

Tagging Effort Analysis

Myfanwy Johnston

July 27, 2016

Power Analysis - or, why we're not doing one

The reason to do a power analysis is to determine whether your sample size will give you a decent chance of detecting an “effect” in an experiment - that is, a decent chance of detecting some minimum deviation from the null hypothesis and the alternative hypothesis. In our case, for example, we'd like to tag enough fish so that a good number of them continue upstream after being tagged. With a normal experimental design, if we assume the true proportion of fish that head upstream is what we've observed in the past (between 10-35%), we could do an exact binomial test and get the sample size we need to make sure we detect 50% or more (an arbitrary number) of our fish heading upstream. According to a quick exact binomial goodness-of-fit power analysis, this would give us a target sample size of between 77 and 140; not helpful.

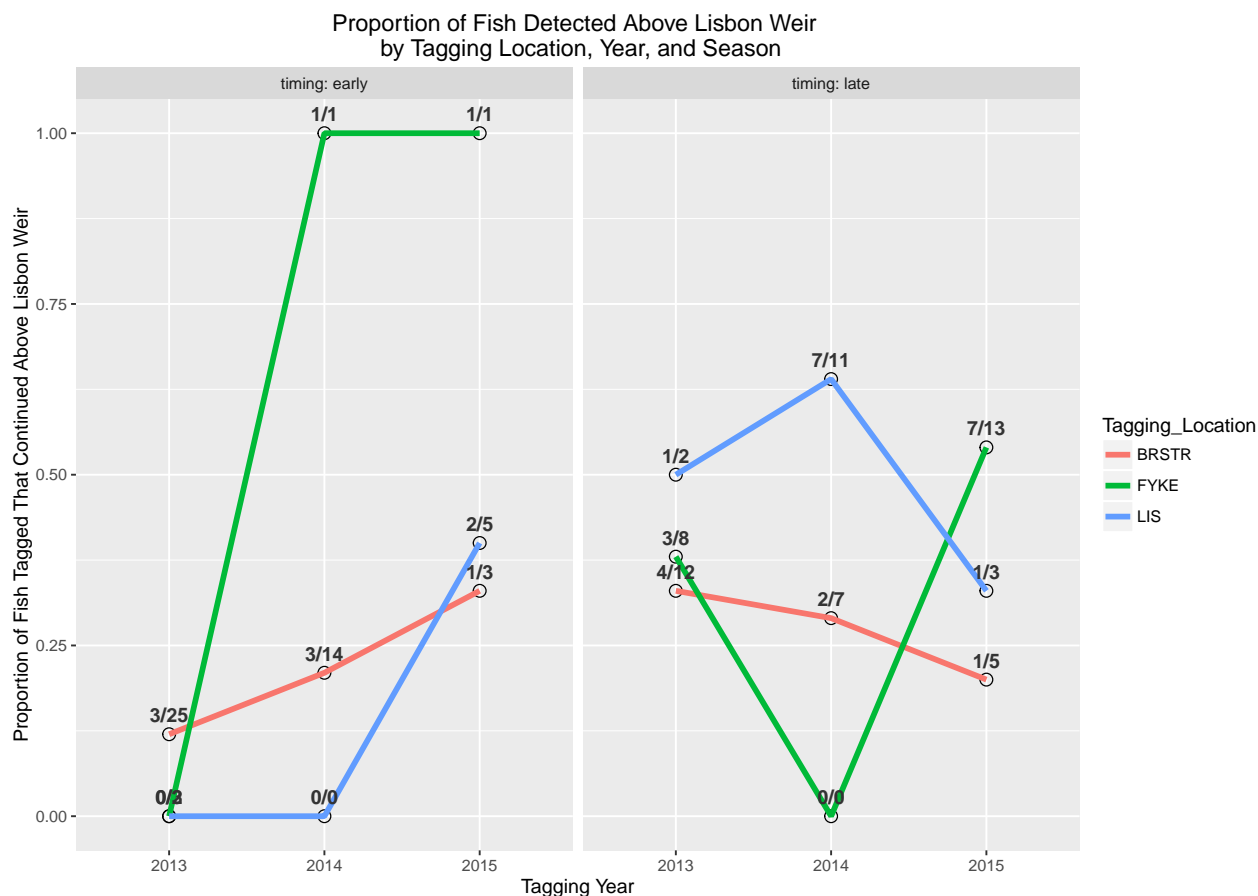
A traditional power analysis doesn't *really* apply in our case anyway, because we do actually detect the “effect” in question at almost 100% efficiency; we know whether every single one of our fish goes either upstream or downstream. So it's not a matter of detecting the true proportion of fish that go in a given direction, it's predicting our fishes' *behavior*, conditional on tagging location and tag date. And the best tool we have available for doing that is the past behavior of our fish.

Does it matter when and where we tag fish?

In the last YETT meeting, I mentioned that tagging location doesn't seem to have as large an effect on whether a fish heads upstream as we previously thought. After looking deeper, that's still true if we just examine the direction a fish heads in right after being tagged: most fish head upstream, regardless of tagging location. That said, if we limit the analysis to the fish that are eventually detected above Lisbon Weir, the story is quite different.

In general, fish tagged later in the season and higher up in the system have a much better chance of being subsequently detected upstream of Lisbon Weir.

The graph below shows the proportion of fish tagged at each location that continued above Lisbon Weir across years, faceted by whether they were tagged “early” in the season (prior to October 20) or “late” in the season (after October 20). We don't have great data from the Fyke, so take that trendline with a huge grain of salt. The thing to notice is that while seasonal timing doesn't seem to make a difference for sampling downstream (BRSTR), it does make a difference for fish tagged upstream (LIS & the FYKE). The point labels are the relative sample sizes represented - note that there are two points overlapping in early 2013, making it hard to differentiate; for the Fyke, 0/3 fish tagged continued above Lisbon Weir; for LIS, 0/2 fish continued above Lisbon Weir.



Clearing out the noise to find a prior

When we calculate weighted averages and raw averages, we get a range of probabilities for fish ending up above Lisbon Weir, depending on whether you sample them early or late, and upstream or downstream:

Timing	General Tagging Location	Percentage Range of Fish Sampled That Will Continue Upstream
Early	Downstream	17% - 22%
Early	Upstream	13% - 80%
Late	Downstream	27% - 29%
Late	Upstream	46% - 57%

To remain consistent with past years, we have a lot of flexibility, but in general we've always had a slightly smaller sample size early in the season relative to late, and I'm happy to push that a little further to increase our odds of getting fish going upstream. This means we can tag about 17 fish before October 20 (early) and 33 fish after October 20 (late). Totals for each sampling location will be:

- BRSTR: between 13-36 fish
- FYKE: between 1 and 23 fish (obviously we have no control over this)
- LIS: between 4 and 16 fish.

For planning purposes and to give us the maximum chance of those fish heading upstream, I think we should move the fish allocated to the FYKE to Lisbon. Adjusting for BRSTR's minimum, we should sample 13-14

fish at BRSTR and 35-36 fish upstream, total.

As for how we should break that up in the tagging season, I ran a lot of different combinations through a paired binomial sampling distribution simulation. I'm showing the code for the functions I set up, with the final numbers I chose, so people can try it for themselves or pick it apart and find any mistakes:

```
# total N early: 17 between upstream and downstream
# total N late: 33 between upstream and downstream

set.seed(1)
simPessimistic <- function(UpstreamEarly, UpstreamLate) {
  # UpstreamEarly = sample size of fish tagged at either LIS or in the FYKE, before October 20
  # UpstreamLate = sample size of fish tagged at either LIS or in the FYKE, after October 20.

  early <- rbinom(1000, size = UpstreamEarly, prob = 0.13) +
    rbinom(1000, size = 17-UpstreamEarly, prob = 0.17)
  # 1000 is the number of random binomial distributions simulated
  # size = UpstreamEarly
  # prob = lower range of the probability of being detected abulis if tagged early upstream,
  # based on previous years' data
  # second prob = lower range of the probability of being detected abulis if tagged early
  # downstream, based on previous years' data

  late <- rbinom(1000, size = UpstreamLate, prob = 0.46) +
    rbinom(1000, size = 33-UpstreamLate, prob = 0.27)
  # size = UpstreamLate
  # prob = lower range of the probability of of being detected abulis if tagged late upstream
  # second prob = lower range of the probability of being detected abulis if tagged late downstream
  ss <- early + late
  frac = sum(ss > 18)/length(ss) # where 18 is the minimum number of fish we want detected above LIS.
  return(frac) # this is the fraction of the results of the simulations that = 18 or greater.
}

simPessimistic(UpstreamEarly = 9, UpstreamLate = 28)
```

```
## [1] 0.293
```

```
simOptimistic <- function(UpstreamEarly, UpstreamLate) {
  # This is the exact same function as above, but with the probabilities changed to the upper
  # limits of the range

  early <- rbinom(1000, size = UpstreamEarly, prob = 0.80) +
    rbinom(1000, size = 17-UpstreamEarly, prob = 0.22)

  late <- rbinom(1000, size = UpstreamLate, prob = 0.57) +
    rbinom(1000, size = 33-UpstreamLate, prob = 0.29)

  ss <- early + late
  frac = sum(ss > 18)/length(ss)
  return(frac)
}

simOptimistic(UpstreamEarly = 9, UpstreamLate = 28)
```

```
## [1] 0.993
```

What the simulations say

Pessimistically, if we tag 7 fish early at BRSTR and 6 fish late at BRSTR, then 9 fish early upstream and 28 fish late upstream, then we will have about a 30% chance of getting at least 18 fish detected above Lisbon Weir.

Optimistically, we will have about a 99% chance of getting at least 18 fish detected above Lisbon Weir (and likely a lot more).

Summary

I think if we want the best possible chance of achieving all our goals, we should sample ~7 fish BRSTR before October 20 and try to get at least 9 fish tagged upstream (between the FYKE and at Lisbon) before October 20. After October 20, we should sample another 6 BRSTR and get the rest (28 fish) upstream. We have flexibility in that framework, but I'd like to get at least 20 fish sampled upstream late in the season; less than that and even the optimistic outlook starts getting bad.

I know that's waaaay too much information and I could have just told you the tagging plan, but I'd love it if someone could point out anywhere I'm using bad logic. Let me know if you have any questions!