Group SBS, Sheet 03, Exercise 03

November 19, 2021

Contributors

Binoy Chacko (chacko@uni-potsdam.de), Sreyas Sony (sony@uni-potsdam.de), Dinesh Kumar (kumar@uni-potsdam.de) Sanika Nair (nair@uni-potsdam.de)

Solution

Statistical modeling

A statistical model associated to that statistical experiment is a pair of

$$(\mathcal{X}, \mathcal{A}, (\mathbb{P}_\theta)_{\theta \in \Theta})$$

where $\mathcal{X}$ is the sample space, $(\mathbb{P}_\theta)_{\theta \in \Theta}$ is a family of probability measures on $(\mathcal{X}, \mathcal{A})$, where $\mathcal{A}$ is the $\sigma-$algebra Θ is the parameter set.

In the given statistical experiment, humans can have one of the three genotypes AA, Aa, aa with corresponding probabilities $(1-p)^2, 2p(1-p)$ and p^2.

$\mathcal{X} = \{AA, Aa, aa\}$

Θ or $p \in (0, 1)$

The distribution is unknown and we are assuming that it is a multinomial distribution because in every trial we are getting one of the three possibilities. For a multinomial distribution the individual probabilities should sum to one

$$\begin{aligned}\sum P(AA) + P(Aa) + P(aa) &= (1-p)^2 + 2p(1-p) + p^2 \\ &= (p+1-p)^2 \\ &= 1\end{aligned}$$

Let X_1, X_2, X_3 be the random variables that denotes the number of times getting the genotype AA, Aa, aa in n i.i.d. trials respectively.

Then,

$$\begin{aligned}X_1 &\sim Binomial(n, (1-p)^2) \\ X_2 &\sim Binomial(n, 2p(1-p)) \\ X_3 &\sim Binomial(n, p^2)\end{aligned}$$

Then,

$$X(X_1, X_2, X_3) \sim Multinominal(n, (p_1, p_2, p_3))$$

where $p_1 = (1-p)^2, p_2 = 2p(1-p)$ and $p_3 = p^2$ are the respective probabilities

Therefore, our statistical model for n independent trials of the experiment is

$$\Big(\mathcal{X} = \{AA, Aa, aa\},$$
$$\mathcal{A} = \{\phi, \{AA\}, \{Aa\}, \{aa\}, \{AA, Aa\}, \{AA, aa\}, \{Aa, aa\}, \mathcal{X}\},$$
$$Multinomial\left(n, (p_1, p_2, p_3)_{p_i \in (0,1)}\right)\Big)$$

Now, we have

$$\begin{aligned}P(X_1 = x, X_2 = y, X_3 = z) &= \binom{n}{x, y, z}(p_1^x p_2^y p_3^z) \\ &= \binom{N}{x, y, z}((1-p)^{2x} 2p(1-p)^y p^{2z})\end{aligned}$$

The likelihood function is given by

$$\begin{aligned}
L(X_1 = x, X_2 = y, X_3 = z|p) &= \left(\binom{n}{x, y, z}((1-p)^{2x}(2p(1-p))^y p^{2z})\right) \\
&= \left(\binom{n}{x, y, z}(1-p)^{2x+y} p^{2z+y}\right.
\end{aligned}$$

Taking log-likelihood is given by

$$\begin{aligned}
log\big(L(X_i|p)\big) &= log\left(\binom{n}{x, y, z}\right) + log\Big((1-p)^{2x+y}\Big) + log\Big(p^{2z+y}\Big) \\
&= log\left(\binom{n}{x, y, z}\right) + (2x+y)log(1-p) + (2z+y)log(p)
\end{aligned}$$

Taking the derivative w.r.t p

$$\begin{aligned}
\frac{\mathrm{d}log(L(X_i|p))}{\mathrm{d}p} &= \frac{-(2x+y)}{1-p} + \frac{2z+y}{p} = 0 \\
&- 2xp - 2yp + 2z + y - 2zp - py = 0 \\
2(x+y+z)p &= y + 2z \\
\hat{p} &= \frac{y+2z}{2(x+y+z)} = \frac{y+2z}{2n}
\end{aligned}$$

In order to check if $\hat{p}$ maximizes the likelihood, taking the second derivative of $\log(L(X_i|p))$, w.r.t. p,

$$\begin{aligned}
\frac{\partial^2}{\partial p^2}\log(L(p)) &= \frac{\partial}{\partial p}\left[\frac{-2x-y}{1-p} + \frac{2z+y}{p}\right] \\
&= -\left[\frac{(2x+y)}{(1-p)^2} + \frac{2z+y}{p^2}\right]
\end{aligned}$$

Substituting $\hat{p}$ in p,

$$\begin{aligned}
&= -\frac{2x+y}{\left(1 - \dfrac{2z+y}{2(x+y+z)}\right)^2} - \frac{2z+y}{\left(\dfrac{2z+y}{2(x+y+z)}\right)^2} \\
&= -\frac{(2(x+y+z))^2}{2x+y} - \frac{(2(x+y+z))^2}{2z+y} \\
&= -\left[\frac{4(x+y+z)^2}{2x+y} + \frac{4(x+y+z)^2}{2z+y}\right] < 0
\end{aligned}$$

Thus, $\hat{p}$ maximizes the likelihood.

Therefore the estimate of $\hat{p}$ is

$$\hat{p} = \frac{y + 2z}{2n}$$

Therefore the estimates of p_1, p_2, p_3 are

$$\begin{aligned}
\hat{p_1} &= \left(1 - \hat{p}\right)^2 \\
\hat{p_2} &= 2\left(\hat{p}\right)\left(1 - \hat{p}\right) \\
\hat{p_3} &= \hat{p}^2
\end{aligned}$$