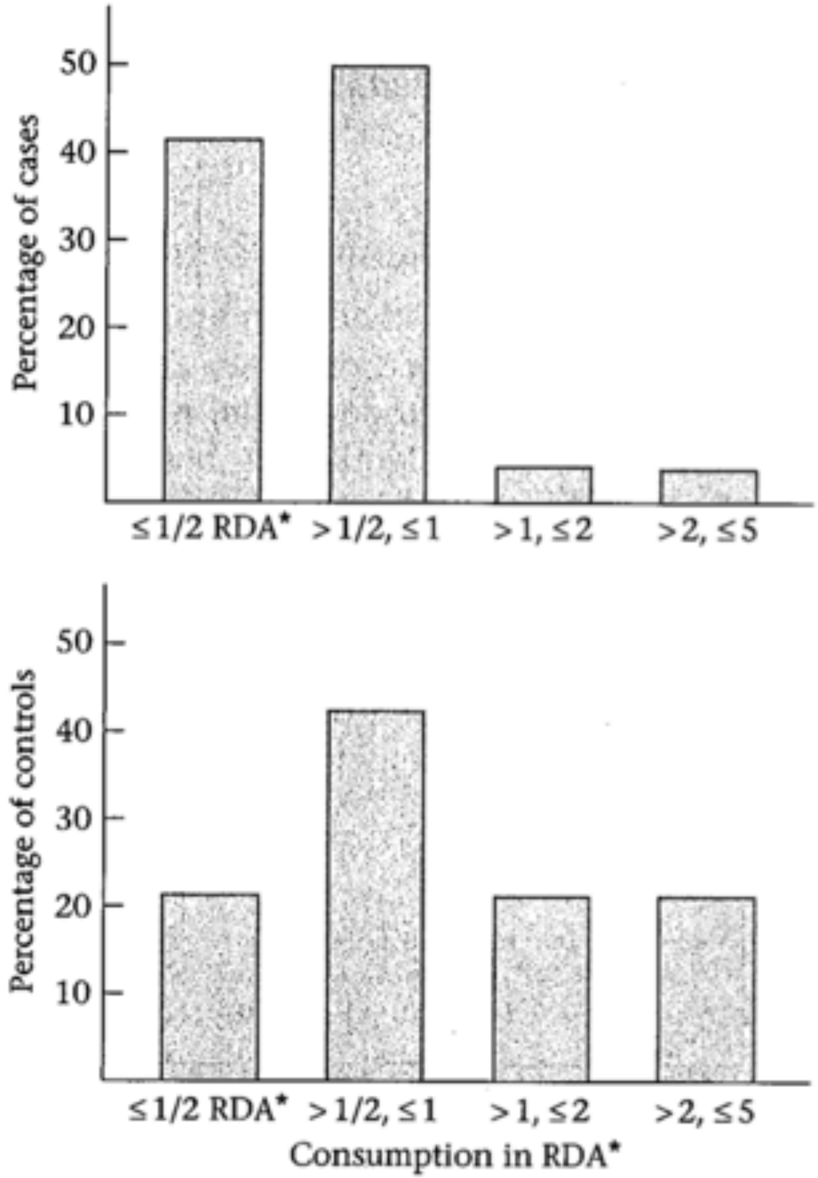
Descriptive Statistics

Descriptive Statistics

- Data we are facing today
  - Scale of data has become larger and larger
  - Dimensionality increased
- Descriptive statistics is the first step statisticians do with their data
  - Understand your data
  - Draw some hypothesis

Daily vitamin-A consumption among cancer cases and controls



*RDA = Recommended Daily Allowance.

3

Common statistical terms

- Data
  - Variables
    - A characteristic that is observed or manipulated
    - Can take on different values
  - Values

Statistical terms (cont.)

- Independent variables
  - Precede dependent variables in time
  - Are often manipulated by the researcher
  - The treatment or intervention that is used in a study
- Dependent variables
  - What is measured as an outcome in a study
  - Values depend on the independent variable

Statistical terms (cont.)

- Parameters
  - Summary data from a population
- Statistics
  - Summary data from a sample

6

Populations

- A population is the group from which a sample is drawn
  - e.g., headache patients in a doctor's office; automobile crash victims in an emergency room
- In research, it is not practical to include all members of a population
- Thus, a sample (a subset of a population) is taken

Random samples

- Subjects are selected from a population so that each individual has an equal chance of being selected
- Random samples are representative of the source population
- Non-random samples are not representative
  – May be biased regarding age, severity of the condition, socioeconomic status etc.

8

Descriptive statistics (DSs)

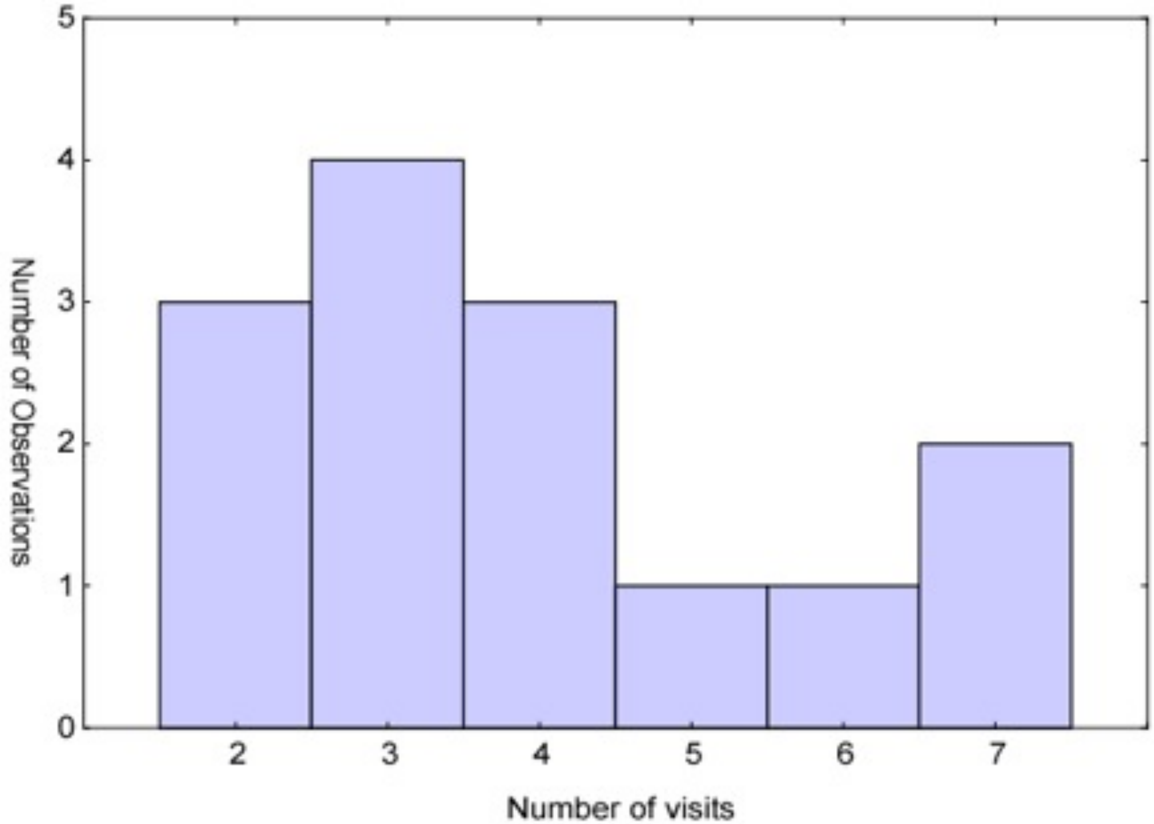
- A way to summarize data from a sample or a population
- DSs illustrate the *shape*, *central tendency*, and *variability* of a set of data
  - The shape of data has to do with the frequencies of the values of observations
  - Central tendency describes the location of the middle of the data
  - Variability is the extent values are spread above and below the middle values
    - a.k.a., Dispersion

Hypothetical study data

- Distribution provides a summary of:
  - Frequencies of each of the values
    - 2 – **3**
    - 3 – **4**
    - 4 – **3**
    - 5 – **1**
    - 6 – **1**
    - 7 – **2**
  - Ranges of values
    - Lowest = **2**
    - Highest = **7**

Case #	Visits
1	7
2	2
3	2
4	3
5	4
6	3
7	5
8	3
9	4
10	6
11	2
12	3
13	7
14	4

Frequency distributions are often depicted by a histogram



11

Measures of central tendency

- Mean (a.k.a., *average*)
  - The most commonly used DS
- To calculate the mean
  - Add all values of a series of numbers and then divided by the total number of elements

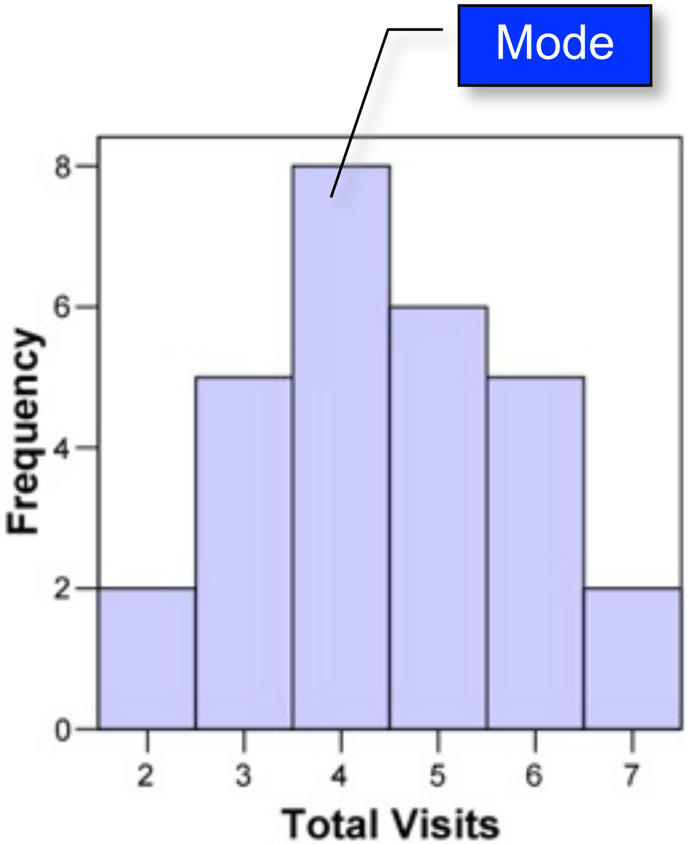
Formula to calculate the mean

- Mean of a sample $$\overline{X} = \frac{\Sigma X}{n}$$
- Mean of a population $$\mu = \frac{\Sigma X}{N}$$

13

Measures of central tendency (cont.)

- Mode
  - The most frequently occurring value in a series
  - The modal value is the highest bar in a histogram



14

Measures of central tendency (cont.)

- Median
  - The value that divides a series of values in half when they are all listed in order
  - When there are an odd number of values
    - The median is the middle value
  - When there are an even number of values
    - Count from each end of the series toward the middle and then average the 2 middle values

Measures of central tendency (cont.)

- Each of the three methods of measuring central tendency has certain advantages and disadvantages
- Which method should be used?
  - It depends on the type of data that is being analyzed
  - e.g., categorical, continuous, and the level of measurement that is involved

Levels of measurement

- There are 4 levels of measurement
  - Nominal, ordinal, interval, and ratio

1. Nominal
   - Data are coded by a number, name, or letter that is assigned to a category or group
   - Examples
     - Gender (e.g., male, female)
     - Treatment preference (e.g., manipulation, mobilization, massage)

Levels of measurement (cont.)

2. Ordinal
   - Is similar to nominal because the measurements involve categories
   - However, the categories are ordered by rank
   - Examples
     - Pain level (e.g., mild, moderate, severe)
     - Military rank (e.g., lieutenant, captain, major, colonel, general)

Levels of measurement (cont.)

- Ordinal values only describe order, not quantity
  - Thus, severe pain is not the same as 2 times mild pain
- The only mathematical operations allowed for nominal and ordinal data are counting of categories
  - e.g., 25 males and 30 females

19

Levels of measurement (cont.)

3. Interval
   – Measurements are ordered (like ordinal data)
   – Have equal intervals
   – Does not have a true zero
   – Examples
     • The Fahrenheit scale, where 0° does not correspond to an absence of heat (no true zero)
     • In contrast to Kelvin, which does have a true zero

20

Levels of measurement (cont.)

4. Ratio
   - Measurements have equal intervals
   - There is a true zero
   - Ratio is the most advanced level of measurement, which can handle most types of mathematical operations

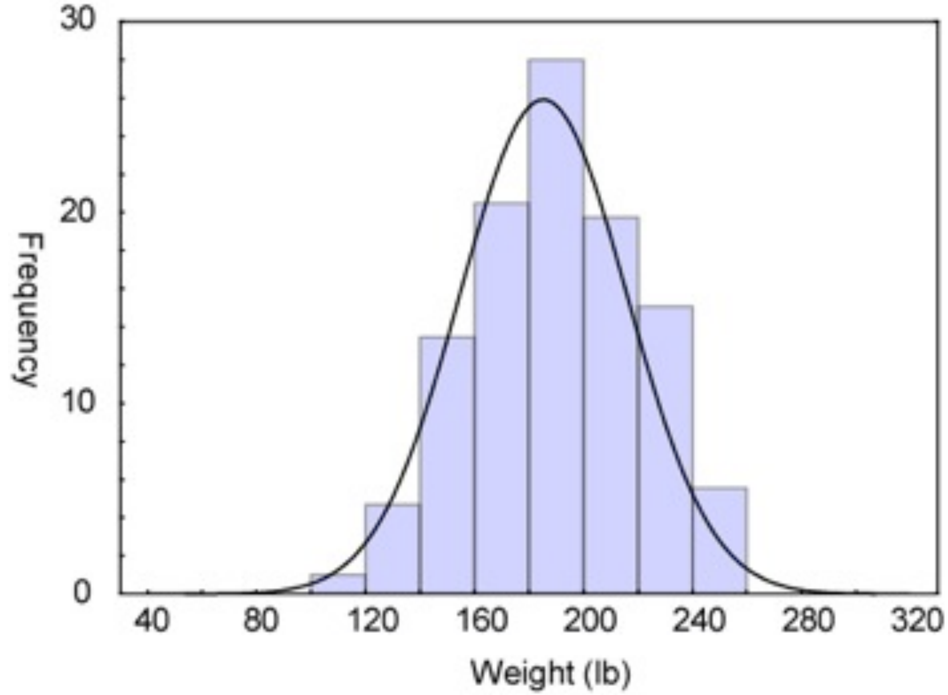
Levels of measurement (cont.)

Measurement scale	Permissible mathematic operations	Best measure of central tendency
Nominal	Counting	Mode
Ordinal	Greater or less than operations	Median
Interval	Addition and subtraction	Symmetrical – Mean Skewed – Median
Ratio	Addition, subtraction, multiplication and division	Symmetrical – Mean Skewed – Median

The shape of data

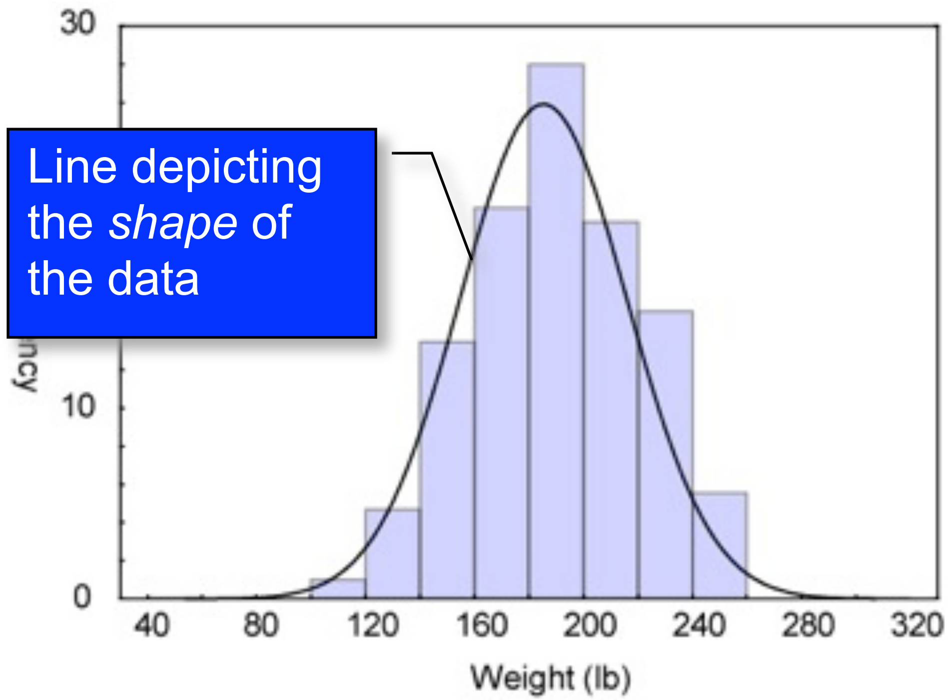
- Histograms of frequency distributions have shape
- Distributions are often symmetrical with most scores falling in the middle and fewer toward the extremes
- Many biological data are symmetrically distributed and form a *normal curve* (a.k.a, bell-shaped curve)

The shape of data (cont.)



24

The shape of data (cont.)

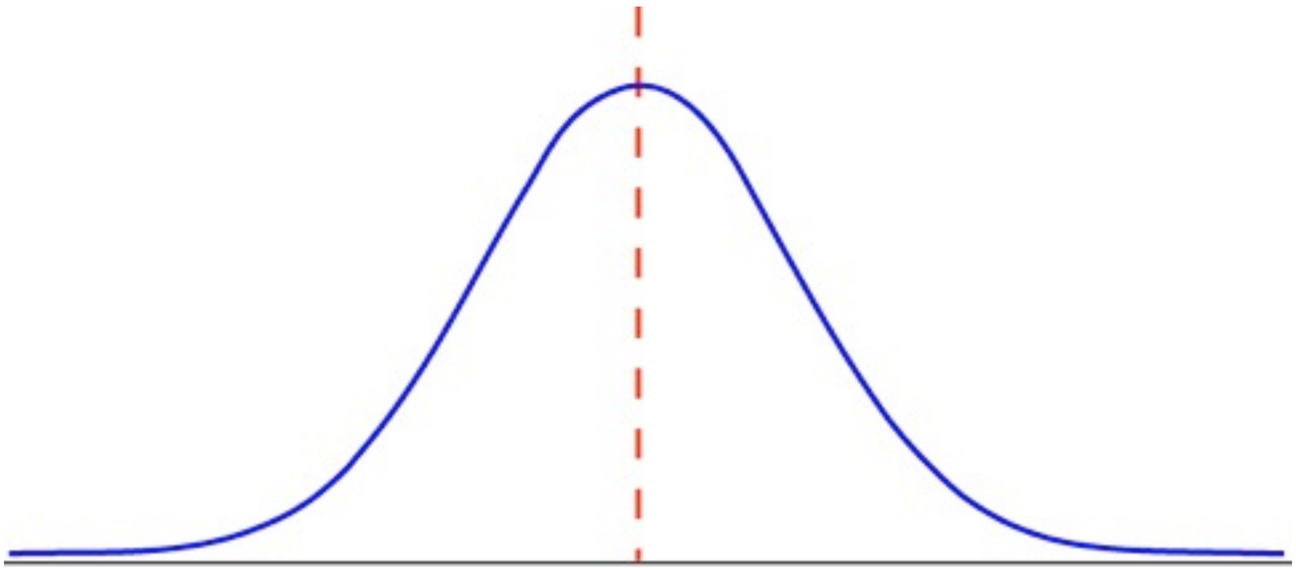


24

The normal distribution

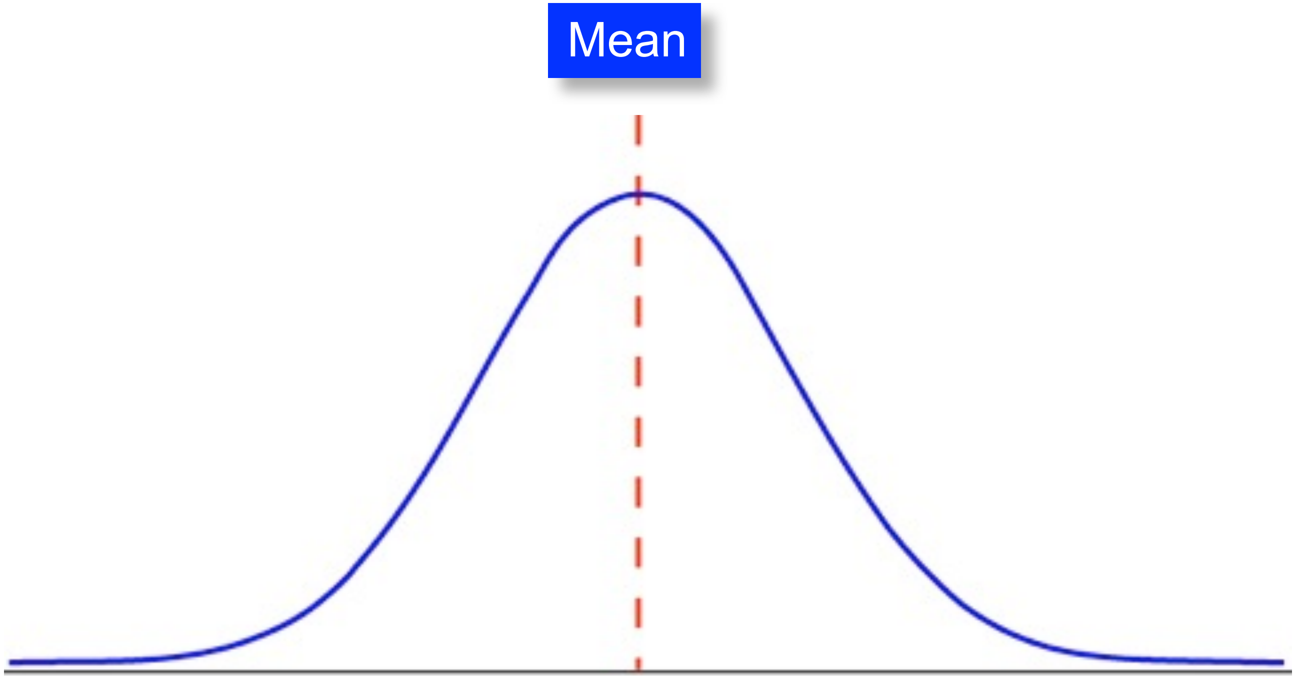
- The area under a normal curve has a *normal distribution* (a.k.a., Gaussian distribution)
- Properties of a normal distribution
  - It is symmetric about its mean
  - The highest point is at its mean
  - The height of the curve decreases as one moves away from the mean in either direction, approaching, but never reaching zero

The normal distribution (cont.)

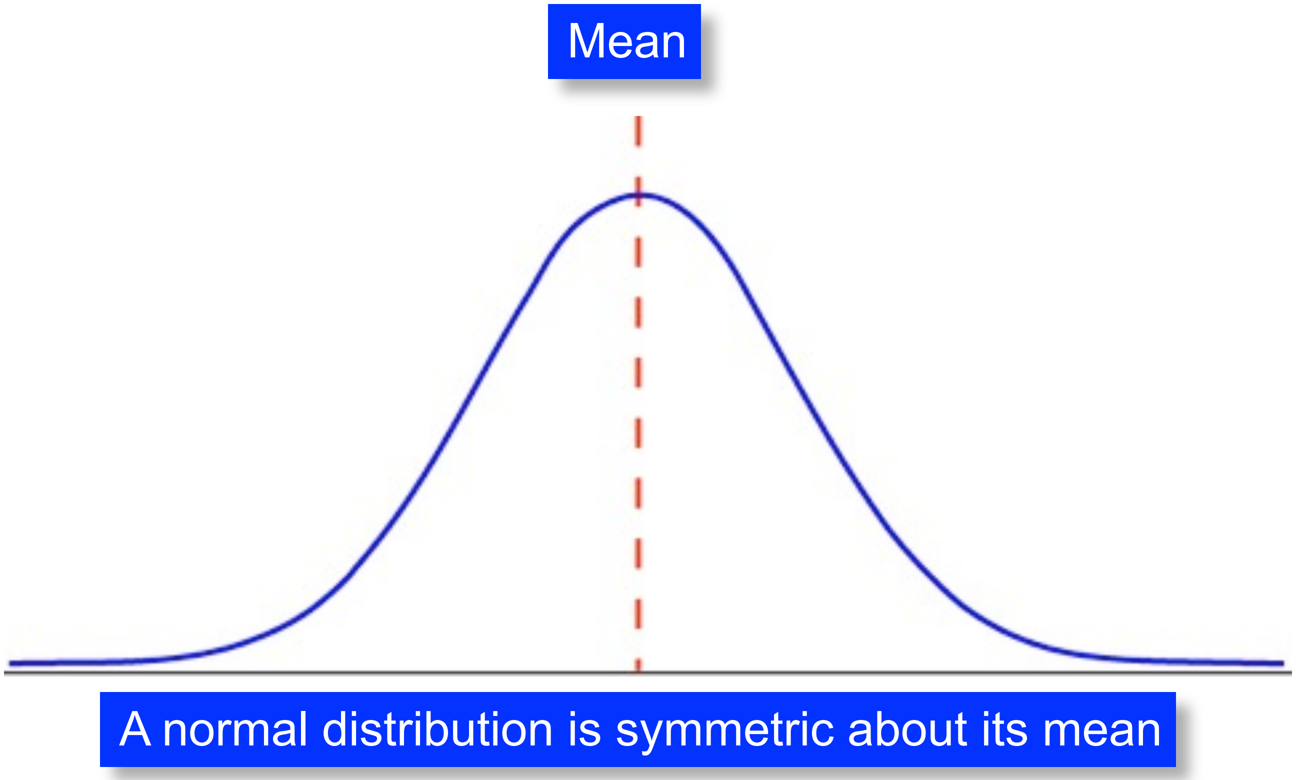


26

The normal distribution (cont.)



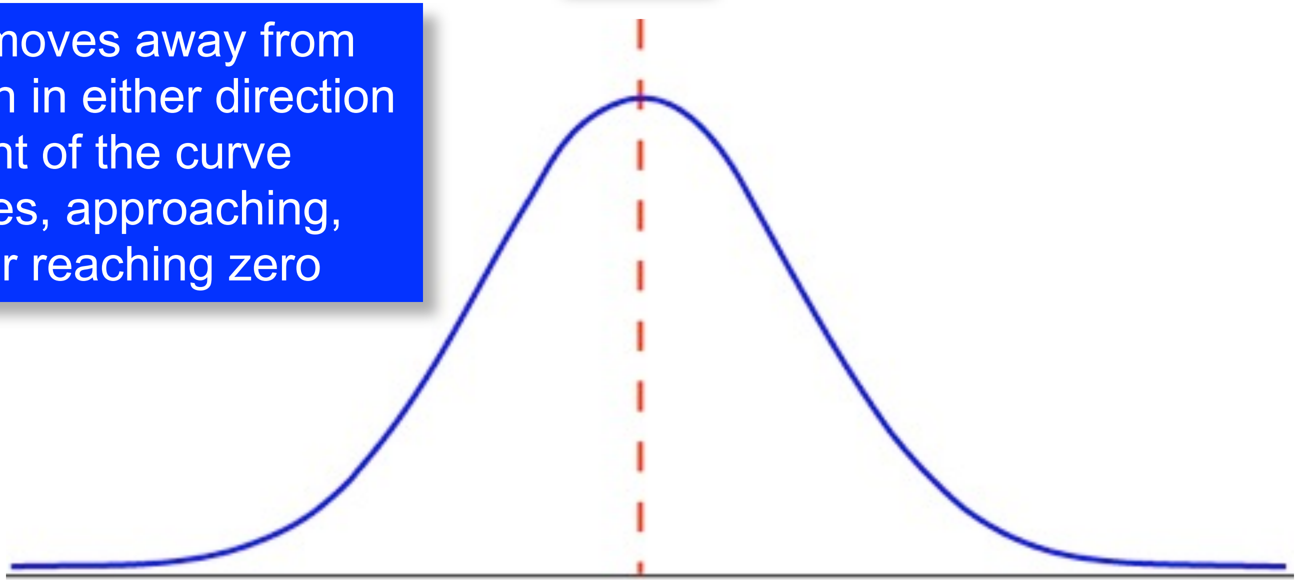
The normal distribution (cont.)



The normal distribution (cont.)

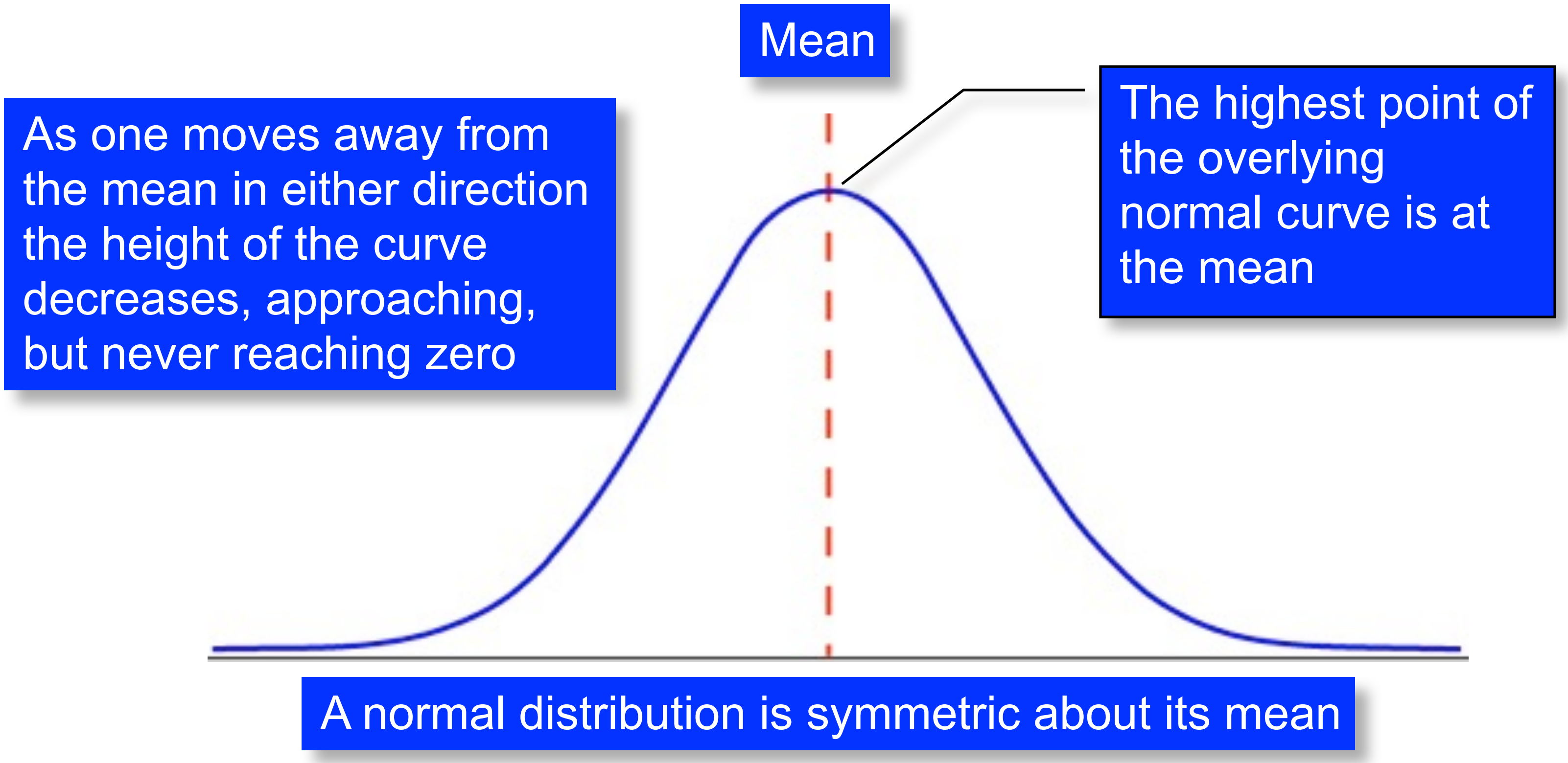
Mean

As one moves away from the mean in either direction the height of the curve decreases, approaching, but never reaching zero

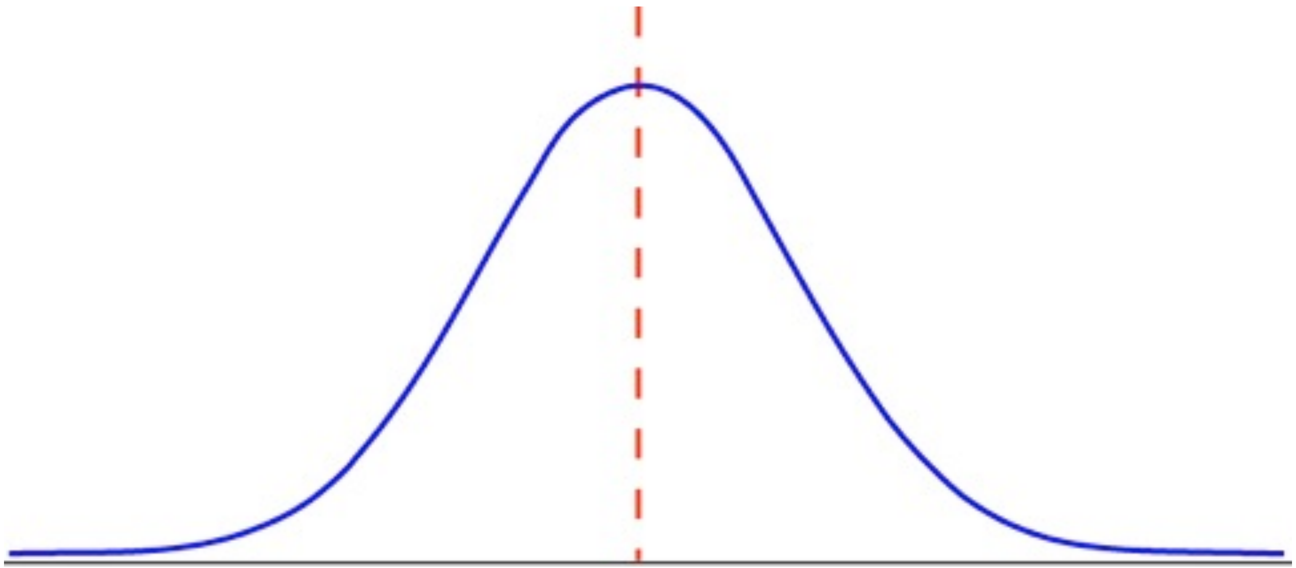


A normal distribution is symmetric about its mean

The normal distribution (cont.)

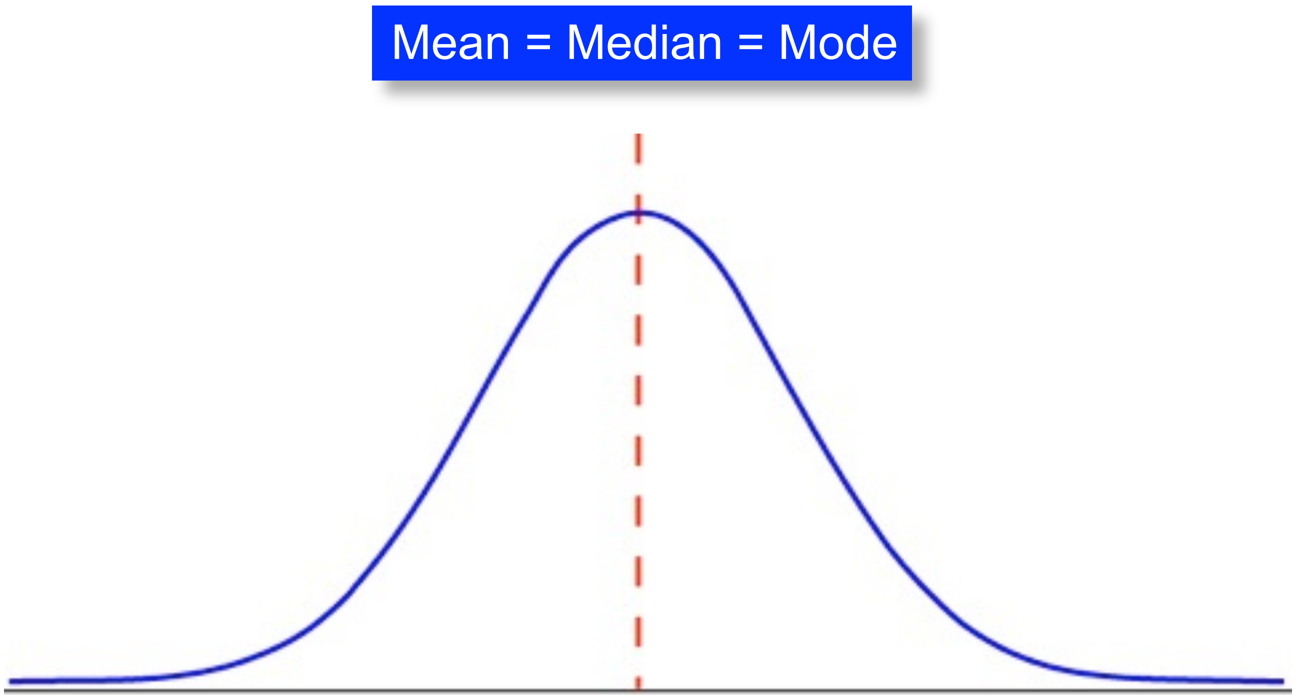


The normal distribution (cont.)



27

The normal distribution (cont.)



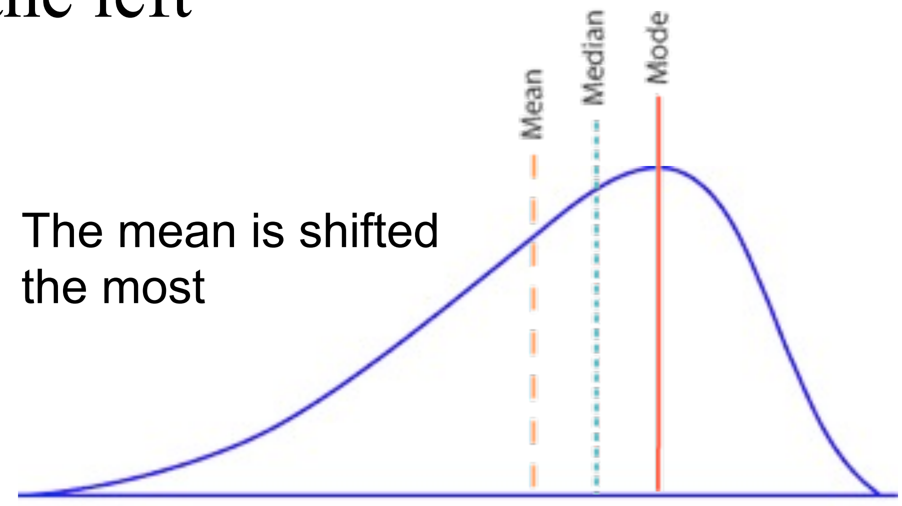
27

Skewed distributions

- The data are not distributed symmetrically in skewed distributions
  – Consequently, the mean, median, and mode are not equal and are in different positions
  – Values are clustered at one end of the distribution
  – A small number of extreme values are located in the limits of the opposite end

Skewed distributions (cont.)

- Skew is always toward the direction of the longer tail
  - Positive if skewed to the right
  - Negative if to the left



Negatively skewed distribution

29

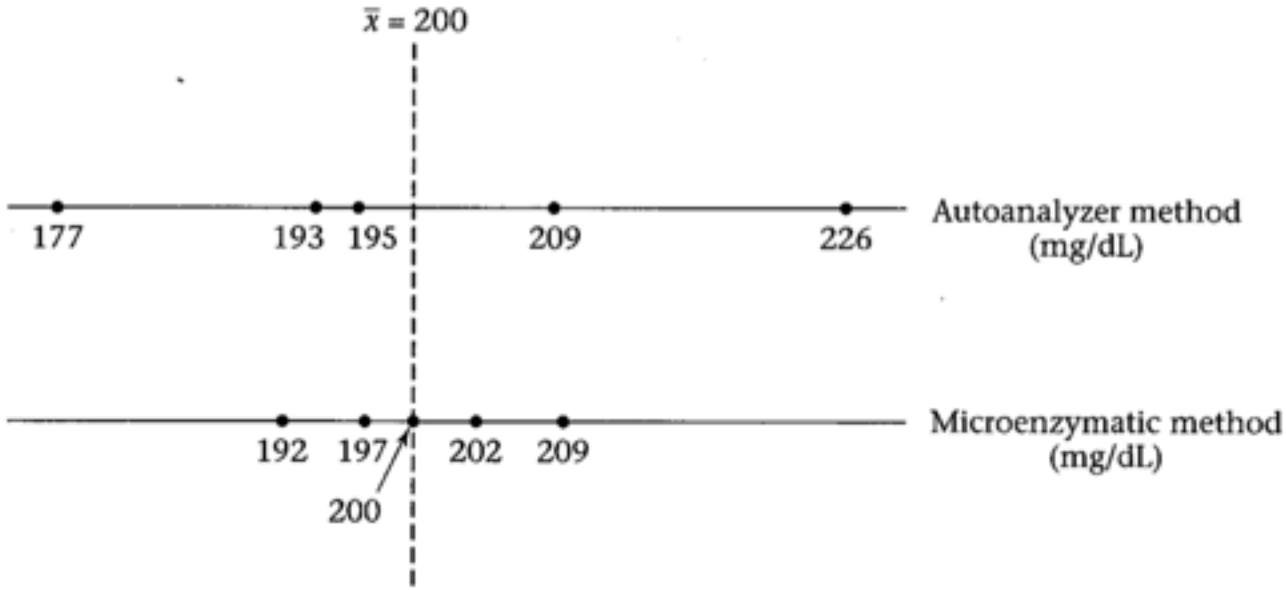
Skewed distributions (cont.)

- Because the mean is shifted so much, it is not the best estimate of the average score for skewed distributions
- The median is a better estimate of the center of skewed distributions
  - It will be the central point of any distribution
  - 50% of the values are above and 50% below the median

Measures of Spread

- Cholesterol measurement
  - Variability or spread of the data

Figure 2.4 **Two samples of cholesterol measurements on a given person using an Autoanalyzer and a Microenzymatic measurement technique**



31

Two-number Summary – Variance and Standard Deviation

- Range
  - the difference between the largest and smallest observations in a sample.
- Variance $$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$
- Standard deviation $$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$
- SD is the average amount of spread in a distribution of scores

32

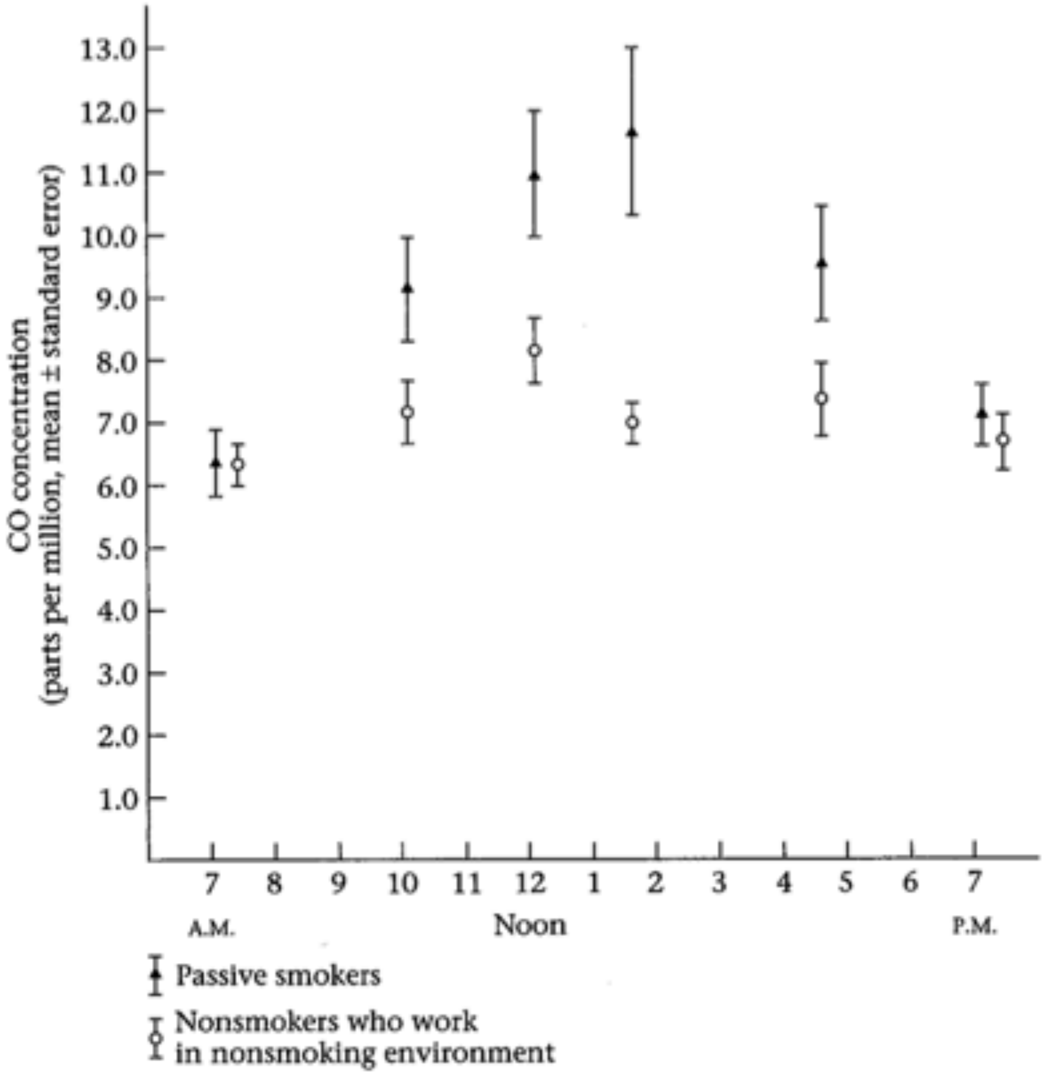
- Example: Pulmonary Disease

The relationship between passive smoking and pulmonary function.

As supporting evidence, the CO concentration in working environments of passive smokers and of nonsmokers.

  – If the relative CO concentration changed over the course of the day.

Figure 2.2 **Mean carbon-monoxide concentration (± standard error) by time of day as measured in the working environment of passive smokers and nonsmokers who work in nonsmoking environments**



Source: Reproduced with permission of *The New England Journal of Medicine*, *302*, 720–723, 1980.

Case Study: The Scottish Heart Health Study (SHHS)

- Scotland's annual mortality rate from coronary heart disease (CHD) is one of the highest in the world.
- Establishment of Cardiovascular Epidemiology Unit at University of Dundee
- Objectives of SHHS:
  - To establish the levels of CHD risk factors in a cross-sectional sample of Scottish men and women aged 40-59.
  - To determine the extent to which the geographical variation in CHD can be explained in terms of the geographical variation in risk factor levels.
  - To assess the relative contribution of the established risk factors, as well as some more recently described ones.

Case Study: The Scottish Heart Health Study (SHHS) cont.

- Subjects sampled from 22 of the 56 mainland Scottish local government districts.
- From each district, an equal number of people were selected in the four age/sex groups: male 40-49, female 40-49, male 50-59, and female 50-59.
- A questionnaire and an invitation to a local clinic.
  - Questionnaire include questions like age, sex, marital status, employment, past medical history, exercise, diet and smoking…
- In clinic visit, height, weight, blood pressure were recorded and a 12-lead electrocardiogram was administered. A blood sample was taken from which serum total cholesterol, fibrinogen, and several other biochemical variables were measured.
- The sample size is 10,359 (5123 men and 5236 women).
- The data set comprised 315 variables totally.

Case Study: The Scottish Heart Health Study (SHHS) cont.

- Prevalence data are not ideal to demonstrate causality.
- SHHS was designed as a two-phase study.
  - The cross-sectional baseline study.
  - Follow-up cohort study of several years' duration.
- The follow-up study
  - Death registration certificates collected
  - Hospital records
  - 8 years

Types of Variables

- Qualitative variables
  - Categorical variables
    - Binary variables
  - Ordinal variables (ordered categorical variables)
    - Responses like poor/satisfactory/good
- Quantitative variables
  - Discrete
  - Continuous

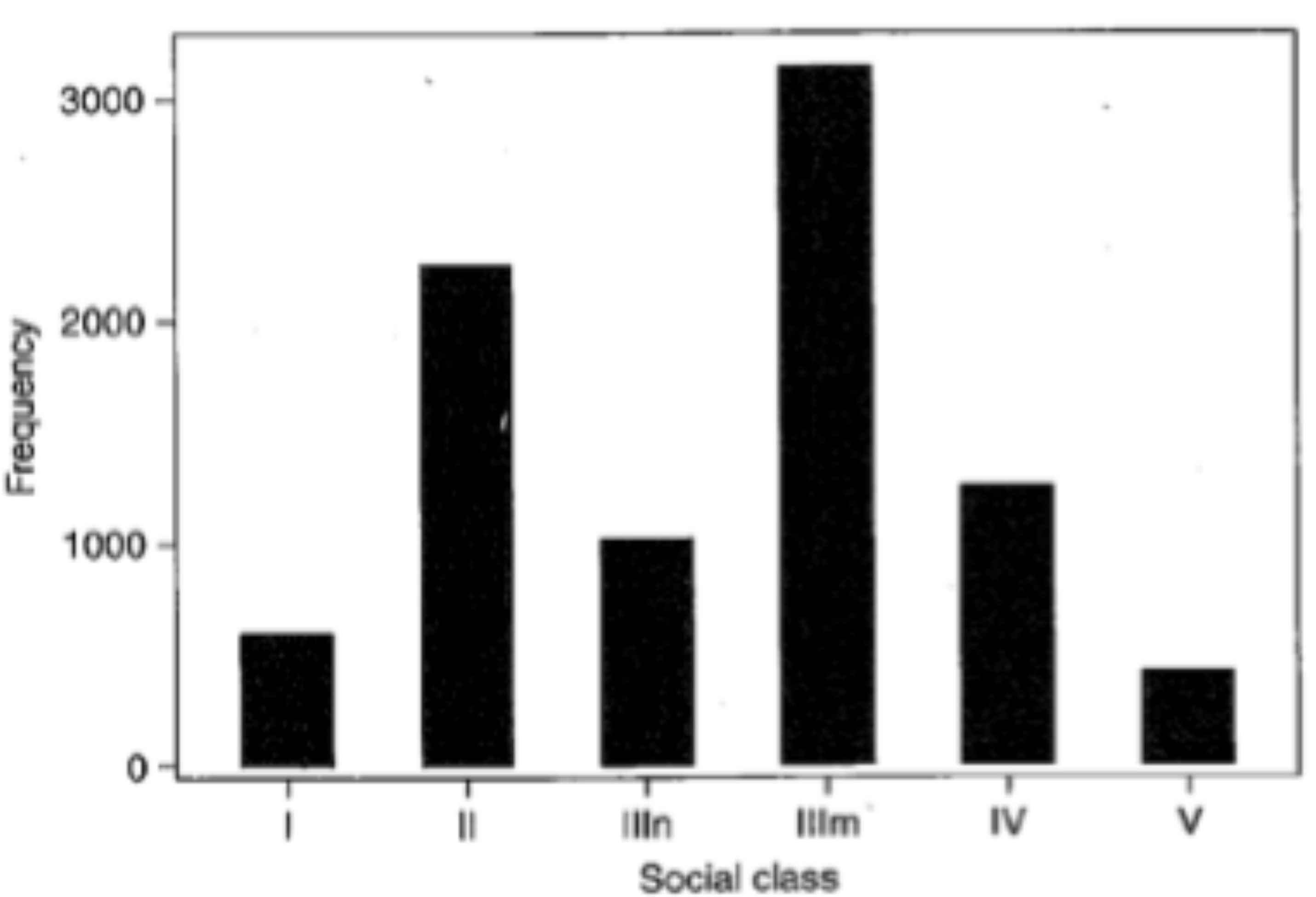
Tables and Charts

- For a single qualitative variable
  - Frequency table, bar chart and pie chart

Table 2.1. Occupational social class in the SHHS.

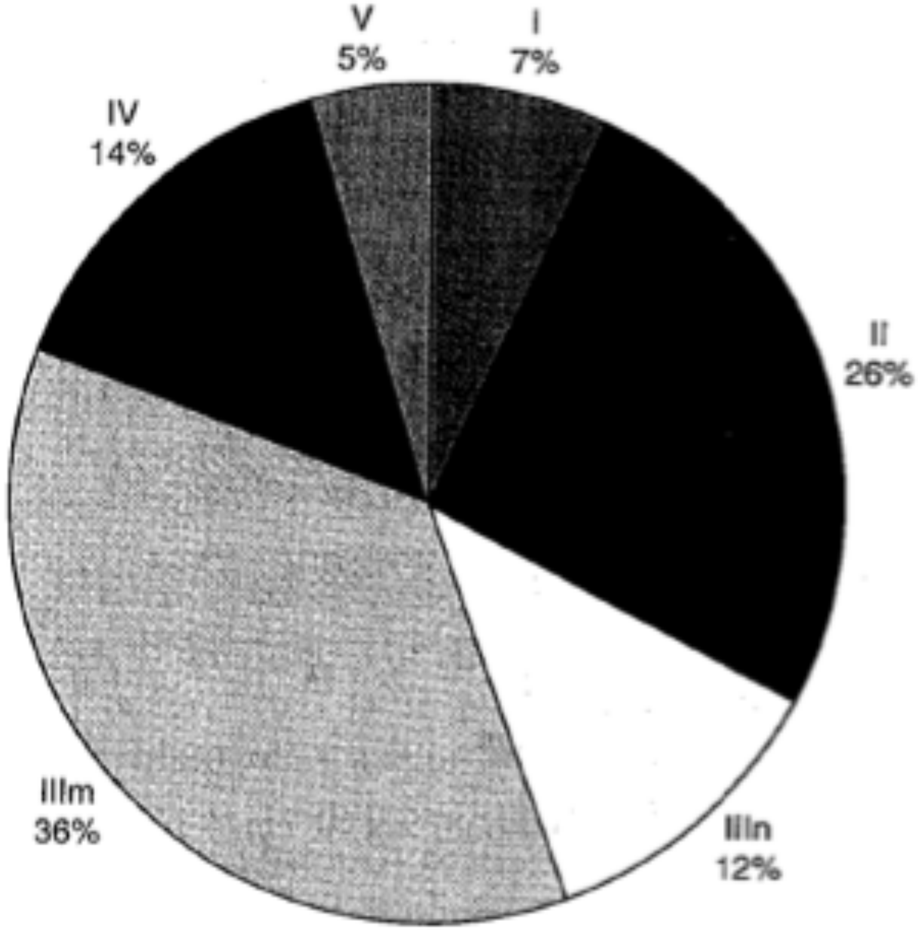
Social class		Number	(%)
I	nonmanual, professional	592	(7)
II	nonmanual, intermediate	2254	(26)
IIIn	nonmanual, skilled	1017	(12)
IIIm	manual, skilled	3150	(36)
IV	manual, partially skilled	1253	(14)
V	manual, unskilled	415	(5)
Total		8681	

Bar Chart



40

Pie Chart



Pie chart for occupational social class in the SHHS.

41

Comparing Two Qualitative Variables

Table 2.2. Social class by prevalent CHD status in the SHHS.

Social class	Prevalent CHD Yes (%)	No	Total
I	100 (16.9)	492	592
II	382 (17.0)	1872	2254
IIIn	183 (18.0)	834	1017
IIIm	668 (21.2)	2482	3150
IV	279 (22.3)	974	1253
V	109 (26.3)	306	415
Total	1721 (19.8)	6960	8681

Bar Chart

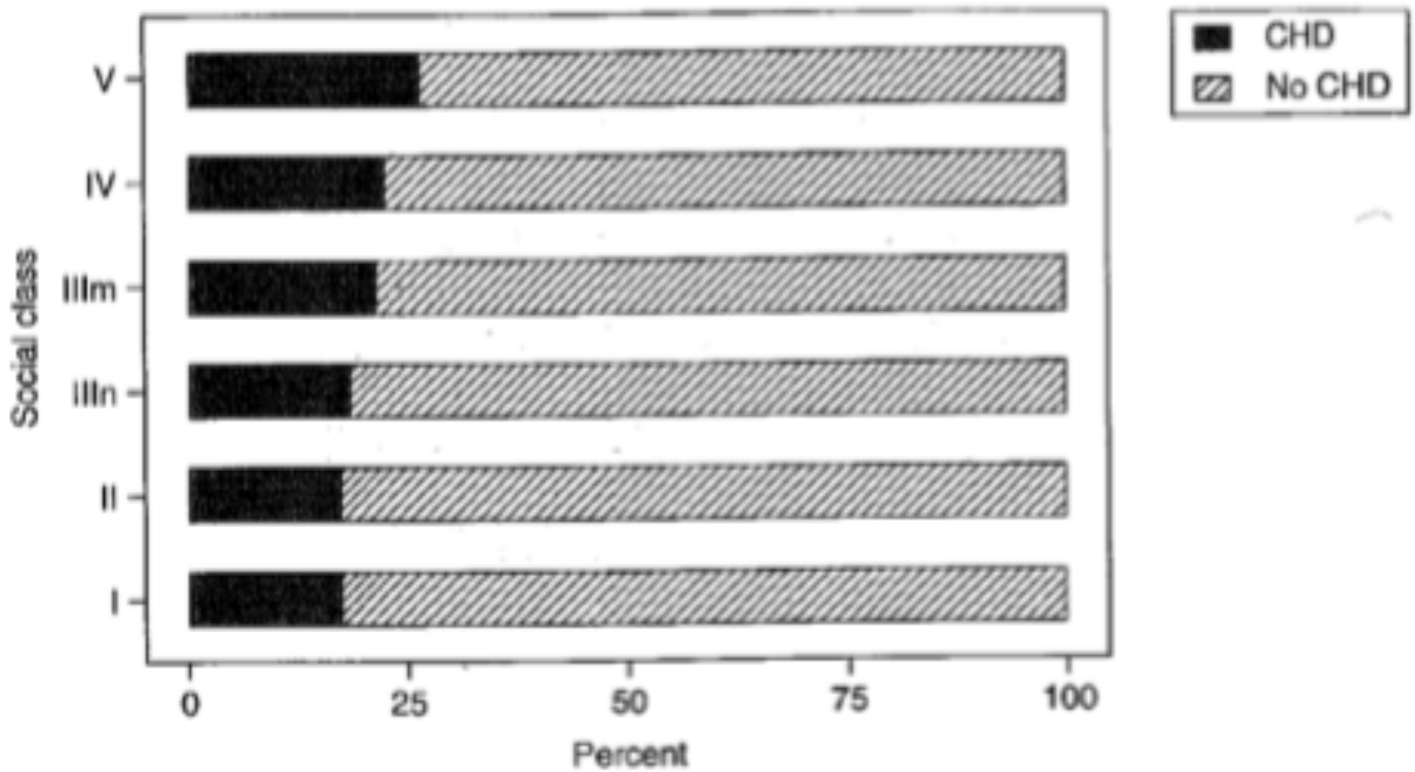


Figure 2.3. Bar chart for occupational social class in the SHHS, showing percentage by CHD status.

43

How to Make Good Tables

- Each table should be self-explanatory.
- Each table should have an attractive appearance.
- The rows and columns should be arranged in a natural order.
- Numbers are easier to compare when the table has a vertical orientation.
- Tables should have consistent appearance throughout the report.

Table 2.3. Minimum, median and maximum values for selected variables in developed countries, 1970.

Variable	Minimum	Median	Maximum
Gross national product per person	1949	4236	6652
Population per km^2	1.6	77.2	324.2
Cigarette consumption per person per year	630	2440	3810
Infant mortality per 1000 births	11.0	18.2	29.6

Source: Cochrane, A.L. *et al.* (1978), *J. Epidemiol. Comm. Health*, 32, 200–205.

Table 2.4. Minimum, median and maximum values for selected variables in developed countries, 1970.

	Gross national product per person	Population density per km^2	Cigarette consumption per person/per year	Infant mortality rate per 1000 births
Maximum	6652	324.2	3810	29.6
Median	4236	77.2	2440	18.2
Minimum	1949	1.6	630	11.0

Table 2.5. Smoking habit by diagnosis group by sex.

Sex/smoking habit	CHD diagnosis group: Diagnosed	Control	Undiagnosed	Total
Males				
Solely cigarette	125 (8%)	1175 (77%)	226 (15%)	1526
Solely cigar	25 (7%)	313 (83%)	40 (10%)	378
Solely pipe	9 (8%)	80 (72%)	22 (20%)	111
Mixed smokers	39 (8%)	379 (75%)	89 (17%)	507
All smokers	198 (8%)	1947 (77%)	377 (15%)	2522
Females[a]				
Solely cigarette	105 (6%)	1410 (78%)	297 (16%)	1812
Solely cigar	0	12 (100%)	0	12
Mixed smokers	2 (8%)	20 (77%)	4 (15%)	26
All smokers	107 (6%)	1442 (78%)	301 (16%)	1850

[a] No females smoked a pipe.

Table 2.6. Diagnosis group by previous smoking habit by sex for nonsmokers.

CHD diagnosis group	Males: Ex-smokers	Males: Never-smokers	Females: Ex-smokers	Females: Never-smokers
Diagnosed	142 (11%)	45 (4%)	58 (6%)	75 (4%)
Control	993 (76%)	888 (83%)	767 (80%)	1581 (78%)
Undiagnosed	164 (13%)	143 (13%)	137 (14%)	375 (18%)

Table 2.7. Men and women classified by self-declared smoking habit and diagnosis group.

Sex/smoking habit	CHD diagnosis group: Diagnosed		Undiagnosed		Control	
Males						
Solely cigarettes	125	(32%)	226	(33%)	1175	(31%)
Solely cigars	25	(7%)	40	(6%)	313	(8%)
Solely pipes	9	(2%)	22	(3%)	80	(2%)
Mixed smokers	39	(10%)	89	(13%)	379	(10%)
Ex-smokers (of any)	142	(37%)	164	(24%)	993	(26%)
Never-smokers (of any)	45	(12%)	143	(21%)	888	(23%)
Total	385	(100%)	684	(100%)	3828	(100%)
Females						
Solely cigarettes	105	(44%)	297	(37%)	1410	(37%)
Solely cigars	0		0		12	(0%)
Mixed smokers	2	(1%)	4	(0%)	20	(1%)
Ex-smokers (of any)	58	(24%)	137	(17%)	767	(20%)
Never-smokers (of any)	75	(31%)	375	(46%)	1581	(42%)
Total	240	(100%)	813	(100%)	3790	(100%)

Note: Percentages less than 0.5% are given as 0%.

Source: Woodward, M. and Tunstall–Pedoe, H. (1992), *Eur. Heart J.*, 13, 160–165.

Descriptive Techniques for Quantitative Variables

- Numerical summarization
  - More important in report writing
    - Economical in space
- Pictorial shape investigation
  - More important in initial exploration
    - Many analytical techniques are only suitable for data of a certain shape