Løsningsforslag ST2301 øving 3

Kapittel 1

Exercise 11

Et utvalg på 100 individer trekkes fra en populasjon med tilfeldig parring. Det ble observert $n_{AA} = 63$ individer av genotype AA, $n_{Aa} = 27$, og $n_{aa} = 10$. Lag et 95 % konfidensintervall for frekvensen til A. Hva må du anta?

Svar:

Finner først frekvensen av A i utvalget,

$$\hat{p}_A = \frac{2n_{AA} + n_{Aa}}{2n} = \frac{2 \cdot 63 + 27}{200} = 0.765$$

Et estimat for variansen til $\hat{p}_A$ er gitt ved

$$\hat{\sigma}^2 \approx \frac{\hat{p}_A(1 - \hat{p}_A)}{2n} = \frac{0.765(1 - 0.765}{200} \approx 8.99 \cdot 10^{-4}$$
$$\hat{\sigma} \approx 0.0300$$

Et tilnærma 95 % konfidensintervall for p_A blir dermed

$$\hat{p}_A - 1.96\hat{\sigma} \leq p_A \leq \hat{p}_A + 1.96\hat{\sigma}$$
$$0.765 - 1.96 \cdot 0.0300 \leq p_A \leq 0.765 + 1.96 \cdot 0.0300$$
$$0.7062 \leq p_A \leq 0.8238$$

For å konstruere konfidensintervallet ble det antatt at $\hat{p}_A$ er tilnærma normalfordelt med forventning p_A og varians σ^2, der

$$\hat{\sigma}^2 \approx \frac{p_A(1 - p_A)}{2n}.$$

Siden p_A er ukjent (det er den det skulle lages konfidensintervall for), måtte vi bruke et estimat, der $\hat{p}_A$ selv inngår. Tilnærminga av binomisk fordeling til normalfordeling fungerer bra så lenge p_A ikke er for nær 0 eller 1, og så lenge utvalget er stort nok (et vanlig krav er at $np_A > 5$ og $n(1 - p_A) > 5$, som er oppfylt her).

Exercise 12

Har et utvalg på 100 individer, og observerer 10 individer med genotype aa. Hvis vi antar tilfeldig parring, hvordan kan vi finne et 95% konfidensintervall for frekvensen av A?

Svar:

For å lage konfidensintervall for frekvensen av A trenger vi først et punktestimat $\hat{p}_A$ for frekvensen. Vi antar at vi kan bruke normaltilnærming til binomisk fordeling, slik at et tilnærma 95% konfindensintervall er gitt ved formelen $\hat{p}_A \pm 1.96\hat{\sigma}$. Som estimat for variansen kan vi f.eks bruke

$$\hat{\sigma}^2 = \frac{\hat{p}_A(1-\hat{p}_A)}{2n}$$

Det er bare utvalgsfrekvensen av genotype aa som er kjent. Vi kjenner ikke antallet av genotype AA eller Aa, og kan derfor ikke bruke maximum likelihood-estimatoren $\hat{p}_A = \frac{n_{AA}+n_{Aa}}{2n}$ å finne $\hat{p}_A$. Vi må i stedet basere estimatet på den frekvensen som er oppgitt, nemlig $\hat{P}_{aa} = 10/100$. Det gir

$$\begin{aligned} 1-\hat{p}_A &= \sqrt{\hat{P}_{aa}} \\ \hat{p}_A &= 1-\sqrt{0.1} \approx 0.684 \end{aligned}$$

Estimatet for $\hat{\sigma}$ blir

$$\begin{aligned} \hat{\sigma} &= \sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{2n}} \\ &= \sqrt{\frac{(1-\sqrt{0.1})\sqrt{0.1}}{200}} \\ &\approx 0.0329 \end{aligned}$$

Et tilnærma 95% konfidensintervall for p_A blir dermed

$$\begin{aligned} \hat{p}_A - 1.96\hat{\sigma} \leq p_A \leq \hat{p}_A + 1.96\hat{\sigma} \\ 0.684 - 1.96 \cdot 0.0329 \leq p_A \leq 0.684 + 1.96 \cdot 0.0329 \\ 0.620 \leq p_A \leq 0.748 \end{aligned}$$

Exercise 13

Vi trekker 200 individer fra en diploid populasjon, og finner følgende antall individer av hver genotype:
Test hypotesen at dette utvalget kommer fra en populasjon med Hardy-Weinbergandeler.

n_{AA}	n_{Aa}	n_{aa}
89	57	54

Svar:

Her kan vi bruke χ^2-testen som står beskrevet i boka. Først må vi finne forventa antall av hver genotype, som er betegna med N i boka. Da trenger vi et estimat for allelfrekvensen.

$$\begin{aligned}\hat{p}_A &= \frac{2n_{AA} + n_{Aa}}{2n} \\ &= \frac{2 \cdot 89 + 57}{400} \\ &= 0.5875\end{aligned}$$

Dette gir følgende forventa antall av hver genotype (eks $N_{AA} = p_A^2 \cdot 200$):

N_{AA}	N_{Aa}	N_{aa}
69.03	96.94	34.03

Testobservatoren χ^2 blir

$$\begin{aligned}\chi^2 &= \frac{(n_{AA} - N_{AA})^2}{N_{AA}} + \frac{(n_{Aa} - N_{Aa})^2}{N_{Aa}} + \frac{(n_{aa} - N_{aa})^2}{N_{aa}} \\ &= \frac{(89 - 69.03)^2}{69.03} + \frac{(57 - 96.94)^2}{96.94} + \frac{(54 - 34.03)^2}{34.03} \\ &\approx 33.952\end{aligned}$$

Antall genotyper er $k = 3$, og antall uavhengig estimerte genfrekvenser er $m = 1$, så antall frihetsgrader er $k - m - 1 = 1$. Ifølge tabell er

$$P(\chi_1^2 > 33.952) < 0.001$$

Dvs vi forkaster hypotesen om at utvalget kommer fra en populasjon med Hardy-Weinbergandeler.

Exercise 14

Vi trekker 100 individer fra en diploid populasjon og finner følgende antall genotyper ved et locus med to allel:

	n_{BB}	n_{Bb}	n_{bb}
n_{AA}	0	25	0
n_{Aa}	25	0	25
n_{aa}	0	25	0

Bruk en 3×3 χ^2-test for å teste om genotypene ved disse lociene er fordelt uavhengig av hverandre. Se om det er mulig å lage et estimat for koplingsulikevekten D_{AB}. Er det en motsetning mellom de to konklusjonene? Hvorfor, eventuelt hvorfor ikke?

Svar:

Finner først estimater for genfrekvensen til A og B:

$$\begin{aligned}\hat{p}_A &= \frac{2n_{AA} + n_{Aa}}{2n} \\ &= \frac{2 \cdot 25 + 50}{200} \\ &= 0.5\end{aligned}$$

$$\begin{aligned}\hat{p}_B &= \frac{2n_{BB} + n_{Bb}}{2n} \\ &= \frac{2 \cdot 25 + 50}{200} \\ &= 0.5\end{aligned}$$

Forventa antall av genotype $AABB$ er

$$\begin{aligned}N_{AABB} &= p_A^2 p_B^2 \cdot 100 \\ &= 0.5^2 \cdot 0.5^2 \cdot 100 \\ &= 6.25\end{aligned}$$

Tilsvarende utregninger gir forventa antall for de andre genotypene, som kan summeres i en tabell:

	N_{BB}	N_{Bb}	N_{bb}
N_{AA}	6.25	12.5	6.25
N_{Aa}	12.5	25	12.5
N_{aa}	6.25	12.5	6.25

Testobservatoren blir (slår sammen de som har lik observert og Forventa antall)

$$\begin{aligned}\chi^2 &= 4 \cdot \frac{(n_{AABB} - N_{AABB})^2}{N_{AABB}} + 4 \cdot \frac{(n_{AaBB} - N_{AaBB})^2}{N_{AaBB}} + \frac{(n_{AaBb} - N_{AaBb})^2}{N_{AaBb}} \\ &= 4 \cdot \frac{(0 - 6.25)^2}{6.25} + 4 \cdot \frac{(25 - 12.5)^2}{12.5} + \frac{(0 - 25)^2}{25} \\ &= 100\end{aligned}$$

Her er det $k = 9$ genotyper, og $m = 2$ uavhengig estimerte genfrekvenser, så antall frihetsgrader blir $k - m - 1 = 6$. Har fra tabell at

$$P(\chi_6^2 > 100) < 0.001$$

Dvs genotypene ved de to lociene er ikke uavhengig fordelte.

Et estimat for koplingsulikevekten er

$$\begin{aligned}\hat{D}_{AB} &= \hat{P}_{AB} - \hat{p}_A\hat{p}_B \\ &= 0.25 - 0.5 \cdot 0.5 \\ &= 0\end{aligned}$$

Dette resultatet skyldes symmetrien i modellen. Det betyr at selv om man ikke finner en koplingsulikevekt, kan det være avhengighet mellom lociene.

Alternativ til χ^2-testen

Dersom fordelingen av genotypene ved de to lociene er uavhengige, så er $P_{AABB} = P_{AA}P_{BB}$, $P_{AABb} = P_{AA}P_{Bb}$, osv. Finner genotypefrekvensene for hvert locus

$$\hat{P}_{AA} = \frac{n_{AABB} + n_{AABb} + n_{AAbb}}{100} = \frac{0 + 25 + 0}{100} = 0.25$$

$$\begin{aligned}\hat{P}_{aa} &= \hat{P}_{BB} = \hat{P}_{bb} = 0.25 \\ \hat{P}_{Aa} &= \hat{P}_{Bb} = 0.5\end{aligned}$$

Dette gir for eksempel

$$\begin{aligned}\hat{P}_{AA}\hat{P}_{BB} &= 0.25 \cdot 0.25 = 0.0625 \\ \hat{P}_{AABB} &= 0 \neq \hat{P}_{AA}\hat{P}_{BB}\end{aligned}$$

Dvs fordelingen av genotyper er ikke uavhengig mellom lociene.

Oppgave, EM-algoritmen

Anta at vi trekker et utvalg på n individ fra en populasjon i Hardy-Weinberg-likevekt og at AA-homozygoter ikke kan skilles fra heterozygoter, type Aa.

1. Hvilken fordeling har det observerte antallet homozygoter av type aa, n_{aa}?
2. Hva blir sannsynlighetsmaksimeringsestimatet (SME) av P_{aa}? Og av allelfrekvensen p_a?
3. Anta at vi i stedet ønsker å estimere p_a ved hjelp av EM-algoritmen. Hva blir forventningsverdien til de manglende dataene n^*_{AA} og n^*_{Aa} betinga på observerte data og gitt forrige estimat av p_a (E-steget i algoritmen)?
4. Hva blir estimatet av p_a basert på de "fullstendige" dataene (n^*_{AA}, n^*_{Aa}, n_{aa}) (M-steget)?
5. Vis at estimatet av p_a konvergerer mot sannsynlighetsmaksimeringsestimatet av p_a, altså at algoritmen fungerer i denne situasjonen.

Svar:

1. Simultanfordelingen av n_{AA}, n_{Aa} og n_{aa} er den multinomiske fordelingen, med parametre n, P_{AA}, P_{Aa} og P_{aa}. Marginalfordelingen av n_{aa} er den binomiske fordelingen, med parametrene n og P_{aa}.

2. For å finne SME for P_{aa} må man maksimere likelihood-funksjonen for n_{aa} gitt P_{aa}. Har at n_{aa} er binomisk fordelt, dvs

$$L(n_{aa}|P_{aa}, n) = \binom{n}{n_{aa}} P_{aa}^{n_{aa}} (1 - P_{aa})^{n - n_{aa}}$$

Det er best å maksimere log-likelihood-funksjonen $\ln L$, dvs

$$\ln L(n_{aa}|P_{aa}) = \ln \binom{n}{n_{aa}} + n_{aa} \ln P_{aa} + (n - n_{aa}) \ln(1 - P_{aa})$$

For å maksimere denne m.h.p. P_{aa} deriverer vi m.h.p. P_{aa} og setter lik 0.

$$\begin{aligned}
\frac{\partial}{\partial n_{aa}} \ln L(n_{aa}|P_{aa}, n) &= 0 \\
\frac{n_{aa}}{\hat{P}_{aa}} - \frac{n - n_{aa}}{1 - \hat{P}_{aa}} &= 0 \\
n_{aa}(1 - \hat{P}_{aa}) &= (n - n_{aa})\hat{P}_{aa} \\
(n - n_{aa} + n_{aa})\hat{P}_{aa} &= n_{aa} \\
\hat{P}_{aa} &= \frac{n_{aa}}{n}
\end{aligned}$$

SME for P_{aa} er altså andelen aa-individer i utvalget. Siden dataene n^*_{AA} og n^*_{Aa} mangler, blir SME for p_a roten av estimatet for P_{aa}.

$$\hat{p}_a = \sqrt{\hat{P}_{aa}}$$

3. "Expectation"-steget. Betinget på $\hat{p}_a$, n og de observerte dataene n_{aa}, er de uobserverte dataene n^*_{AA} og n^*_{Aa} multinomisk fordelt, med parametre $n - n_{aa}$ og

$$
\begin{aligned}
P_{AA}|n, n_{aa}, \hat{p}_a &= \frac{P_{AA}}{P_{AA} + P_{Aa}} \\
&= \frac{(1-\hat{p}_a)^2}{(1-\hat{p}_a)^2 + 2\hat{p}_a(1-\hat{p}_a)} \\
&= \frac{1-\hat{p}_a}{1+\hat{p}_a} \\
P_{Aa}|n, n_{aa}, \hat{p}_a &= \frac{P_{Aa}}{P_{AA} + P_{Aa}} \\
&= \frac{2\hat{p}_a}{1+\hat{p}_a}
\end{aligned}
$$

Dvs den betinga forventningen for n_{Aa} er

$$
\mathrm{E}[n^*_{Aa}|n, n_{aa}, \hat{p}_a] = \frac{2(n-n_{aa})\hat{p}_a}{1+\hat{p}_a}
$$

4. "Maximization"-steget. Dersom man hadde hatt de manglende dataene for n^*_{Aa}, ville SME for $\hat{p}_a$ vÊrt

$$
\hat{p}_a = \frac{2n_{aa} + n^*_{Aa}}{2n}
$$

La foregående estimat av p_a vÊre $\hat{p}'_a$, det f¯r der igjen $\hat{p}''_a$ osv. Det nye estimatet av p_a blir likningen over, men innsatt den betinga forventningsverdien for n^*_{Aa}, gitt de observerte dataene n_{aa} og den foregående parameteren $\hat{p}'_a$, i stedet for n^*_{Aa}. Dvs det nye estimatet av p_a er

$$
\hat{p}_a = \frac{2n_{aa} + \mathrm{E}[n^*_{Aa}|n, \hat{p}'_a]}{2n}
$$

5. Skal vise at algoritmen konvergerer mot SME for $\hat{p}_a$. Når den har konvergert, er $\hat{p}'_a \approx \hat{p}_a$. Setter $\hat{p}'_a = \hat{p}_a$ i uttrykket for $\hat{p}_a$ over for å se om det gir SME

som ble funnet i punkt 2.

$$\begin{aligned}
\hat{p}_a &= \frac{2n_{aa} + \mathrm{E}[n^*_{Aa}|n, \hat{p}_a]}{2n} \\
&= \frac{n_{aa} + \frac{(n-n_{aa})\hat{p}_a}{1+\hat{p}_a}}{n} \\
&= \frac{n_{aa}}{n} + \frac{(n-n_{aa})\hat{p}_a}{n(1+\hat{p}_a)} \\
\hat{p}_a - \hat{p}_a^2 &= \frac{n_{aa}}{n} + \hat{p}_a \frac{n_{aa}}{n} + \hat{p}_a - \hat{p}_a \frac{n_{aa}}{n} \\
\hat{p}_a &= \sqrt{\frac{n_{aa}}{n}}
\end{aligned}$$

Dvs algoritmen konvergerer mot SME for $\hat{p}_a$.