AGAINST ALL ODDS
EPISODE 25 – "TESTS OF SIGNIFICANCE"
TRANSCRIPT

FUNDER CREDITS

Funding for this program is provided by Annenberg Learner.

INTRO

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

Sometimes, when you look at the outcome of a particular study, it can be hard to tell just how noteworthy the results are. For example, if the severe injury and death rates due to car crashes on one state's roads dropped from 4.7% down to 3.8% after enacting a seat belt law…would we think of this as a significant change? It's certainly a lower rate, but how do we know this is due to the seat belt law, and not just chance?

We can clarify results like these by using something called Tests of Significance. This tool is one of the most widely used in statistical inference, as it can tell us if a result is likely due to chance, or if there's something else at work.

Significance testing can be applied in a variety of circumstances. Let's explore how researchers used it to help solve a controversy in a field of study you might not associate with statistics – classic literature.

> **Shakespeare Professor**
> Shall I fly
> Lovers' baits and deceits,
> sorrow breeding?

Pardis Sabeti
"What's in a name?" Well, if you think you've discovered a long-lost Shakespearean sonnet, a name is everything. In 1985, scholar Gary Taylor was conducting research for a new book of the Complete Works of William Shakespeare. While at the Bodleian Library at Oxford University, he came upon a sonnet he had never seen or heard of before. You can imagine his surprise when he saw "William Shakespeare" written underneath it.

> **Gary Taylor**
> What matters about the poem is that at the bottom of it is written 'William Shakespeare.' It's attributed to him. And this is clearly written in the same ink and at the same time as the rest of the poem.

Pardis Sabeti
Obviously, Taylor was excited about his find and eager to include it in the new edition of the Complete Works. But first, he and his colleagues publicized the work to see if any scholars could disprove the validity of the discovery.

It produced a storm of literary controversy. Some scholars were thrilled about the first new Shakespeare find since the 17th century. Others were more skeptical and called the poem "second-rate hack-work" and "a piece of doggerel."

But aside from pondering the stylistic merits of the poem, was there any other way to determine whether Shakespeare did or did not write it? Statistics to the rescue!

A decade earlier, statistician Ron Thisted had done a statistical analysis of Shakespeare's vocabulary. He and his colleague wanted to determine how many words did Shakespeare know, but never use? Until Taylor's 1985 discovery, Thisted had thought this analysis was purely an academic exercise.

> **Ronald Thisted**
> We certainly had no expectation that we would ever be able to put our mathematical description of Shakespeare to any use, after all, there had been no new Shakespeare discovered for several hundred years and there was no prospect of any new Shakespeare being discovered. And so for us, it was – Shakespeare was really an interesting illustration of the basic statistical ideas that we developed.

Pardis Sabeti
Thisted's program provided a detailed, numeric description of Shakespeare's vocabulary. From here, they could address the authorship question. For every work, Thisted could tell how many new words there are that Shakespeare didn't use anywhere else, as well as how many words he had used only one other time, two other times, and so on. Using this model, Thisted predicted that if Shakespeare had written the poem in question, it would have seven unique words in it. However, when they ran the poem through the program, they found that there were ten unique words.

Did this difference reflect random variation within Shakespeare's writing? Or did it indicate that Shakespeare wasn't the author? This is when significance testing can really come in handy.

Thisted began with what we call a null hypothesis – written as H_0 – that basically means nothing unusual is happening. In this case, the null hypothesis was that Shakespeare wrote the poem. Then Thisted formulated an alternative hypothesis, written as H_a. His alternative hypothesis was that someone else besides Shakespeare wrote the poem. That would mean the difference between the observed number of unique words – 10 – and the predicted number of unique words – 7 – wasn't due to chance but rather to another author writing the poem, using his or her own idiosyncratic vocabulary. Researchers aim to reject the null hypothesis with evidence that suggests something more is going on other than random variation.

Ronald Thisted
Now, the question then, is that three-word difference a big difference? Or a small difference? And we – the way to answer that question is to compare the difference to some measure of how much variability there would be from one poem to another of this size. That's the information the standard deviation gives you, and the standard deviation of about 2.6 here says that the difference we've observed is slightly more than one standard deviation on the high side. Which is, in itself, not very different. It's within the range of variability we expect to see.

Pardis Sabeti
Thisted assumed the number of unique words in Shakespeare's poems has an approximately Normal distribution with a mean of 7 and a standard deviation of 2.6, which they calculated from the data they had on all known poems. This density curve illustrates the percent of area that corresponds to a number as extreme as 10, which is 3 away from the mean of 7. Since this difference between the observed and expected is just over one standard deviation, Thisted could expect to find a value as extreme as 10 unique words about 25% of the time. Therefore, Thisted failed to find significant evidence against the null hypothesis that Shakespeare actually wrote the poem. They could not reject H_0, that Shakespeare composed these lines.

Shakespeare Professor
Yet I must vent my lust,
And explain inward pain…

Pardis Sabeti
It's important to understand that this does not mean that Shakespeare definitely wrote the poem. It only fails to give sufficient evidence to the hypothesis that he did not write the poem.

Ronald Thisted
What we could have hoped for, perhaps, was that such a bad poem would have been clearly non-Shakespearean. In which case we could have disproven the Shakespeare hypothesis. But, like a paternity test, it can only rule out…it can't rule in.

Pardis Sabeti
Taylor's own analysis of the stylistic similarities between the newly discovered work and Shakespeare's established canon also failed to disprove the null hypothesis.

Gary Taylor
We have historical evidence that says it's by Shakespeare. And so in a way the burden of proof is on the people who want to say it isn't by Shakespeare, because they have to contradict that early 17th century witness. All that the supporters of Shakespeare's authorship have to do is

to prove in as many ways as they can, as many different tests as they can, that this could be by Shakespeare and therefore there's no strong reason to doubt what the witness says.

Pardis Sabeti
In the absence of literary or statistical evidence against Shakespeare's authorship, the poem was published in Taylor's edition of The Complete Works. Let's take a closer look at how Significance Testing was able to help solve this Shakespearean mystery.

Since we want to work with sample means, let's suppose researchers found a folio of five new poems that were attributed to Shakespeare. Instead of a single count of the number of new words in one poem, we would want to find the average number of new words for the five poems.

We'll say the average number of new words was 8.2 per poem. We know from Thisted's research into Shakespeare's collected works that the mean number of new words in poems of this length is 7. But our sample mean, "*x*-bar*,"* is 8.2. We want to know if, based on this evidence, we can conclude that Shakespeare didn't write these poems.

Our null hypothesis is that Shakespeare did write these poems. In other words, we'll assume that our result of 8.2 new words is nothing more than the normal variation within Shakespeare's writing. To state our null hypothesis numerically, we'll express it in terms of the population mean. Our null hypothesis states that the mean of the whole population from which the sample was drawn equals 7 unique words. We know that is true for all Shakespeare's published works.

The alternative hypothesis states that the population mean is NOT equal to 7. This essentially says that we suspect another author wrote the poems.

Something to keep in mind when setting up a significance test is whether to use a one-sided or two-sided Alternative Hypothesis. In our Shakespeare example, we're using a two-sided alternative hypothesis because a different author might consistently use either more or fewer unique words than Shakespeare. But suppose we suspected the poem was written by a particular author who was known to consistently use more new words than Shakespeare. Then the alternative hypothesis would be one-sided, expressed as H_a: mu > 7

We begin by assuming the null hypothesis is true. Then we find the probability of getting a result as extreme as ours if the null hypothesis really is true. When the probability is small, we reject the null hypothesis and accept the alternative hypothesis as plausible.

How do we find this probability? We go back to the distribution of a sample mean "*x*-bar." Here's the distribution of the number of new words in a Shakespeare poem. It's Normal, with a mean of 7 and a standard deviation of 2.6.

If we form the distribution of "*x*-bar" for samples of 5 poems, the mean is still 7 and the curve is still normal. But because we're now looking for the standard deviation of the sample mean, sigma is calculated by this equation. That means the standard deviation for the sampling distribution is smaller than for the whole population, as you can see when we compare the two curves. The distribution for average number of unique words per poem from our sample of 5 poems is less variable than the distribution for new words in all the individual poems. So the standard deviation is now 2.6 over the square root of 5, or 1.163.

The mean of our five mystery poems is here, at 8.2. We want to find the probability that any sample of five Shakespeare poems would have an "*x*-bar" at least that far, or farther, in either direction from 7.

That's an "*x*-bar" above 8.2, or below 5.8. Remember, we want the probability in both directions because our Alternative Hypothesis is two-sided.

You already know how to do this probability calculation: we standardize the distribution of "*x*-bar" to obtain the familiar *z*. We call *z* the test statistic. The *z*-score of "*x*-bar" is essentially the distance from the hypothesized mean, measured in standard deviations.

The ingredients of *z* are the observed "*x*-bar", the population mean mu given by the null hypothesis, the standard deviation of the population, and the sample size. We just plug in the numbers.

"*x*-bar" is the sample mean of our five poems, 8.2.

Mu is the population mean of 7 for all of Shakespeare's known poems.

The standard deviation is 2.6 and the sample size is 5.

So our *z* is 1.03, a little bit more than one standard deviation away from the mean on the standardized Normal curve. That's our observed value of our test statistic.

The final step in our Test of Significance is to find the probability of a *z* at least this extreme. This probability is called the *p*-value.

To find this *p*-value we use the value of *z* that we just calculated, 1.03. Consulting the *z*-table we can see that the area under the curve below 1.03 is .8485.

This means that .1515 is left in the tail. To find our *p*-value we double this value because we're interested in the area under both tails of the curve. Our final result, .303, is the *p*-value we've been looking for.

So there's a .303 or 30.3% chance that random variation would produce a mean unique word count as far from 7 in either direction as 8.2 is. 30.3% is a pretty good chance – we could expect a result like that almost once every three times. This means we've failed to disprove the null hypothesis and did not find good evidence against Shakespeare's authorship of these new poems.

This example helps illustrate the general rule about *p*-values: Small *p*-values give evidence against the null hypothesis, while large *p*-values fail to reject the null hypothesis.

Think about it…if we had gotten a smaller *p*-value, say .04, that would mean there was only a 4% chance that 5 poems by Shakespeare would produce an average number of unique words this far from his usual 7. That's a pretty small chance, so we would have good reason to suspect that something else was at work – namely, a different author.

As you can imagine, *p*-values can range from the very small – close to zero – to the very large – close to one. So researchers are faced with a dilemma: is there a particular *p*-value that gives enough evidence to reject the null hypothesis? If we have a result of 1 in 1,000 it's pretty easy to see that the null hypothesis is likely wrong. But what about 1 in 100? Or 1 in 50?

Because this is such a common problem, several fixed *p*-values are often used. One of the most common values is .05 or 5%. This would mean that a result would be likely to occur only 5% of the time if the null hypothesis is true. If something is statistically significant at the 5% level, it means that the results produced a *p*-value less than .05, and were therefore significant. Another widely used level is .01. This would mean that a result would be likely to occur only 1 time in 100.

It's important to pick a *p*-value that's appropriate to the situation. For example, if we were doing research on a new cancer therapy, because of the health consequences for the patients, we might decide that a .05 *p*-value wasn't good enough. We might then decide that our results are only significant if they reach a .01 *p*-value or even .001.

You may not have expected your English homework to come creeping into your statistics lessons, but as we've seen here, significance testing can be useful in just about any subject. And though you might think parting with this topic is such "sweet sorrow," don't worry, tests of significance will keep turning up in future modules that use inference.

So stay tuned! I’m Pardis Sabeti for *Against All Odds*.

PRODUCTION CREDITS

Host – Dr. Pardis Sabeti

Writer/Producer/Director – Maggie Villiger

Associate Producer – Katharine Duffy

Editor – 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.