Arjun Biddanda | **Problem Set 2**
Statistical Genetics Reading Group | **7/7/2025**

GWAS Power and Study Design

Sample Size Calculations

Suppose you are conducting a GWAS for quantitative trait YFT (your favorite quantitative trait), and you want to collect $N = 100,000$ individuals.

- At a power threshold of 0.8, what is the minimal standardized effect size (β) detectable for a fully typed variant $r^2 = 1$ at an MAF of:
  - 5%
  - 1%
  - 0.1%
- Your colleague wants to see if a specific missense variant at 0.1% frequency in the population (with an imputation $r^2 = 0.8$), would be detectable in your current GWAS design?. What would be the minimal effect-size detectable for this variant?

♦♦ GWAS Discovery Rate

1. Fixing GWAS power at 0.8 and assuming fully typed variants ($r^2 = 1$), if causal variant effects are drawn from $\beta_{causal} \sim \mathcal{N}(0, \sigma^2)$, what is the *expected fraction of additional discoveries* made in going from $N_1 = 10^4$ to $N_2 = 10^5$ when $\sigma^2 = 2$? (Assume causal allele frequencies are drawn from the Uniform distribution).
2. If $\beta|p \sim N(0, \sigma^2)$, where $\sigma^2 = (p(1-p))^\alpha$ for causal variants (e.g. the effect size is linked to allele frequency), what is the expected fraction of additional discoveries when $\alpha = 0.75$? (assume causal allele frequencies are drawn from the Uniform distribution).

Provide either an analytical solution or a plot reflecting the numerical solution for both scenarios.

♦♦♦ Effect of Tagging Variant LD

The r^2 measure of LD is critical when you do not have direct access to a causal variant and only have a tagging variant. If we assume that $p_{tag} >= p_{causal}$, the expression is:

$$r^2_{max}(tag, causal) = \frac{(1 - p_{tag})(1 - p_{causal})}{p_{tag} p_{causal}}$$

Using the expressions for GWAS power for a quantitative trait, and $r^2_{max}(p_{tag}, p_{causal})$, show that:

$$Power_{tag} \geq Power_{causal} \quad \forall \quad p_{tag} \geq p_{causal}$$

♦♦ Comparing GWAS

Your colleague has an interesting case where YFT is highly prevalent in a different population (lets say population "B")and wants to understand if the same variant is driving this increase in prevalence. Your collaborator has access to $N_B = 50000$ samples, but doesn't have a good sense of whether this would be well-suited to address the idea.

If the effect-size of the causal variant is the same (e.g. $\beta_A = \beta_B$), what would be the difference in frequency of the causal variant required to maintain a power of 80% if the variant in population A is at 0.1% frequency? (Assume $N_A = 100,000$)

GWAS Practical

For a GWAS practical, we will use publicly available summary statistics from the Pan-UKBB project. Specifically we will use data for a dataset on total bilirubin (a key liver enzyme) where GWAS was performed separately across six different ancestries in the UK Biobank. The ancestral groupings are:

- EUR = European ancestry($N = 399286$)
- CSA = Central/South Asian ancestry ($N = 8395$)
- AFR = African ancestry ($N = 6176$)
- EAS = East Asian ancestry ($N = 2549$)
- MID = Middle Eastern ancestry ($N = 1491$)
- AMR = Admixed American ancestry ($N = 933$)

To obtain the data for this exercise run:

```
wget https://pan-ukb-us-east-1.s3.amazonaws.com/sumstats_flat_files/
              biomarkers-30840-both_sexes-irnt.tsv.bgz
```

Genomic Inflation Factor

While many covariates have been accounted for when evaluating genetic effects - it is possible that there are still unobserved confounders present. One way to evaluate the extent of possible confounding has been using the genomic-inflation factor, which is defined as:

$$\hat{\lambda}_{GC} = \text{median}(\hat{\chi}^2)/F^{-1}_{\chi^2_1}(0.5)$$

where $F^{-1}_{\chi^2_1}(0.5) \approx 0.455$ is the expected median of the chi-squared distribution with 1 degree of freedom under the null hypothesis of no association.

1. Estimate $\hat{\lambda}_{GC}$ for each individual population separately using the relationship that $\hat{\chi}^2 = \left(\frac{\hat{\beta}}{s_\beta}\right)^2$.
2. Are the estimates of λ_{GC} significantly > 1 for each population? (Hint: use bootstrapping to re-sample the effect-sizes to construct empirical 95% confidence intervals)
3. ♦ Re-estimate λ using only every 10000^{th} variant to avoid the effect of linkage disequilibrium and evaluate the differences relative the non LD-pruned version. Comment on how LD may violate some of the assumptions of λ_{GC}.

Power & Shared Signals

One of the useful things about the Pan-UKBB project is that it can illustrate how to compare the *power* of GWAS across different study designs and sample-populations.

1. For a start, let us compare the intersection of signals found at a genome-wide threshold $\alpha = 5 \times 10^{-8}$ in the EUR and CSA sample subsets? Plot the fraction of signals that overlap as a function of $\alpha \in \{10^{-4}, 10^{-20}\}$
2. Since we have the allele-frequency of the variants in each ancestry subset, for all shared marginal signals at the detection threshold α, estimate the power to detect each of the shared signals? What is the mean detection power threshold for shared signals at different values of α? (NOTE: assume that $\hat{\beta} \approx \beta_{causal}$ and $r^2 = 1$ for simplicity, though consider how violations of these assumptions might alter results)