Chapter 2 – Describing and Summarizing Data

Types of Data

- **Qualitative or Categorical**
  - **Nominal (table, chart on p.23: race, major, blood type, etc.)**
  - **Ordinal (also table, chart: letter grades on p.26, etc.)**
- **Quantitative**
  - **Discrete (number of planets orbiting a distant star – could even be countably infinite)**
  - **Continuous (your age, not rounded off)**

Histograms, Bar Charts, Area and Probabilities

For bar charts (Fig. 2.1 on p.23), the bars are not contiguous (touching one another) nor do the areas of the strips have a meaning; rather, the *heights* of the rectangles are proportional to the frequency. For histograms (Fig. 2.2 on p.28 and Fig.2.3 on p.30), on the other hand, the bars are contiguous and the *areas* of the rectangles are proportional to the frequencies. As such, notice that the areas for a histogram sum to *one*. This has direct and important applications to *probability* (and percentages and proportions). When it comes to constructing a histogram, we first choose the classes, and sometimes it is wise that these *not* be of the same width (p.33). In these cases, we have some work to do to find the Relative Frequencies, which are then plotted in the histogram (see pp.33-34).

Students should also be familiar with Sections 2.7, 2.8, and 2.9.

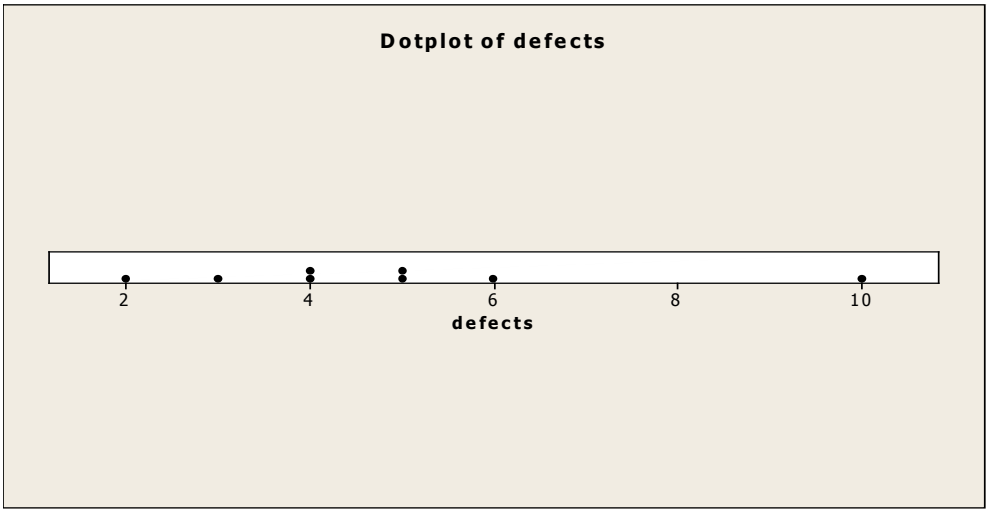
Example 2.4 (discrete) –

Original data: 4, 2, 4, 5, 10, 5, 3, 6

Ordered: 2, 3, 4, 4, 5, 5, 6, 10

Center: *mean* (average) is $(4 + 2 + \ldots + 6)/8 = 39/8 = 4.875$, and the *median* is 4.5 (the average of the 4^{th} and 5^{th} ordered entries); the mean is to the right of the median due to the *right skew* in the data.

Dotplot from Minitab:



We could *group* the data into the following table (p.33):

Class	Class Midpoint	Class Width	Frequency	Relative Frequency	Bar Height
0-2	1.0	3	1	0.125	0.04167
3-4	3.5	2	3	0.375	0.18750
5-6	5.5	2	3	0.375	0.18750
7-11	9.0	5	1	0.125	0.02500

The Bar Height is the 'relative density', and is such that the bar area is equal to the relative frequency. Thus,

Bar Height = Relative Frequency / Class Width

Notice that if we threw away the original data (not a good idea) and only had the above table, we'd obtain the *grouped mean* (pp. 43-4):
(0.125)*(1.0) + (0.375)*(3.5) +(0.375)*(5.5) +(0.125)*(9.0) = 4.625.
We would also find the *grouped median* of 4.5 (the average of 3.5 and 5.5, the 4^{th} and 5^{th} ordered grouped entries).

Dispersion – Variances and Standard Deviations and so on:

- **Sample Variance = s^2 =** $\dfrac{\sum_{j=1}^{n}(x_j-\bar{x})^2}{n-1} = \dfrac{\left(\sum_{j=1}^{n}x_j^{\,2}\right)-(\sum_{j=1}^{n}x_j)^2/n}{n-1}$
- **Population Variance = σ^2 =** $\dfrac{\sum_{j=1}^{N}(x_j-\mu)^2}{N}$
- **Sample Standard Deviation (SD) = s =** $\sqrt{s^2}$
- **Population Standard Deviation (SD) = σ =** $\sqrt{\sigma^2}$
- **Standard Error of the Mean =** $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$
- **Estimated Standard Error of the Mean =** $SE_{\bar{x}} = \dfrac{s}{\sqrt{n}}$
- **Standard Error of the Proportion =** $\sigma_{\pi} = \sqrt{\dfrac{\pi(1-\pi)}{n}}$
- **Estimated Standard Error of the Proportion =** $SE_{\pi} = \sqrt{\dfrac{p(1-p)}{n}}$
- **Analogous formulas exist (and are important) for grouped data (see the text).**

The SD evaluates how much variability there is in a *sample*; The Standard Error evaluates how much expected variability there is in the *estimated parameter*.

***Random Error Rule*: If an element x of a data set is chosen at random and you have to guess its value, your best guess is the mean $\bar{x}$. The difference between your guess and the true value of the element is called the *random error*. If you guess the mean, then the absolute value of your random error |x - $\bar{x}$ | will be roughly the standard deviation s or so.**

For any data set, the following *Empirical Rules* apply (p.49):

Approximately 68% of the measurements are in the interval [$\bar{x}$ - s, $\bar{x}$ + s],

approximately 95% of the measurements are in the interval [$\bar{x}$ - 2s, $\bar{x}$ + 2s],

approximately 99.7% of the measurements are in the interval [$\bar{x}$ - 3s, $\bar{x}$ + 3s]

Sometimes, this approximation can be a bit crude – depending upon the sample size and skewness of the sample – but it's surprisingly accurate.

For the above data (4, 2, 4, 5, 10, 5, 3, 6), $\sum_{j=1}^{n} x_j = 39$, $\sum_{j=1}^{n} x_j^{\,2} = 231$, so

$s^2 = \dfrac{(231) - (39)^2/8}{7} = 5.8392857 = 2.4164614^2$.

Thus, [$\bar{x}$ - s, $\bar{x}$ + s] = [2.46,7.29], and this interval contains 6/8 = 75% of the original data points. Not too far from 68%.

Next, [$\bar{x}$ - 2s, $\bar{x}$ + 2s] = [0.042,9.71], and this interval contains 7/8 = 87.5% of the original data points. Not too far from 95%.

Finally, [$\bar{x}$ - 3s, $\bar{x}$ + 3s] = [-2.37,12.12], and this interval contains 8/8 = 100% of the original data points. Not too far from 99.7%.

Homework #1 (on Chapters 1 & 2) is posted on our Class Website, and is due next Thursday. Next Tuesday, we'll discuss Chapter 3: Probability.