Stat 534: formulae referenced in lecture, week 3, updated:
model-based species comp. analysis and Tweedie distributions
mark recapture, part 1

Model-based species composition analysis:
The goal: Does the species composition (all species) change according to a model, e.g.:
two groups (1930's, 1960') or more: 1 way ANOVA
factorial treatment structures: 2+ way ANOVA
regression model, e.g., linear with year, or polynomial with year
but no mixed models (at least for now)

Putting the pieces together:

- Fit specified model to each species separately
  Using likelihood and a specified distribution
- Fit reduced model without the term of interest (null hypothesis)
  My understanding is this follows the usual R sequential testing approach
  So for a model: species = A + B + C,
  R will compare:

Test of:	Null (H0) model	Full model
A	intercept only	intercept + A
B	intercept + A	intercept + A + B
C	intercept + A + B	intercept + A + B + C

- Collect the change in log likelihood for each comparison and each species
- Add change in lnL across the species
  Has known (asymptotic) distribution when all species are independent
  They're almost certainly correlated
- Use randomization to get a valid test in spite of correlation
  Randomize quantile residuals, one for each species and site
  Keep together all residuals from a site (accounts for species correlation)

Does the total number of individuals matter?
Consider two sites, each with 3 species, Abundances:

Site	1	2	3	Total
A	4	4	32	40
B	8	8	64	80

Proportion of total:

Site	1	2	3	Total
A	0.1	0.1	0.8	1.0
B	0.1	0.1	0.8	1.0

Two situations

- Higher total because of more effort, known effort
  Include log effort, E_i as an offset, this models μ_{ij}/E_i

$$Y_{ij} \sim F(\mu_{ij})$$

$$\begin{aligned}\log \mu_{ij} &= \text{model} + \log E_i \\ \log \mu_{ij} - \log E_{ij} &= \text{model} \\ \mu_{ij}/E_i &= \exp(\text{model})\end{aligned}$$

- Better "catchability", total not known
  Include a site effect in the model

$$\begin{aligned}Y_{ij} &\sim F(\mu_{ij}) \\ \log \mu_{ij} &= \alpha_i + \text{model}\end{aligned}$$

  Estimate α_i, will usually be very close to logTotal
  But uncertainty in model estimates takes account of unknown total

Distributions for continuous data with non-constant variance

- Normal distribution: usually, constant variance for any mean
  - Can write "power of the mean" models
  - $Y_i \sim N(\mu, g(\mu))$, e.g., $g(\mu) = \sigma^2 \mu^k$
  - Allow unequal variance, but distribution always symmetric around the mean
  - Common experience is that distributions, e.g., of tree basal area, are skewed not symmetric
- logNormal distribution: $\log Y_i \sim N(\mu_l, \sigma_l^2)$
  - Skewed
  - Var $Y_i = k\mu^2$
  - constant coefficient of variance: cv $= \sqrt{\text{Var } Y}/\mu = \sqrt{k}$
  - But: 0 can never occur
    Zero-inflated distributions or Hurdle models (both allow zeros, both more complicated)
- Gamma distribution:
  - Very similar to a log-normal (also skewed)
  - But very slightly fewer very large values ("skinnier upper tail")
  - Also doesn't allow 0's
- a Tweedie distribution

Tweedie distribution: more flexible than log normal

- Continuous random variable, Var $= k\mu^p$, p is a parameter to be specified or estimated

- probability density function not especially informative
- normal, Poisson, and Gamma distributions are special cases of the Tweedie
  - p = 0 $\Rightarrow$ normal
  - p = 1 $\Rightarrow$ Poisson
  - p = 2 $\Rightarrow$ Gamma
- Most interesting distributions are those with $1 < p < 2$
  - Skewed distribution for continuous data with additional point mass at 0
    * log normal and Gamma distributions are only for Y > 0
    * "additional point mass at 0": a Tweedie distribution has a non-zero $\mathrm{P}[Y = 0]$
  - Tweedie is a compound Poisson-gamma distribution.
  - For $1 < p < 2$, here's how to simulate a value, Y, from the Tweedie(λ, a, b)
    * simulate $N \sim \text{Poisson}(\lambda)$
    * simulate N independent values of $Y_i \sim \text{Gamma}(a, b)$
    * return $Y = \sum_{i=1}^{N} Y_i$
    * If you want values from a Tweedie with a specified μ, σ^2, and p, use:

$$a = \frac{2-p}{p-1} \quad b = \tfrac{\mu^{1-p}}{(p-1)\sigma^2} \quad \lambda = \frac{\mu^{2-p}}{(2-p)\sigma^2}$$

Mark-recapture analysis

- General population model:

$$N_{t+\Delta t} = N_t + B_t - D_t + I_t - E_t$$

  - N_t: number of individuals in the population at time t
  - Δt: time increment, often 1 year, can be other timespans
  - B_t: # births between t and $t + \Delta t$
  - D_t: # deaths between t and $t + \Delta t$
  - I_t: # immigrants between t and $t + \Delta t$
  - E_t: # emigrants between t and $t + \Delta t$
- With a single population, commonly assume $I_t = E_t = 0$
- And often interested in "how many?": N_t
- Derived quantities that are often of interest:
  - ϕ_t: fraction of N_t who survive the interval, $D_t = (1 - \phi_t)N_t$

- per-capita birth rate, B_t/N_t

Horvitz-Thompson estimator

- Sample survey:
  - Survey design $\Rightarrow$ probability that individual i included in the sample = π_i

$$\widehat{\text{Total}} = \sum \frac{Y_i}{\pi_i}$$

    where the sum is over the individuals included in the sample
  - Example: simple random sample of n individuals from a population of N
  - $\pi_i = n/N$
  - Estimated population total = $\sum \frac{Y_i}{n/N} = N \sum \frac{Y_i}{n} = N\overline{Y}$
- Applied to estimating population size N_t:
  - Known probability of capture for each individual, π_i
  - For now, assume same for all individual, π_{known}
  - $Y_i = 1$ for all individuals caught in the first sample

$$\hat{N}_1 = \sum \frac{1}{\pi_{known}} = \frac{n_1}{\pi_{known}}$$

Lincoln-Petersen estimator $$\hat{N} = \frac{n_1 n_2}{m_2}$$

- n_1: number of individuals released with marks at time 1
- n_2: number of individuals caught at time 2
- m_2: number of individuals caught with marks at time 2
- Intuitive estimator:
  - Assume π is same for 1st and 2nd times
  - and same for marked and unmarked individuals
  - At time 2:
    Caught n_2 individuals
    marked individuals $\Rightarrow \hat{\pi} = m_2/n_1$
    Apply H-T: $\hat{N} = n_2/(m_2/n_1)$

Multinomial model for 2 sampling occasions

- 2 x 2 contingency table for capture events

	Capture time 2		
Capture time 1	Yes	No	Total
Yes	m_2	$n_1 - m_2$	n_1
No	$n_2 - m_2$	?	$N - n_1$
Total	n_2	$N - n_2$	N

- Corresponding capture history table

Time 1	Time 2	# animals	probability
Y	Y	$n_{11} = m_2$	$p_1 \, p_2$
Y	N	$n_{10} = n_1 - m_2$	$p_1 \, (1 - p_2)$
N	Y	$n_{01} = n_2 - m_2$	$(1 - p_1) \, p_2$
N	N	$n_{00} = N - n_1 - n_2 + m_2$	$(1 - p_1) \, (1 - p_2)$

Multinomial distribution: generalization of the binomial to more than 2 outcomes

- Consider an event with 4 possible outcomes:
  red, blue, green, yellow with probabilities π_r, π_b, π_g, π_y
- Data from N total events, probability of observing n_r, n_b, n_g, n_y is:

$$f(n_r,\, n_b,\, n_g,\, n_y \mid N,\, \pi_r,\, \pi_b,\, \pi_g,\, \pi_y) = \frac{N!}{n_r!\, n_b!\, n_g!\, n_y!} \pi_r^{n_r}\, \pi_g^{n_g}\, \pi_b^{n_b}\, \pi_y^{n_y}$$

- log likelihood is: $\text{lnL}(\pi_r,\, \pi_b,\, \pi_g,\, \pi_y \mid N, n_r,\, n_b,\, n_g,\, n_y) =$

  $\log N! - \log n_r! - \log n_b! - \log n_g! - \log n_y! + n_r \log \pi_r + n_g \log \pi_g + n_b \log \pi_b + n_y \log \pi_y$

- Usual set up:
  - N is known.
  - Only need 3 of the 5 quantities: e.g., n_r, n_b, n_g because $n_y = N - n_r - n_b - n_g$
  - And only have to estimate 3 parameters
    e.g., π_r, π_b, π_g because $\pi_y = 1 - (\pi_r + \pi_b + \pi_g)$

Multinomial distribution for 2 capture occasions:

- Different setup from the "usual" multinomial:
  - N no longer known
  - have 3 counts: $n_{11} = m_2$, $n_{10} = n_1 - m_2$, $n_{01} = n_2 - m_2$
  - their probabilities depend on only 2 parameters, π_1 and π_2

- The log likelihood function is: $\mathrm{lnL}(N,\ \pi_1,\ \pi_2 \mid m_2, n_1, n_2)$

$$\begin{aligned} = &\quad \log N! - \log m_2! - \log(n_1 - m_2)! - \log(n_2 - m_2)! - \log(N - n_1 - n_2 + m_2)! \\ + &\quad m_2 \log[\pi_1\, \pi_2] + (n_1 - m_2) \log[\pi_1\, (1 - \pi_2)] + (n_2 - m_2) \log[(1 - \pi_1)\, \pi_2] \\ + &\quad (N - n_1 - n_2 + m_2) \log[(1 - \pi_1)\, (1 - \pi_2)] \end{aligned}$$

- Analytic solutions can be found by solving:

$$\begin{aligned} \frac{\partial lnL}{\partial \pi_1} &= \frac{m_2}{\pi_1} + \frac{n_1 - m_2}{\pi_1} - \frac{n_2 - m_2}{1 - \pi_1} - \frac{N - n_1 - n_2 + m_2}{1 - \pi_1} = 0 \\ \hat{\pi}_1 &= \frac{n_1}{\hat{N}} && (1) \\ \frac{\partial lnL}{\partial \pi_2} &= \frac{m_2}{\pi_2} + \frac{n_2 - m_2}{\pi_2} - \frac{n_1 - m_2}{1 - \pi_2} - \frac{N - n_1 - n_2 + m_2}{1 - \pi_2} = 0 \\ \hat{\pi}_2 &= \frac{n_2}{\hat{N}} && (2) \\ \frac{\partial lnL}{\partial N} &= \frac{\partial \log N!}{\partial N} - \frac{\partial \log(N - n_1 - n_2 + m_2)!}{\partial N} + \log[(1 - \pi_1)\, \pi_2] && (3) \end{aligned}$$

- To evaluate equation (3), remember $\frac{\partial \log \Gamma(N)}{\partial N}$ is the digamma function, $\Psi(N)$:

$$\Psi(N+1) = \frac{\partial \log \Gamma(N+1)}{\partial N} = \frac{\partial \log N!}{\partial N}$$

- Reference books on mathematical functions, e.g., Abramowitz and Stegun (1964) Handbook of Mathematical Functions gives

$$\Psi(N+1) \approx \log(N+1) - \frac{1}{2(N+1)} - \frac{1}{12(N+1)^2} + \frac{1}{120(N+1)^4} - \frac{1}{252(N+1)^6} + \cdots \approx \log N$$

- Using this approximation in (3) and simplifying gives, after some algebra:

$$\hat{N} = \frac{n_1\, n_2}{m_2}$$