If you create a scatterplot of a bunch of data points, the points you plot may or may not form a straight line.

Usually they won't, of course.

The Pearson correlation coefficient, affectionately known as 'r', gives a way of measuring just how close to a straight line the data points are.

Here's the formula: $r = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_i - \bar{x})}{(s_x)} \cdot \frac{(y_i - \bar{y})}{(s_y)}$

Here's how to think about it:

r is the mean of the products of the z-scores for X and Y.

If L is a list of points, then X is the list of x-coordinates and
Y is the list of y-coordinates.

z-scores are coming up on the next page.

After that, you'll have all you need for the formula for r!

A z-score is a deviation score divided by the standard deviation.

It measures how many standard deviations a score is from the mean.

$(x_i - \bar{x})$ and $(y_i - \bar{y})$ are deviation scores.

s_x and s_y are the standard deviations of X and Y.

$\frac{(x_i - \bar{x})}{(s_x)}$ is an X z-score, and $\frac{(y_i - \bar{y})}{(s_y)}$ is a Y z-score.

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_i - \bar{x})}{(s_x)} \cdot \frac{(y_i - \bar{y})}{(s_y)}$$

The formula is adding up the products of the X and Y z-scores and then dividing by n - 1, because we're dealing with sample data.

Nevertheless, we are still finding a kind of mean.

If we were dealing with population data, we'd divide by n.

After you perform all these calculations to find r, you will get some real number between -1 and +1.

In other words, $-1 \le r(L) \le +1$.

Values close to ±1 indicate strong linear correlation.

Values close to 0 indicate weak linear correlation.

```
def X(L): return [x for (x, y) in L]

def Y(L): return [y for (x, y) in L]

def zscores(L): return [deviation/stdev(L) for deviation in deviations(L)]

def r(L):
    Z = [zx*zy for (zx, zy) in zip(zscores(X(L)), zscores(Y(L)))]
    if sample: return adjusted_mean(Z)
    else: return mean(Z)
```

'zip' in Python is a function that 'zips' two lists together to create a list of ordered pairs.

'Z' is a list of the products of the z-scores for X and Y.