1. Data Types
- nominal data – values fall into unordered categories
  e.g. gender, blood type

- ordinal data – values fall into ordered categories (but differences have no meaning)
  e.g. disease severity class (p. 9)
  Likert scale in surveys

- ranked data – ordinal data where observations are arranged from lowest to highest (or v.v)
  e.g. leading causes of death (p. 10)
  AP football rankings

- discrete data – numbers represent measurable quantities, and are restricted to only specified values. (often integers)
  e.g. family size, number of TB cases

- continuous data – numbers represent measurable quantities where possible values are not restricted.
  e.g. weight, elapsed time

One limiting factor for continuous data is the accuracy, hence they are often rounded to the nearest gm, hr, etc.

2. Tables

2.1 Frequency distributions

- Definition – set of classes along with numerical count of obs (i.e. frequency) in each class.

- Classes should be distinct and non-overlapping, and cover complete range of values. Ideally, the class intervals are of equal width. The number of classes should be appropriate for the data set.
  e.g. cholesterol levels for US males p. 12
  data range from 80 to 400
  use 8 classes , each of width 40
2.2 Relative frequency table

- Definition – set of classes with proportional counts in each class.
- Cumulative relative frequency table – set of classes with proportions of total observation less than or equal upper limit of each class.
  e.g. cholesterol data p. 14

3. Graphs

3.1 Bar chart

- For nominal or ordinal data, categories are on $X$ axis and frequencies are represented by vertical bars.
  e.g. cigarette consumption, p. 16

3.2 Histograms

- Histogram is like a bar chart, except $X$ axis represents discrete or continuous data. The frequency is represented by the area of the bar (which has the same shape as using height of bar provided widths are equal). Histogram could also have relative frequency on $Y$ axis, in which case total area is 100%.
  e.g. cholesterol levels, p. 17

3.3 Frequency polygon

- Graph where $X$ observation is the midpoint of the class interval and the $Y$ observation is the frequency (or relative frequency).
  e.g. p. 18
- These are easy to superimpose.
  e.g. p. 19
- For cumulative frequency polygons, the $Y$ value is cumulative frequency (or cum rel freq.) and corresponding $X$ values are upper limits of class intervals.
  e.g. p. 20
- The cumulative frequency polygons are useful for finding percentiles, e.g. 50$^{th}$ percentile, say $P_{50}$, (also called the median) which is a number $\geq 50\%$ of observations and $\leq 50\%$ of observations.

3.4 Scatter plots (read)

3.5 Stem and leaf plots (added material)
• These plots are easy to create and are now widely used to get a quick look at the distribution of the data.

• Mechanics are:
  1) leading digit(s) give the stem
  2) trailing digit(s) give the leaf
  3) list range of possible stems in a column
  4) enter leaf for each observation
  5) typically, order leaves for each stem.

• Example

• Advantages:
  1) Stems usually easier to find than class intervals,
  2) leaves give numerical data lost in histograms, and
  3) could also have side-by-side plots to compare two data sets.

3.6 Box plots

• Box plots are now widely used in the literature to compare multiple data sets.

• Mechanics: (more precise definitions follow in next chapter)
  1) Find median, denoted $M$ (or $P_{50}$), which divides data in half,
  2) find first and third quantiles, denoted $Q_1$ and $Q_3$ (or $P_{25}$ and $P_{75}$), which divide the two halves again, giving quarters,
  3) form a box from $Q_1$ and $Q_3$, with $M$ in the middle,
  4) find interquantile range, $IQR = Q_3 - Q_1$,
  5) find adjacent values, which are most extreme observations that are not beyond 1.5 x IQR from the quantiles, $Q_1$ and $Q_3$,
  6) mark adjacent values with lines on box plot, and
  7) all data points beyond the adjacent values are outliers, and are marked by circles.

• Example
3.7 Two-way Scatter Plot (read)

3.8 Line Graphs (read)