The key factors affecting the quality of an inference you might make relate to violations in our thinking about the sampling process and the model for the population. Obviously, if we cannot coherently define the population, then any "inference" that we make to the population will be similarly vaguely defined.

A violation of our understanding of how the sampling process worked would result in our having collected data that did not represent the population in the way that we thought it would. This would affect our inference in that the inference we would make would apply not to the entire population, but to a specific selection of the population. This phenomenon is sometimes referred to as **selection bias**because the quantities that you estimate are biased toward the selection of the population that you *did* sample.

A violation of the model that we posit for the population could result in us estimating the wrong relationship between features of the population or underestimating the uncertainty of our estimates. For example, if it's true that penguins can influence what color hats other penguins wear, then that would violate the assumption of independence between penguins. This would result in an increase in the uncertainty of any estimates that we make from the data. In general, dependence between units in a population reduce the "effective sample size" of your dataset because the units you observe are not truly independent of each other and do not represent independent bits of information.

A final reason for a difference between our estimate from data and the truth in the population is **sampling variability**. Because we randomly sampled penguins from the population, it's likely that if we were to conduct the experiment again and sample another three penguins, we would get a different estimate of the number of penguins with turquoise hats, simply due to random variation in the sampling process. This would occur even if our description of the sampling process were accurate and our model for the population were perfect.

In most cases, differences between what we can estimate with data and what the truth is in the population can be explained by a combination of all three factors. How big a role each plays in a given problem can be difficult to determine sometimes due to a lack of information, but it is usually worth putting some thought into each one of these factors and deciding which might be playing a dominant role. That way, one may be able to correct the problem, for example, in future studies or experiments.

NOTE: Part of this reading was taken from *The Art of Data Science* by Peng & Matsui.