Disproving Neutral Models With Baby Names

Michael Jordan

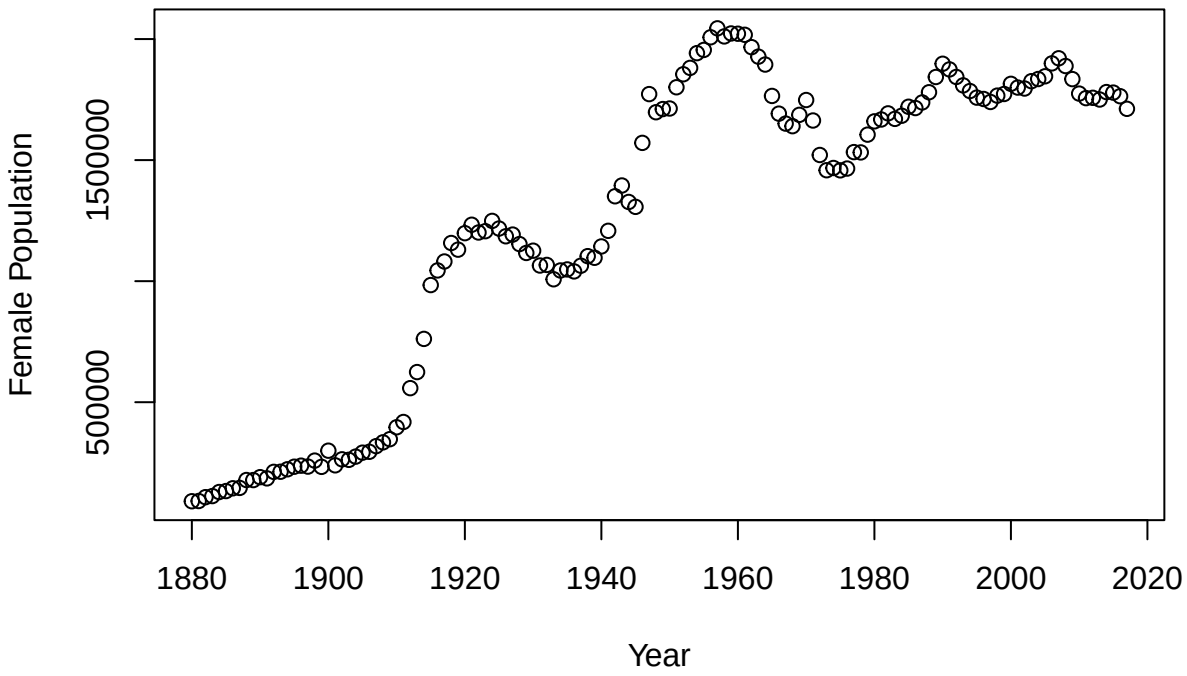
2023-12-05

In this week's module, we reviewed the proposal that so-called "neutral" models, in which trait inheritance is stochastic, could explain outcomes that evolutionary biology has long held are due to the pressure of selective fitness. If neutral models can explain the data, why bother with a complicated theory when a simple stochastic process will do? Occam's razor, etc.

Of course, fitting a model is not in-and-of-itself proof positive of a particular causal mechanism. Numerous studies, such as the Lenski study of *e. coli* evolution, have demonstrated that traits do not develop randomly in a population, but instead develop in response to selective pressure operating on the population. Nevertheless, neutral models have come into vogue in recent years and proponents have claimed that they can explain outcomes in a wide variety of processes that common sense tells us should involve some degree of selective pressure.

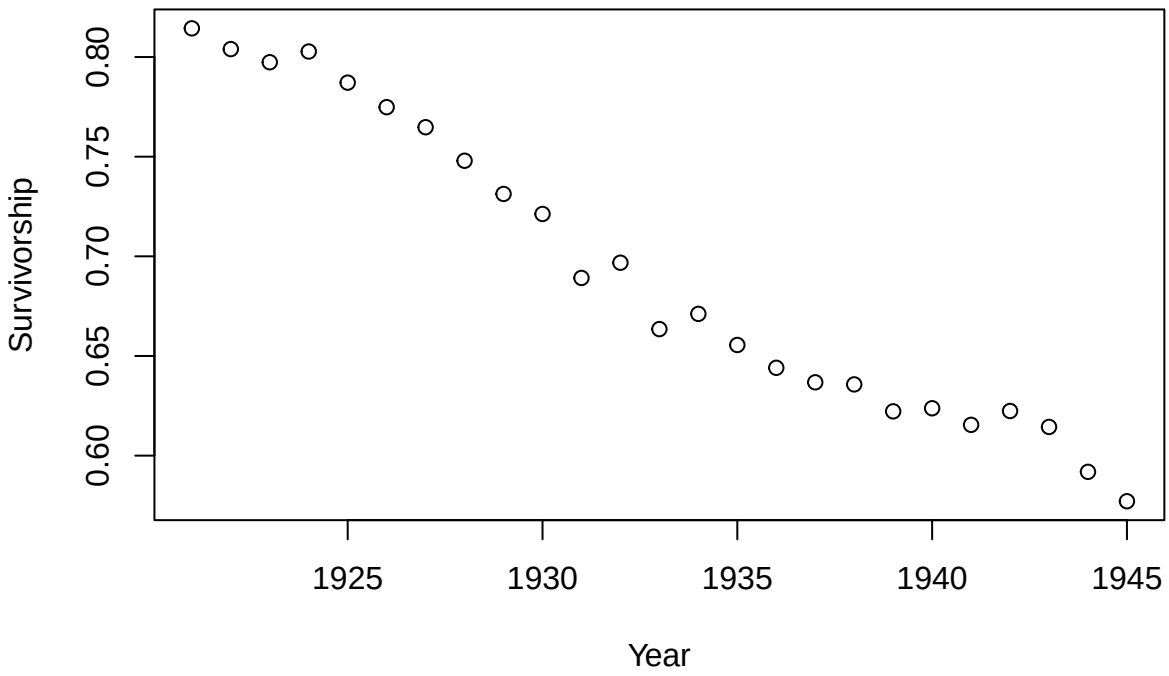
In this practical, we were tasked with proposing and falsify a simple neutral model using just such a process: naming newborn babies. Do new parents typically pluck a name for their child out of a hat (as a neutral model might suggest)? Or, are some names more heritable than others based on factors such as fads, cultural traditions, social movements, etc. (i.e., do they come under selective pressure)? To test this, we will take a data set of over one hundred years of baby names in the US, describe a parameter of the data set, and compare this parameter in the real world data to the results of simulations of name inheritance that use a neutral model. By comparing the results of the simulations of the neutral model to the real world data, we can evaluate whether the real world data could plausibly have been the result of the neutral model.

First, let's explore our real world data. We'll examine only female names to avoid confusion around names that occur for both sexes.

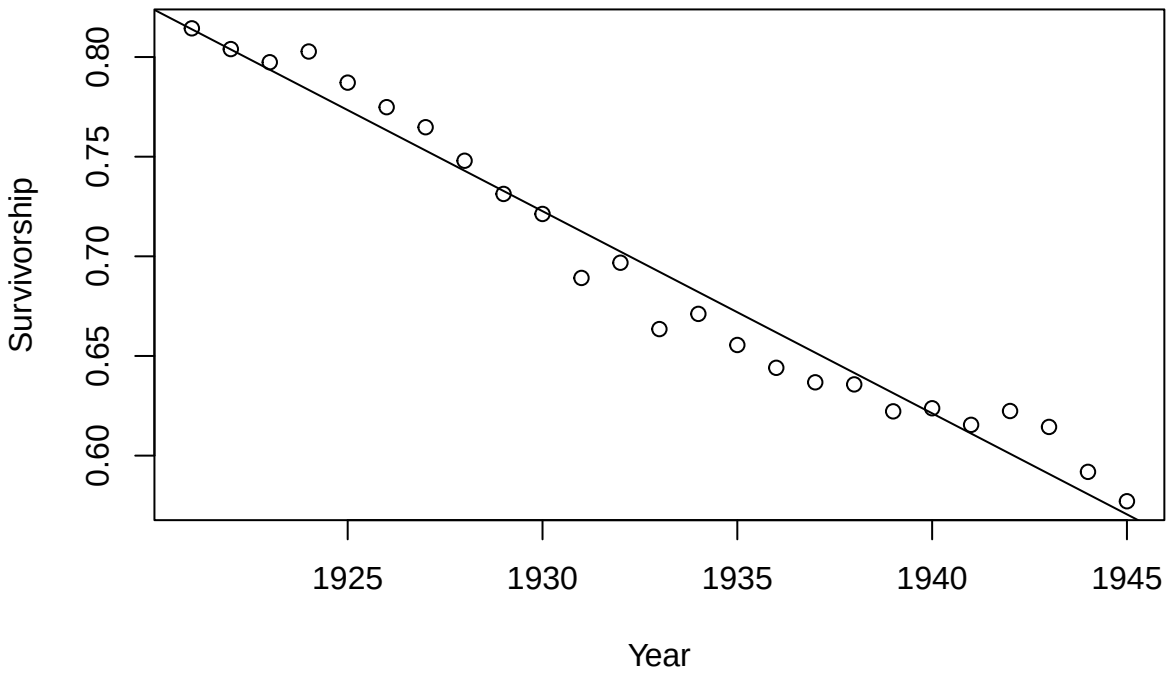


As expected, population varies over time. We could use any time period for our simulation. But, let's start at 1920. It's a bit arbitrary, but it's around a time when population was stable, it's in the modern era (not sure how accurate 19th century record keeping was), it will allow us to test over a long time period, and frankly a somewhat lower initial name count is preferable to keep our loops from being too computationally costly (I'd rather this thing not run for hours).

Next we need to decide what population parameter we'd like to test. One dynamic that might be interesting to look at is survivorship. Among female names in 1920, what percentage were inherited in the following years? Let's look at that in the real data over a 25 year period (I'd like to do more years, but this is again so that my computer doesn't explode).



The relationship looks more or less linear, so let's fit a linear model to it and extract the slope.

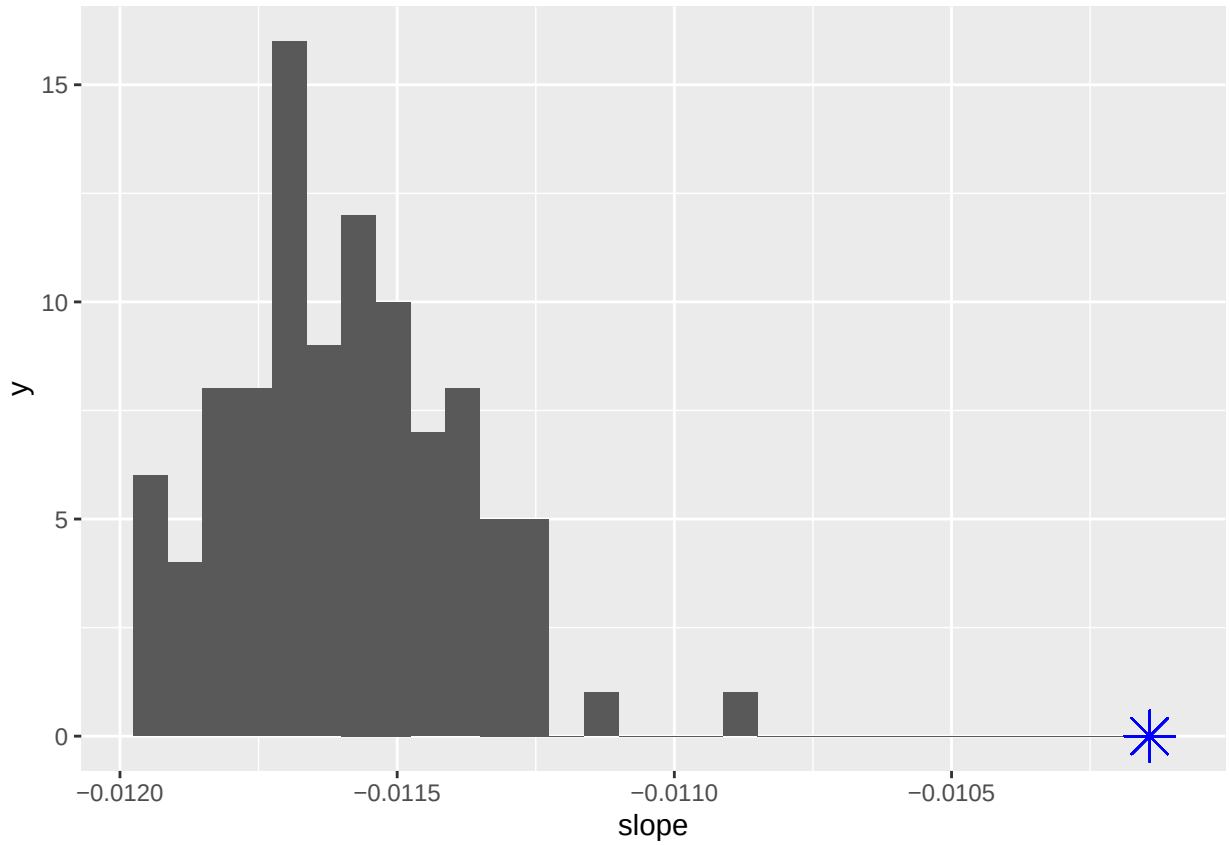


The R-squared of the model is great, 0.967, so a linear model fits the data reasonably well. The slope of the model is -0.01.

The slope of a linear model fit on survivorship over time might be an interesting test. If there is selective pressure on names, we might expect survivorship of a neutral model to decrease at a faster or slower rate than we see in the actual population. It's a reasonable supposition... let's go with it!

Now, we need to pick a neutral model to simulate. We'll keep it simple. We only care about survivorship of names in 1920, so we'll begin with all names in 1920 in proportion to their popularity that year. To simulate name inheritance in 1921, we'll randomly sample a number of names from the 1920 name pool equal to the 1921 real world population. We'll continue to sample in the same way, iteratively sampling from the name pool in the prior year, for 25 years following 1920. This is a very simple and of course unrealistic model (it assumes that parents have knowledge of all names in the prior year). But, it is neutral since inheritance is purely stochastic and not impacted by selective pressure.

So, we're all set. Now all we need to do is run our simulation 100 times, fit a linear model to each run, extract the slope from each model, and plot a histogram of slopes to construct a distribution of slopes to ascertain the range of slopes our neutral model is expected to produce. This is analogous to taking a sample from a population, estimating the sample distribution based on the count and standard deviation of the sample, and evaluating a claim about a test statistic based on whether or not it falls within the hypothetical sample distribution. The main difference is there here we actually derive our sample distribution out of simulation runs! If the slope of the model fit on real world survivorship falls within the histogram of slopes of simulation runs of our neutral model, then we will believe that the neutral model could explain the real world data. If it does not, we will say that it is unlikely that the neutral model could have produced the real world data.

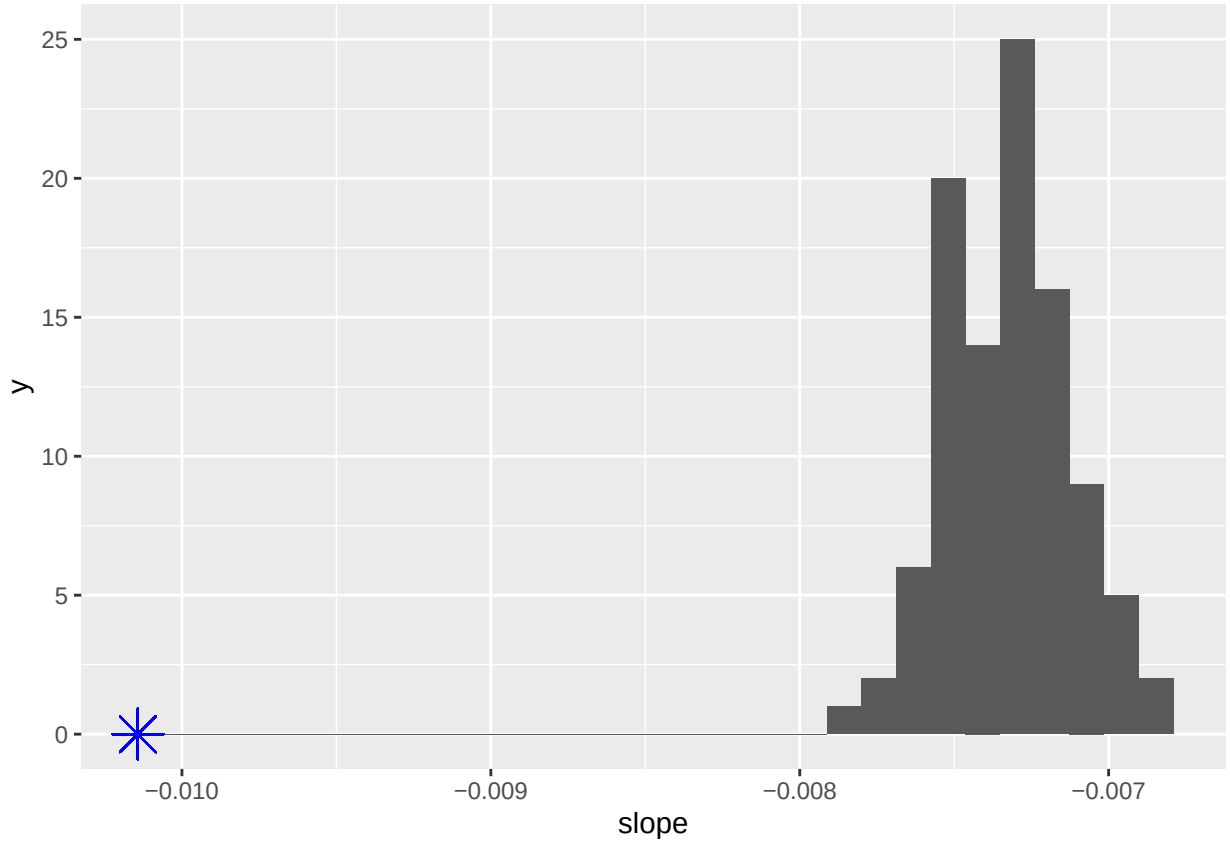


What do you know? The real world slope fell far outside of the histogram of simulation runs! Specifically, it was far to the right of the simulations, meaning that the neutral model produced results in which survivorship declined much faster than we see in reality. What could explain that?

One possibility is that parents might have a preference for some names over others (i.e., that selective pressure acts on name inheritence!). For example, perhaps parents tend to prefer rare names and avoid common names. Who wants to be like everyone else? At some point, how many Olivias can there be?

To test this, instead of simulating a truly neutral model, let's simulate a model in which some names are fitter than others. This is arbitrary, but let's say that the top 10% of names by frequency in any given year get a 10% downgrade in their likelihood of being inherited and the bottom 10% get a 10% boost. The model will otherwise be identical. We'll still start in 1920, sample each year from the names of the previous run, and run it for 25 years across 100 simulation runs.

What happens if we run a model like this?



As you can see, the simulated slopes of the weighted model are much closer to the actual slope than they were in the original un-weighted model. In fact, the mean of the slopes in the un-weighted model is 0.52 times farther from the actual slope than it is in the weighted model.

This suggests that baby names are not inherited on a "neutral" basis, but are inherited on the basis of some kind of selective pressure. Although, we would have to more work to tease out what rules of inheritence best describe the data.