solvingEquation

Jennifer Semple

4/1/2020

2.3.1 Classical statistics for classical data

Proof that the mean of the Poisson distribution maximises the log-likelihood: From before we know that the liklihood (written here as L) is a multiplication of all the inidividual probabilities:

$$L(\lambda, x = (k_1, k_2, k_3...)) = \prod_{i=1}^{100} f(k_i)$$

f(k) is simply the Poisson density function:

$$f(k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

So if we put those together and take the log of both sides, we get:

$$\log(L(\lambda, x)) = \log(\prod_{i=1}^{100} \frac{e^{-\lambda}\lambda^k}{k!})$$

We know that the product log of a product ($\prod$) is the same as the sum ($\sum$) of a log, we can rewrite it as:

$$\log L = \sum_{i=1}^{100} \log(\frac{e^{-\lambda}\lambda^k}{k!})$$

We can also break up the fraction, again using the log rules of $\log(a * b) = log(a) + log(b)$ and $\log(\frac{a}{b}) = \log(a) - log(b)$:

$$\log L = \sum_{i=1}^{100} (\log(e^{-\lambda}) + \log(\lambda^k) - \log(k!))$$

Now we can get rid of the powers using $\log(a^b) = b\log(a)$. Also $\log(e) = 1$ because this is the natural log.

$$\log L = \sum_{i=1}^{100} (-\lambda + k\log(\lambda) - \log(k!))$$

Now we want to break apart the sum by extracting terms that do not depend on k.The final term does not depend on lambda, so it is just a constant:

$$logL = -100\lambda + \log\lambda(\sum_{i=1}^{100} k_i) + const.$$

To get the maximum of a function we want the derivative of the function to be equal to 0:

$$\frac{d}{d\lambda}\log L = \frac{d}{d\lambda}(-100\lambda + \log\lambda(\sum_{i=1}^{100} k_i) + const.) = 0$$

Using the derivative rules of $\frac{d}{dx}ax = a$ and $\frac{d}{dx}\log(x) = \frac{1}{x}$, and derivative of a constant is 0, we get:

$$-100 + \frac{1}{\lambda}\sum_{i=1}^{100} k_i = 0$$

$$100 = \frac{1}{\lambda}\sum_{i=1}^{100} k_i$$

Multiply by $\frac{\lambda}{100}$:

$$\lambda = \frac{1}{100}\sum_{i=1}^{100} k_i = \overline{k}$$

So the λ paramter is the same as the mean ($\overline{k}$).

Likelihood for the binomial distribution

$$f(\theta \mid n, y) = f(y \mid n, \theta) = \binom{n}{y} \theta^y (1-\theta)^{(n-y)}$$

To avoid large-number multiplications we take the log of both sides (log likelihood):

$$\log f(\theta \mid n, y) = \log(\binom{n}{y} \theta^y (1-\theta)^{(n-y)})$$

We break up the product using the log rule $\log(ab) = \log(a) + \log(b)$:

$$\log f(\theta \mid n, y) = \log \binom{n}{y} + \log \theta^y + \log(1-\theta)^{(n-y)})$$

We bring down the exponents using the $\log(a^b) = b \log(a)$ rule:

$$\log f(\theta \mid n, y) = \log \binom{n}{y} + y \log \theta + (n-y) \log(1-\theta)$$

This is the formula used in the text.

Hardy Weinberg equilibrium

We have two alleles in a population, N and M. Frequency of M is p Frequency of N is 1-p, which we will call q.

Since both copies of an allele are distributed independantly, the freuency or probability of the different combinations of two alleles are:

$$P_{MM} = p \times p = p^2, P_{NN} = q \times q = q^2, P_{MN} = p \times q + q \times p = 2pq$$

As data we normally observe the counts of the different genotypes (n_{MM}, n_{NN}, n_{MN}) and the total number of individuals $S = n_{MM} + n_{NN} + n_{MN}$

Using the mutinomial formula we can write the likelihood of the data given the expected probabilities of $p^2, q^2, 2pq$:

$$P(n_{MM}, n_{MN}, n_{NN} \mid p) = \binom{S}{n_{MM}, n_{MN}, n_{NN}} (p^2)^{n_{MM}} \times (2pq)^{n_{NM}} \times (q^2)^{n_{NN}}$$

To get the log likelihood we take the log of both sides:

$$L(p) = \log(\binom{S}{n_{MM}, n_{MN}, n_{NN}} (p^2)^{n_{MM}} \times (2pq)^{n_{NM}} \times (q^2)^{n_{NN}})$$

Separating the terms with $\log(ab) = \log(a) + \log(b)$ rule:

$$L(p) = \log(\binom{S}{n_{MM}, n_{MN}, n_{NN}}) + \log((p^2)^{n_{MM}}) + \log((2pq)^{n_{NM}}) + \log((q^2)^{n_{NN}})$$

We bring forward the exponants with $\log(a^b) = b\log(a)$ rule, and simplify the middle term with the $\log(ab) = \log(a) + \log(b)$ rule:

$$L(p) = \log(\binom{S}{n_{MM}, n_{MN}, n_{NN}}) + 2n_{MM}\log(p) + n_{NM}\log(2) + n_{NM}\log(p) + n_{NM}\log(q) + 2n_{NN}\log(q)$$

We can gather some of the terms:

$$L(p) = \log(\binom{S}{n_{MM}, n_{MN}, n_{NN}}) + (2n_{MM} + n_{NM})\log(p) + n_{NM}\log(2) + (n_{NM} + 2n_{NN})\log(q)$$

To maximise the log likelihood we must take the derivative and set it to 0:

$$0 = \frac{d}{dp}\log(\binom{S}{n_{MM}, n_{MN}, n_{NN}}) + \frac{d}{dp}(2n_{MM} + n_{NM})\log(p) + \frac{d}{dp}n_{NM}\log(2) + \frac{d}{dp}(n_{NM} + 2n_{NN})\log(q)$$

The first and third term do not depend on p, so their derivative is 0, and we can substitute back in $q = 1 - p$:

$$0 = 0 + \frac{d}{dp}(2n_{MM} + n_{NM}) \log(p) + 0 + \frac{d}{dp}(n_{NM} + 2n_{NN}) \log(1 - p)$$

Using the $\frac{d}{dx} \ln(x) = \frac{1}{x}$ rule (remember, log without an explicit base in maths is usually ln):

$$0 = \frac{2n_{MM} + n_{NM}}{p} + \frac{n_{NM} + 2n_{NN}}{1 - p}$$

We multiply both sides by $p(1 - p)$:

$$0 = (2n_{MM} + n_{NM})(p - 1) + (n_{NM} + 2n_{NN})p$$

Simplifying:

$$0 = 2n_{MM}p + n_{NM}p - 2n_{MM} - n_{NM} + n_{NM}p + 2n_{NN}p$$

$$0 = 2n_{MM}p + 2n_{NM}p + 2n_{NN}p - 2n_{MM} - n_{NM}$$

$$0 = (2n_{MM} + 2n_{NM} + 2n_{NN})p - 2n_{MM} - n_{NM}$$

$$2n_{MM} + n_{NM} = (2n_{MM} + 2n_{NM} + 2n_{NN})p$$

$$p = \frac{2n_{MM} + n_{NM}}{2n_{MM} + 2n_{NM} + 2n_{NN}}$$

$$p = \frac{2n_{MM} + n_{NM}}{2S}$$

$$p = \frac{2n_{MM} + n_{NM}}{2S}$$

$$p = \frac{n_{MM} + \frac{1}{2}n_{NM}}{S}$$

So the most likely p given the data, is the proportion of homozygotes carrying that allele plus half the proportion of heterozygotes