Stat 534: formulae referenced in lecture, week 15, part 1: Resource selection

Concept

- Do individuals use resources proportional to their availability in the environment?
  - Simplest case: discrete resources, an example

	Food item					
	A	B	C	D	E	F
Availability	0.1	0.4	0.01	0.09	0.05	0.35
Use	0.01	0.01	0.10	0.09	0.20	0.60

  - Preferentially use C, E F
- resource could be food or type of habitat
  or time of day (less common)

Why interesting?

- Different preferences $\Rightarrow$ mechanism for coexistence
- Management:
  - Platte River Recovery Implementation program is building nesting habitat for piping plovers and least terns
  - What should it look like?
  - Want to build "preferred" habitat
  - that has good nesting success
- Discovery:
  - Rare species - model habitat preference from extant data
  - focus search on other locations with preferred habitat
  - *Braya humilis* story
- Gear selectivity
  - Different ways of catching fish (seine nets, other nets, electrofishing) are selective
  - capture probability depends on size and other fish characteristics
  - need to adjust for gear selectivity to compare data from different methods

Early examples

- Scott 1920: considered 1st to quantify selection

$$\frac{\text{prey \# / fish stomach} \times \text{time unit}}{\text{\# in plankton haul / area}} = \frac{\text{consumption rate}}{\text{density}}$$

- Savage 1931: 1st to compare habitat use to habitat availability
- Ivlev 1961: 1st to construct a measure of strength of habitat selection
  - electivity index for resource i:

    $$E_i = \frac{O_i - \hat{\pi}_i}{O_i + \hat{\pi}_i}$$

    O_i proportion used, $\hat{\pi}_i$ proportion available
  - -1 = Resource never used,
    0 = used in proportion to availability,
    1 = Resource always used
- Now, various other indices
  - Lechowicz 1982 Oecologia evaluates 7 indices, not a complete list
  - differ in how O_i and $\hat{\pi}_i$ are combined
  - Different numeric values
  - but most indices ranked gypsy moth preference for tree species very similarly
- Two general types of indices
  - ad hoc, e.g. Ivlev or log odds ratio
  - probabilistic: $\propto$ P[next resource is of type I]
    - Chesson's index

      $$\frac{O_i/\hat{\pi}_i}{\sum O_i/\hat{\pi}_i}$$

I find the literature extremely confusing

- concepts are muddled
  - how to interpret a particular measure:
    - Is it a ratio or a log odds ratio or something else?
  - what unit is being described by P[used]
  - what is the reference group?

        * Johnson (1961)'s scales
        * different choices of reference group
        * 1st order: entire range of the species
        * 2nd order: home range of an individual or group
        * 3rd order: resource use w/i a home range
        * 4th order: use of resources at a site

- multiple types of data
- multiple sampling designs
- multiple statistical models
    - don't always align with data type, sampling design, and intended concept

These notes are an overview of the issues, as I see them. No definitive answers.

Key resources:

- Manly et al.'s book, Resource Selection by Animals: statistical design and analysis for field studies, 2nd ed. 2002.
- Keating and Cherry 2004, Journal of Wildlife Management 68:774-789
    - popularized the logistic regression approach

Data: names as used by Manly et al.

- SP-A:
    - available units sampled or censussed
    - used units randomly sampled
- SP-B:
    - available units sampled or censussed
    - **unused** units randomly sampled
- SP-C:
    - used and unused units independently randomly sampled

Design: Again, Manly et al.'s names

- I: population level - all animals in study area

  - classify animal locations
  - GIS analysis of area $\Rightarrow$ availability

- II: individual animals
  - e.g. marked or radio collared
  - availability as for design I (GIS)

- III: also individual animals
  - multiple used and unused for each animal
  - e.g., based on individual home range or feeding sites of each individual

- Each study has one combination of data and design

- II & III $\rightarrow$ resource selection per individual
  - enable a 2nd stage analysis of sex or age differences

Choice of unit, 2 examples

- bird nest in a tree
  - Q: does that species have preferred tree species?
  - unit = tree
  - what domain is available? not used?
  - if only 1 nest per pair, only population design (I)
  - if multiple nests for a single individual/pair, this is II or III

- GPS collar on a deer or tag on fish
  - Location every 15 minutes
  - use to get habitat every 15 minutes
  - what is available? not used?
  - could do population or individual analysis (II or III)

Statistical models for resource selection

- Notation:
  - X: habitat characteristic(s), discrete or continuous
  - Z: 1/0, used or not used

- Densities:

- used observations: $f(X \mid Z = 1)$
- not used observations: $f(X \mid Z = 0)$
- available observations: $f(X)$

- Resource selection function (RSF)

$$w(x) = \frac{f(X \mid Z = 1)}{f(X)} = \frac{\text{p[} X \text{ in used sample]}}{\text{p[} X \text{ available]}}$$

  - relative probability, bounds are $(0, \infty)$

- Resource selection probability function (RSPF)
  - $\text{P}[Z = 1 \mid X] = \pi(X)$: P[probability that a unit with X is used]
  - or, proportion of the population of available units in category X that are used
  - NOT $w(x) = f(X \mid Z = 1)$

Why the scale of "available" matters

- Scenario 1: sample 1000 available items, 100 used items
- used items are a subsample of those available

	A	B	C	D	E	F
1000 available	100	400	10	90	50	350
100 used	1	1	10	8	20	60

  - w(E) = (20/100) / (50/1000) = 4
  - P[Z=1 | E] = 20 / 50 = 0.4
  - w(C) = (10/100) / (10/1000) = 10
  - P[Z=1 | C] = 10/10 = 1

- Now sample 10000 available

	A	B	C	D	E	F
10000 available	1000	4000	100	900	500	3500
100 used	1	1	10	8	20	60

  - w(E) = (20/100) / (500/10000) = 4
  - P[Z=1 | E] = 20 / 500 = 0.04
  - w(C) = (10/100) / (100/10000) = 10
  - P[Z=1 | C] = 10/100 = 0.1

Connection between RSF and RSPF

- Bayes rule

$$P[Z=1 \mid X] = \frac{P[Z=1 \& X=x]}{P[X=x]} = \frac{P[X=x \mid Z=1]\, P[Z=1]}{P[X=x]}$$

- RSPF = $w(x) \times (N_{used}/N_{avail})$

What about used, not used (SP-C) data?

- Have $f(X \mid Z=1)$ and $f(X \mid Z=0)$
- Want $f(X)$
- Can't just add $f(X \mid Z=1) + f(X) \mid Z=0)$!
- Can add joint distributions: $f(X, Z=1) + f(X, Z=0)$

$$f(X) = f(X \mid Z=1) f(Z=1) + f(X \mid Z=0) f(Z=0)$$

- Same issues with the fraction of available units that are used
- Numerical examples
  - The data:

	A	B	C	D	E	F
100 not used	10	40	1	9	5	35
100 used	1	1	10	8	20	60

  - Assume P[used] = P[$Z=1$] = 0.1:

	A	B	C	D	E	F
$f(x)$	0.091	0.361	0.019	0.089	0.065	0.375
$w(x)$	0.11	0.028	5.26	0.89	3.08	1.6

  - Assume P[used] = P[$Z=1$] = 0.01:

	A	B	C	D	E	F
$f(x)$	0.01	0.40	0.011	0.090	0.052	0.35
$w(x)$	0.10	0.025	9.17	0.90	3.88	1.7

- Interested in describing relative use of different habitats
  - but $w(x)$ depends P[$Z=1$]
  - i.e., on # available and # used
  - Same issues for P[$Z=1 \mid X$]

One solution: odds of use

- Turns out that f[$Z=1 \mid X=x$] / f[$Z=1 \mid X=x$] does not depend on P[$Z=1$]!

- The algebra:

$$\frac{f[Z=1 \mid X=x]}{f[Z=0 \mid X=x]} = \frac{f[X \mid Z=1]\, P[Z=1]/f[X]}{f[X \mid Z=0]\, P[Z=0]/f[X]} = \frac{f(x \mid Z=1)}{f(x \mid Z=0)}$$

- Connections to epidemiology
  - odds of use

$$\frac{P[Z=1 \mid X=x]}{P[Z=0 \mid X=x]} = \frac{P[Z=1 \mid X=x]}{1-P[Z=1 \mid X=x]} = \text{the odds of use given } x$$

  - enumerating used / not used is a case-control design
- estimate log odds of use by logistic regression

$$\text{logit } P[Z=1 \mid X=x] = \log \frac{P[Z=1 \mid X=x]}{P[Z=0 \mid X=x]} = \mathbf{X}\boldsymbol{\beta}$$

- Keeting and Cherry 2004 JWM is the key reference on this approach

Back to use/available data

- Can you use logistic regression?
- What can you learn using a logistic regression?
- Tempting to assume that available = not used
  - Not necessary - here's why
- The model

$$\begin{aligned} Z_i &\sim \text{Bern}(\pi_i) \\ \text{logit } \pi_i &= \beta_0 + \beta_1 X_i = X b_i \\ \pi_i &= \frac{e^{Xb_i}}{1+e^{Xb_i}} \end{aligned}$$

- Divide available into used (Z=1) and not used (Z=0)
- lnL = $\sum_{used} \log \pi_i + \sum_{not\ used} \log(1-\pi_i)$

$$\begin{aligned} lnL &= \sum_{used} \log \left[\frac{e^{Xb_i}}{1+e^{Xb_i}}\right] + \sum_{not\ used} \left[\log \frac{1}{1+e^{Xb_i}}\right] \\ &= \sum_{used} \log e^{Xb_i} - \sum_{used} \log(1+e^{Xb_i}) - \sum_{not\ used} \log(1+e^{Xb_i}) \\ &= \sum_{used} Xb_i - \sum_{avail} \log(1+e^{Xb_i}) \end{aligned}$$

- Not a likelihood for a standard logistic model
- But it is approximately the likelihood for a heterogeneous Poisson process

$$\begin{aligned} Z_i &\sim \text{Pois}(\pi_i) \\ \log \pi_i &= \beta_0 + \beta_1 X_i = Xb_i \end{aligned}$$

  - Imagine the study area broken into many small grid cells
    - have X_i for each grid cell
    - π_i is now the average number of "used" in that grid cell
    - grid cells are small so π_i very small
    - the only possible values for Z_i are 0 or 1

$$\begin{aligned} P[Z = 0 \mid X_i] &= \frac{e^{-\pi_i}\pi_i^0}{1} = e^{-\pi_i} \\ P[Z = 1 \mid X_i] &= \frac{e^{-\pi_i}\pi_i^1}{1} = \pi_i e^{-\pi_i} \\ lnL &= \sum_{used} \log\left(\pi_i e^{-\pi_i}\right) + \sum_{not\ used} \log e^{-\pi_i} \\ &= \sum_{used} \log \pi_i - \sum_{used} \log e^{-\pi_i} - \sum_{not\ used} \log e^{-\pi_i} \\ &= \sum_{used} \log \pi_i - \sum_{avail} \pi_i \\ &\approx \sum_{used} \log \pi_i - \sum_{avail} \log(1 + \pi_i) \\ &= \sum_{used} \log e^{Xb_i} - \sum_{avail} \log\left(1 + e^{Xb_i}\right) \\ &= \sum_{used} Xb_i - \sum_{avail} \log\left(1 + e^{Xb_i}\right) \end{aligned}$$

- The Poisson version is exactly the same log likelihood used by the MaxEnt approach to species distribution modeling