Mistaken assumptions

Statistical inference requires assumptions, often starting with a claim like "X follows a Binomial distribution" or "X and Y have a linear relationship." This is unfortunate and limiting: there are many cases when we just don't know the distribution of the data or the form of a function. One major agenda in modern statistics is to reduce these assumptions to their barest form.

This is often done by making models that are infinitely flexible in some way: lines of best fit that bend and wiggle, categorical assignments with no limit on the number of categories, probability mass functions with no pre-specified shape or generating mechanism, or procedures that require not specific distributions ("Binomial", "Normal") but only abstract properties ("non-decreasing", "symmetric", "independent", "exchangeable"). We won't have the mathematical firepower to dig into these yet, but spectacular examples of this type of analysis include

- generalized estimating equations
- doubly-robust causal effect estimation
- causal effect estimation with many invalid instrumental variables

The common thread here is the ability to say, "Even if my method is wrong about some part of the data generating mechanism, it will yield results that are reliable in some specific sense."

In this session, we will explore the principle that sometimes, a statistical method must tolerate when its assumptions are mistaken.

Estimating allele-count variance with assortative mating

We return to the example from the "efficiency" lecture: variance of allele counts. But, let's generate the data a little differently. Let's assume that the population in question has a high degree of assortative mating. Assortative mating means that people with similar allele counts tend to have kids together. Instead of binomial allele counts, which require each chromosome to be independent, the chromosomes tend to be similar, and the allele counts no longer follow such a simple set of probability calculations. Specifically, the allele counts will have more 0's and 2's and fewer 1's.

Learning objectives

- Via simulation, compare two different methods to estimate the variance of a binomial random variable. Determine which method is more accurate in the presence of assortative mating.
- Make a judgement about which variance estimator to use.

Exercises

Repeat the tasks from the efficiency lecture (copied below) but with assortative mating. Type your answers in a Word document and your code in an R script and send them by email as `LAST_FIRST_assortative_mating_variance.docx` and `LAST_FIRST_assortative_mating_variance.R`. Here are the (modified) questions.

1. Simulate allele counts for 1 locus in 1000 people so that people with a reference allele on one chromosome are more likely to have it on the other chromosome, and people with the alternate allele on one chromosome are more likely to have the alternate allele on the other chromosome. There are different ways to do this, but in order to keep the class in sync, please use the simulation code from this starter script.

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. Recall the estimation methods from the efficiency lecture: > 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.

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. Which estimation method is closer to the truth on average? Previously, we used the formula for the variance of a binomial random variable, $Np(1-p)$. Since we changed the generating mechanism, this formula is no longer correct. Instead, use the function in the starter script to compute the true variance.

5. Each of 9 demo datasets in this folder contains simulated allele counts for 9 loci across 1000 people. Each sample may or may not have any assortative mating. I don't know which ones do or don't; the decision was made randomly with 50% probability of each. For each dataset, decide which method to use. Explain your decision.

Footnote

Assortative mating is common in humans. I'm not sure if it shows up this obviously, but for fancier types of analysis, it makes a big difference!