Statistical and Population Genetics

Hanbin Lee

2019-07-18

Contents

Chapter 1

Prerequisites

Elementary probability theory and linear algebra will be sufficient. However, topics regarding theoretical genetics requires knowledge about stochastic processes and measure theory.

Reference on advanced mathematics can be found in Rudin (1986), Durrett (2019) and Durrett (2008).

Chapter 2

Introduction

As a mathematics undergraduate, theorem-proof structure was much easier to understand. However, most statistical & population genetics articles were less familiar to me in their organization.

This page aims to list theorems and methods regarding statistical & population genetics in a theorem-proof structure so that I can readily access when I need them.

Chapter 3

Linkage Disequilibrium Score Regression (LDSC)

Linkage Disequilibrium Score Regression (LDSC) is a popular method in statistical genetics with a wide range of application. It is used to measure confounding due to population structure, computing genetic correlation and heritability. Most of all, the method can be performed soley on summary statistics rather than the full data which reduces memeory and computation requirements tremendously.

3.1 Univariate LDSC

The original from of LDSC (here we call it Univariate LDSC) can be used to estimate SNP heritability and confounding effects of population structure (Bulik-Sullivan et al., 2015). I refer the supplementary notes of Bulik-Sullivan et al. (2015) for the notations.

Definition 3.1 (LD score)**.** Define the **LD Score** of variant j as

$$l_j = \sum_{k=1}^{M} r_{jk}^2$$

Then for a homogenous sample without population structure, the following holds.

Theorem 3.1 (LDSC without structure)**.** *In a sample without population structure*

$$\mathbb{E}[\chi_j^2] = \frac{Nh_g^2}{M} l_j + 1$$

holds.

If a population structure (measured by the fixation index F_{ST}) exists, a non-zero constant is added to .

Theorem 3.2 (General LDSC)**.** *In a sample with population structure*

$$\mathbb{E}[\chi_j^2] = \frac{Nh_g^2}{M} l_j + 1 + aNF_{ST}$$

holds.

Proof.

$$\begin{aligned} \mathrm{Var}[\hat{\beta}_j] &= \mathbb{E}[\mathrm{Var}[\hat{\beta}_j|X]] + \mathrm{Var}[\mathbb{E}[\hat{\beta}_j|X]] \\ &= \mathbb{E}[\mathrm{Var}[\hat{\beta}_j|X]] + 0 \end{aligned}$$

The second term vanishes since $\mathbb{E}[\hat{\beta}_j|X] = 0$.

Inspecting the first term,

$$\begin{aligned}
\mathrm{Var}[\hat{\beta}_j|X] &= \frac{1}{N^2}\mathrm{Var}[X_j^T\phi|X] \\
&= \frac{1}{N^2}X_j^T\mathrm{Var}[\phi|X]X_j \\
&= \frac{1}{N^2}\left(\frac{h_g^2}{M}X_j^T X X^T X_j + N(1-h_g^2)\right)
\end{aligned}$$

□

Chapter 4

Mendelian Randomization

Mendelian Randomization (MR) is a special case of the Instrumental Variable (IV) which is widely used in econometrics. MR employs genetic variants (e.g. SNP) as IVs since genotypes are automatically randomized in the gametic phase of cell division.

4.1 Simple MR

4.2 Two-sample MR

4.3 MR-Egger

4.4 Weighted Median MR

4.5 Modal Based MR

Bibliography

Bulik-Sullivan, B. K., , Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., Daly, M. J., Price, A. L., and Neale, B. M. (2015). LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, 47(3):291–295.

Durrett, R. (2008). *Probability Models for DNA Sequence Evolution (Probability and Its Applications)*. Springer.

Durrett, R. (2019). *Probability: Theory and Examples (Cambridge Series in Statistical and Probabilistic Mathematics)*. Cambridge University Press.

Rudin, W. (1986). *Real and Complex Analysis (Higher Mathematics Series)*. McGraw-Hill Education.