For starters - pick up the file pebmass.PDW from the H:Drive.

Put it on your G:/Drive and open this sheet in PsiPlot.

Rock property assessment
What attributes might you use to describe a rock?....
.... grain size, porosity, composition (percent quartz, orthoclase, ...), dip, etc.

How are these different attributes obtained?

How reliable are the values that are reported?
Data Collection
Concepts and terminology -

Specimen - a part of a whole or one individual of a group - an observation or measurement.
Sample - several specimens (measurements)
Population - all members of the group, all possible specimens from the group

The attributes derived from the sample are referred to as statistics.
The attributes derived from the entire population of specimens are referred to as parameters.

Consider your grade in a class -
Let's say that your semester grade is based on the following 4 test scores.
85, 80, 70 and 95
What is your grade for the semester?
Your grade is the average of these 4 test scores or 82.5
The average is often used to represent the most likely value to be encountered in a sample or population.

**Is the average grade of 82.5 a statistic or a parameter?**

Since the entire population of grades for the student consists of just those 4 test scores, that average is a **parameter**.

---

**On page 113 (Chapter 7) Waltham lists the masses (in grams) of 100 pebbles taken from a beach.**

The average mass of these pebbles is **350.18 grams**

This average is a **... statistic**

---

**Computation of the mean or average**

\[
\overline{m} = \frac{1}{N} \sum_{i=1}^{N} m_i
\]

In this equation
- \(m_i\) is the mass of pebble \(i\)
- \(N\) is the total number of specimens
- \(i\) ranges from 1 to \(N\)
- \(\overline{m}\) = the average mass of all the pebbles in the sample.
If we draw smaller samples at random from our original sample of 100 specimens and then compute their averages, we begin to appreciate that the statistical average is only an estimate of the population average. Recall that the mean estimated from 100 samples was 350.18.

<table>
<thead>
<tr>
<th>Specimen</th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample 3</th>
<th>Sample 4</th>
<th>Sample 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>352</td>
<td>359</td>
<td>383</td>
<td>394</td>
<td>401</td>
</tr>
<tr>
<td>2</td>
<td>374</td>
<td>330</td>
<td>399</td>
<td>407</td>
<td>422</td>
</tr>
<tr>
<td>3</td>
<td>323</td>
<td>339</td>
<td>333</td>
<td>350</td>
<td>369</td>
</tr>
<tr>
<td>4</td>
<td>372</td>
<td>365</td>
<td>369</td>
<td>363</td>
<td>462</td>
</tr>
<tr>
<td>5</td>
<td>342</td>
<td>355</td>
<td>342</td>
<td>354</td>
<td>458</td>
</tr>
<tr>
<td>6</td>
<td>242</td>
<td>305</td>
<td>253</td>
<td>348</td>
<td>439</td>
</tr>
<tr>
<td>7</td>
<td>434</td>
<td>356</td>
<td>392</td>
<td>375</td>
<td>379</td>
</tr>
<tr>
<td>8</td>
<td>454</td>
<td>393</td>
<td>311</td>
<td>341</td>
<td>433</td>
</tr>
<tr>
<td>9</td>
<td>345</td>
<td>314</td>
<td>265</td>
<td>435</td>
<td>370</td>
</tr>
</tbody>
</table>

Average: $\bar{x} = 355.1\text{g}$

Other measures of the most common value in a population include the median and the mode. If we sort measured values (for example pebble mass) in increasing order, from the lightest pebble to the heaviest and look at the mass of the center pebble, that value is the median mass.

The Median

In the sample (1,2,3,4,5) - 3 is the center or median value of the sample.

In the sample 1,2,3,4 we have an even number of samples and in this case the median is taken as the average of the middle two values or 2.5.
At right, the pebble mass has been sorted in ascending order from the lightest to heaviest pebbles in the sample.

There are an even number of specimens in this sample, so the median must be determined from the average of specimens 50 and 51

\[ \text{Median} = \frac{352 + 353}{2} = 352.5 \text{ g} \]

The mode is the value that occurs most frequently. For example, in the following sample, \(1, 2, 2, 3, 3, 3, 4, 5\), 3 occurs most frequently and would be the mode of this sample.

In the sample of rock masses, 283, 331, 338, 355 and 403 all occur 3 times. We cannot define a single mode. The data are poly-modal

Refer to your handout and use PsiPlot to generate descriptive statistics of the pebble masses
The histogram - a graphical display of the distribution of values - in this case - the pebble mass data from Waltham.

The appearance of the histogram will vary depending on the specified range you use to subdivide the sample values.

You might group your samples into 25 gram ranges extending from 226 through 250, 251 through 275 and so on.

Histogram constructed using 50 gram intervals -

The median is 352.5 grams and the mean 350.18 grams.
Based on either of these representations of mass distribution, would you say that this beach deposit is well sorted?

Variation of pebble mass

The histogram reveals a mass distribution characterized by a certain range of values - 224 and 454 grams - a 230 gram range.

Refer to your handout and construct a histogram of pebble masses.
The distribution of masses from beach B is similar in shape to that from our first beach, but its range is much smaller. Both samples have the same mean.

The pebbles on beach B are much better sorted than those on beach A. There is not as much variation in pebble mass on beach B.

Sample B is better sorted than our first sample. The mass distribution from beach C has the same range as distribution B and nearly the same mean (348 grams), but its shape is very different. This distribution is much more irregularly distributed across the range - i.e. there's not a preferred value.

So we need some other measure of the distribution to define the qualities of well and poorly sorted more quantitatively.
A parameter that describes the spread or dispersion in the values of a population is its variance.

\[ \sigma^2 = \langle (\text{mass} - \text{average mass})^2 \rangle \]

Note the brackets indicate that we are taking the mean of this quantity.

\( \sigma^2 \) is used to represent the population variance.

The equivalent statistic used to quantify the spread or dispersion in the values of a sample is the sample variance.

The sample variance is computed in the following way:

\[ s^2 = \frac{1}{N} \left( \sum_{i=1}^{N} (m_i - \bar{m})^2 \right) \]

\( s^2 \) represents the sample variance.

The standard deviation of the sample values is a statistic that is also often used to describe the degree of variation in values of a sample.

\[ s = \sqrt{\frac{1}{N} \left( \sum_{i=1}^{N} (m_i - \bar{m})^2 \right)} \]

The standard deviation is just the square root of the variance.
Standard deviation = 32.37  Standard deviation = 23.83
While these two distributions have similar means, the one on the right is better sorted and the standard deviation - not the range reveals this difference.

The standard deviation describes geological differences in the sample that are not apparent in the mean or range.

One sample is better sorted - has smaller standard deviation than the other, which is less sorted and has higher standard deviation.

The sample variance is considered to be an underestimate of the population variance or actual variance of the parent population.

To compensate for that, and to obtain an estimate of the population variance which is considered more accurate, the sample variance is corrected to form an “unbiased” estimate of the population variance.
The following equation is used to compute the unbiased estimate of the population variance -

\[ s^2 = \frac{N}{N-1} \hat{s}^2 \]

or equivalently

\[ s = \sqrt{\frac{1}{N-1} \sum (m_i - \overline{m})^2} \]

Refer back to worksheet generated descriptive statistics for the pebble masses and note their standard deviation and variance.

**Probability**

Probability can be thought of as describing the fraction of the time that a specific value or range of values is observed out of the total number of observations or specimens in a sample.

Just as with the average and the standard deviation, there is a distinction between the probabilities observed in a sample and those of the parent population.
But the idea is that the probabilities observed in the sample give one an estimate of the probability or likelihood of occurrence in the parent population.

Probabilities are used to make predictions.

The probability that a specimen will have a certain value or range of values is expressed as the fraction of the total specimens that have that value or range of values.

**Spreadsheet interactions**

Compute the frequency of occurrence for each range of masses used to construct the histogram.

Note that the probability of individual occurrences may vary. The probability that you will get a pebble with a mass of 242 grams is one in a hundred.

The probability of picking up a pebble that has a 283 gram mass is 3 in 100 (0.03), etc.
The probability that a pebble will have a mass somewhere in the range 300 to 350 will be 35 out of 100 or 0.35.

These probabilities are the probabilities that individual values in a sample will fall in a 50 gram range, and thus represent the integral of individual probability over the range.
Probability distributions with a shape similar to the above example are quite common. They are nearly symmetrical and values near the average are more probable than those further from the average.

Distributions of this type are often referred to as Gaussian or normal distributions.

\[ p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \quad \text{Population} \]

And we can estimate the probability distribution of the parent population from statistical estimates of the mean and variance.

\[ p(x) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} e^{-(x-\overline{x})^2/2\hat{\sigma}^2} \quad \text{Statistic} \]

This expression is often simplified by substituting \( Z \) for \((x-\overline{x})/\sigma\). \( Z \) is referred to as the standardized variable or the standard normal deviate, and the Gaussian distribution is rewritten as -

\[ p(x) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} e^{-z^2/2} \]

Note that \((x-\overline{x})/\sigma\) represents the number of standard deviations the value \( x \) is from the mean value.
Thus a value $x$ corresponding to a $z$ of 2 would be located two standard deviations from the mean in the positive direction.

Using the pebble mass statistics, $<x>=350.18$ and $s=48$, the $z$ of 2 implies $x = 446$ grams.

Before you leave - hand in the histogram you constructed with your name in the title

For Tuesday

Carefully read through sections 7.5 through 7.8