Assignment 04

PUBH 8878

**Requirements**

* Show complete mathematical work for any derivations.
* Submit well-documented **R code** with clear comments and a fixed seed (set.seed(8878)).
* Interpret results in a **genetic/biological** context where applicable.
* **Submit the rendered PDF only** (do not submit the source .qmd).

A helpful vignette on using cmdstanr is available at <https://mc-stan.org/cmdstanr/articles/cmdstanr.html>.

## Problem 1: Population Substructure and Allele Frequency Estimation (30 pts)

Let there be diploid individuals sampled at random from a population made of subpopulations . In subpopulation , the allele frequency is , and let be the fraction of sampled individuals from subpopulation . The “global” allele frequency we want to estimate is the mixture average

Let be the counts of genotypes in the sample of size . The standard estimator of the allele frequency is

**(a)** Show that in the presence of population substructure, is unbiased.

*Hint:* Let be the -dosage for individual (). Under HWE within , . Use LOTUS: .

**(b)** What is under *random mixture sampling* (i.i.d. individuals from the mixture)? Assume each subpopulation is in HWE. Compare to the case with no substructure ().

*Hint:* Use the law of total variance on :  
, with and .

Define

and note .

**(c)** Now consider a *stratified* sample: take exactly individuals from subpopulation (; write ). Maintain HWE within subpopulations. Derive under this design and compare it to your answer in (b). State which design yields the larger variance, and by how much, in terms of .

*Hint:* Write with . Use independence across strata and .

## Problem 2: Population Substructure, LD, and Association Testing (40 pts)

Let be the genotype dosage (0/1/2 copies of the effect allele) at a tag SNP and at a causal SNP. The causal SNP has effect size on quantitative trait . The observed effect from simple regression of on is .

**Convenience (scaling):** Work with **standardized** genotypes

With this scaling, exactly, where .

Assume individuals are sampled i.i.d. from a mixture with , . Within each subpopulation : - HWE holds at each locus, - LE (no within- LD) holds between and .

Define

### Part A (10 pts): Correlation induced by population structure

Let the allele frequencies at the tag and causal SNPs be and in subpopulation . Show that

and express , , and in terms of .

*Hint:* Law of total covariance:  
. Under LE, the first term is 0. Use .

### Part B (10 pts): Bias from ignoring structure in the trait

Suppose population structure also affects the trait mean: . Consider the model with . Show that the naïve regression of on (without structure covariates) is biased:

Under the assumptions above, prove that , and give the bias in terms of .

*Hint:* .

### Part C (20 pts): Brief interpretation

In a few sentences each:

1. Explain why can arise even if **within** each subpopulation there is no LD. What feature of the mixture induces it?
2. Give a sign‑consistent example: if subpopulations with larger also have larger , what is the expected direction of the naïve bias?
3. Name **two** standard strategies to mitigate both components of bias (structure‑induced and trait mean differences) in practice.

## Problem 3: Bayesian Analysis (30 pts)

### Part A (10 pts): Beta–Binomial conjugacy

1. With prior and data successes out of , write the posterior distribution for .
2. Compute the posterior mean and a **central 95% credible interval** in R using qbeta. Compare to the MLE . Briefly interpret the *shrinkage*.
3. **Sensitivity:** repeat with priors and . Summarize how the posterior mean and width change across priors, and why.

### Part B (10 pts): Beta–Binomial in Stan

1. **Write a Stan model** to estimate the allele frequency from Binomial data with a prior. Use .
2. **Run the model** in R using cmdstanr. **Check convergence** and effective sample size; report and bulk ESS for .
3. **Summarize** the posterior mean and a central 95% credible interval. Compare to your analytical result from Part A.

You will need to install cmdstanr and cmdstan if you haven’t already. Please follow installation instructions at <https://mc-stan.org/cmdstanr/>

Boilerplate is provided below. You will need to set eval to TRUE to run the code when knitting the final document:

library(cmdstanr)  
  
stan\_beta\_binomial <- "  
data {  
 int<lower=0> n;  
 int<lower=0, upper=n> x;  
 real<lower=0> a;  
 real<lower=0> b;  
}  
parameters {  
 real<lower=0, upper=1> p;  
}  
model {  
 // TODO: prior on p  
 // Example: p ~ beta(a, b);  
 // TODO: likelihood  
 // Example: x ~ binomial(n, p);  
}  
generated quantities {  
 real logit\_p = logit(p);  
 int x\_rep = binomial\_rng(n, p);  
}  
"  
  
library(cmdstanr)  
set.seed(8878)  
  
writeLines(stan\_beta\_binomial, con = "beta\_binomial.stan")  
mod\_bb <- cmdstan\_model("beta\_binomial.stan")  
fit\_bb <- mod\_bb$sample(  
 data = list(n = 27, x = 11, a = 4, b = 18),  
 seed = 8878, chains = 4, parallel\_chains = 4,  
 iter\_warmup = 1000, iter\_sampling = 1000  
)  
  
# TODO: check convergence and summarize posterior

### Part C (10 pts): ABO blood group frequencies in Stan (missing AB phenotype)

We observe phenotype counts in a population sample where **AB individuals are not sampled**:



Under HWE with allele frequencies , the **unconditional** phenotype probabilities are

Because AB is missing, the **observed** category probabilities are the renormalized values

1. **Write a Stan model** that estimates with prior and a **Multinomial** likelihood on using .
2. **Run the model** and check convergence (report and ESS).
3. **Prior sensitivity:** re‑run with for . Summarize how posterior means and credible intervals change with , and why.

Boilerplate is provided below:

stan\_abo\_missing\_ab <- "  
data {  
 int<lower=0> n\_A;  
 int<lower=0> n\_B;  
 int<lower=0> n\_O;  
 vector<lower=0>[3] alpha;  
}  
transformed data {  
 int N = n\_A + n\_B + n\_O;  
 int y[3] = { n\_A, n\_B, n\_O };  
}  
parameters {  
 simplex[3] p;  
}  
transformed parameters {  
 // Unconditional phenotype probabilities under HWE:  
 real PrA = square(p[1]) + 2 \* p[1] \* p[3];  
 real PrB = square(p[2]) + 2 \* p[2] \* p[3];  
 real PrAB = 2 \* p[1] \* p[2];  
 real PrO = square(p[3]);  
  
 // Observed (AB excluded): renormalize by (1 - PrAB)  
 simplex[3] q;  
 {  
 real denom = 1 - PrAB;  
 q[1] = PrA / denom;  
 q[2] = PrB / denom;  
 q[3] = PrO / denom;  
 }  
}  
model {  
 // TODO: prior on allele frequencies  
 // Example: p ~ dirichlet(alpha);  
 // TODO: likelihood for observed counts  
 // Example: y ~ multinomial(q);  
}  
generated quantities {  
 // Unconditional phenotype probabilities (optional checks)  
 vector[4] phen\_prob = [PrA, PrB, PrAB, PrO]';  
}  
"  
  
set.seed(8878)  
writeLines(stan\_abo\_missing\_ab, con = "abo\_missing\_ab.stan")  
mod\_abo <- cmdstan\_model("abo\_missing\_ab.stan")  
fit\_abo <- mod\_abo$sample(  
 data = list(n\_A = 725, n\_B = 258, n\_O = 1073, alpha = c(1, 1, 1)),  
 seed = 8878, chains = 4, parallel\_chains = 4,  
 iter\_warmup = 1000, iter\_sampling = 1000  
)