AGAINST ALL ODDS
EPISODE 29 – "INFERENCE FOR TWO-WAY TABLES"
TRANSCRIPT

FUNDER CREDITS

Funding for this program is provided by Annenberg Learner.

INTRO

Pardis Sabeti
Hi, I'm Pardis Sabeti and this is *Against All Odds*, where we make statistics count.

And as you can see, today I'm taking a break from our studio to tell you about my day job. This is the Broad Institute in Cambridge, Massachusetts, where I have a small research team investigating an ancient biological battle – the non-stop evolutionary arms race between our bodies and the infectious microorganisms that try to invade and inhabit them.

The good news is that we humans have an array of new high tech tools at our disposal in the battle between the bugs and us. These machines, for instance, are the latest generation of genome sequencers. They allow us to sequence out the letters that code the genomes of both humans and our microbial enemies.

In my research I'm using the data these machines provide to give me and my colleagues clues to new ways to battle some of our most dangerous and pernicious diseases – diseases that we in the west rarely encounter.

One of the deadliest is Lassa fever. Like the more notorious tropical disease, Ebola, it's caused by a virus, and fells its victims with hemorrhagic fever. Throughout West Africa, thousands of people die of Lassa fever every year.

In fact, one of the most rewarding things I've done in my career is to work with the dedicated doctors, nurses, and scientists on the staff of the teaching hospital in Irrua, Nigeria.

Irrua is situated in a region where Lassa is especially rife. Some of the hospital staff have even died of the disease they are trying to treat, yet they persist.

But what's surprising is that many tens of thousands more throughout the region are exposed to the virus without getting sick. This suggests these people have some sort of resistance to the virus. And it's the source of this resistance that I want to track down.

Our work is still at its early stages, but one of the models for what I hope to uncover is another tropical disease that I study, and that kills and sickens millions every year, especially children – malaria. Because, with malaria, we already know about an important source of resistance to the disease. It's a genetic mutation that is better known for the harm it does than the good, and in fact we have already encountered it in *Against All Odds* – it's sickle cell anemia.

As we discovered in the module on binomial distributions, if a child inherits two copies of the sickle cell mutation from his parents, he suffers from the debilitating

and sometimes deadly effects of sickle cell anemia. If the child inherits only one copy of the gene, they're unaffected by the disease, but even though most kids in the United States who are sickle cell carriers would never know it, he or she is protected against malaria. And it is in fact this protective effect that is responsible for the sickle cell mutation becoming so prevalent – it is one of the winners in the arms race I was talking about.

And it is statistics that can reveal it. I've borrowed this data from research done by a friend of mine. He and his colleagues looked at the genotypes of 315 children with severe malaria. Since each child inherits two hemoglobin genes, one from each parent, they examined 630 genes in these sick kids. They found seven instances of the HbS sickle cell hemoglobin gene, while the other 623 were the normal HbA hemoglobin gene. The idea was to quantify whether children who came down with malaria were less likely to have inherited the protective sickle cell version of the gene rather than the normal version, as compared to the general population.

Here is a Two-Way table made from their data. Remember from our module about one city's happiness survey, a two-way table is a great way to display categorical data. In this case, the top row tells you the genes they found in children with severe malaria. The bottom row tells you the genes they found in a control group of newborn babies. Intuitively, we would expect to find the protective version of the gene less frequently in the children sick with malaria than in the control group. After all, if they were protected they likely wouldn't have come down with the disease. And in fact, you can see that HbS was inherited by the kids who caught malaria only 1.11% of the time, whereas in the control group, made of the general population, HbS was inherited 8.66% of the time. Is that difference larger than we would expect by chance? Is it statistically significant?

Our null hypothesis is that there is no association between contracting malaria and having the HbS sickle cell gene. The alternative hypothesis is that there is in fact an association between contracting malaria and having the protective HbS sickle cell gene. Basically, we want to know if there's sufficient evidence that the status of the two variables Malaria/General Population and HbS/HbA are linked.

We can compute what the expected counts in our two-way table would be if our variables really are independent, as the null hypothesis states. Use this equation: The row total times the column total divided by the grand total. Let's add those numbers to our table. Now we can see that if there were no relationship between having the gene and coming down with the disease we'd expect to find 37.9 HbS genes in the children with malaria. But in reality there were only 7 HbS genes in that group. Is that difference between 7 and 37.9 enough to tell us that there is a relationship between our two categorical variables?

The next step in our analysis is to use the chi-square test to figure out if that difference is significant. The chi-square statistic is a measure of how far the observed counts in the table are from the expected counts. Chi-squared is the sum of that term for each of the cells in the table. For our sickle cell/malaria table, chi-squared equals 41.263. Using software or a table, we look up what *p*-value corresponds to our chi-squared value of 41.263. Here it's exceedingly low, essentially zero. We have very strong evidence that there is a relationship between our variables and we can reject the null hypothesis. This gives support to our research hypothesis that the HbS sickle cell variant of the hemoglobin gene does protect against malaria.

I plan to use a similar statistical technique on the Lassa data we're collecting, with the goal of tracking down what we suspect is the genetic source of resistance to the disease. It's just one of many statistical tools I use in my work, including one that allows me to sift through the human genome seeking mutations that have become more common recently – by which I mean in the last few thousand years or so. This means they've likely been selected for by evolution for some reason – perhaps, like the sickle cell mutation, because they're providing some protection against an infectious disease. If we can identify these protective mutations in the case of Lassa fever, we may get valuable clues to potential new treatments.

It was only in 2003 – at the cost of several billion dollars and the work of hundreds of researchers – that the first read of the human genome was made. In the years since, the genome sequencing technology has gone through several generations, to the point today where the machines like those at the Broad can read out billions of letters of the genetic code, from thousands of samples, in a matter of hours, and at a cost of only a few thousand dollars per genome. As you can imagine, and as I hope you've come to appreciate from this course, one of the best ways to make any sense from this outpouring of information is through picturing the data.

Here's an example of such a picture that is becoming increasingly common in the genomics literature. They're called Manhattan plots because of their likeness to the Manhattan skyline. Whereas in our sickle cell study we only looked at one position in the genome, here we're plotting the result of our association analysis for each point along the 23 human chromosomes. The vertical lines plot the places that researchers have found evidence of an association with a particular human disease.

In this way researchers are hunting for genes underlying diseases that don't have an obvious cause like an infectious agent; complex human illnesses ranging from cancer through diabetes, to hypertension and a host of others.

I'm Pardis Sabeti for *Against All Odds*. See you next time!

PRODUCTION CREDITS

Host – Dr. Pardis Sabeti

Writer/Producer/Director – Graham Chedd

Producer – Maggie Villiger

Associate Producer – Katharine Duffy

Editors – Dave Berenson
--Seth Bender

Director of Photography – Dan Lyons

Sound Mix – Richard Bock

Animation – Jason Tierney

Title Animation – Jeremy Angier

Web + Interactive Developer – Matt Denault / Azility, Inc.

Website Designer – Dana Busch

Production Assistant – Kristopher Cain

Teleprompter – Kelly Cronin

Hair/Makeup – Amber Voner

Music
DeWolfe Music Library

Based on the original Annenberg/CPB series *Against All Odds*,
Executive Producer Joe Blatt

Annenberg Learner Program Officer – Michele McLeod

Project Manager – Dr. Sol Garfunkel

Chief Content Advisor – Dr. Marsha Davis

Executive Producer – Graham Chedd

FUNDER CREDITS

Funding for this program is provided by Annenberg Learner.

For information about this, and other Annenberg Learner programs, call 1-800-LEARNER, and visit us at www.learner.org.