In addition to simply describing the frequency distribution of a set of data we can also make use of the following.

1. The **EMPIRICAL RULE**: If the data set displays a frequency distribution that is somewhat symmetric, unimodal, without much skew, and approximately equal length tails then the following is true:

   a) about 68% of the observations are within 1 standard deviation of the mean, i.e. fall between

   $$\bar{x} - s \quad and \quad \bar{x} + s = \bar{x} \pm s$$

   b) about 95% of the observations are within 2 standard deviations of the mean, i.e. fall between

   $$\bar{x} - 2s \quad and \quad \bar{x} + 2s = \bar{x} \pm 2s$$

   c) > 99% of the observations are within 3 standard deviations of the mean, i.e. fall between

   $$\bar{x} - 3s \quad and \quad \bar{x} + 3s = \bar{x} \pm 3s$$
The Empirical Rule is useful for summarizing datasets that display approximately Normal shapes.

**EXAMPLE:** 200 observations of the cranial capacity of skulls of modern male Caucasians (in$^3$)

![Quantiles and Moments](image)

<table>
<thead>
<tr>
<th>Cranial</th>
<th>Moments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantiles</td>
<td>Moments</td>
</tr>
<tr>
<td>Mean</td>
<td>86.8050</td>
</tr>
<tr>
<td>Std Dev</td>
<td>2.9861</td>
</tr>
<tr>
<td>Std Error Mean</td>
<td>0.2112</td>
</tr>
<tr>
<td>Upper 95% Mean</td>
<td>87.2214</td>
</tr>
<tr>
<td>Lower 95% Mean</td>
<td>86.3887</td>
</tr>
<tr>
<td>N</td>
<td>200.0000</td>
</tr>
<tr>
<td>Sum Weights</td>
<td>200.0000</td>
</tr>
</tbody>
</table>

What is the approximate proportion of skulls from modern male Caucasians with cranial capacities between 84 and 90 in$^3$?

$84 \approx \bar{x} - s = 86.81 - 2.98$ and $90 \approx \bar{x} + s = 86.81 + 2.98$, so approximately 68% of the observations should fall between 84 and 90 in$^3$.

Similarly, approximately 95% of the observations fall between 81.5 and 92.5 in$^3$.

What is the value of the 97.5$^{th}$ percentile, i.e. the value above 97.5% of the observations?
The middle 95% of the data falls between 81.5 and 92.5 in\(^3\). That leaves the remaining 5% outside of that range. If the distribution is symmetric, 2.5% should fall below 81.5 in\(^3\) and 2.5% should fall above 92.5 in\(^3\). Hence, 92.5 in\(^3\) is the approximate 97.5\(^{th}\) percentile of the data.

\[ \bar{x} = 86.8 \quad \bar{x} + 2s = 92.5 \]

2. **PERCENTILES (QUANTILES)**

**Defn:** For any particular number, \(r\), between 1 and 100, the \(r^{th}\) PERCENTILE is the value such that \(r\) percent of the observations in the dataset fall at or below that value.

In a smoothed histogram, the \(r^{th}\) percentile is the value that divides the total area under the smoothed curve into 2
parts: to the left is $r$ percent of the area and to the right is $(100-r)$ percent of the area.

3. **MEASURES OF RELATIVE STANDING (Z-SCORES)**

**Defn:** The Z-SCORE for a particular observation ($x_i$) in a dataset is

$$z = \frac{x_i - \text{mean}}{\text{standard deviation}}.$$  

It tells us how many standard deviations the observation is from its mean. Z-scores are called standardized scores.
Z-scores are useful for comparing different observations within the same dataset as well as different observations in different datasets!!

**EXAMPLE**: Biodiversity in Caves. In counties in the coterminous U.S.A. for which at least one subterranean species is known: Terrestrial species average 5.2 species per county with a standard deviation of 3.3. Aquatics species average 5.8 species with a s.d. of 4.6.

A county is extensively studied and 7 terrestrial species and 7 aquatic species were found. How unusual are these findings?
EXAMPLE Is a man who is 6’2” taller relative to other men than a women who is 5’11” relative to other women?

Intuition?

Now, suppose
For adult men: $\mu_M = 69$ inches, $\sigma_M = 2.4$ inches.
For adult women: $\mu_F = 65$ inches, $\sigma_F = 2.5$ inches.