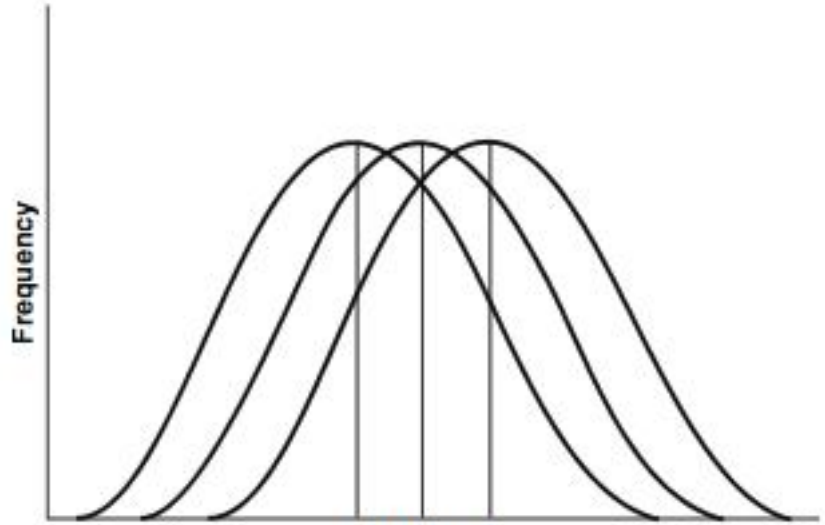
A second property of frequency distributions is variation or dispersion

This is the spread of a distribution out from its central value.

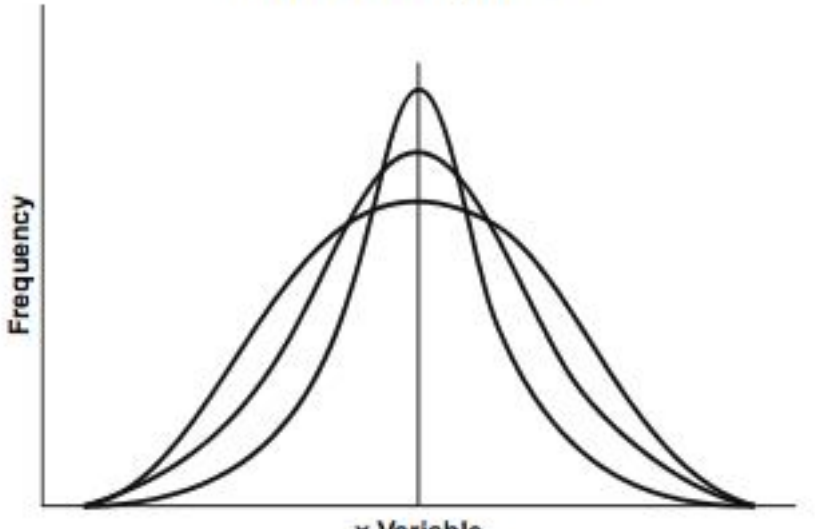
Measures of dispersion

- The variation or dispersion is the spread of a distribution out from its central value.
- Some of the measures of dispersion that we use in epidemiology are the range, variance, and the standard deviation.
- The dispersion of a frequency distribution is independent of its central location.

Three curves identical in shape with different central locations

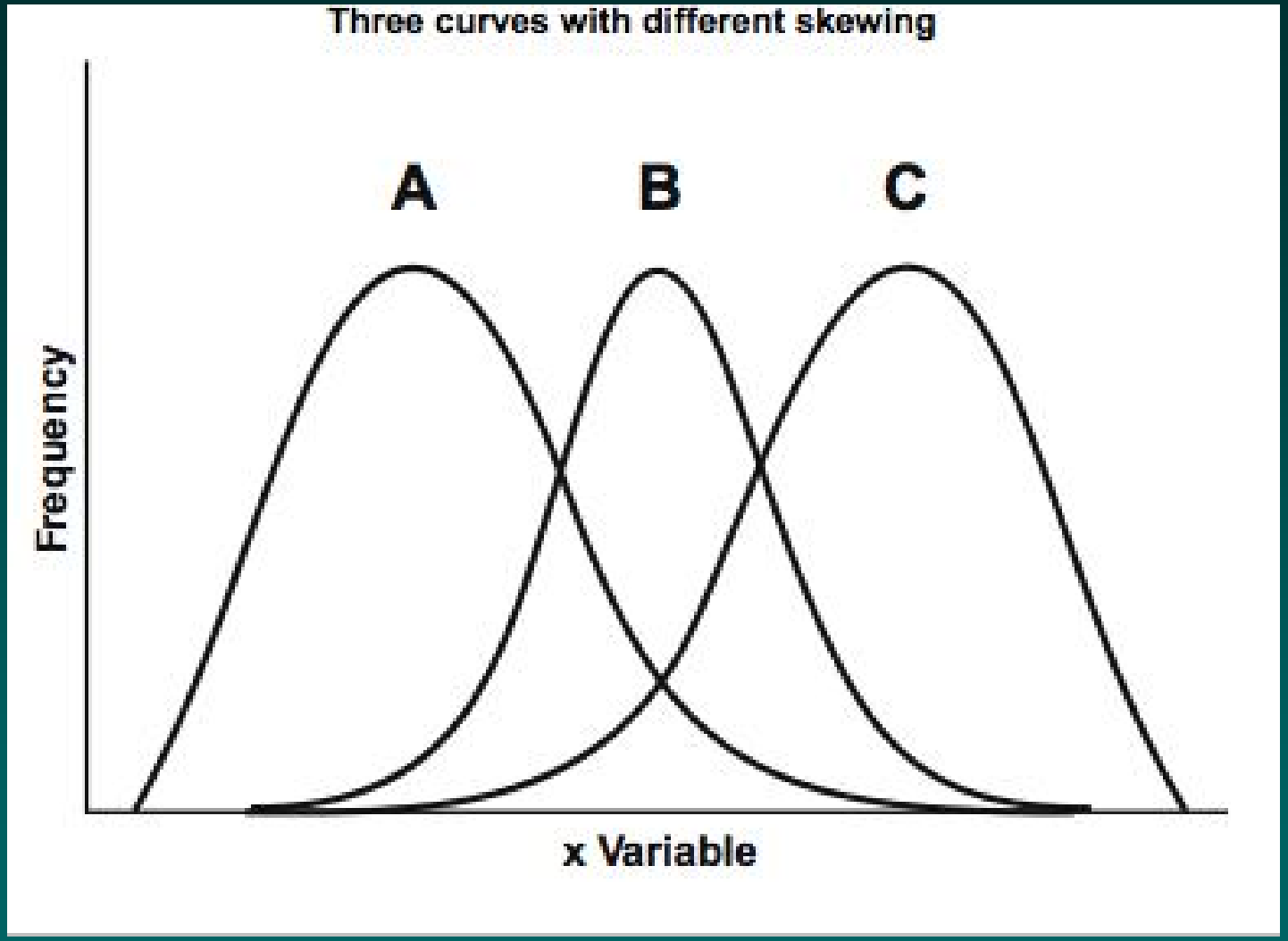


Three curves with same central location
but different dispersion



The shape of a distribution

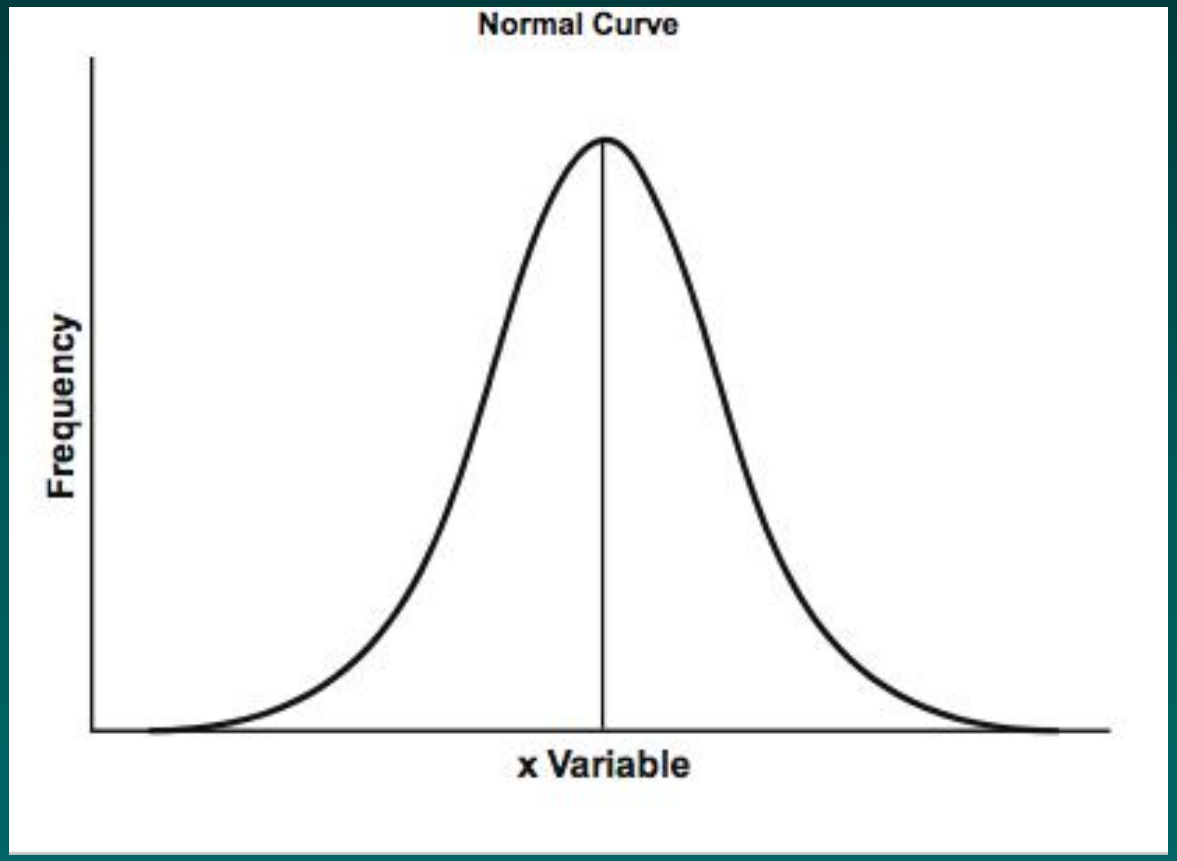
- A third property of a frequency distribution is its shape.
- Frequency distributions of some characteristics of human populations tend to be symmetrical.
- A distribution that is asymmetrical is said to be skewed.
- A distribution that has the central location to the left and a tail off to the right is said to be "positively skewed" or "skewed to the right."
- A distribution that has the central location to the right and a tail off to the left is said to be "negatively skewed" or "skewed to the left."


Three curves with different skewing
A
B
C
Frequency
x Variable

The Normal distribution

- The symmetrical clustering of values around a central location that is typical of many frequency distributions is called the normal distribution.
- The bell-shaped curve that results when a normal distribution is graphed is called the normal curve.
- This common bell-shaped distribution is the basis of many of the tests of inference that we use to draw conclusions or make generalizations from data.
- To use these tests, our data should be normally distributed, that is, should show a normal curve if graphed.

The Normal curve



The Arithmetic Mean

- The arithmetic mean is the arithmetic average and is commonly called simply “mean” or “average.”

Example:
The BMI of 6 market women was 29, 31, 24, 29, 30, and 25 kg/m^2, respectively

- To calculate the numerator, sum the individual observations: 29 + 31 + 24 + 29 + 30 + 25 = 168
- For the denominator, count the number of observations: n = 6
- To calculate the mean, divide the numerator (sum of observations) by the denominator (number of observations): 168/6 = 28.0

Therefore, the mean BMI was 28.0 kg/m^2

The Arithmetic Mean

- We use the arithmetic mean more than any other measure of central location because it has many desirable statistical properties.

- One such property is the centering property of the mean. if we subtract the mean from each observation, the sum of the differences is 0.

- Although the mean is often an excellent summary measure of a set of data, the data must be approximately normally distributed, because the mean is quite sensitive to extreme values that skew a distribution.

The centering property of the mean

Value minus Mean	Difference
24 – 28.0	–4.0
25 – 28.0	–3.0
29 – 28.0	+1.0
29 – 28.0	+1.0
30 – 28.0	+2.0
31 – 28.0	+3.0
168 – 168.0 = 0	–7.0 + 7.0 = 0

The Arithmetic Mean

•For example, in the last example, if the largest value of the six values were 131 instead of 31, the mean would change from 28.0 to 44.7

29, 31*, 24, 29, 30, 25

29, 131*, 24, 29, 30, 25

• As a result of a single extremely large value, the mean is much larger than all values in the distribution except that extreme value.

•Because the mean is so sensitive to extreme values, it is a poor summary measure for data that are severely skewed in either direction.

The Median

Another common measure of central location is the median.

Median means middle, and the median is the middle of a set of data that has been put into rank order.

Specifically, it is the value that divides a set of data into two halves, with one half of the observations being larger than the median value, and one half smaller.

For example, suppose we had the following set of systolic blood pressures (in mm/Hg): 110, 120, 122, 130, 180 :
the median is 122

The Median

In contrast to the mean, the median is not influenced to the same extent by extreme values.

110, 120, 122, 130, 180*
110, 120, 122, 130, 1800*

The median is preferred over the mean as a measure of central location for data skewed in one direction or another, or for data with a few extremely large or extremely small values.

Summary: central tendency

Measures of central location are single values that summarize the observed values of a continuous variable.

The most common measure of central location is the arithmetic mean, what most people call the average.

The arithmetic mean is most useful when the data are normally distributed. It represents the center of gravity of a set of data.

Unfortunately, the arithmetic mean is quite sensitive to extreme values, that is, it is pulled in the direction of extreme values.

Summary: central tendency

The median represents the middle of the set, with half the observations below and half the observations above the median value.

When a set of data is skewed or has a few extreme values in one direction, the median is the preferred measure of central location.

The mode is simply the most common value. While every set of data has one and only one arithmetic mean and median, a set of data may have one mode, no mode, or multiple modes.

As a measure of central location, the mode is useful if we are interested in knowing which values are most popular.

Measures of dispersion

We use a measure of dispersion to describe how much spread there is in the distribution.

Several measures of dispersion are available.

Range, Minimum Values, and Maximum Values

The range of a set of data is the difference between its largest (maximum) and smallest (minimum) values.

In the statistical world, the range is reported as a single number, the difference between maximum and minimum.

In the epidemiologic community, the range is often reported as “from (the minimum) to (the maximum),” i.e., two numbers.

Range

Example

Find the minimum value, maximum value, and range of the following data: 29, 31, 24, 29, 30, 25 1.

Arrange the data from smallest to largest. 24, 25, 29, 29, 30, 31
Identify the minimum and maximum values: Minimum = 24, Maximum = 31

Calculate the range: Range = Maximum-Minimum = 31–24 = 7.

Thus the range is 7.

Percentiles, Quartiles and Interquartile range

We can think of the maximum value of a distribution as the value in a set of data that has 100% of the observations at or below it.

When we consider it in this way, we call it the 100 th percentile.

From this same perspective, the median, which has 50% of the observations at or below it, is the 50 th percentile.

The pth percentile of a distribution is the value such that p percent of the observations fall at or below it.

The most commonly used percentiles other than the median are the 25 th percentile and the 75 th percentile.

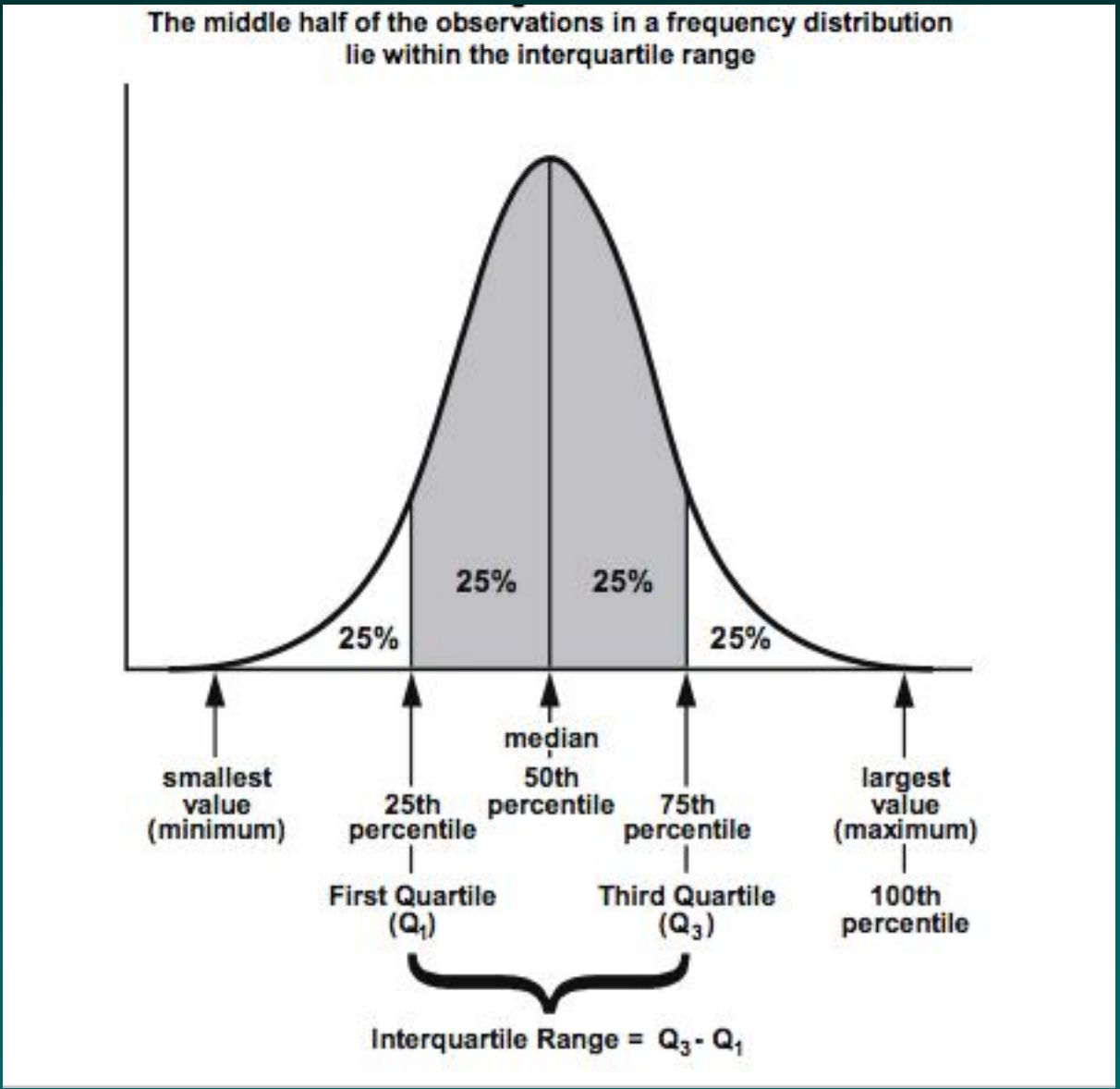
Percentiles, Quartiles and Interquartile range

The 25th percentile demarcates the first quartile, the median or 50th percentile demarcates the second quartile, the 75 th percentile demarcates the third quartile, and the 100th percentile demarcates the fourth quartile.

The interquartile range represents the central portion of the distribution, and is calculated as the difference between the third quartile and the first quartile.

This range includes about one-half of the observations in the set, leaving one-quarter of the observations on each side.

Percentiles, Quartiles and Interquartile range



Variance and standard deviation

If we subtract the mean from each observation, the sum of the differences is 0.

This is the basis of the variance and standard deviation.

For these measures we square each difference to eliminate negative numbers.

We then sum the squared differences and divide by n–1 to find an "average" squared difference.

This "average" is the variance.

We convert the variance back into the units we began with by taking its square root.

The square root of the variance is called the standard deviation.

Variance and standard deviation

Value minus Mean	Difference	Difference Squared
24 – 28.0	–4.0	16
25 – 28.0	–3.0	9
29 – 28.0	+1.0	1
29 – 28.0	+1.0	1
30 – 28.0	+2.0	4
31 – 28.0	+3.0	9
168 – 168.0 = 0	–7.0 + 7.0 = 0	40

$$\text{Variance} = \frac{\text{sum of squared differences}}{n-1} = 40/5 = 8$$

$$\text{Standard deviation} = \sqrt{8} = 2.83$$

Variance and standard deviation

The variance and standard deviation are measures of the deviation or dispersion of observations around the mean of a distribution.

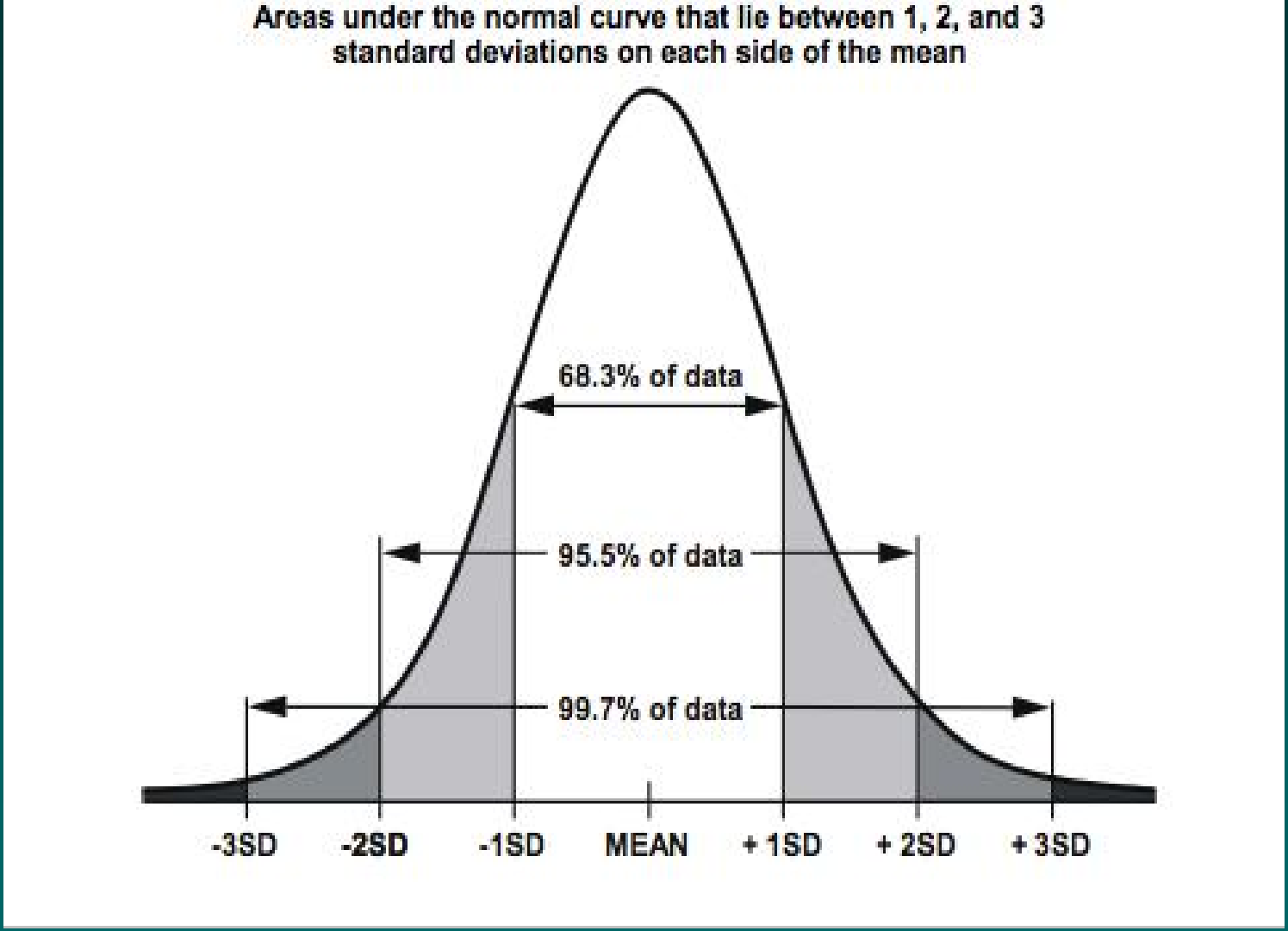
Variance is the mean of the squared differences of the observations from the mean.

To illustrate the relationships of the standard deviation and the mean to the normal curve, consider data which are normally distributed.

68.3% of the area under the normal curve lies between the mean and ± 1 standard deviation, that is, from 1 standard deviation below the mean to 1 standard deviation above the mean.

Also, 95.5% of the area lies between the mean and ± 2 standard deviations, and 99.7% of the area lies between the mean and ± 3 standard deviations. Further, 95% of the area lies between the mean and ± 1.96 standard deviations.

Means and standard deviation



Summary: measures of dispersion

In summary, measures of dispersion quantify the spread or variability of the observed values of a continuous variable.

The simplest measure of dispersion is the range from the smallest value to the largest value.

The range is obviously quite sensitive to extreme values in either or both directions.

For data which are normally distributed, the standard deviation is used in conjunction with the arithmetic mean.

The standard deviation reflects how closely clustered the observed values are to the mean.

Summary: measures of dispersion

For normally distributed data, the range from ‘minus one standard deviation’ to ‘plus one standard deviation’ represents the middle 68.3% of the data. About 95% of the data fall in the range from −1.96 standard deviations to +1.96 standard deviations.

For data which are skewed, the interquartile range is used in conjunction with the median.

The interquartile range represents the range from the 25 th percentile (the first quartile) to the 75th percentile (the third quartile), or roughly the middle 50% of the data.