Efficiency

Background

In today's demo, you are motivated by the need to make efficient use of limited samples. This need is central in many applications. One broad example is late-stage clinical trials: pre-planned studies including rare outcomes such as heart attacks can require tens of thousands of participants in order to observe enough rare events to make conclusions. Another example is genomics: a small number of RNA sequencing samples circa 2020 could easily cost thousands of dollars, and some important biological signals are present in low-abundance RNA's that are hard to quantify. A third example is in predicting election outcomes. The ground truth, meaning the elections themselves, happens infrequently; no amount of money can buy more data.

This aspect of statistics is EXTREMELY heavily studied, and there are some common techniques that automatically inherit justifications about being the most powerful or most efficient option. These include maximum likelihood estimation, which chooses parameters to maximize the probability of the observed data, and Bayesian inference, which imposes certain prior assumptions on the unknown parameters and then rigidly follows the rules of mathematical probability. There are also competing techniques with other virtues that you might use sometimes, even though they are not the most efficient. In these situations, what are you gaining or losing with one method or the other?

Learning objectives

- Define the variance of a random variable.
- Via simulation, compare two different methods to estimate the variance of a binomial random variable. Determine which method is more efficient.

Exercises

In human genetics, samples often contain people of multiple ancestries, and separating them out can be interesting and useful for anthropology and medical studies. This is usually best done using genetic data themselves. Some methods for doing this attempt to operate in a way that gives equal weight to each locus. (A **locus** is a spot on the genome; the plural of **locus** is **loci**.) In other words, loci with high variance are not supposed to influence the results more than loci with low variance. This example is about estimating genotype variance at some genetic loci, so you can later assign them the right weight.

To define that more precisely, we need more terms from genetics. At each locus, there can be different DNA sequences from person to person, or different sequences on the two chromosomes within a person. These different sequences are called **alleles**. We will deal with the total number of copies of a reference allele at each locus, which is 0, 1, or 2 for each person. Call this allele count X. Then the expected value $E[X]$ is the **population allele frequency**, which is what you would get if you genotyped everybody and averaged the results.

This lecture deals with the variance, defined as $Var[X] = E[(X - E[X])^2]$. In this demo, you will compare two existing methods for estimating variance.

Method 1 is the R function `var`, which uses the formula $\frac{(X_1-\bar{X})^2+...+X_N-\bar{X})^2}{N-1}$. It plugs in the sample mean $\bar{X}$ in place of the population allele frequency $E[X]$, and it averages the squared deviations across each person in the sample. The denominator we'll use is N-1, not N, for reasons beyond the scope of this course.

Method 2 is to compute $\bar{X}(2 - \bar{X})/2N$, where $\bar{X}$ is the sample mean. This method is derived from mathematical properties of the binomial distribution (coin flips). It requires certain assumptions (look up "Hardy-Weinberg equilibrium"), but it is more efficient than method 1.

Type your answers in a Word document and your code in an R script and email them to Eric as `LAST_FIRST_binomial_variance.docx` and `LAST_FIRST_binomial_variance.R`.

1. Simulate allele counts for 1 locus in 1000 people using the function `rbinom`, which simulates binomial random variables. Store the result in a dataframe. Each person should receive a value of 0, 1, or 2.
2. Estimate the variance across the 1000 people by method 1 and method 2. Store the results in a separate dataframe with one row and two columns.
3. Using `lapply` or a `for` loop, repeat this 2000 times. To store the results, extend the second dataframe to have 2000 rows and 2 columns.
4. Plot the results as a scatterplot (method 1 vs method 2) and a pair of histograms (method 1 vs method 2).
5. Which estimation method is closer to the truth on average? The variance of a binomial random variable with N trials and probability p is $Np(1 - p)$. If you didn't get to see why in the probability lecture, or if you want a refresher, ask and we can work through it at the board.
6. What allele frequency did you choose? Wrap all your code in a function that takes an allele frequency as input. Repeat the test at 3 very different allele frequencies and comment on the results.

Footnotes

Population allele frequency is usually described as being between 0 and 1, not 0 and 2, but I'll ignore that here. For today, you may assume you have a simple random sample that is representative of "everybody". You may assume that people only share alleles by chance (there are no family members). In real studies, you have to be more careful about family members, whose genotypes are not independent. Given current trends, you also have to make specific plans to get substantial representation of non-European ancestries in most large human genetics studies. This has a big effect on allele frequencies.