Pitfalls of using absolute numbers in correlations and t-tests

You're working on Census data with integrated deaths from the Influenza data set . While working on the hypothesis that age is a deciding factor in mortality you decide to run a correlation on the columns that represents the older population and the column with deaths. The correlation test returns a very coefficient. Hurray! You've basically proven your hypothesis. But have you?

Let's look at some aspects of the analysis here and review it from a more critical perspective:

- You should know that a correlation coefficient above > 0.8 is extremely suspicious. Anything above this is almost a perfect correlation, so whenever you see a value like this in a correlation test, start digging because something is wrong. In the Influenza project scenario, the coefficient is so high because naturally, the larger the population, the larger the death count. This goes without saying, it's nothing we don't know.

- The main culprit here is the use of absolute numbers. Using them is absolutely wrong, because the states are of different sizes, hence we can't compare them. That's where the use of normalised numbers comes in. This is important because when you're using the absolute numbers you have one mean and variance and when you use the normalised numbers, they change completely! And essentially you are comparing the states against each other because the numbers are in the same vector (column). So, Excel calculates the mean and variance in these vectors and bases the correlation calculation on that. You can see for yourself how much they vary if you simply calculate the mean and variance for the columns with absolute numbers and for the columns with normalised numbers. You'll see that the variance in the absolute number columns is much much greater - because the numbers are very very different due to the different state sizes. And that's really important!

- To better understand why you're ultimately comparing the states, it's a very good idea to delve deeper into the formula of the Pearson correlation test. Here's a great resource with examples. The Pearson formula is built upon the mean - in the upper part of the equation the base is taking the mean out of _every value_ in the vector (column). So naturally, when using absolute numbers, we'll end up with a mean that not representative at all! Can you say that the mean for 65+ population for Alaska is the same for Maine? Definitely not. That's why it's vital to use normalised numbers when we're handling entities that have different sizes, otherwise we're distorting the result big time.

What to do in order to fix your test:

- You need to create the relevant columns with normalised numbers - one for % of vulnerable population and one for death rate (all deaths/total population)

- Recalculate the correlation and fit the new interpretation.

Please take the time to go deep into the concept of the relationship between the mean and the correlation formula. As well as the use of normalised numbers when using normalised numbers as opposed to absolute ones.