Prosjekt: Korrelasjon mellom genotyper i to lokus

Veileder: Øyvind Bakke

Bakgrunn: I GWA-studier (genomwide association studies) er det av interesse å undersøke om det er sammenheng mellom en bestemt sykdom og genotype. Genotypen i m (mange hundre tusen) lokus, som hvert består av ett basepar på et kromosom, kartlegges, både hos friske individer og hos individer med den aktuelle sykdommen. Genotypen kodes som 0, 1 eller 2, som svarer til genotyper hhv. aa, aA og AA (to homozygoter og én heterozygot).

For å undersøke en slik sammenheng, er én mulighet å gjøre en hypotesetest for hvert lokus. Da må imidlertid signifikansnivået for hver enkelt test, α_{local} settes kraftig ned for å holde sannsynligheten for minst én type I-feil begrenset til et valgt nivå (f.eks. $\alpha = 0{,}05$). Det er mange måter å gjøre dette på. Den mest kjente er Bonferroni-korreksjon, der signifikansnivået for den enkelte testen reduseres til α/m.

En annen metode er å estimere og utnytte korrelasjonsstrukturen mellom testobservatorene for hvert lokus. Ideelt burde man ha brukt hele simultanfordelingen, men man kan få en konservativ skranke for α_{local} ved bare å bruke simultanfordelingene til to og to nabolokus.

Problem: Anta at to nabolokus, f.eks. nr. i og $i+1$ har simultan sannsynlighetsfordeling p_{jk} for genotype j på lokus i og genotype k på lokus $i+1$, der $j = 0,\ 1,\ 2$ og $k = 0,\ 1,\ 2$, og $\sum_{jk} p_{jk} = 1$. Definer marginalsannsynlighetene $p_{j\cdot} = \sum_k p_{jk}$ og $p_{\cdot k} = \sum_j p_{jk}$.

For en av flere mulige testobservatorer, $\text{CATT}_{1/2}$, er den asymptotiske korrelasjonen under nullhypotesen mellom denne testobservatoren i lokus i og i lokus $i+1$

$$\rho_{i,i+1} = \frac{\frac{1}{4}(p_{11} - p_{1\cdot}p_{\cdot 1}) + \frac{1}{2}(p_{12} - p_{1\cdot}p_{\cdot 2}) + \frac{1}{2}(p_{21} - p_{2\cdot}p_{\cdot 1}) + p_{22} - p_{2\cdot}p_{\cdot 2}}{\sqrt{\left(\frac{1}{4}p_{1\cdot} + p_{2\cdot} - (\frac{1}{2}p_{1\cdot} + p_{2\cdot})^2\right)\left(\frac{1}{4}p_{\cdot 1} + p_{\cdot 2} - (\frac{1}{2}p_{\cdot 1} + p_{\cdot 2})^2\right)}}.$$

For å undersøke hvor godt korreksjonsmetoden for multippel testing fungerer, er det av interesse å kunne simulere genotyper slik at korrelasjonen $\rho_{i,i+1}$ mellom to nabolokus får en ønsket verdi.

1. Hvordan kan p_{jk}-ene velges for å få en gitt $\rho_{i,i+1}$, $-1 \le \rho_{i,i+1} \le 1$?

La oss døpe om p_{jk} til $p_{jk}(i, i+1)$, og vi definerer tilsvarende simultanfordelingen $p_{jk}(i+1, i+2)$ for genotypene i lokus $i+1$ og $i+2$ osv. Merk at marginalfordelingen for lokus $i+1$ må være den samme om vi betrakter simultanfordelingen til i og $i+1$ eller simultanfordelingen til $i+1$ og $i+2$, osv.

2. Hvis vi har flere nabolokus, i, $i+1$, $i+2$ osv. – er det da mulig å velge simultanfordelingene for to og to nabolokus slik at $\rho_{i,i+1}$, $\rho_{i+1,i+2}$ får ønskede verdier? Hvilke begrensninger fins det eventuelt?

Forkunnskaper: Grunnkurs i statistikk (ST1101/TMA4240/TMA4245). I tillegg er det en fordel med noe programmeringserfaring i MATLAB, R eller annet.

Opplæring: Veileder vil gi mer utfyllende informasjon.