hw2

Jiahao Tian

2023-01-29

Chapter 2

Question 1

- Show $\sqrt{\frac{n_r}{n}}$ is mle of p_r.

$$
\begin{aligned}
n_r &= number\ of\ people\ with\ genotype\ rr \\
n_R &= number\ of\ people\ with\ genotype\ Rr\ and\ RR \\
p(_{RR}) + p(_{Rr}) &= p_R^2 + 2p_R p_r \\
p(_{rr}) &= p_r^2
\end{aligned}
$$

- Find likelihood and log-likelihood, then take a derivative set to equal 0.

$$L(p_R, p_r) = k(p_R^2 + 2p_R p_r)^{n_R}(p_r^2)^{n_r},\ k\ is\ constant\ here$$

$$Becuase\ under\ HWE: \quad n = n_r + n_R \quad\quad p_r + p_R = 1$$

$$L(p_r) = k(1 - p_r^2)^{(n-n_r)}(p_r^2)^{n_r}$$

$$log(L(p_r)) \propto (n - n_r)log(1 - p_r^2) + n_r log(p_r^2)$$

$$\triangledown log(L(p_r)) \propto -\frac{(n - n_r)2p_r}{1 - p_r^2} + \frac{2n_r}{p_r}$$

$$Set\ to\ equal\ 0$$

$$\frac{2n_r}{p_r} = \frac{2p_r(n - n_r)}{1 - p_r^2}$$

$$2n_r - 2p_r^2 n_r = 2np_r^2 - 2p_r^2 n_r$$

$$\hat{p}_r = \sqrt{\frac{n_r}{n}}$$

- Check second derivative is less than 0.

$$\nabla^2 log(L(p_r)) = -\frac{2(np_r^4 + (n-3n_r)p_r^2 + n_r)}{p_r^2(p_r^2-1)^2} < 0$$

Question 5

- Observed data likelihood

$$\begin{aligned} L^O(p_{AB}, p_{Ab}, p_{aB}, p_{ab}) \propto\ & (p_{AB}^2)^{n_{AABB}} (p_{Ab}p_{AB})^{n_{AABb}} (p_{Ab}^2)^{n_{AAbb}} \\ & (p_{AB}p_{aB})^{n_{AaBB}} (2p_{AB}p_{ab} + 2p_{Ab}p_{aB})^{n_{AaBb}} (p_{Ab}p_{ab})^{n_{Aabb}} \\ & (p_{aB}^2)^{n_{aaBB}} (p_{aB}p_{ab})^{n_{aaBb}} (p_{ab}^2)^{n_{aabb}} \end{aligned}$$

- Complete data likelihood

$$\begin{aligned} L^C(p_{AB}, p_{Ab}, p_{aB}, p_{ab}) \propto\ & (p_{AB}^2)^{n_{AABB}} (p_{Ab}p_{AB})^{n_{AABb}} (p_{Ab}^2)^{n_{AAbb}} \\ & (p_{AB}p_{aB})^{n_{AaBB}} (2p_{AB}p_{ab})^{n_{AB|ab}} (2p_{Ab}p_{aB})^{n_{Aa|Bb}} (p_{Ab}p_{ab})^{n_{Aabb}} \\ & (p_{aB}^2)^{n_{aaBB}} (p_{aB}p_{ab})^{n_{aaBb}} (p_{ab}^2)^{n_{aabb}} \end{aligned}$$

- The two double heterozygous n_{AaBb} are "missing data" due to ambiguity.

First step, Expectation:

- Assume an initial value of $p_{AB}, p_{Ab}, p_{aB}, p_{ab}$, and estimate the "missing data" as proportions of the total people of n_{AaBb} of double heterozygous.

$$\begin{aligned} & \textit{Expected to see}: \\ & p(AB|ab) = 2p_{AB}p_{ab} \\ & p(Ab|aB) = 2p_{Ab}p_{aB} \end{aligned}$$

$$\begin{aligned} E(n_{AB|ab} \mid p_{AB}^m, p_{Ab}^m, p_{aB}^m, p_{ab}^m) &= \frac{p_{AB}^m p_{ab}^m}{p_{AB}^m p_{ab}^m + p_{Ab}^m p_{aB}^m} n_{AaBb} \\ E(n_{Ab|aB} \mid p_{AB}^m, p_{Ab}^m, p_{aB}^m, p_{ab}^m) &= \frac{p_{Ab}^m p_{aB}^m}{p_{AB}^m p_{ab}^m + p_{Ab}^m p_{aB}^m} n_{AaBb} \\ {n_{AB|ab}}^m &= \frac{p_{AB}^m p_{ab}^m}{p_{AB}^m p_{ab}^m + p_{Ab}^m p_{aB}^m} n_{AaBb} \end{aligned}$$

- So the number of people with genotype AB|ab from our "guess" is as above.
- Where "m" is the current step.
- Then need to construct a function Q and maximize function Q to get a new estimator .s.t $(P_{AB})^{m+1}$ of the parameter. Then use the new "guess" of estimator $(P_{AB})^{m+1}$ to create new guess at what the value of n_{AaBb} are and then give the value of n_{AaBb} to construct a new function Q and maximize the new function Q to obtain the results.

- Function $Q = E_{n_{AB|ab}, n_{Aa|Bb} | p^m_{AB}, p^m_{Ab}, p^m_{aB}, p^m_{ab}} [log(complete\ data\ likelihood)]$
- Complete data log-likelihood

$$
\begin{aligned}
log(L^C(p_{AB}, p_{Ab}, p_{aB}, p_{ab})) \propto & \left[2n_{AABB} + n_{AABb} + n_{AaBB} + n_{AB|ab}\right] \ln p_{AB} + \\
& \left[n_{AABb} + 2n_{AAbb} + n_{Aa|Bb} + n_{Aabb}\right] \ln p_{Ab} + \\
& \left[n_{AaBB} + n_{Aa|Bb} + 2n_{aaBB} + n_{aaBb}\right] \ln p_{aB} + \\
& \left[n_{AB|ab} + n_{Aabb} + n_{aaBb} + 2n_{aabb}\right] \ln p_{ab}
\end{aligned}
$$

- Function Q

$$
\begin{aligned}
Q = & E_{n_{AaBb}, n_{Aa|Bb} | p^m_{AB}, p^m_{Ab}, p^m_{aB}, p^m_{ab}} [log(L^C(p_{AB}, p_{Ab}, p_{aB}, p_{ab}))] \\
& \propto \left[2n_{AABB} + n_{AABb} + n_{AaBB} + E(n_{AB|ab} \mid p^m_{AB}, p^m_{Ab}, p^m_{aB}, p^m_{ab})\right] \ln p_{AB} + \\
& \left[n_{AABb} + 2n_{AAbb} + E(n_{Aa|Bb} \mid p^m_{AB}, p^m_{Ab}, p^m_{aB}, p^m_{ab}) + n_{Aabb}\right] \ln p_{Ab} + \\
& \left[n_{AaBB} + E(n_{Aa|Bb} \mid p^m_{AB}, p^m_{Ab}, p^m_{aB}, p^m_{ab}) + 2n_{aaBB} + n_{aaBb}\right] \ln p_{aB} + \\
& \left[E(n_{AB|ab} \mid p^m_{AB}, p^m_{Ab}, p^m_{aB}, p^m_{ab}) + n_{Aabb} + n_{aaBb} + 2n_{aabb}\right] \ln p_{ab}
\end{aligned}
$$

Second step, Maximization:

- An initial estimate p^m_{AB} is put into the right hand side to give an updated estimated p^{m+1}_{AB} on the left hand side.

$$
\begin{aligned}
p^{m+1}_{AB} &= \frac{2n_{AABB} + n_{AABb} + n_{AaBB} + n^m_{AB|ab}}{2n} \\
p^{m+1}_{AB} 2n &= 2n_{AABB} + n_{AABb} + n_{AaBB} + \frac{p^m_{AB} p^m_{ab}}{p^m_{AB} p^m_{ab} + p^m_{Ab} p^m_{aB}} n_{AaBb}
\end{aligned}
$$

- Implement this EM algorithm on the mosquito data.

$$
\begin{aligned}
& From\ the\ table\ 'n = 40' \\
& n_{AA} = 25,\ n_{Aa} = 16,\ n_{aa} = 0 \\
& n_{BB} = 27,\ n_{Bb} = 13,\ n_{bb} = 0 \\
& \\
& Then\ we\ can\ get\ '2n = 80' \\
& n_A = 64,\ n_a = 16 \\
& n_B = 67,\ n_b = 13 \\
& \\
& p_A = 0.8,\ p_B = 0.8375,\ p_a = 0.2,\ p_b = 0.1625 \\
& \\
& Our\ initial\ starting\ point : p^m_{AB} = p_A * p_B,\ so\ on \\
& each\ iteration\ will\ update\ n_{AaBb}
\end{aligned}
$$

- Plug above information back to the equation of p^{m+1}_{AB}.

$$p_{AB}^{m+1} = \frac{38+5+8+n_{AB|ab}^{m}}{2n}$$

$$p_{AB}^{m+1}2n = 51 + \frac{p_{AB}^{m}p_{ab}^{m}}{p_{AB}^{m}p_{ab}^{m} + p_{Ab}^{m}p_{aB}^{m}}n_{AaBb}$$

```
geno = function(pAB0, pab0, pAb0, paB0, n_iter) {

nAABB = 19
nAABb = 5
nAaBB = 8
nAaBb = 8
nAAbb = 0
nAabb = 0
naabb = 0
naaBb = 0
naaBB = 0
n = nAABB + nAABb + nAaBB + nAaBb

pAB = rep(0, n_iter)
pab = rep(0, n_iter)
pAb = rep(0, n_iter)
paB = rep(0, n_iter)

pAB[1] = pAB0
pab[1] = pab0
pAb[1] = pAb0
paB[1] = paB0

for(i in 1:n_iter) {
  nABabm = ((pAB[i] * pab[i]) / (pAB[i] * pab[i] + pAb[i] * paB[i])) * nAaBb
  nAbaBm = ((pAb[i] * paB[i]) / (pAb[i] * paB[i] + pAB[i] * pab[i])) * nAaBb

  pAB[i+1] = (2 * nAABB + nAABb + nAaBB + nABabm) / (n * 2)
  pAb[i+1] = (nAABb + 2 * nAAbb + nAbaBm + nAabb) / (n * 2)
  paB[i+1] = (nAaBB + nAbaBm + 2 * naaBB + naaBb) / (n * 2)
  pab[i+1] = (nABabm + nAabb + naaBb + 2 * naabb) / (n * 2)
}

list(pAB = pAB, pab = pab, pAb = pAb, paB = paB)

}
```

```
set.seed(22)
pAB0 = 0.67
pab0 = 0.33
pAb0 = 0.134
paB0 = 0.1675
geno(pAB0, pab0, pAb0, paB0, 5)
```

```
## $pAB
## [1] 0.6700000 0.7282840 0.7269082 0.7264383 0.7262742 0.7262166
##
```

```
## $pab
## [1] 0.33000000 0.09078404 0.08940822 0.08893827 0.08877424 0.08871655
##
## $pAb
## [1] 0.13400000 0.07171596 0.07309178 0.07356173 0.07372576 0.07378345
##
## $paB
## [1] 0.1675000 0.1092160 0.1105918 0.1110617 0.1112258 0.1112834
```

Chapter 4

Question 1

- H_0: HWE holds in the population.
- H_a: HWE does not hold in the population.

$$\begin{aligned} Phenotype &: MM \quad MN \quad NN \\ Genotype &: M/M \quad M/N \quad N/N \\ (Observed)\ Number &: 119 \qquad 76 \qquad 13 \end{aligned}$$

$$\hat{p}_M = \frac{2 * 119 + 76}{2 * 208} = 0.755$$

$$\hat{p}_N = \frac{2 * 13 + 76}{2 * 208} = 0.245$$

$$\begin{aligned} &Expected\ value \\ n_{MM} &= 208 * 0.755^2 = 118.57 \\ n_{NM} &= 208 * 2 * 0.755 * 0.245 = 76.95 \\ n_{NN} &= 208 * 0.245^2 = 12.49 \end{aligned}$$

$$\begin{aligned} \chi^2 &= \sum \frac{(O-E)^2}{E} \\ &= \frac{(119-118.57)^2}{118.57} + \frac{(76-76.95)^2}{76.95} + \frac{(13-12.49)^2}{12.49} \\ &= 0.0341 \end{aligned}$$

- With DF = 1, and at $\alpha = 0.05$, $0.034 < \chi^2_{1,0.05} = 3.841$. We do not reject H_0. Thus the population is in HWE.

Question 2

- H_0: HWE holds in the population.
- H_a: HWE does not hold in the population.

$$\begin{aligned} Female &: \\ & t/t = 63 \;\; t/y = 55 \;\; y/y = 12 \\ Male &: \\ & t = 74 \quad\; y = 38 \end{aligned}$$

- Phenotypic frequencies

$$\begin{aligned} Female &: \\ & \hat{q}_{fy} = \frac{12}{130} = 0.092; \;\; \hat{q}_{ft} = \frac{63}{130} = 0.485; \;\; \hat{q}_{fty} = \frac{55}{130} = 0.423 \\ Male &: \\ & \hat{q}_{my} = \frac{38}{112} = 0.339; \;\; \hat{q}_{mt} = \frac{74}{112} = 0.661 \end{aligned}$$

- Gene counting (both male and female)

$$\hat{p}_t = \frac{2 * 63 + 55 + 74}{(63 + 55 + 12) * 2 + 74 + 38} = 0.685$$

$$\hat{p}_y = \frac{55 + 2 * 12 + 38}{(63 + 55 + 12) * 2 + 74 + 38} = 0.315$$

- LRT
- nmt = number of male is genotype t, nft = number of female is genotype t/t, etc.

$$\begin{aligned} 2log\frac{L(\hat{q})}{L(\hat{p})} &= 2log\left(\frac{\hat{q}_{fy}^{nfy} * \hat{q}_{ft}^{nft} * \hat{q}_{fty}^{nfty} * \hat{q}_{mt}^{nmt} * \hat{q}_{my}^{nmy}}{(\hat{p}_t^2)^{nft} * (\hat{p}_y^2)^{nfy} * (2\hat{p}_t\hat{p}_y)^{nfty} * (\hat{p}_t)^{nmt} * (\hat{p}_y)^{nmy}}\right) \\ &= 2log(1.254) \\ &= 0.1966 \end{aligned}$$

- With DF = 1, and at $\alpha = 0.05$, $0.1966 < \chi^2_{1,0.05} = 3.841$. We do not reject H_0. Thus the population is in HWE.