Stat 534: formulae referenced in lecture, week 1: species composition analysis

Poisson probability mass function (pmf), $Y \sim \text{Pois}(\lambda)$, for x a non-negative integer:

$$f(y \mid \lambda) = P(Y = y \mid \lambda) = \frac{e^{-\lambda}\lambda^y}{y!}$$

Properties of Poisson distributions:

Mean	$\text{E}\, Y = \lambda$
Variance	$\text{Var}\, Y = \lambda$
	$\text{P}[\text{ Y} = 0\ ] = e^{-\lambda}$

One species *Vanellus vanellus*, northern lapwing
1930's: 10, 11, 12, 10, 8, 6, 5, 3, 5, 4
1960's: 25, 17, 20, 4, 7, 18, 27, 18, 18, 10

Log likelihood for n independent Poisson observations:

$$\text{L}(\lambda \mid y) = \frac{e^{-\lambda}\lambda^y}{y!}$$

$$\text{lnL}(\lambda \mid (y_1, y_2, \cdots y_n) = -n\,\lambda + \log\lambda \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} y_i!$$

Estimating λ: find the value of λ that maximizes the lnL

$$\begin{array}{ll} \frac{d\,lnL}{d\,\lambda} & = -n + \frac{\sum_{i=1}^{n} Y_i}{\lambda} = 0 \\ \text{mle} & \hat{\lambda} = \sum_{i=1}^{n} Y_i / n = \overline{Y} \end{array}$$

called the maximum likelihood estimator (mle) of θ

How precise is $\hat{\lambda}$? Two parts:

- For the statisticians, mostly a reminder
  - As sample size, $n \to \infty$, Var $\hat{\theta} \to -1/I$

$$\begin{aligned} I &= \text{E}\left(\frac{d\,lnL(\theta)}{d\,\theta}\right)\left(\frac{d\,lnL(\theta)}{d\,\theta}\right) \text{ evaluated at } \theta \\ &= \text{E}\,\frac{d^2\,lnL}{d\,\theta^2}\,|_{\lambda=\theta} \end{aligned}$$

  - I is called information, Fisher information or expected information
  - Cramer-Rao lower bound: The mle of θ has the smallest variance of any possible estimator
    - When the model is correct

- Property of the population
- All the math depends on some assumptions about $f(\theta)$, "the regularity conditions"

- For the applied statistician / biologist
  - Observed information
$$H = \frac{d^2\, lnL}{d\,\theta^2} \text{ evaluated at } \hat{\theta}$$
  - Property of the sample and the probability model
  - Easily (usually) calculated by software

- For any mle, $\hat{\theta}$: Var $\hat{\theta} \approx -1/H$.

- When more than one parameter, H is the Hessian matrix (will see examples later)

- For a Poisson distribution:
  - $H = \frac{d^2\, lnL}{d\,\lambda^2}\,|_{\lambda=\hat{\lambda}} = \frac{-1\sum_i Y_i}{\lambda^2}\,|_{\lambda=\hat{\lambda}} = \frac{-1\sum_i Y_i}{\hat{\lambda}^2}$
  - Var $\hat{\lambda} = -1/H = \frac{\hat{\lambda}}{n}$

Confidence interval for λ:

- Asymptotic normality
  - $\left(\hat{\theta} - z_p\sqrt{\text{Var } \hat{\theta}},\;\; \hat{\theta} + z_p\sqrt{\text{Var } \hat{\theta}}\right)$
  - p =1 - (1-coverage)/2, e.g. $p = 0.975$ for a 95% confidence interval
  - $z_{0.975} = 1.96$
  - normal quantiles are symmetric so $z_{1-p} = -z_p$.

- Profile likelihood: will see soon

Q: Did the abundance change more than you would expect from random variation?
A hypothesis test: H_o: same λ both periods, H_a: two λ's, one for each period

Use a likelihood ratio test. Two ways to set this up:

1. based on the hypothesis statements:
   H_o: one group of observations, all with one λ
   $\text{LnL}_o = -20\,\lambda + \log\lambda \sum_{i=1}^{20} Y_i - \sum_{i=1}^{20} Y_i!$
   H_a: two groups of observations, 1930's with λ_1, 1960's with λ_2
   $\text{LnL}_a = \left[-10\,\lambda_1 + \log\lambda_1 \sum_{i=1}^{10} Y_i - \sum_{i=1}^{10} Y_i!\right] + \left[-10\,\lambda_2 + \log\lambda_2 \sum_{i=11}^{20} Y_i - \sum_{i=11}^{20} Y_i!\right]$

2. based on a model:

$$\begin{aligned} Y_i &\sim \text{Pois}(\lambda_i) \\ \log \lambda_i &= \beta_0 + \beta_1 X_i \end{aligned}$$

   $X_i = 0$ if observation i is in 1930's group, and $X_i = 1$ if in 1960's group.

Notice the relationship between the two approaches

- $\exp \beta_0 = \lambda_1$
- $\beta_1 = \log \lambda_2 - \log \lambda_1$, so $\exp \beta_1 = \lambda_2 / \lambda_1$
- $\beta_1 = 0 \Leftrightarrow \lambda_2 = \lambda_1$. Expresses H_o
- We choose to put a model on $\log \lambda$ because $\lambda \geq 0$

Construct a test using $D = -2\,(\text{LnL}_o - \text{LnL}_a)$

- $D = 0$: both models fit the data equally well
- $D >> 0$: H_a fits a lot better than H_o
- When H_o true, asymptotically $D \sim \chi^2_{df}$
  - df = difference in number of parameters between the models
  - Applies to any likelihood comparison
  - when based on the same data (e.g., same # observations)
  - Assumes large samples (asymptotic) but commonly applied to any sample size
- Here, H_a has 2 parameters, H_o has 1 parameter, so df = 2-1 = 1
  - 0.95 quantile of a χ^2_1 distribution $= 3.84 = 1.96^2$.
  - 0.975 quantile of a normal distribution = 1.96
  - Two useful numbers to remember.

Notice that:

- $\sum_{i=1}^{20} Y_i!$ can be ignored - cancels out when LnL subtracted
- The only way the data enters into the likelihood is through $\sum_i Y_i$.
  - Sufficient statistic: how the data enters the lnL. $\sum_i Y_i$
  - $\sum_i Y_i$ is the sufficient statistic for the Poisson distribution
  - Variability between observations is ignored

  - But depends on distribution:

- Normal distributions have 2 parameters, mean and variance
  - LnL has two sufficient statistics: $\sum_i Y_i$ and $\sum_i Y_i^2$

Binomial distribution:

- 2 parameters, N = # trials, and π = probability of a "success" on any single trial.
- P[# successes $= y] = \binom{N}{x}\pi^x(1-\pi)^{(N-x)} = \frac{N!}{(N-x)!x!}\pi^x(1-\pi)^{(N-x)}$
- E Y $= N\pi$
- $\hat{\pi} = \frac{Y}{N}$
- Var $\hat{\pi} = \frac{\hat{\pi}\,(1-\hat{\pi})}{N}$

Comparison of Binomial and Poisson distributions

- Imagine N increasing but mean # successes staying constant:

$$Y \sim Bin(N, \pi),\ \ N \to \infty,\ \ \pi \to 0,\ \ N\pi = constant$$

- Distribution of $Y \to$ Poisson

Negative binomial distribution:

- Y =# successes before getting r failures
  - Two parameters: r = # failures and π = P[success on a single trial]
  - Two parameters: μ = mean # events, r = overdispersion parameter, could be continuous
- E $Y = \mu$
- Var $Y = \mu + \mu^2/r$, or $\mu + \alpha\mu^2$
- pmf:

$$\mathrm{P}(Y = y) = \frac{\Gamma(r+y)}{y!\,\Gamma(r)}\left(\frac{r}{r+\mu}\right)^r\left(\frac{\mu}{r+\mu}\right)^y$$

- $\Gamma(n) = (n-1)!$ when n integer
- Var $Y \geq$ E Y, equal only when $r = \infty$ or $\alpha = 0$
- $r = \infty$ or $\alpha = 0 \Rightarrow Y \sim \mathrm{Pois}(\mu)$

Estimating μ and r:

- Observe n values: $y_1,\ y_2,\ y_3 \cdots,\ y_n$
- LnL($\mu,\ r \mid \{y\}$ =
  $\sum \log \Gamma(r + y_i) - \sum \log \Gamma(y_i + 1) - n \log \Gamma(r) + n\, r \log \left(\frac{r}{r+\mu}\right) + \sum y_i \left(\frac{\mu}{r+\mu}\right)$
- Derivatives are ugly:
  - derivative of $\log \Gamma(x)$ is the digamma function, $\approx\ \log x$
  - results are transcendental equations, have terms with r and terms with $\log r$
  - Generally no analytical solution, so no equations giving mle's for μ or r
- Need to use numeric maximization
  common theme in this course, so don't worry about your calculus