**Preliminary Exam Questions: Dr. Dinsmore**

**Marty Simonson**

**April 9, 2020**

**3. Below is a 2x2 contingency table showing the observed breeding status (B=breeding, NB = non-breeding) for a hypothetical fish in a given year. Furthermore, there are two classes of fish in each breeding category: fish attempting to breed for the first time (First), and experienced breeders (Exp).**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | B | NB | Total | % B |
| First | **110** | **45** | 155 | 71% |
| Exp | **92** | **63** | 155 | 59% |
| Total | 202 | 108 | 310 | 65% |

1. **Create a partition of the above table into two possible 2x2 tables (which must sum to the numbers in the above table) that constitute a good example of Simpson’s Paradox. (note: there is no unique solution to this).**

|  |  |  |  |
| --- | --- | --- | --- |
| **TABLE 1** | B | NB | %B |
| First | 108 | 43 | 72% |
| Exp | 12 | 3 | 80% |

|  |  |  |  |
| --- | --- | --- | --- |
| **TABLE 2** | B | NB | %B |
| First | 2 | 2 | 50% |
| Exp | 80 | 60 | 57% |

The two tables are set up such that they could be two independent samples of the same population, which together sum to the bolded part of the table above. Simpson’s paradox is demonstrated here because in the partitioned tables the experienced breeders are observed in the breeding state more often than first time breeders. When the data are aggregated, however, we see that, overall, the first-time breeders are observed in the breeding state more often. In Table 1, we observe that 72% of the first-time breeders fall into the breeding status, while 80% of experienced breeders fall into the breeding category, and in Table 2, similarly, shows that 50% of first-time breeders are in the breeding state whereas 57% of the experienced breeders are observed in the breeding state. This appears to be a factor of sample size within the first-time and experienced breeder groups!

1. **State Simpson’s Paradox in its general form. Why does it raise concern regarding the analysis of observational ecological data, even under a model-based data analysis?**

Simpson’s paradox states that the presence or absence of a trend, when data is pooled, disappears or arises after data is partitioned into groups, which is essentially a manifestation of confounding factors and human perspectives. This is concerning in observational and ecological data, because stratification of data by groups can be entirely subjective to human bias (e.g., “healthy” vs. “sick”) and equal sample sizes among groups can be difficult to obtain when count data that is at the mercy of relative abundances and detection abilities. Testing pooled analysis vs. stratified analysis is important to ensure that a conclusion or result from pooled data is not contrary or different than a closer look at the data split among groups.

1. **Why are we nearly immune to Simpson’s Paradox when using a well-designed and highly replicated experiment?**

Simpson’s Paradox is handled many ways using current statistical methods. The best approach comes in well-designed and well-replicated experiments or observational studies. The fundamentals of frequentist statistics are aligned toward reduced Type I error and (ideally) fewer false positive conclusions. Equal sample sizes are critical to making appropriate comparisons between groups, often setting a minimum or maximum number of individuals measured for a trait will improve comparisons, assuming the individuals are randomly drawn from within their sub-population.

Statistical design that is robust to Simpson’s Paradox can include a priori groupings based on quantifiable traits (e.g., length class) or clearly distinct taxa classifications. For example, a landscape-level study on how forests in the Great Plains serve as bird habitat might be interested in looking more at functional groups of species and thus aggregate data into larger groups of ground- and tree-nesting birds, regardless of species. Randomization of treatment within a system can help avoid uneven sample size from data collection. Bayesian network analysis is another powerful tool to use literature and expert knowledge to identify nodes that are data gaps and subsequently develop a robust sample design ahead of collecting data.

Statistics is the story that encodes causal relationships and explains natural phenomena, and Simpson’s Paradox can confound those relationships. However, knowledge about the study organism or system, careful consideration of confounding factors, and adequate replication will reduce or eliminate the effect of Simpson’s Paradox.