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 n diploid individuals sampled at random from a population made of subpopulations $k = 1, \dots, K$. In subpopulation k, the allele A frequency is p_k, and let w_k be the fraction of sampled individuals from subpopulation k ($\sum_k w_k = 1$). The "global" allele frequency we want to estimate is the mixture average

$$p \;=\; \sum_k w_k p_k.$$

Let n_{AA}, n_{Aa}, n_{aa} be the counts of genotypes AA, Aa, aa in the sample of size n. The standard estimator of the allele frequency is

$$\hat{p} \;=\; \frac{2n_{AA} + n_{Aa}}{2n}.$$

(a) Show that in the presence of population substructure, $\hat{p}$ is unbiased.

Hint: Let D_i be the A-dosage for individual i ($D_i \in \{0, 1, 2\}$). Under HWE within k, $D_i \mid Z_i = k \sim \text{Bin}(2, p_k)$. Use LOTUS: $\mathbb{E}[D_i] = \sum_k w_k \mathbb{E}[D_i \mid Z_i = k]$.

(b) What is $\text{Var}(\hat{p})$ under *random mixture sampling* (i.i.d. individuals from the mixture)? Assume each subpopulation is in HWE. Compare to the case with no substructure ($K = 1$).

Hint: Use the law of total variance on D_i:
$\mathrm{Var}(D_i) = \mathbb{E}\{\mathrm{Var}(D_i \mid Z)\} + \mathrm{Var}\{\mathbb{E}(D_i \mid Z)\}$, with $\mathrm{Var}(D_i \mid Z = k) = 2p_k(1-p_k)$ and $\mathbb{E}(D_i \mid Z = k) = 2p_k$.

Define

$$\bar{p} = \sum_k w_k p_k \tag{1}$$

$$\mathrm{Var}_w(p_k) = \sum_k w_k (p_k - \bar{p})^2, \tag{2}$$

and note $\mathbb{E}[p_k(1-p_k)] = \bar{p}(1-\bar{p}) - \mathrm{Var}_w(p_k)$.

(c) Now consider a *stratified* sample: take exactly n_k individuals from subpopulation k ($\sum_k n_k = n$; write $w_k = n_k/n$). Maintain HWE within subpopulations. Derive $\mathrm{Var}(\hat{p})$ 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 $\mathrm{Var}_w(p_k)$.

Hint: Write $\hat{p} = \sum_k w_k \hat{p}_k$ with $\hat{p}_k = \frac{1}{2n_k} \sum_{i:Z_i=k} D_i$. Use independence across strata and $\mathrm{Var}(\hat{p}_k) = \frac{p_k(1-p_k)}{2n_k}$.

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

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

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

$$\tilde{X} = \frac{X - \mathbb{E}[X]}{\sqrt{\mathrm{Var}(X)}}, \qquad \tilde{X}_c = \frac{X_c - \mathbb{E}[X_c]}{\sqrt{\mathrm{Var}(X_c)}}.$$

With this scaling, $\beta_{\mathrm{obs}} = r\,\beta_c$ exactly, where $r = \mathrm{Corr}(\tilde{X}, \tilde{X}_c) = \mathrm{Corr}(X, X_c)$.

Assume individuals are sampled i.i.d. from a mixture with $\Pr(Z = k) = w_k$, $\sum_k w_k = 1$. Within each subpopulation k: - HWE holds at each locus, - LE (no within-k LD) holds between X and X_c.

Define

$$\bar{p} = \sum_k w_k p_k, \quad \bar{p}_c = \sum_k w_k p_{c,k} \tag{3}$$

$$\mathrm{Var}_w(p_k) = \sum_k w_k (p_k - \bar{p})^2 \tag{4}$$

$$\mathrm{Cov}_w(p_k, p_{c,k}) = \sum_k w_k (p_k - \bar{p})(p_{c,k} - \bar{p}_c) \tag{5}$$

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

Let the allele frequencies at the tag and causal SNPs be p_k and $p_{c,k}$ in subpopulation k. Show that

$$r = \frac{\mathrm{Cov}(X, X_c)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(X_c)}},$$

and express $\mathrm{Cov}(X, X_c)$, $\mathrm{Var}(X)$, and $\mathrm{Var}(X_c)$ in terms of $\{w_k, p_k, p_{c,k}\}$.

Hint: Law of total covariance:
$\mathrm{Cov}(X, X_c) = \mathbb{E}[\mathrm{Cov}(X, X_c \mid Z)] + \mathrm{Cov}(\mathbb{E}[X \mid Z], \mathbb{E}[X_c \mid Z])$. Under LE, the first term is 0. Use $\mathbb{E}[X \mid Z = k] = 2p_k$.

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

Suppose population structure also affects the trait mean: $\mathbb{E}[Y \mid Z = k] = \mu_k$. Consider the model $Y = \mu_Z + \beta_c X_c + \varepsilon$ with $\mathbb{E}[\varepsilon] = 0$. Show that the naïve regression of Y on X (without structure covariates) is biased:

$$\hat{\beta}_{\text{naïve}} \approx \frac{\beta_c \,\mathrm{Cov}(X, X_c) + \mathrm{Cov}(X, \mu_Z)}{\mathrm{Var}(X)} = r\,\beta_c + \underbrace{\frac{\mathrm{Cov}(X, \mu_Z)}{\mathrm{Var}(X)}}_{\text{bias}}.$$

Under the assumptions above, prove that $\mathrm{Cov}(X, \mu_Z) = 2\,\mathrm{Cov}_w(p_k, \mu_k)$, and give the bias in terms of w_k, p_k, μ_k.

Hint: $\mathrm{Cov}(X, \mu_Z) = \mathrm{Cov}(\mathbb{E}[X \mid Z], \mu_Z) = \mathrm{Cov}(2p_Z, \mu_Z)$.

Part C (20 pts): Brief interpretation

In a few sentences each:

1. Explain why $r \neq 0$ 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 p_k also have larger μ_k, what is the expected direction of the naïve bias?

3. Name **two** standard strategies to mitigate both components of bias (structure-induced r and trait mean differences) in practice.

Problem 3: Bayesian Analysis (30 pts)

Part A (10 pts): Beta–Binomial conjugacy

1. With prior Beta$(4, 18)$ and data $x = 11$ successes out of $n = 27$, write the posterior distribution for p.

2. Compute the posterior mean and a **central 95% credible interval** in R using `qbeta`. Compare to the MLE $\hat{p} = 11/27$. Briefly interpret the *shrinkage*.

3. **Sensitivity:** repeat with priors Beta$(1, 1)$ and Beta$(8, 32)$. 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 p from Binomial data with a Beta$(4, 18)$ prior. Use $x = 11, n = 27$.

2. **Run the model** in R using `cmdstanr`. **Check convergence** and effective sample size; report $\hat{R}$ and bulk ESS for p.

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**:

- $n_A = 725$
- $n_B = 258$
- $n_O = 1073$

Under HWE with allele frequencies $\mathbf{p} = (p_A, p_B, p_O)$, the **unconditional** phenotype probabilities are

$$\Pr(A) = p_A^2 + 2p_Ap_O, \quad \Pr(B) = p_B^2 + 2p_Bp_O, \quad \Pr(AB) = 2p_Ap_B, \quad \Pr(O) = p_O^2.$$

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

$$q_A = \frac{\Pr(A)}{1 - \Pr(AB)}, \quad q_B = \frac{\Pr(B)}{1 - \Pr(AB)}, \quad q_O = \frac{\Pr(O)}{1 - \Pr(AB)}.$$

1. **Write a Stan model** that estimates (p_A, p_B, p_O) with prior Dirichlet$(1, 1, 1)$ and a **Multinomial** likelihood on (n_A, n_B, n_O) using (q_A, q_B, q_O).

2. **Run the model** and check convergence (report $\hat{R}$ and ESS).

3. **Prior sensitivity:** re-run with $\alpha = k\,(0.26, 0.09, 0.65)$ for $k \in \{1, 10, 100\}$. Summarize how posterior means and credible intervals change with k, 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
)
```