Questions for Discussion

1.1

> The plots above were made by simulating data from Gaussian distributions. Describe some pairs of real world quantities that might be a) highly correlated, b) highly anti-correlated, c) uncorrelated, d) measured in the same units, but with a much larger variance in x than in y.

- A: A person's income and the value of their home might be highly correlated.
- B: The (human) population in a region and the number of species of flora and fauna that can be found there might be highly anti-correlated.
- C: The color of someone's hair and their IQ might be uncorrelated.
- D: The heights of humans and dinosaurs can be measured in the same units, but the heights of dinosaurs may have a much larger variance than that of humans.

1.2

> The variances and covariances are nice summary statistics for distributions of data. For Gaussian distributions they pretty much capture all the relevant information (i.e., with the means, variances and covariances for a set of data points, you pretty much know exactly what the scatter plots look like). However, for real world data they might leave off some very important information.
>
> Imagine distributions of points shaped like the letters U or V or M or W. What would you expect the covariance to be in each case? (Trick question). Why might that make using the covariance as a summary statistic problematic?

I'd expect the covariance of a scatter plot in the shapes of U, V, M, or W to be close to zero because—although the distribution is far from random—there isn't a clear *linear* relationship, and covariance measures the strength of *linear* relationships. This means that a lot of important information is missing from the covariance and that we shouldn't assume we know everything about a distribution just from its summary statistics.

2.1

> What do you see as the advantage/disadvantage of using the correlation coefficient versus the covariance?

The advantage of using the correlation coefficient is that it's normalized to fall within the range -1 to 1 (inclusive) which can make it easier to interpret and use as a basis of comparison with other distributions. The disadvantage is that you lose any notion of any absolute measure of spread, which can be more useful in certain situations.

3.1

> Do you think that the small correlation we found in the previous cell is going to be statistically significant? Why or why not? Does the contrast between the last two plots affect your opinion? What about the difference between the correlation values?

No, I don't think the small correlation from the previous cell is going to be statistically significant because it's pretty obvious that there's a strong linear relationship between the observed and expected counts (plot 1) and a lack thereof in the residuals (plot 2). The difference between the correlation values are dramatic (i.e., 0.998 v. -0.013).