Extra exercise 23

Planning of sampling to detect disease, or freedom of disease. The following information is provided:

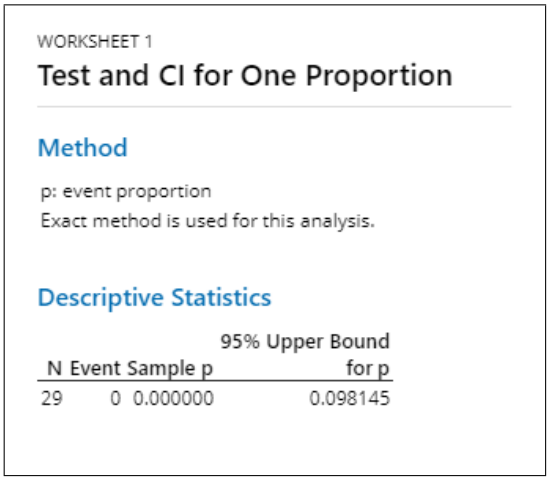
- targeted minimal prevalence $p_{\min} = 0.1$,
- confidence level 95%, corresponding to error level $\alpha = 0.05$,
- finite population size $N = 1000$.

Part 1)
The sample size calculations based on binomial and hypergeometric distributions, corresponding to sampling with replacement (or an assumed infinite population size) and sampling without replacement from a finite population use the formulas from the lecture (slide 12L−11):

$$\text{binomial:} \quad n = \frac{\ln(\alpha)}{\ln(1-p_{\min})} = \frac{\ln(0.05)}{\ln(1-0.1)} = 28.4, \quad \text{take } n = 29,$$

$$\text{binomial:} \quad n = (1-\alpha^{1/D})\left(N - \frac{D-1}{2}\right) = (1-0.05^{1/100}) \cdot \left(1000 - \frac{100-1}{2}\right) = 28.05, \quad \text{take } n = 29,$$

where we in the second calculation also used $D = N \cdot p_{\min} = 1000 \cdot 0.1 = 100$. Both calculations indicate a required sample size of $n = 29$. Clearly the approximation by an infinite population has minimal implications for this situation. Because both formulas are easily available, there seems little point in bothering about a guideline for when that approximation may be used. The Freecalc calculator gives the same result, in both settings. We can confirm our binomial calculation by analyzing one proportion with all outcomes negative in the `1 Proportion` menu with an exact binomial confidence interval:


WORKSHEET 1

Test and CI for One Proportion

Method

p: event proportion

Exact method is used for this analysis.

Descriptive Statistics

N	Event	Sample p	95% Upper Bound for p
29	0	0.000000	0.098145

The upper bound of the 95% confidence interval is at 0.098, just below the desired $p_{\min} = 0.1$. The reason it is not exactly 0.10 is the discrete nature of the distribution — we cannot observe 28.4 negative fish.

Part 2)
Whether the calculation needs to be adjusted to account for the distribution of fish in cages, depends essentially on the assumptions one is willing to make about the disease spread dynamics. If the disease was able to easily spread from one cage to another, then the separation into cages is of no real importance for the sample size calculation. In practice, one would probably still want to take

samples from several cages, as a precaution against making a statement based on a single cage when that, for whatever reason, might not be representative for the entire population.

It it was considered possible/plausible that the disease could exist in one cage (at a prevalence of at least $p_{\min}$) but not in others, the situation is different and may potentially require each cage to be tested individually. Apart from each population size dropping down to 100, taking the required sample size from the hypergeometric model down to $n = 24.7$ (when inserting $N = 100$ and $D = 10$ into the formula above), the calculation is the same. In practice, it may be unreasonably conservative to employ a testing strategy, where each cage is considered as a separate population. What is known about the origin of the disease potentially becomes crucial. For example, if it was considered to spread through water, then one might want to focus on the cages with the most extreme locations within the site. On the other hand, for a vertically transmitted disease the origin of the fish becomes the key issue, and one might want to test subpopulations from different origins.

Another important consideration (for both parts) is which fish to sample. It is not so easy to randomly sample fish from a population held in cages, and perhaps it is not desirable after all because for diseases that affect fish survival it may be better to sample fish that are dead or moribond. It will however be more complicated to translate results for such targeted sampling into confidence statements about the prevalence in the population, unless one simply adopts the conservative notion that the targeted sampling cannot reduce the chance of finding diseased fish, and then goes by the (above) statements corresponding to random sampling.