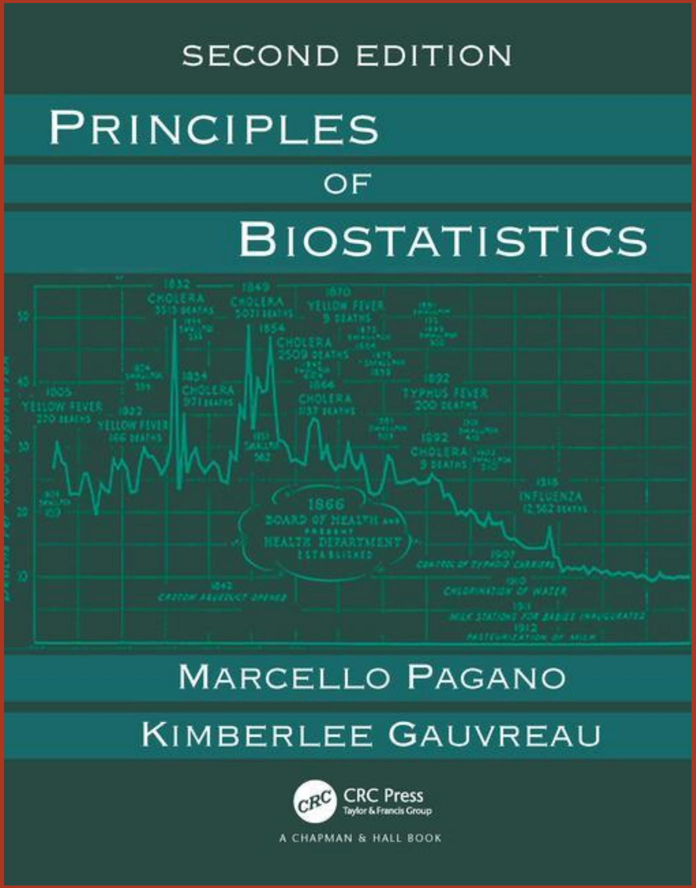
IT3030 - Biostatistics

Week#1 (3/03/2020)

Textbook & Grading

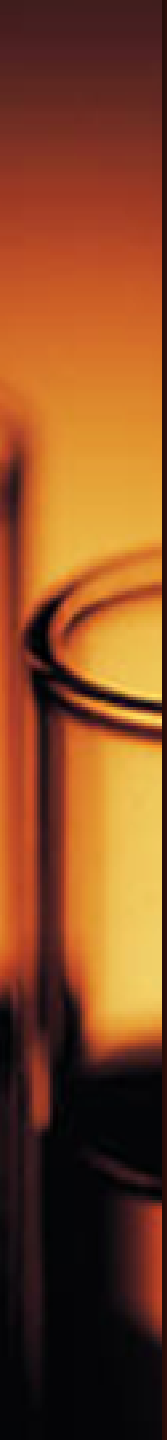
- Principles of Biostatistics by Pagano & Gauvreau, 2nd edition.
- Quizzes (25%), open-book tests.
- Two midterm exams (25%x2) and one final exam (20%), close-book tests.



滄海書局代理~~

PPT slides go by **Week#**

- PPT slides are to be uploaded and available in CGU E-Learning Platform.



Important Dates

- **4/14**: first midterm exam (7^{th} week)
- **5/19**: second midterm exam (12^{th} week)
- **6/30**: final exam (18^{th} week)

Office hours

- Wednesday 1:10~4:00 pm (9F教學資源中心主任辦公室)
- Or by appointment

Chapter 1 - Introduction

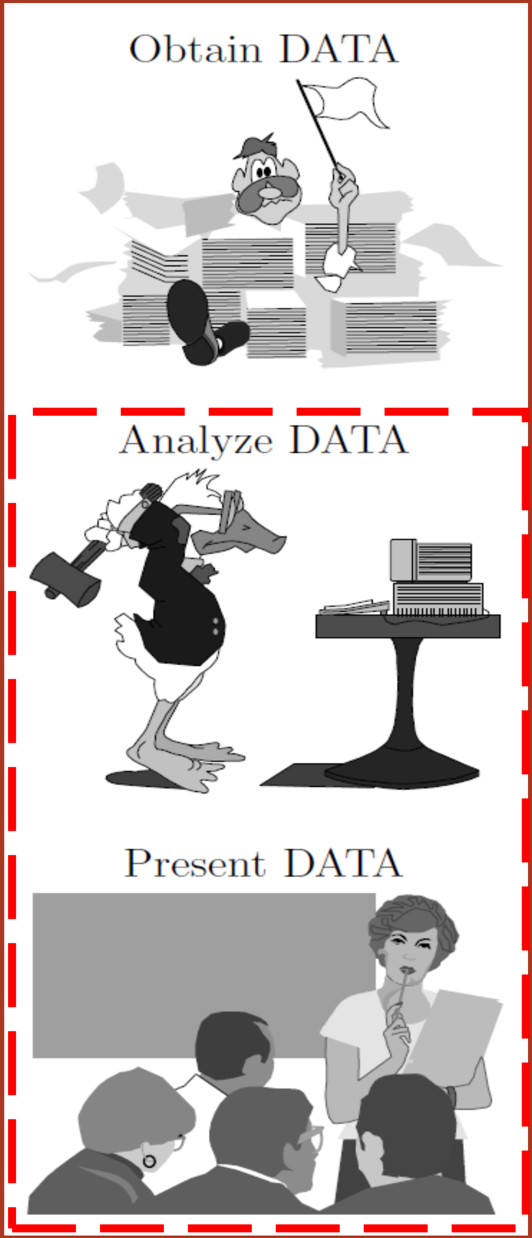
Definition

- **Biostatistics** (a hybrid word made from **biology** and **statistics**; sometimes referred to as **biometry** or **biometrics**)
- The application of **statistics** to a wide range of topics in **biology**. (Wikipedia)

The Science of Biostatistics

- (1) the **design of biological experiments**, especially in medicine and agriculture (crops & livestocks);
- (2) the **collection**, **summarization**, and **analysis of data** from those experiments;
- (3) the **interpretation** of, and **inference（推論)** from, the results.

Chapter 2
Data Presentation



What does a statistician do?

Introduction

- Between the **raw data** and the **reported results** of the study lies some ***intelligent*** *and* ***imaginative manipulation*** of the numbers, carried out using the methods of ***descriptive statistics* (描述性統計)**.

Cont'd

- ***Descriptive statistics*** (as opposed to ***inferential statistics*** 推論性統計) are a means of organizing and summarizing observations.
- They provide us with **an overview of the general features** of a set of data.
- In short, they are various methods of **displaying** a set of data.

2.1 Types of Numerical Data

❶ Nominal Data

❷ Ordinal Data

❸ Ranked Data

❹ Discrete Data

❺ Continuous Data

➊ Nominal (記名的) Data

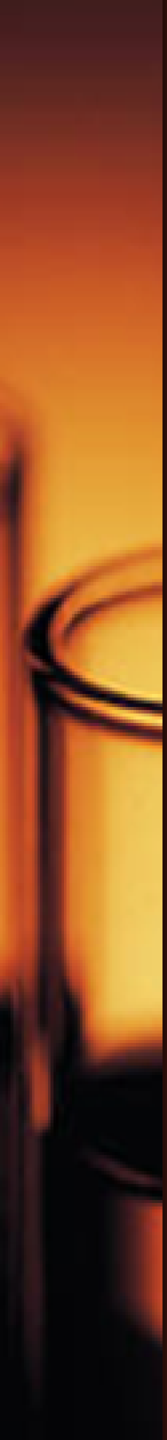
- It is still a numerical data (a code in the form of a number).
- The numeric values **do not** represent magnitude or order at all.
- In a certain way they simply act like "**labels**", representing certain **class** or **category**.
- For example, use "1" for males and "0" for females.
- Computation on nominal data is totally meaningless, e.g., the average of male and female is 0.5

Cont'd

- Nominal data that take on only **two** values – such as male and female – are said to be **dichotomous** or **binary**.
- In general, of course, nominal data can have more than two values, such as using 1 for blood type "O", 2 for type "A", 3 for type "B" and 4 for type "AB", and so on.

Categorical Data

- In general non-numerical comparing with "nominal".
- A set of data is said to be categorical if the values or observations belonging to it *can be sorted* according to category.
- Each value is chosen from a set of **non-overlapping** categories. (e.g., blood type "A", "B", "O", "AB", etc.)
- "Nominal" = "Numerical & Categorical".



❷ Ordinal Data

- The **order** is important.
- For example, injuries may be classified according to their level of ***severity***: 1=fatal, 2=severe, 3=moderate, 4=minor.
- In general the **magnitude is not important**. That is, the severity difference between 1 and 2 is not the same as between 2 and 3.

Cont'd

- In the example from previous slide, a small scale means the most severe. It can, however, be the other way around too. For example, 4=fatal and 1=minor.
- Many **clinical trials** (experimental study involving human subjects) would involve data like this. (See next slide)

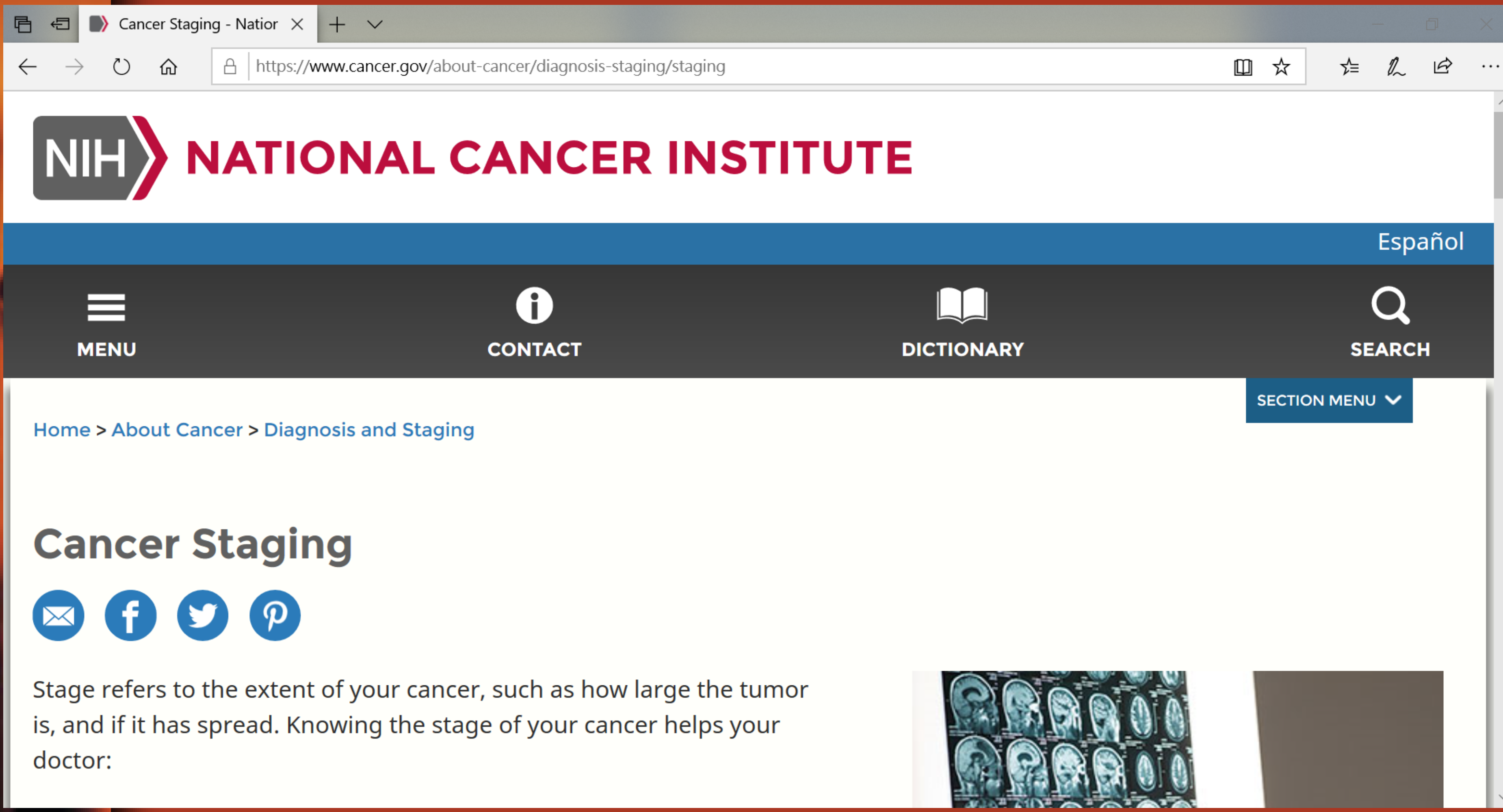
TABLE 2.2

Eastern Cooperative Oncology Group's classification of patient performance status

Status	Definition
0	Patient fully active, able to carry on all predisease performance without restriction
1	Patient restricted in physically strenuous activity but ambulatory and able to carry out work of a light or sedentary nature
2	Patient ambulatory and capable of all self-care but unable to carry out any work activities; up and about more than 50% of waking hours
3	Patient capable of only limited self-care; confined to bed or chair more than 50% of waking hours
4	Patient completely disabled; not capable of any self-care; totally confined to bed or chair

Get a feeling for those vocabulary that may involve health care and biostatistics (the **descriptive** part).

Cancer Staging



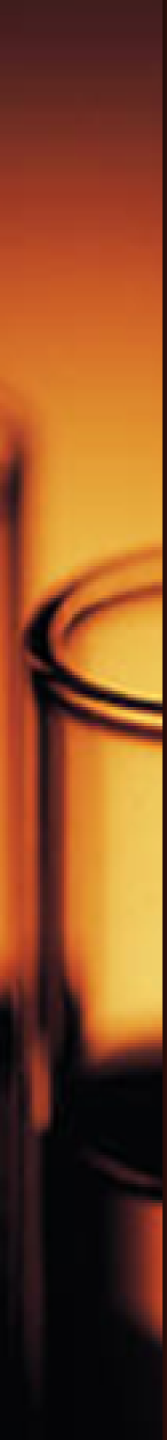
Stage	What it means
Stage 0	Abnormal cells are present but have not spread to nearby tissue. Also called carcinoma in situ (原位癌), or CIS. CIS is not cancer, but it may become cancer.
Stage I, Stage II, and Stage III	Cancer is present. The higher the number, the larger the cancer tumor and the more it has spread into nearby tissues.
Stage IV	The cancer has spread to distant parts of the body.

The TNM Staging System

- The T refers to the size and extent of the **main tumor**. The main tumor is usually called the primary tumor.
- The N refers to the number of nearby **lymph nodes** that have cancer.
- The M refers to whether the cancer has **metastasized**. This means that the cancer has spread from the primary tumor to other parts of the body.

Primary tumor (T)

- **TX**: Main tumor cannot be measured.
- **T0**: Main tumor cannot be found.
- **T1, T2, T3, T4**: Refers to the size and/or extent of the main tumor. The higher the number after the T, the larger the tumor or the more it has grown into nearby tissues. T's may be further divided to provide more detail, such as T3a and T3b.



Regional lymph nodes (N)

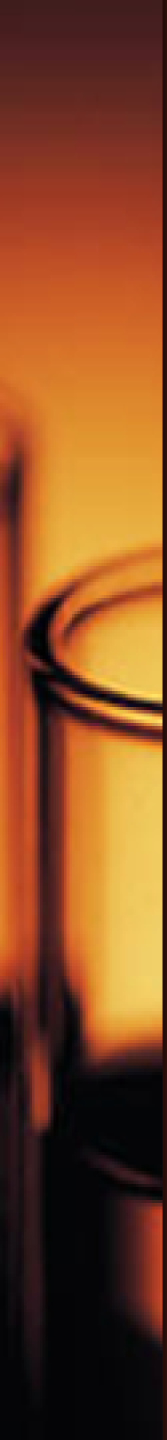
- **NX**: Cancer in nearby lymph nodes cannot be measured.
- **N0**: There is no cancer in nearby lymph nodes.
- **N1, N2, N3**: Refers to the number and location of lymph nodes that contain cancer. The higher the number after the N, the more lymph nodes that contain cancer.

Distant metastasis (M)

- MX: Metastasis cannot be measured.
- M0: Cancer has not spread to other parts of the body.
- M1: Cancer has spread to other parts of the body.

❸ Ranked Data

- This is **similar to “ordinal” data**.
- For example, ranking all departments according to their size (employee count) from top to bottom, we have {1, 2, 3, 4, 5, 6}.
- Still, we disregard the magnitudes of the observations and consider only their **relative positions**.



❹ Discrete Data

- **Both ordering and magnitude are important.**
- Numbers represent actual ***measurable*** quantities rather than mere labels.
- Restricted to taking on only specified values – often **integers** or **counts**. No intermediate values are possible.
  - Number of new cases of tuberculosis (肺結核) reported in the US during a one-year period.
  - Number of beds in a particular hospital

Cont'd

- The outcome of **an arithmetic operation performed on discrete values** is not necessarily discrete itself.
- For example, one woman has given birth three times, and the other has given birth twice. It makes sense to say that the average number of birth for these two women is 2.5, which is not an integer.

❺ Continuous Data

- Data representing measureable quantities that are not restricted to taking on certain specified values (such as integers).
- Time, **serum cholesterol (膽固醇)** level of a patient, the concentration of a pollutant, the temperature, body weight and height, etc.

Summary

- Ordinal data are often easier to handle than discrete or continuous data.
- Thus, ***conversion from discrete/continuous to ordinal*** is often seen in many data analysis work when there are too many values to handle.
- We shall see this in a moment when we introduce the frequency table next.

2.2 Tables

- Frequency Distributions
- Relative Frequency
- *Absolute* vs *Cumulative* Representations

Introduction

- Once we have data collected, we need to “know” what these data “reveal”, or what they may tell us about.
- It would be better if the data can be “**summarized**”. (Keep in mind, though, data details might be lost when being summarized.)
- A table is perhaps the simplest means of summarizing a set of *observations* and can be used in all types of numerical data.

Frequency Distributions

- **Summarize the amount of measurements over a series of ranges.**
- For nominal and ordinal data:
  - A set of classes or categories each with a numerical count

Cont'd

- For discrete or continuous data:
  - **Break down the range of values into a series of distinct, non-overlapping intervals**, so that the new representation could be ***more informative*** than the raw data. (**Some details, as we said, might be lost upon this conversion.**)

Relative Frequency

- Similar to a regular frequency table, and use the proportion (**percentage** %) of values.
- The relative frequencies for all intervals in a table sum to 100%.
- Useful for comparing sets of data that contain ***unequal numbers*** **of observations**.

Cumulative Frequency

- Both frequency and relative frequency tables can have an additional "**cumulative**" column that helps in better "visualizing" and "interpreting" these tabulated data.
- Both frequency and relative frequency tables can be easily represented by figures.

Example

- Considering the following dataset, an example of **nominal** data for the cause of death upon 100 victims:

1	5	3	1	2	4	1	3	1	5
2	1	1	5	3	1	2	1	4	1
4	1	3	1	5	1	2	1	1	2
5	1	1	5	1	5	3	1	2	1
2	3	1	1	2	1	5	1	5	1
1	2	5	1	1	2	3	4	1	1
1	1	2	1	1	2	1	1	2	3
3	3	1	5	2	3	5	1	3	4
1	1	2	4	5	4	1	5	1	5
5	1	1	5	1	1	5	1	1	5

1. Motor vehicle, 2. Drowning, 3. House fire, 4. Homicide, 5. Other

- The data from previous slide can be used in generating the following table, which would be **much more informational** than the dataset itself.

```
. tab accident

    acc_lab |      Freq.     Percent        Cum.
------------+-----------------------------------
   Motor Ve |         48       48.00       48.00
   Drowning |         14       14.00       62.00
   House Fi |         12       12.00       74.00
   Homicide |          7        7.00       81.00
      Other |         19       19.00      100.00
------------+-----------------------------------
      Total |        100      100.00

. label define acclab 1 "Motor vehicle" 2 "Drowning" 3 "House fire"
> 4 "Homicide" 5 "Other"
```

(by a software called “Stata”)

Table 1.1: Frequencies of serum cholesterol levels

Cholesterol level (mg/100 ml)	(Absolute) Frequency	Cumulative Frequency	(Absolute) Relative Frequency (%)	Cumulative Relative Frequency (%)
80-119	13	13	1.2	1.2
120-159	150	163	14.1	15.3
160-199	442	605	41.4	56.7
200-239	299	904	28.0	84.7
240-279	115	1019	10.8	95.5
280-319	34	1053	3.2	98.7
320-360	9	1062	0.8	99.5
360-399	5	1067	0.5	100.0
Total		1067		100.0

Continuous Ch levels are divided into 8 non-overlapped categories (groups).

Head counts within each range.

This is surely more realistic to understand (to get an idea about these measurements) than reading 1,067 cholesterol levels.

2.3 Graphs

1. Bar Charts
2. Histograms
3. Frequency Polygons
4. One-Way Scatter Plots
5. Box Plots
6. Two-Way Scatter Plots
7. Line Graphs

1. Bar Charts

- A popular type of graph to display a frequency distribution for **nominal** or **ordinal** data.

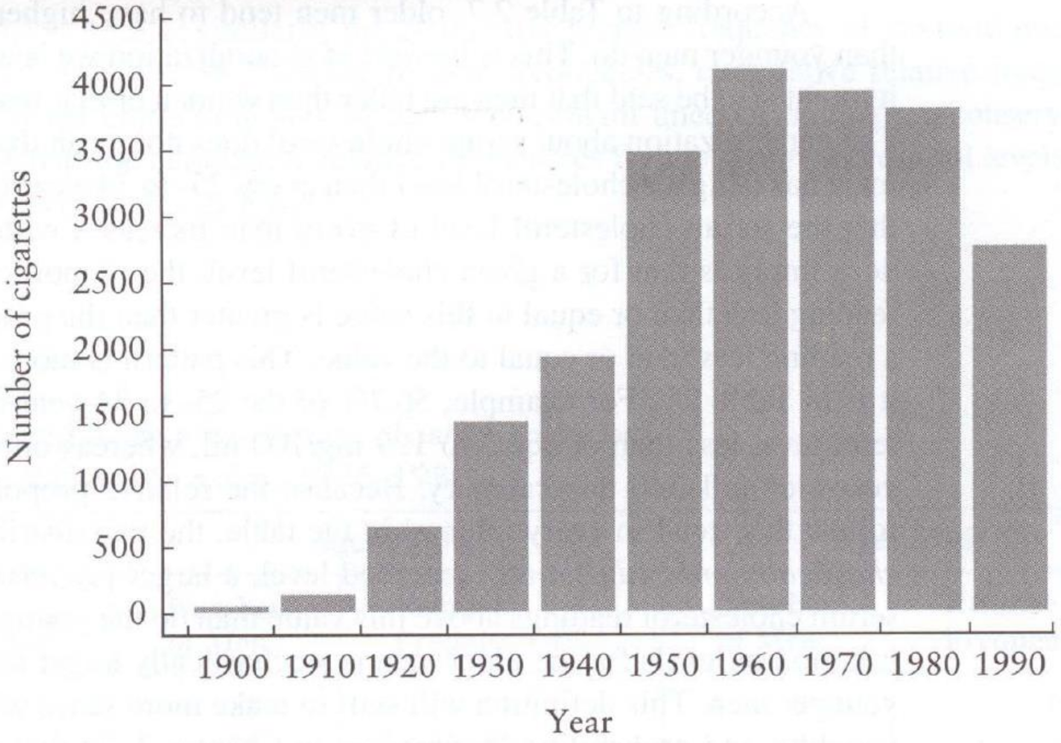


FIGURE 2.1

Bar chart: Cigarette consumption per person 18 years of age or older, United States, 1900–1990

2. Histograms

- Whereas a bar chart is a pictorial representation of a frequency distribution for either nominal or ordinal data, a histogram depicts a frequency distribution for **discrete** or **continuous** data.
- Labels on the horizontal axis are no longer the category it represents. Instead, it is the **true boundary** between these intervals.

Histogram
(absolute frequencies)

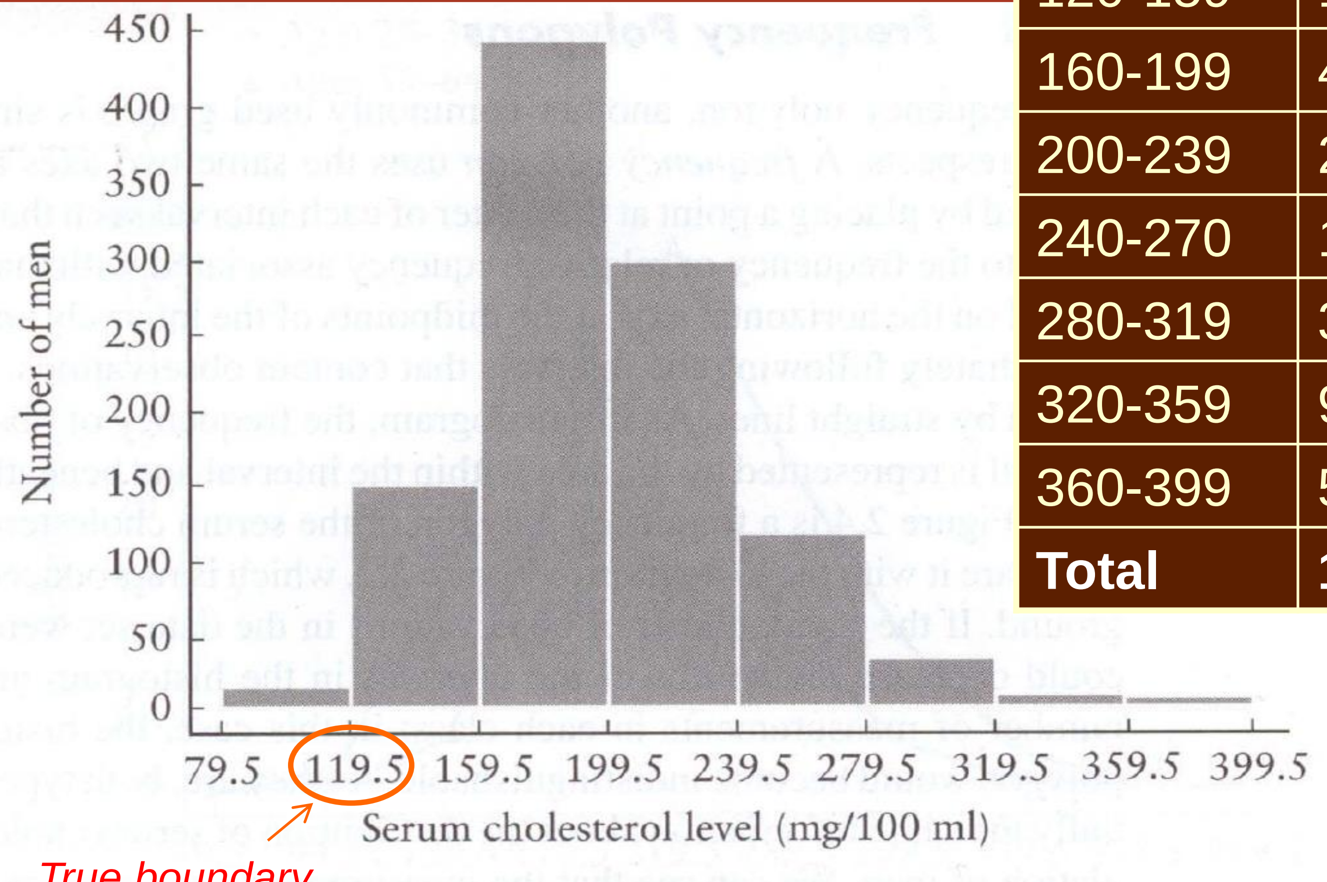
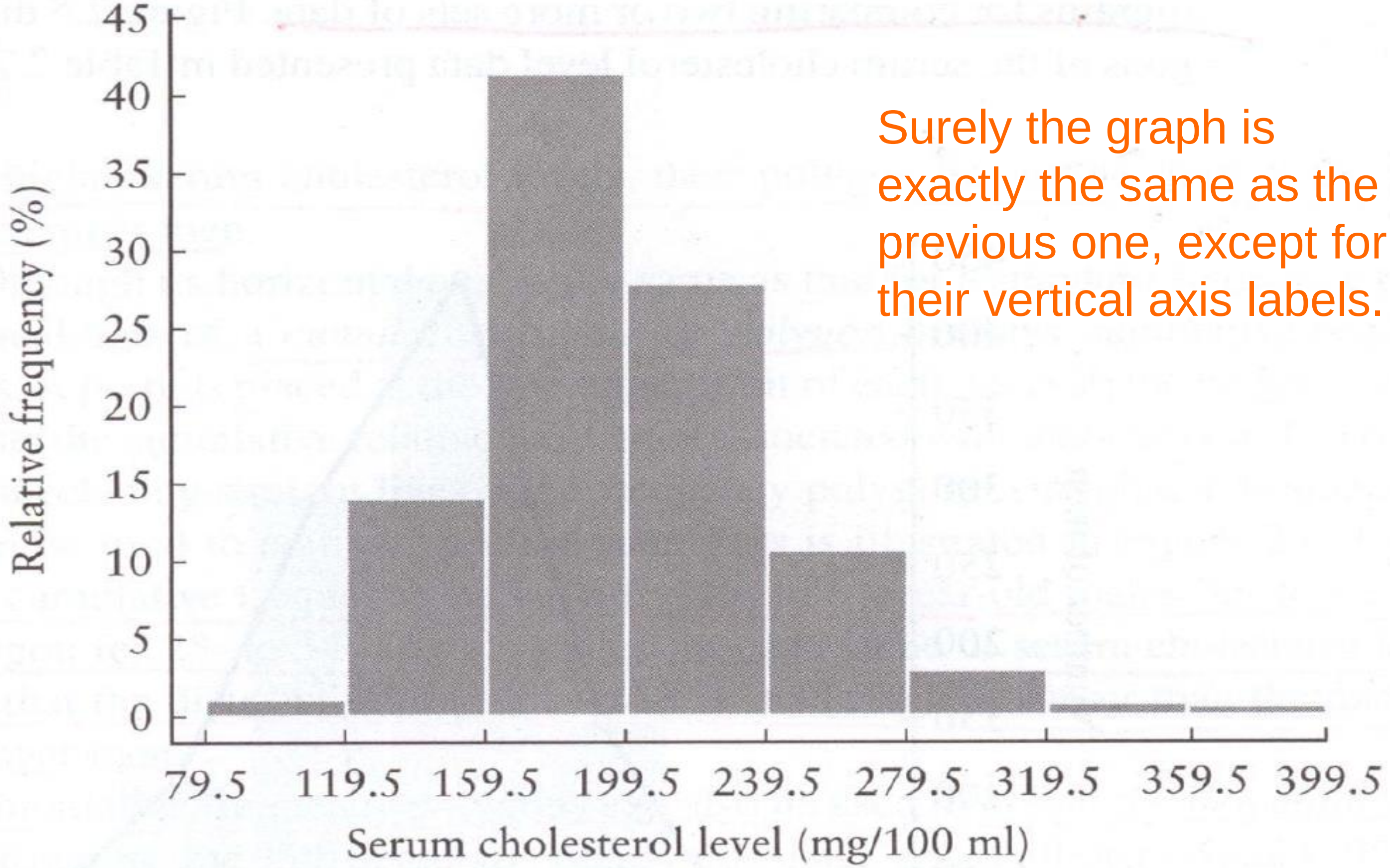


FIGURE 2.2
Histogram: Absolute frequencies of serum cholesterol levels for 1067 U.S. males, aged 25 to 34 years, 1976–1980

Cholesterol Level	Number of Men
80-119	13
120-159	150
160-199	442
200-239	299
240-270	115
280-319	34
320-359	9
360-399	5
Total	**1067**

Histogram
(relative frequencies)



Surely the graph is exactly the same as the previous one, except for their vertical axis labels.

FIGURE 2.3

Histogram: Relative frequencies of serum cholesterol levels for 1067 U.S. males, aged 25 to 34 years, 1976–1980

3. Frequency Polygons

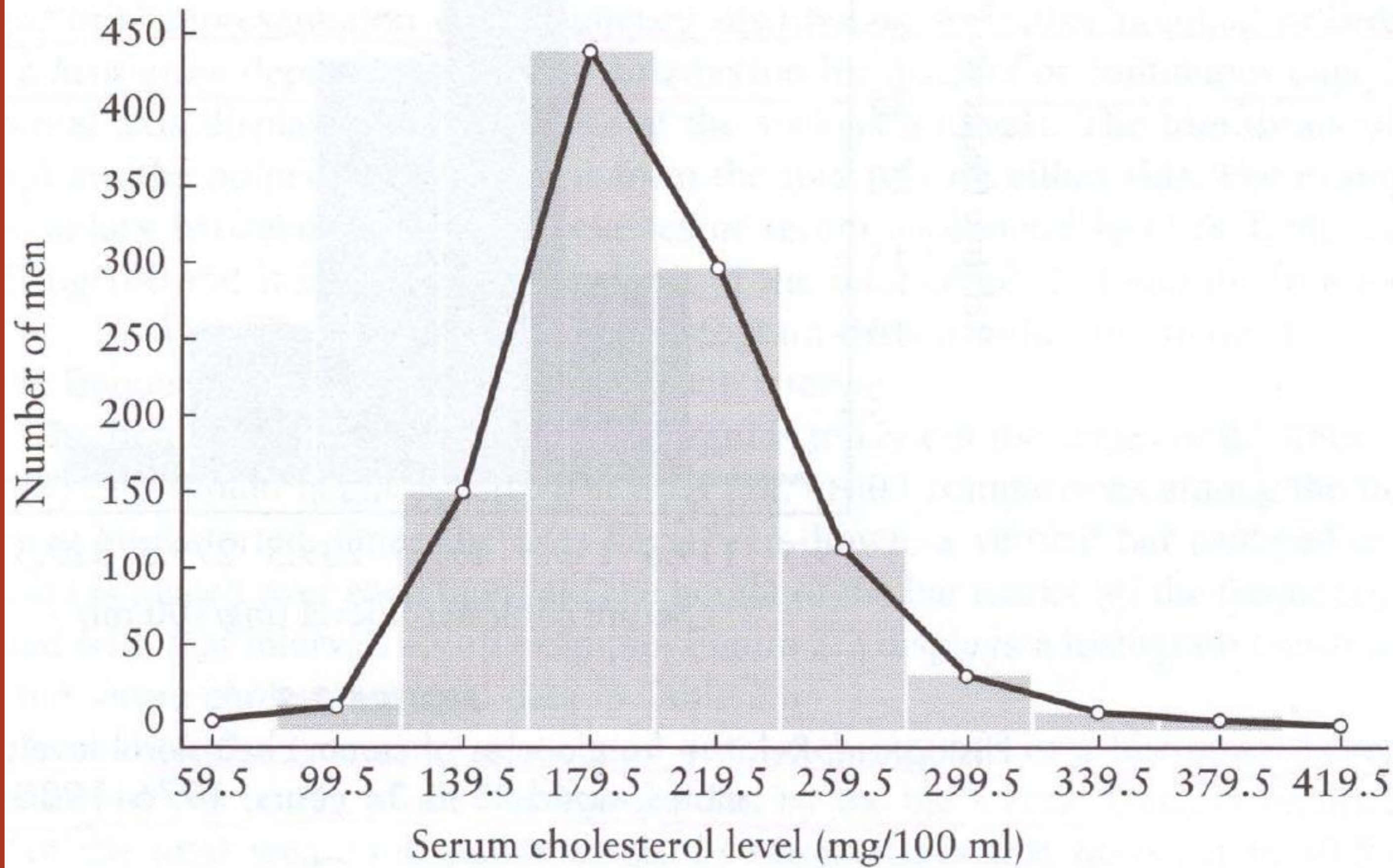


FIGURE 2.4
Frequency polygon: Absolute frequencies of serum cholesterol levels for 1067 U.S. males, aged 25 to 34 years, 1976–1980

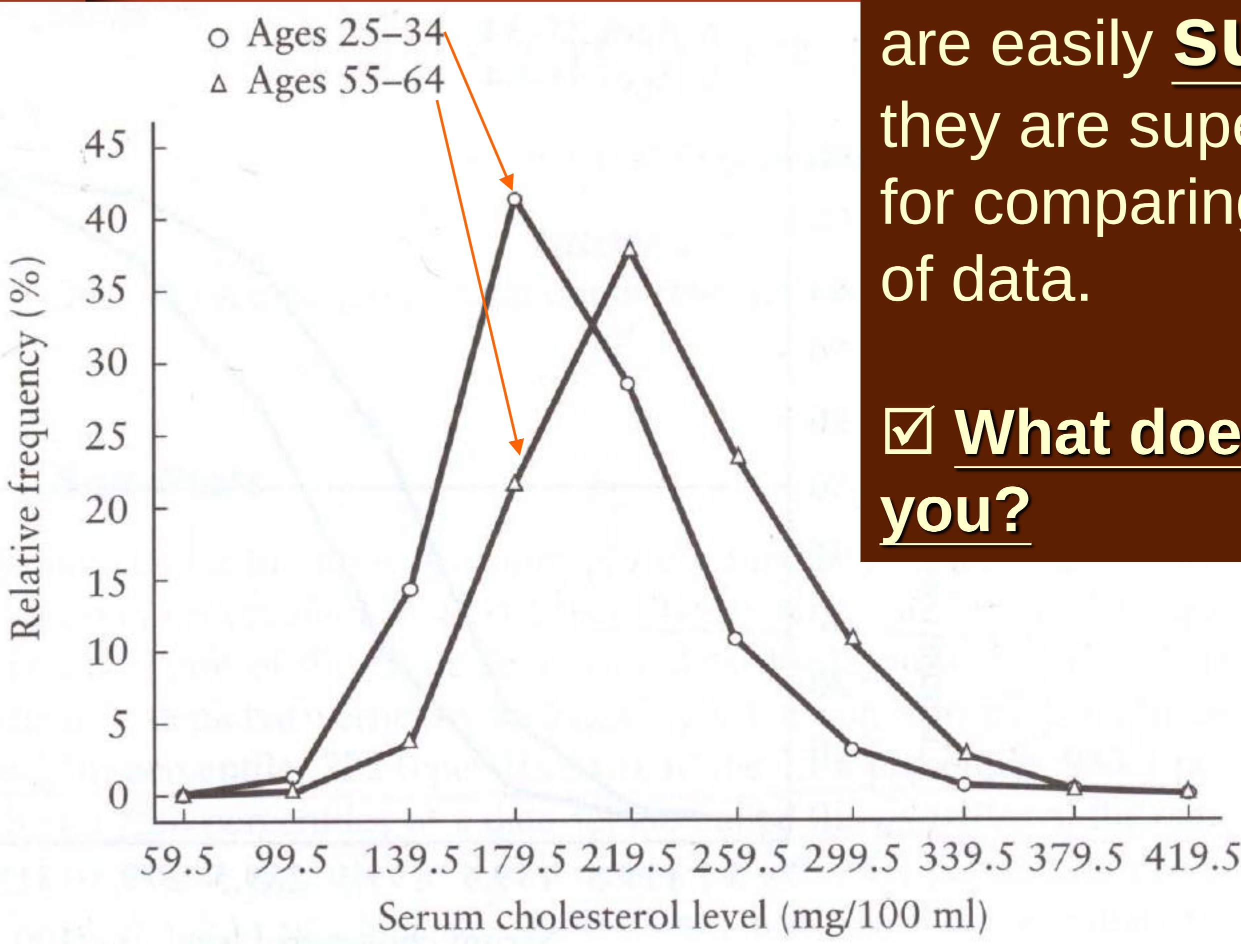


FIGURE 2.5
Frequency polygon: Relative frequencies of serum cholesterol levels for 2294 U.S. males, 1976–1980

Because frequency polygons are easily **superimposed**, they are superior to histograms for comparing two or more sets of data.

☑ **What does Figure 2.5 tell you?**

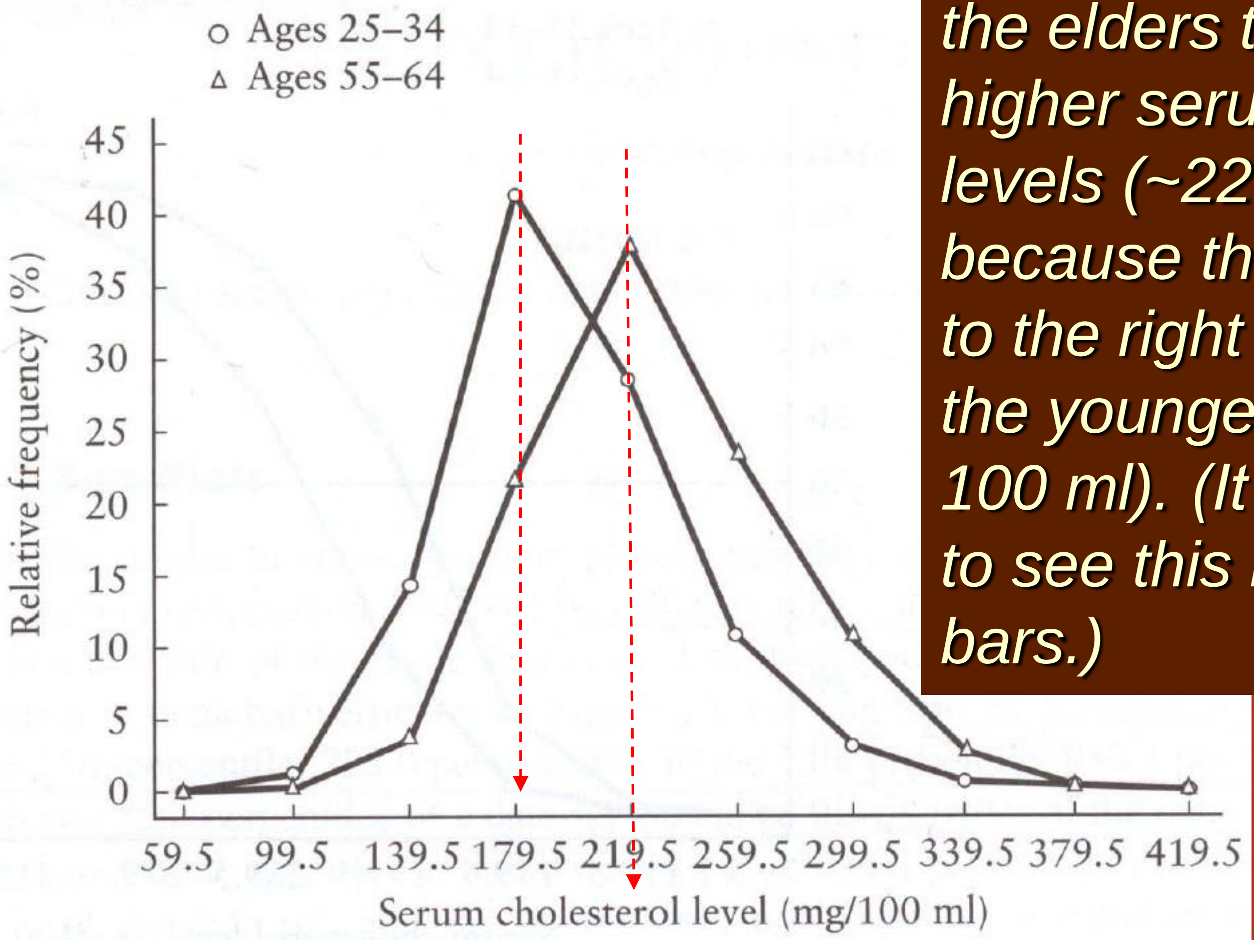


FIGURE 2.5

Frequency polygon: Relative frequencies of serum cholesterol levels for 2294 U.S. males, 1976–1980

Answer:

the elders tend to have higher serum cholesterol levels (~220 mg/100 ml) because their polygon shifts to the right of the polygon for the younger men (~180 mg/ 100 ml). (It wouldn't be easy to see this if using regular bars.)

Percentiles

- **95th percentile** : the value that is greater or equal to 95% of the observations and less than or equal to the remaining 5%.
- Some other often-used percentiles include:
  - – **75th percentile**, also referred as the 3rd quartile or **Q3**.
  - – **50th percentile**, also referred as the 2nd quartile, or **Q2**, which is equivalent to **median (中位數)**.
  - – **25th percentile** , also referred as the 1st quartile or **Q1**.

Cont'd

- These percentiles do not necessarily fall onto one of the observations. There is often some **rounding (四捨五入)** or **interpolation (內插)** involved.

The dataset {1, 3, ***6***, 7, 9} has a Q2=6. It lies on the 3rd value of these observations.

The dataset {1, 3, ***6***, ***7***, 9, 14} **may** have a Q2 from an interpolation of the 3rd and 4th observations. In other words, Q2 does not lie on any of these observations. (We may enforce it to lie on one of the observations, though.)

Cont'd

- **There is no standard definition of percentile**. All definitions yield similar results when the number of observations is large.
- *When percentiles* **need to** *land on one particular observation*, one definition usually given in texts is that the *p*-th percentile of *N* ordered values is obtained by first calculating the rank

$$n = (N/100)*p + \tfrac{1}{2},$$

rounding to the nearest integer, and taking the value that corresponds to that rank. (Wikipedia)

Example 1

The dataset {1, 3, **6**, 7, 9} has a Q2=6. It lies on the 3rd value of these observations.

- Find Q2 (50th percentile):

rank $n = (N/100)*p + ½ = (5/100)*50 + ½ = 3.0 \sim$ **3**.
Thus the **3rd** observation “6” is this Q2.

- Find Q1 (25th percentile):

rank $n = (N/100)*p + ½ = (5/100)*25 + ½ = 1.75 \sim$**2**.
Thus the **2nd** observation “3” is this Q1.

Example 2

The dataset {1, 3, ***6***, ***7***, 9, 14} may have a Q2 from an interpolation of the 3rd and 4th observations, which is 6.5.

When Q2 needs to land on one of the observations:

- Find Q2 (50th percentile):

rank n = (N/100)*p + ½ = (6/100)*50 + ½ = 3.5 ~ **4**.
Thus the **4th** observation "7" is this Q2.

- Find Q3 (75th percentile):

rank n = (N/100)*p + ½ = (6/100)*75 + ½ = 5.0 ~**5**.
Thus the **5th** observation "9" is this Q3.

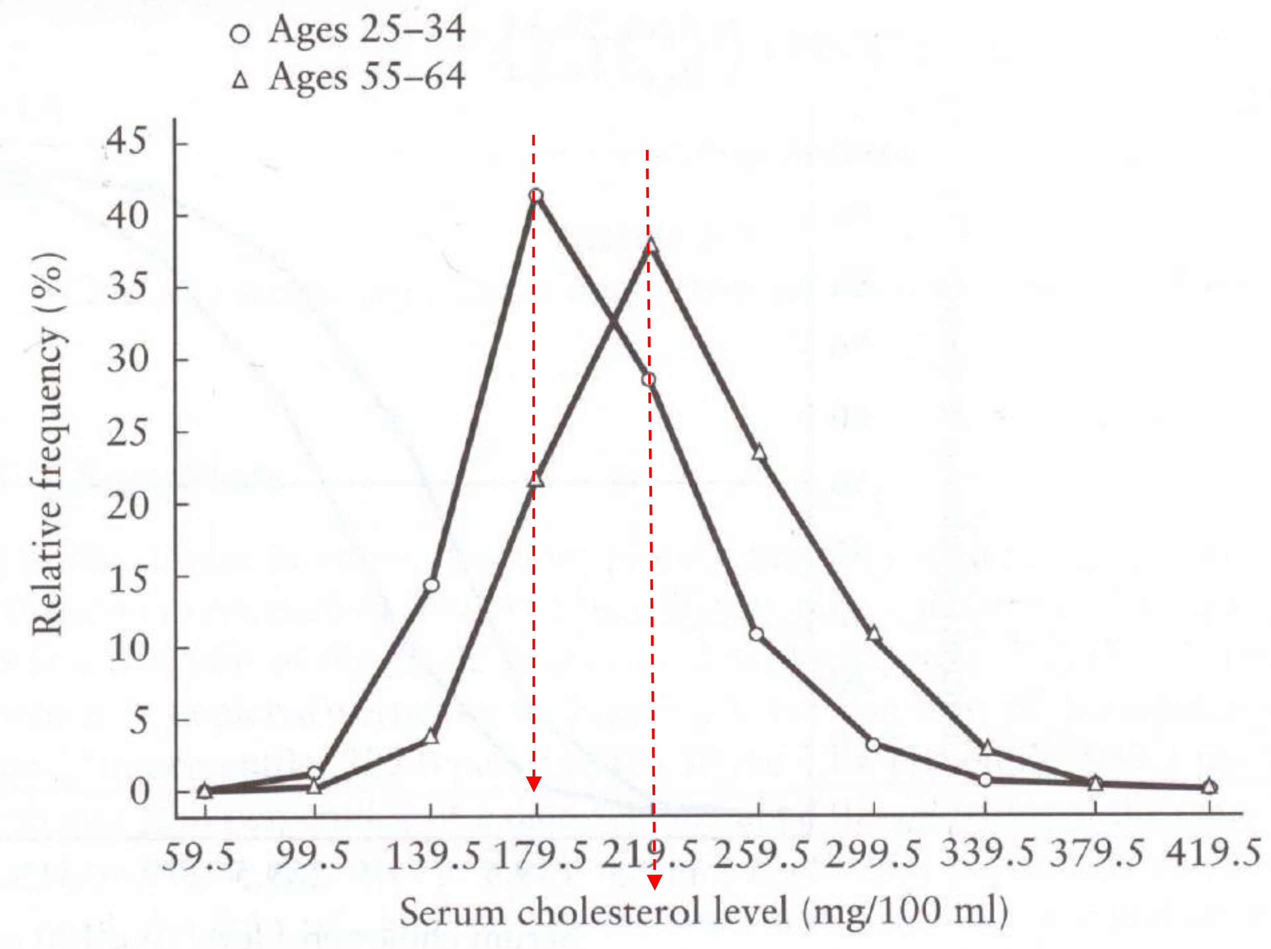


FIGURE 2.5

Frequency polygon: Relative frequencies of serum cholesterol levels for 2294 U.S. males, 1976–1980

1. Recall earlier we had serum cholesterol levels for two different age groups of US males in terms of relative frequencies.
2. As mentioned, we can also show <u>*cumulative*</u> ones.

Cumulative Frequency Polygons

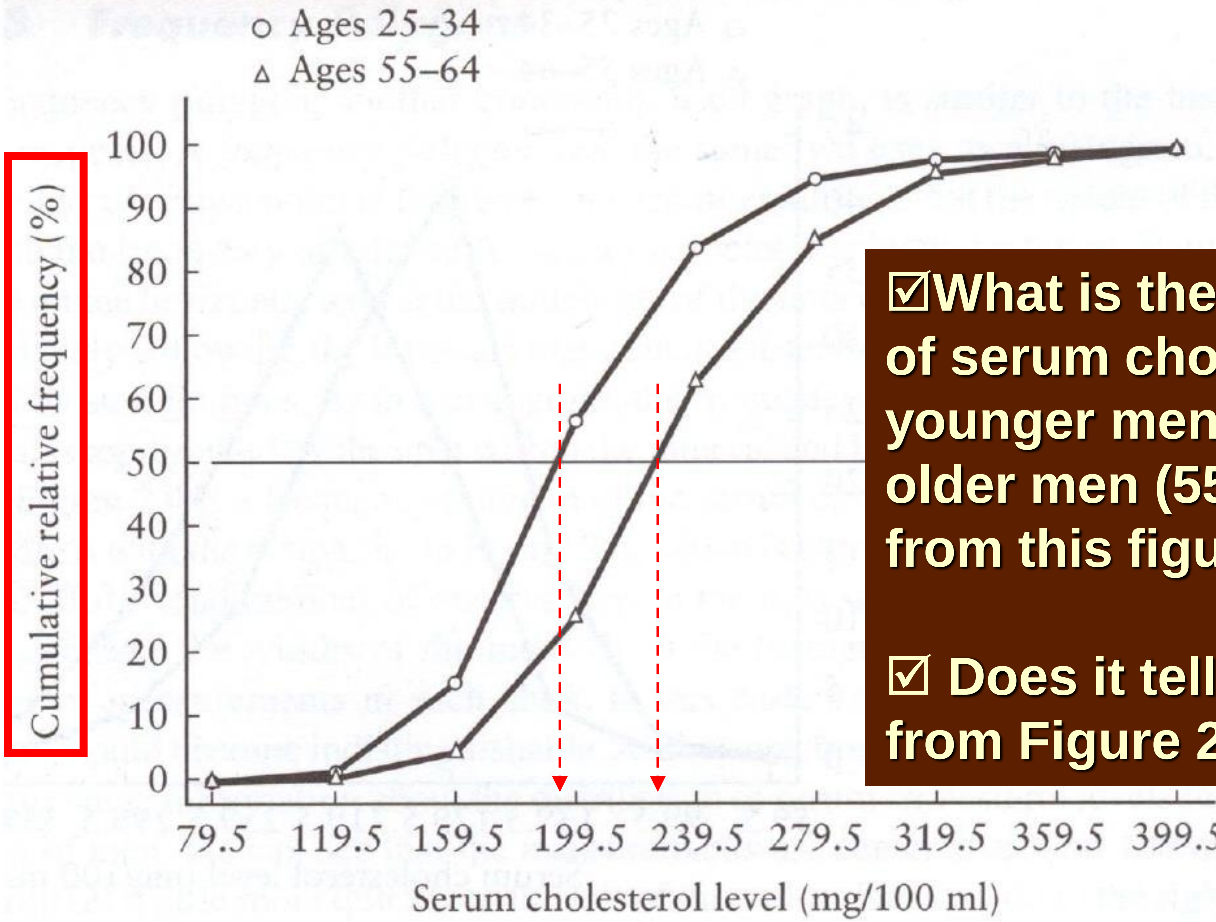


FIGURE 2.6
Cumulative frequency polygon: Cumulative relative frequencies of serum cholesterol levels for 2294 U.S. males, 1976–1980

☑What is the 50th percentile of serum cholesterol of younger men (ages 25-34) and older men (55-64), judging from this figure?

☑ Does it tell you more than from Figure 2.5?

4. One-Way Scatter Plots

- Used to summarize both discrete or continuous data.
- Display the **relative position** of each data point.
- No information is lost, but might be **too crowded** to view.
- Figure 2.7 shows the death rates from the 50 states and Washington DC of USA, from as low as 391.8 in Alaska, to a high of 1214.9 in DC.

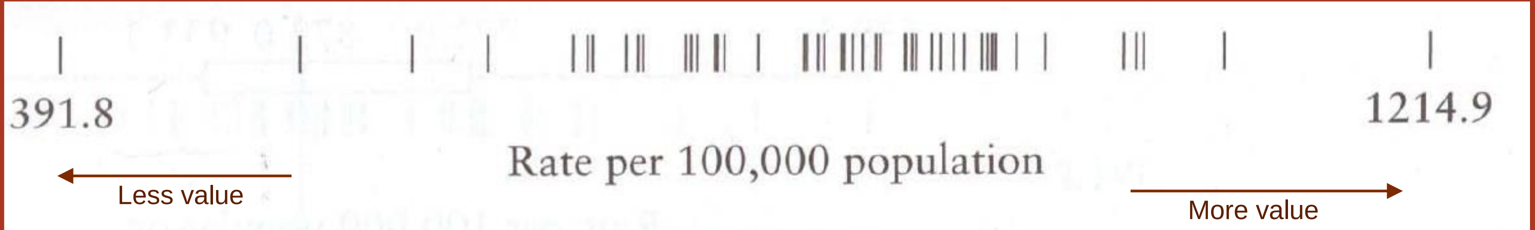
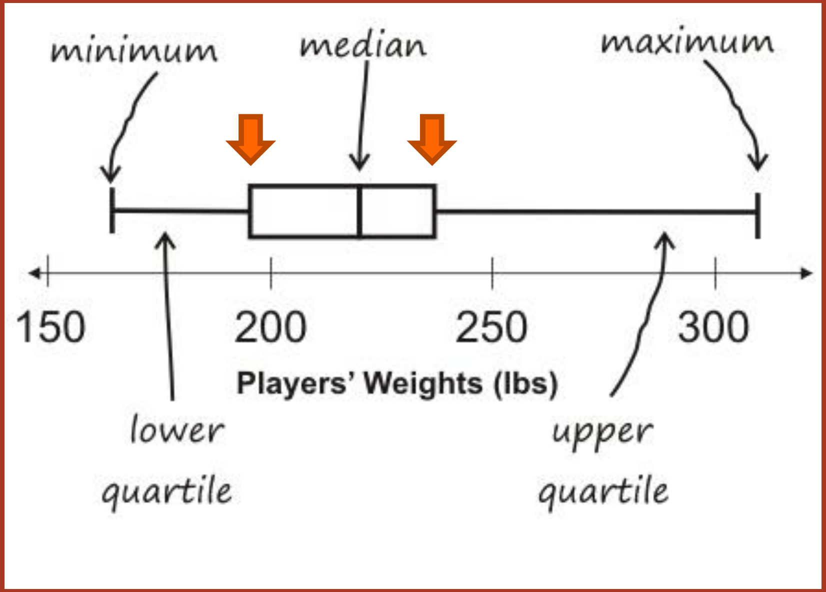


FIGURE 2.7
One-way scatter plot: Crude death rates for the United States, 1992

5. Box Plots

- Similar to one-way scatter plot for using a single axis. Instead plotting all observations, it displays only a **summary** of the data.
- This is done by drawing a box showing the 25th percentile to the 75th percentile as the two edges of the box.

*A box plot is also known as a **whisker** plot.*



Box Plots – cont'd

- It also features with two adjacent values (*minimum* and *maximum* shown in previous slide), which are the most extreme observations in the data set that are not more than, for example, 1.5 times the **box height** beyond either quartile.
- In some texts, observations between 1.5 and 3 times of box height are called **mild outliers**, and beyond 3 times are called **extreme outliers**. (see next slide)

Figure 2.8

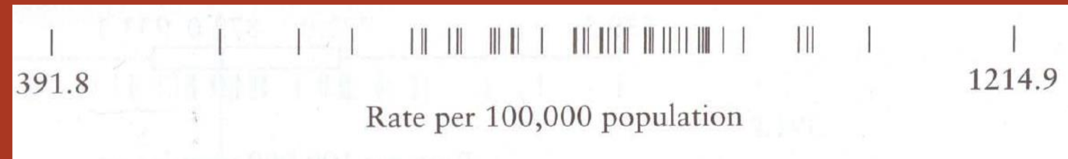
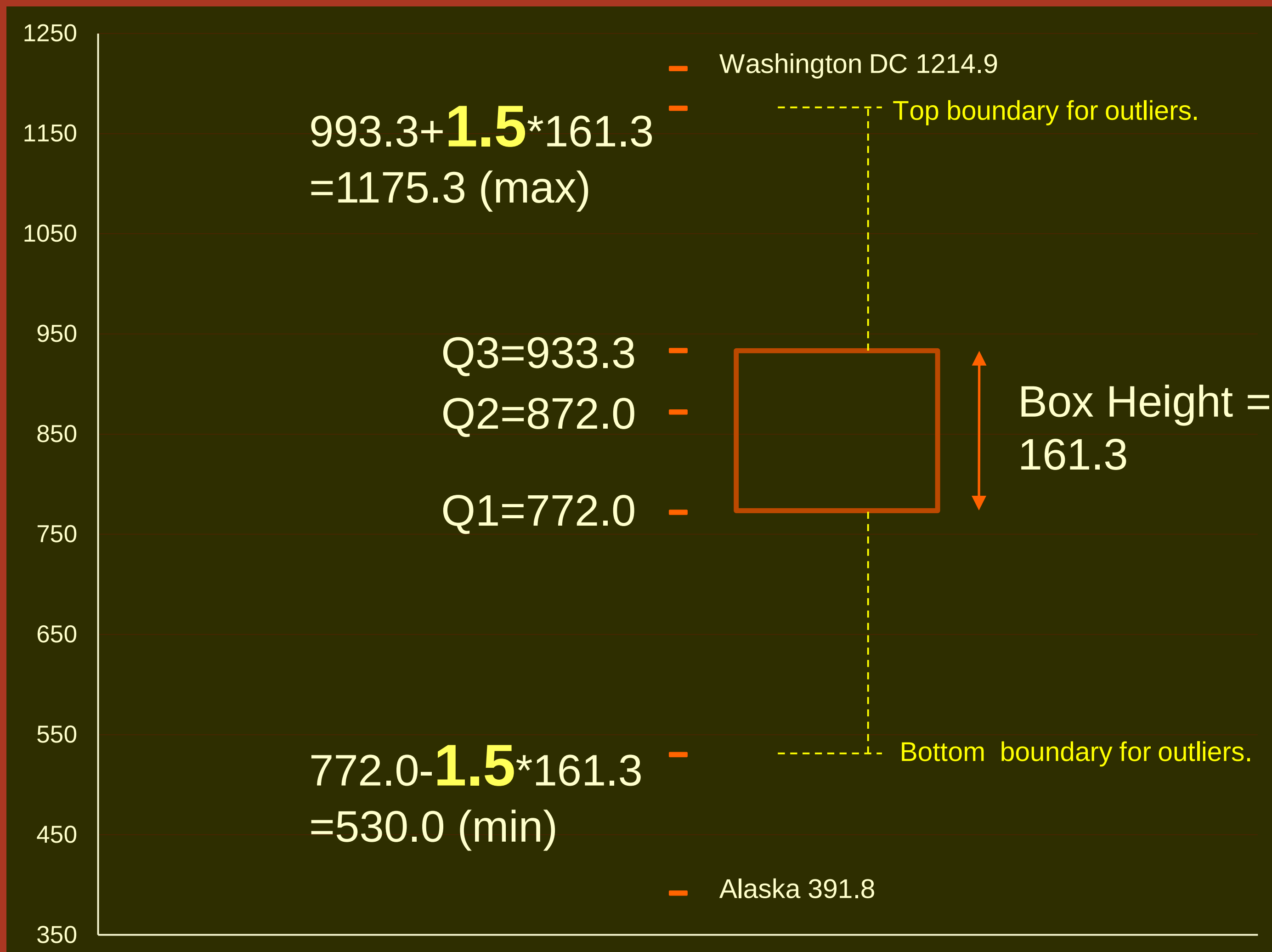


FIGURE 2.7
One-way scatter plot: Crude death rates for the United States, 1992



One-Way Scatter Plot + Box Plots

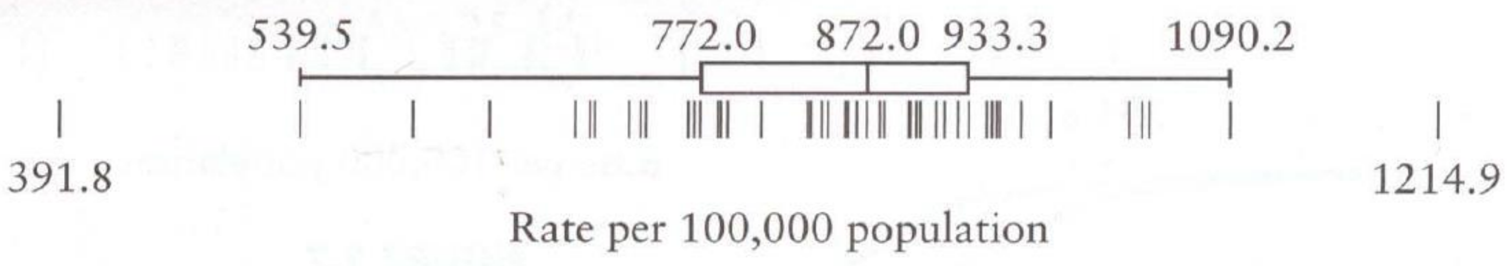


FIGURE 2.9
One-way scatter plot and box plot: Crude death rates for the United States, 1992

- The two extreme values 539.5 and 1090.2 (whiskers) are the most extreme without going into the outlier region.
- It is clear that the lowest one (Alaska) and highest one (Washington DC) are both *extreme outliers*.

Outlier

- An outlier is a data point that is **not typical** (or **atypical**) of the rest of the values.
- In fairly symmetric data sets, the adjacent values should contain approximately **95%** to **99%** of the measurements. [This is, in general, a standard to define whether a random variable is “normally” distributed, as we will review later.]

6. Two-Way Scatter Plots (up) and 7. Line Graphs (bottom)

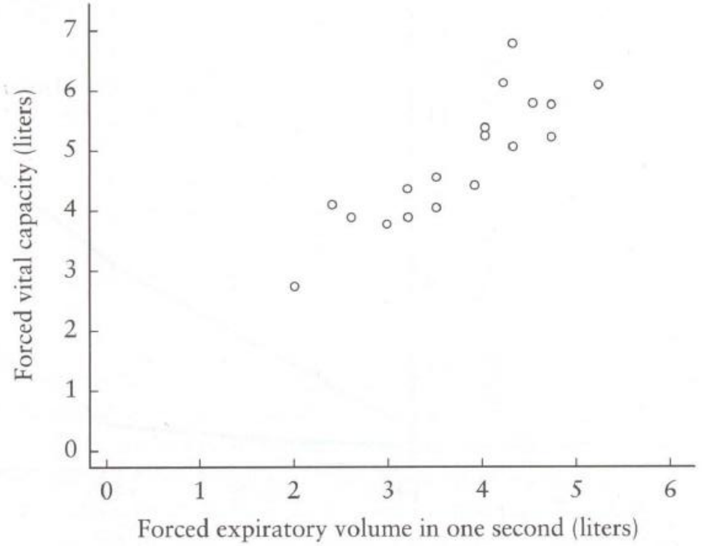
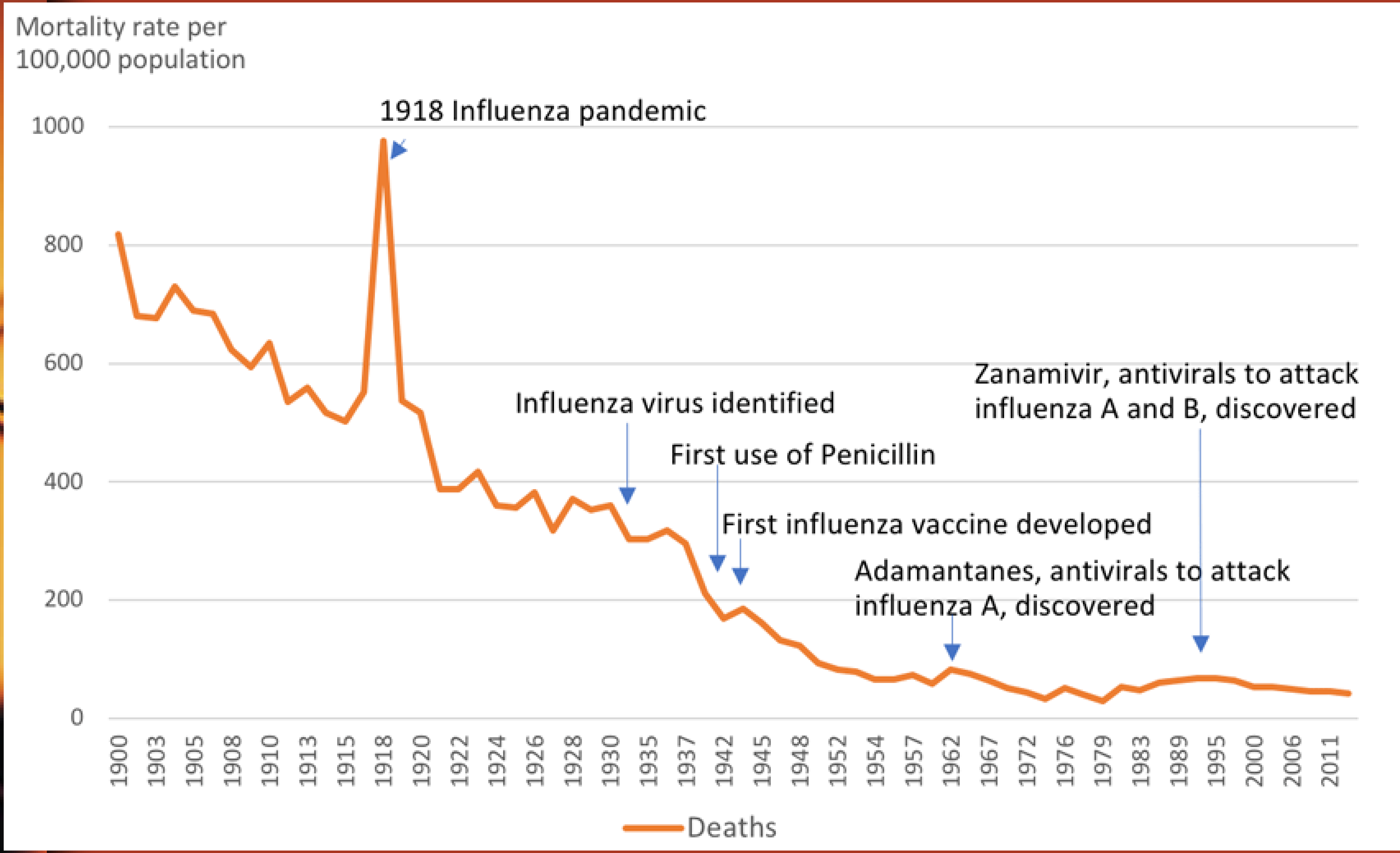


FIGURE 2.10
Two-way scatter plot: Forced vital capacity versus forced expiratory volume in one second for 19 asthmatic subjects

FIGURE 2.11
Line graph: Reported rates of malaria by year, United States, 1940–1989

Note the **log scale** on the vertical axis of Figure 2.11; this scale allows us to depict a large range of observations while still showing the variations among the smaller values.

A “Line Graph” Example



Infectious disease mortality rate in the United States 1900-2014, with a timeline of medical advances. (Source: CDC)

List of statistical packages (wiki)

- public domain / open source / freeware
- retail (commercial)
  - **SAS** (originally *Statistical Analysis System*)
  - **Stata** (hybrid of *Statistics & Data*)
  - **SPSS** (originally *Statistical Package for the Social Sciences*, later modified to read *Statistical Product and Service Solutions*)
  - **MATLAB**