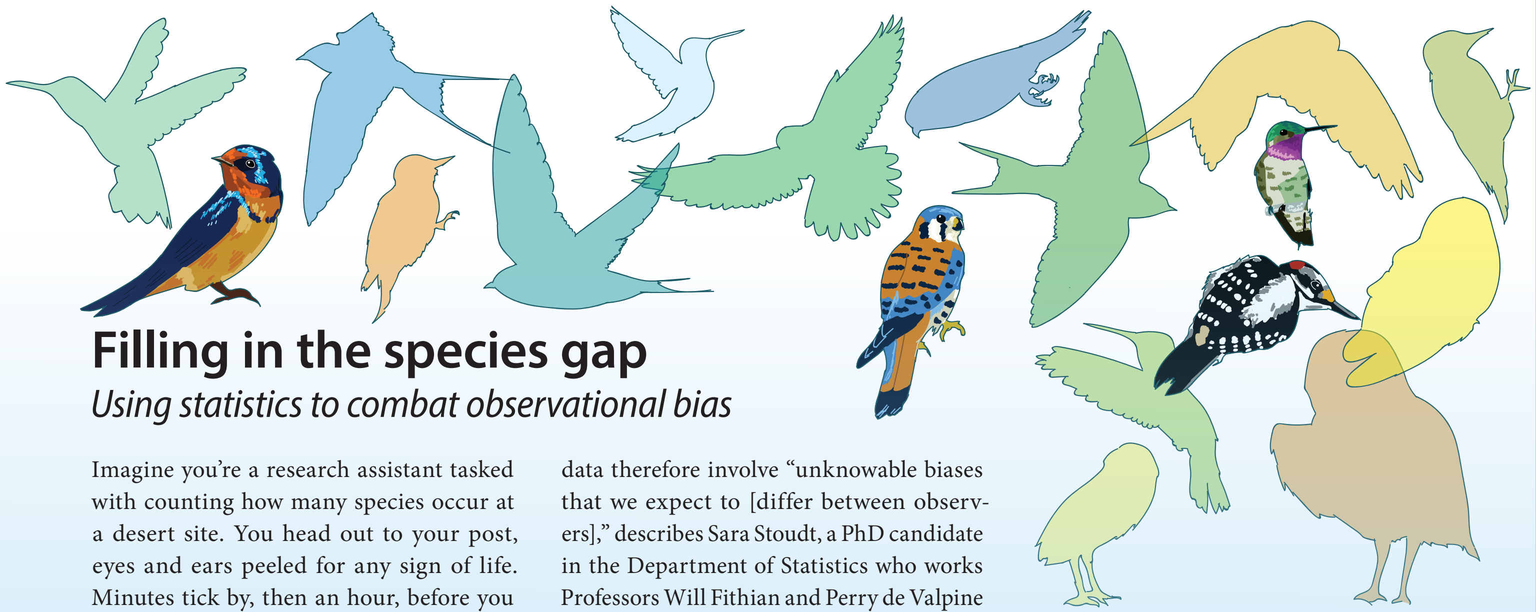

Filling in the species gap

Using statistics to combat observational bias

Imagine you're a research assistant tasked with counting how many species occur at a desert site. You head out to your post, eyes and ears peeled for any sign of life. Minutes tick by, then an hour, before you finally spot a lone American kestrel soaring overhead. "One species", you record, before moving on. Unfortunately, you overlooked an inconspicuous lesser nighthawk nesting motionless on the ground nearby, skewing your record of desert wildlife.

Accurately estimating species diversity is essential to assessing ecosystem health, natural resources, and the impacts of climate change. It may even seem like a simple task on the surface. What might be surprising, however, is that it requires solving a slew of statistical conundrums due to the complex sources of bias involved when humans try to observe organisms in the wild. "One of the things we can always say is happening is detection error," says Kelly Iknayan, a postdoc in the Department of Environmental Science, Policy, and Management working with Professor Steven Beissinger to assess how climate change has altered avian diversity in California deserts. Beyond finite observation time, our inherent biases mean that we notice large, conspicuous species at the expense of small, more camouflaged species. Those that hide from humans or are highly mobile are also prone to be undercounted. The collected data therefore involve "unknowable biases that we expect to [differ between observers]," describes Sara Stoudt, a PhD candidate in the Department of Statistics who works Professors Will Fithian and Perry de Valpine to study statistical ambiguities in ecology.

To correct for these biases, Iknayan and Stoudt develop statistical models of how the true quantities they care about—such as how often kestrels occupy a particular desert site—are related to biased human observations at that site. By estimating the probability that a species will be detected at a site given that it is known to exist there, these models provide estimates of species diversity that compensate for missed observations. Additionally, for a group of related species, "we might expect them to have shared characteristics about how detectable they are so we can use information from multiple species to help us model that behavior, " explains Iknayan. Letting detection probabilities inform each other in this way, a strategy known as borrowing strength, can be particularly advantageous when data are limited for some species of interest. Some finesse is required with this modeling choice, however, as you need expertise on which species behave similarly enough to be statistically informative of each other. Borrowing strength may not yield the best estimates for species with disparate behavioral patterns, which are better modeled individually.

An even bigger challenge is that some species will be missed altogether, which guarantees that estimates of diversity will be too low. To compensate, Iknayan explains that you can look at how many new species are encountered with increasing amounts of observational effort. That number will plateau at some point, and comparing that plateau to how many species you observed gives an estimate of how many were missed.

Even studying a single species demands nuanced modeling decisions. Stoudt has been investigating the issue of model identifiability, which arises when two different hypotheses—such as whether kestrels occur at a particular desert site five percent versus fifteen percent of the time—are equally consistent with the data and therefore cannot be distinguished. As Stoudt explains, this is not an issue of quantity of data: "If I gave you infinite data collected in the way you're collecting it, could you [tell the hypotheses apart]?" If you can, your model is identifiable, meaning your hypotheses can be proven or refuted. But if your model isn't identifiable, how can you make accurate scientific conclusions? Stoudt has found that there are actually varying degrees of identifiability, which arise from incorporating different kinds of assumptions into single-species models. Her work develops guidelines for ecologists on how to collect data to achieve the strongest form of identifiability.

The challenges Stoudt and Iknayan face are characteristic of modern ecology, which is driven by increasingly large and complex datasets. Ecologists now need just as much ingenuity and discernment in interpreting data as in designing experiments. With the advent of citizen science platforms like eBird and iNaturalist, which enable the general public to record and share species observations of their own, researchers monitoring biodiversity now have access to data collected on a larger scale—and in a completely new way—than ever before. With it will come both new opportunities and statistical challenges.

Clara Wong-Fannjiang *is a graduate student in electrical engineering and computer sciences.*

DESIGN: ALISON NGUYEN