1. Measures of central tendency

1.1 Mean

- Let \( x \) denote variable of interest, \( n \) the sample size, and \( x_i \) the \( i^{th} \) observation, \( i = 1, \ldots, n \).
- The sample mean (or \( x \)-bar) is
  \[
  \bar{x} = \frac{1}{n} \sum x_i
  \]
  e.g. p. 39. \( x = \text{FEV} \) (forced expiration volume)
  \[
  \bar{x} = \frac{(x_1 + \cdots + x_{13})}{13} = \frac{(2.30 + \cdots + 3.38)}{13} = 2.95
  \]
- If \( x \) is dichotomous, let 1 represent "success" and 0 a "failure".
  Then \( \bar{x} \) = proportion of successes, \( p \)
  e.g. p. 40. \( p = 0.615 \) male

Notes: 1) mean is sensitive to outliers
   2) the mean of a population is denoted \( \mu \).

1.2 Median

- Median, denoted \( M \), is the “middle” value in ordered data. Let data be ordered from low to high.
  If \( n \) is odd, \( M \) is the middle value, i.e. the \( (n+1)/2 \) largest value.
  If \( n \) is even, \( M \) is the average of the middle two,
  \[
  M = \left[ \frac{n}{2} \text{obs} + \left( \frac{n}{2} + 1 \right) \text{obs} \right] / 2
  \]
  e.g. ordered FEV data, \( n = 13 \).
  2.15, 2.25, 2.30, 2.60, 2.68, 2.75, 2.82, 2.85, 3.00, 3.38, 3.50, 4.02, 4.05
  \( M = \)
  - Median is resistant to outliers, i.e. it's "robust"

1.3 Mode

- Mode is the value that occurs most frequently. If there are two peak frequencies, the data are bimodal, if more its multimodal.
2. Measures of dispersion

2.1 Range

- Range is difference between the largest and the smallest observation.
  e.g. for FEV, range = 4.05 - 2.15 = 1.90 liters
  sulphur dioxide summary on p. 45

- It is very sensitive to extreme observations.

2.2 Interquartile range, IQR.

- Recall that quantiles divide the data into quarters. There are several different procedures for finding them. The book (P&G) defines the $k^{th}$ percentile, denoted $P_k$, as follows:

  1) order the data
  2) if $nk/100$ is an integer,
     $$P_k = [(nk/100)\, obs + (nk/100+1)\, obs]/2$$
  3) if $nk/100$ is not an integer, find integer $j$ just less than $nk/100$.
     $$P_k = (j+1)\, \text{observation}.$$ 

- The first and third quantiles are
  $$Q_1 = P_{25} \quad Q_3 = P_{75}.$$ 
  e.g. FEV data
  2.15, 2.25, 2.30, 2.60, 2.68, 2.75, 2.82, 2.85, 3.00, 3.38, 3.50, 4.02, 4.05
  for $k = 25$, $nk/100 = (13)(25)/100 = 3.25$
  hence $j = 3$ and $Q_1 = P_{25} = 4^{th} \, obs = 2.60$
  for $k = 75$, $nk/100 = 9.75$
  hence $j = 9$ and $Q_3 = P_{75} = 10^{th} \, obs = 3.38$.

- The interquartile range is
  $$\text{IQR} = Q_3 - Q_1.$$ 
  e.g.
  \[
  \text{IQR} = 3.38 - 2.60 = 0.78 \, \text{liters}.
  \]

- The IQR is an excellent descriptive tool, i.e. range of middle 50% of data. However, it is not widely used in statistical inference.
2.3 Variance and standard deviation

- Consider the 'deviation' of an observation from the mean, i.e. the $i^{th}$ deviation is

$$d_i = x_i - \bar{x}$$

We could show that $\sum d_i = 0$.

- Therefore, let’s use squared deviations. The variance, $s^2$, is defined to be the "average" squared deviation, i.e.

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

The reason for dividing by $(n-1)$ rather than $n$ will be clearer later. e.g. for FEV data.

- The formula above is often called the definitional formula. An equivalent computational formula given in most basic texts is

$$s^2 = \frac{\sum x^2 - n\bar{x}^2}{n-1}$$

or

$$s^2 = \frac{\sum x^2 - (\sum x)^2 / n}{n-1}$$

- Variance is in squared units. The standard deviation, $s$, is the square root of the variance.

  e.g. for FEV $s = \sqrt{0.39} = 0.62$ liters.

  The standard deviation will be given some intuitive interpretation in the next section.

- Notes:
  1) the variance and standard deviation for populations are denoted as $\sigma^2$ and $\sigma$.
  2) both measures are sensitive to outliers
  3) both $s^2$ and $s$ will be widely used subsequently in statistical inference.
3. Coefficient of variation

- It may not make sense to compare $s$ across samples from different populations. Such comparisons may be made using the coefficient of variation, defined as

$$CV = \frac{s}{\bar{x}} \times 100\%$$

This is also known as the "relative variability".

- The CV is not used for statistical inference. However, it's an easy measure for experimenters to understand, and is used frequently.

4. Grouped data (read)

- If data are grouped in a grouped frequency table, the midpoints, denoted $m_i$, of the class intervals are used to represent the individual points. Formulas for the grouped mean and variance are available.

5. Chebychev's inequality and the empirical rule

- When the data are roughly symmetric (i.e. not badly skewed) and unimodal, (which conditions are often true) the empirical rule states that:
  1) about 67% of obs are in interval $\bar{x} \pm s$
  2) about 95% of obs are in interval $\bar{x} \pm 2s$
  3) about 99.7% (i.e. almost all) of obs are in interval $\bar{x} \pm 3s$.

  e.g. for FEV, the interval $\bar{x} \pm s$ is $2.95 \pm 0.62$ or $(2.33, 3.57)$, which contains 8 of 13, or 62% of obs. The interval $\bar{x} \pm 2s$ or $(1.71, 4.19)$ contains 100% of data.

- A more general rule which applies to any distribution is the Chebychev inequality which states that, for $k > 1$, at least $(1 - k^{-2})$ of observations are within $k$ standard deviations of the mean.
e.g. let $k = 2$. Then at least $(1 - \frac{1}{4}) = .75$ of obs are in interval $\bar{x} \pm 2s$. For FEV data, the interval contains 100% of data, hence the rule is true in that case.

6. Further Applications (read. This has nice examples)