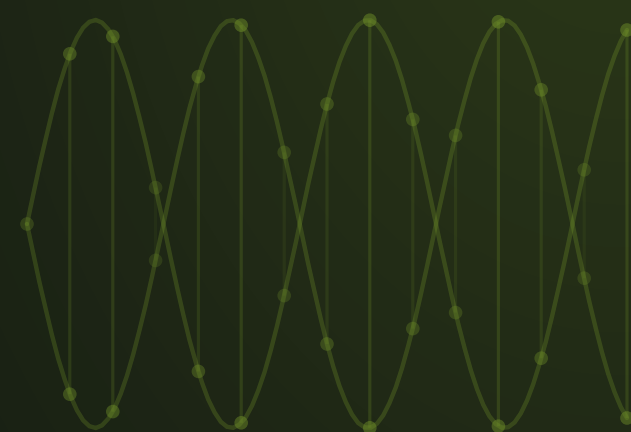


Lesson: p-values, and why genetics uses 0.00000005



Statistics

Mathematics

Genetics

Age 16+

Advanced

A free statistics lesson with one coin per student. The class runs six hundred studies of something that is definitely not happening, gets about thirty-eight discoveries, and works out from there why genetics needs a p-value of 0.00000005 rather than 0.05.

Free to read, use and adapt. No account, nothing to buy, nothing collected from anyone.

Full lesson, figures and updates: <https://www.mygenelog.com/updates/lesson-the-threshold>

What this is. A ready-to-run statistics lesson built on real genetics, free to use and adapt in any school, college or university. Nothing to buy from us, no account, no data collected from anyone.

- Level: from about age 16 through postgraduate. The activity is identical at every level; the questions at the end are graded, and the last set is genuinely hard.
- Time: 45–60 minutes. The activity itself takes about fifteen.
- You need: one coin per student, or a phone. A board. A calculator for the last part. That is all — there is nothing to order and nothing to spend.
- It covers: p-values, the multiple testing problem, family-wise error, Bonferroni correction, why genome-wide significance is 5×10^{-8} , why the only cure for a hard threshold is more data, and how a scientific convention actually gets decided.

Start here: six out of one hundred and sixty-six

By 2002, genetics had a problem it did not want to talk about.

Researchers had spent a decade picking genes that looked like plausible suspects for a disease, testing whether a variant in that gene was commoner in patients than in healthy people, and publishing when the answer came out below the usual bar of $p < 0.05$. More than 600 positive associations had been reported this way.

Then four researchers sat down and checked which of them had held up. Of the 166 associations that had been tested three or more times by independent groups, the number that had been consistently replicated was six.

Six. Out of a hundred and sixty-six.

Nobody had cheated. Every one of those studies had used the same threshold that introductory statistics teaches, applied correctly. The threshold was the problem, and the next hour is about why.

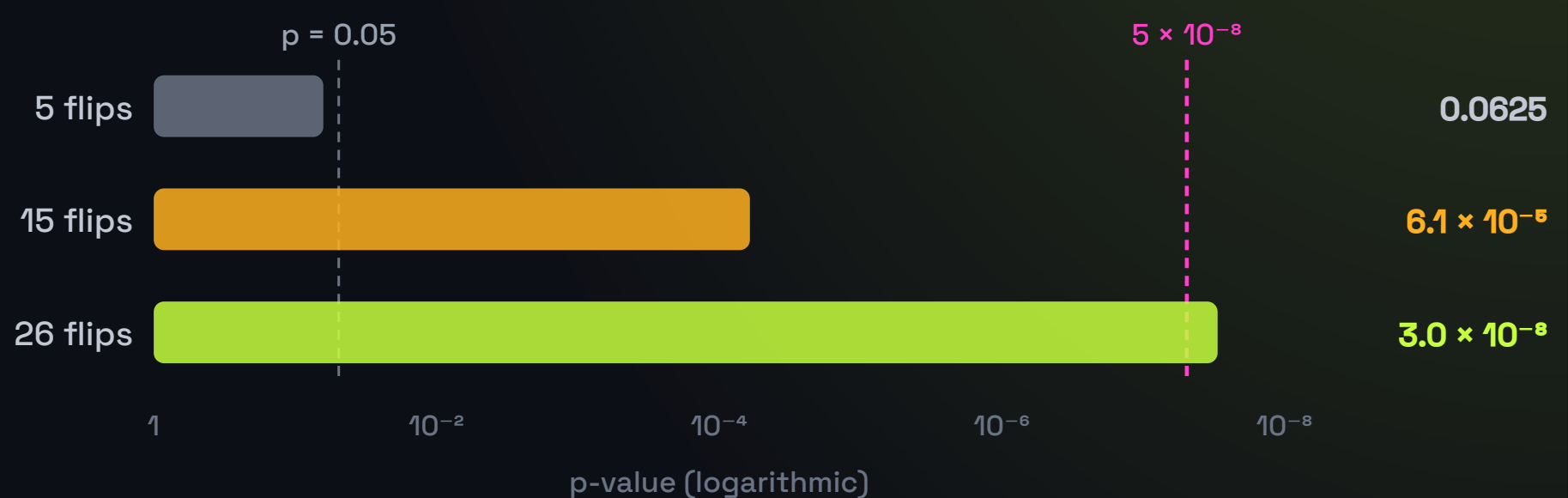
Do this

1. Take a prediction first, before anything is flipped. Ask: “Suppose we test a million positions in the genome, and suppose not one of them affects the disease at all. Nothing is happening anywhere. How many will come out significant at $p < 0.05$?” Take three guesses and write them on the board. Do not say whether any of them is close.
2. Define the study. One study is five flips of a coin. The result is “significant” if all five come out the same — five heads or five tails. Work out the probability with the class rather than telling them: $2 \div 25 = 2/32 = 0.0625$, which is about the conventional 0.05.
3. Say out loud what the control is. This whole experiment is a control condition. The coins are fair, nothing is influencing them, and there is definitively nothing to find. Every “discovery” the class is about to make is false by construction — that is the design, not a flaw in it. This point does the work of the entire lesson and it is worth stopping on.
4. Run twenty studies each. Every student flips five times, records H or T, marks the run if all five match, and repeats twenty times. It takes about ten minutes and it is noisy. Let it be noisy.
5. Count the class total on the board. Two numbers only: how many studies the class ran altogether (students $\times$ 20), and how many came out “significant”. A class of 30 runs 600 studies and should expect $600 \div 16 = 37$ or 38 discoveries of absolutely nothing.
6. Now go back to the prediction on the board. A million tests at $p < 0.05$ gives 50,000 false discoveries. Almost nobody guesses a number that large, and the class has just produced the small version of it with their own hands.
7. Fix it, and watch what fixing it costs. If you run 600 tests and want to be 95% confident that not one false discovery gets through, divide the threshold by the number of tests: $0.05 \div 600 = 8.3 \times 10^{-5}$. Ask the class: can any five-flip study reach that? It cannot — the smallest p-value five flips can produce is 0.0625. No amount of luck can rescue a study that is too small.
8. Finish with the real number. How many flips would it take? Solve $2 \div 2^n \leq 8.3 \times 10^{-5} \rightarrow n = 15$. And to clear the threshold real genetics uses, 5×10^{-8} ? $n = 26$. Twenty-five flips gets to 5.96×10^{-8} and misses.

Nothing about anybody's body is collected in this lesson, and nothing should be. This activity uses coins on purpose. Do not ask students for their own genetic results, their health, or their family's — not as an extension, not as homework, not anonymously. Count a room if you need a number; never write a name list. If a student volunteers something about their own genetics, that is their business and not the class's data.

What the class just proved

smallest p-value a run of identical flips can reach →



Exact, not drawn by eye: a run of n identical flips has probability 2^{-n} . Five flips lands at 0.0625 and does not quite clear the 0.05 line. To clear 5×10^{-8} you need 26 — 25 gets you to 5.96×10^{-8} , which misses. Those extra 21 flips are what “more data” means.

The bar for one study is not the bar for a thousand studies, and it never was. What $p < 0.05$ actually promises is: if nothing is happening, a result this strong turns up about one time in twenty. Run one test and that is a reasonable risk. Run a million and it is a guarantee of fifty thousand wrong answers.

This is the whole of the multiple testing problem, and the class has just generated it from first principles rather than being told about it.

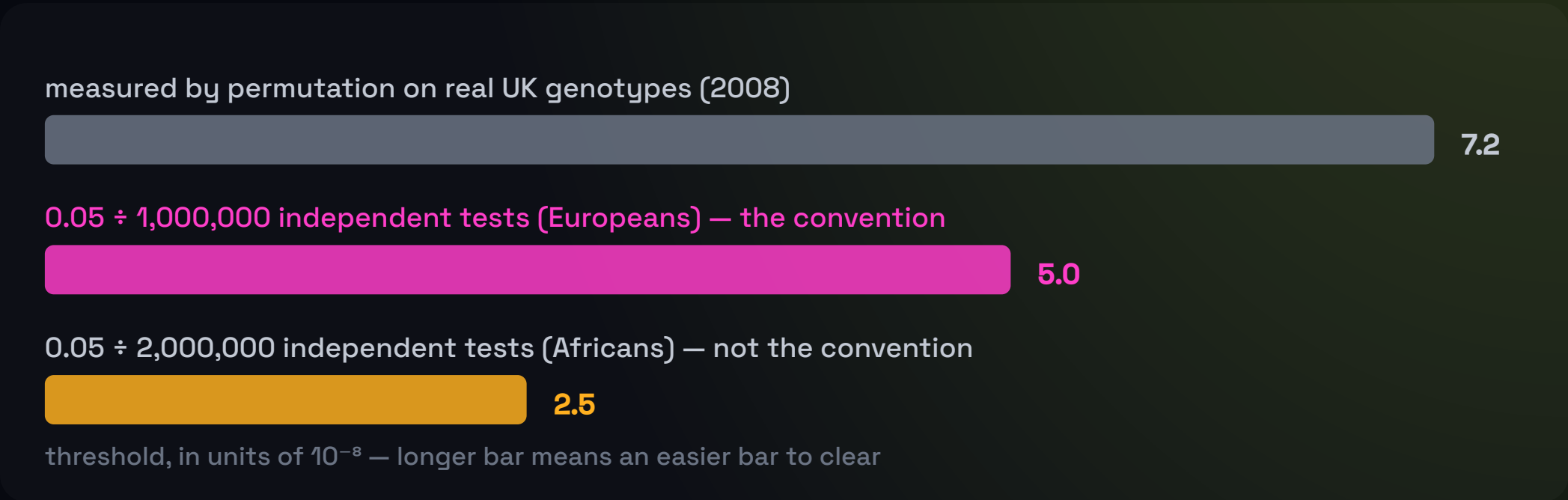
Where 5×10^{-8} came from

So genetics needed a number. In 2008 two groups went and found one, in two completely different ways, and this is the part students almost never get told.

One group counted. Because nearby positions in the genome are inherited together, three million tested positions are nowhere near three million independent tests. Working from the International HapMap data, they estimated the real testing burden at about one million independent tests in Europeans — and about twice that in Africans, whose genomes carry more variation and less of it travelling in blocks. Divide

0.05 by a million and you get exactly 5×10^{-8} .

The other group measured. They took real genotypes from the Wellcome Trust Case-Control Consortium, shuffled the disease labels at random so that any signal found had to be noise, recorded how strong the strongest noise got, and repeated that at increasing marker densities until the answer stopped moving. Their extrapolated threshold was 7.2×10^{-8} .



The same question, asked three ways, in the same year. The field settled on the middle one and applies it to everybody — which is the argument at the end of this lesson.

Two methods, one theoretical and one empirical, landing within a factor of 1.5 of each other. The field rounded to 5×10^{-8} and has used it ever since.

Worth saying plainly to a class: this is what a scientific convention looks like from the inside. Not handed down, not proved — argued out, twice, and then agreed on because agreeing was better than everybody choosing their own.

Real data: what the bar actually stops

This site collects genetic associations automatically from the public GWAS Catalog and refuses anything weaker than 5×10^{-8} . Here is what that filter has done, counted from its own logs.

Counted on 8 September 2026. The catalogue has grown since, and these numbers are left as they were: they describe one run of the collector, and a figure replaced by today's would not match the proportions drawn from it.

3,510 association records, as they came out of the filter



- **282** published — 8.0%, and every one of them cleared 5×10^{-8}
- **1,669** refused for a p-value weaker than 5×10^{-8}
- **1,559** refused for something else — unclear alleles, no gene, already here

Counted from this site's own collector logs, not estimated. Nearly half of what it looked at was refused on the p-value alone.

Of 3,510 association records examined, 1,669 — 47.5% — were refused on the p-value alone. Another 1,559 were refused for other reasons: the allele data was not clean, no gene was mapped, the record was already here. 282 became pages.

Nearly half. Every one of those refused records is a published, peer-reviewed result sitting in a public database, and this site will not print it, because the threshold the class just derived says it might be a coin coming up heads five times.

The 653 variants this site does carry come from 182 different papers, and every single one of them cleared 5×10^{-8} .

The argument that is still open

Give the class the figure above again and let them fight about it, because researchers are still doing exactly that.

The threshold is one number applied to everybody, and the burden is not the same for everybody. If a European-ancestry study really is running a million independent tests and an African-ancestry study is running two million, then 5×10^{-8} is the right bar for one of them and too lenient for the other. Using the stricter 2.5×10^{-8} for African-ancestry studies would be more correct and would also make it harder to publish findings from the populations that genetics has already studied least. There is no answer to that which is purely statistical.

And the number was calculated for a technology that is being replaced. A million independent tests describes genotyping chips reading common variants. Whole-genome sequencing reads rare ones too, in far greater numbers, and some argue the threshold should move to 5×10^{-9} or lower. Others answer that a threshold strict enough to be safe is strict enough to make real effects unfindable at any sample size anyone can afford.

Both arguments are live. A class with the figure and twenty minutes has the same

information the people arguing have.

Questions

To warm up

1. The class ran a number of studies of something that definitely was not happening. How many “discoveries” would you have expected before you started? How many did you get?
2. Why was it important that the coins were fair? What would have been wrong with the lesson if one of them had been weighted?
3. A newspaper reports that a gene has been linked to being good at maths, with $p < 0.05$. What is the first question you should ask?

Core

1. Explain, in one sentence a 12-year-old would understand, why a threshold that is fine for one test is useless for a million.
2. The class corrected its threshold to $0.05 \div (\text{number of studies})$. Why did that make every single one of the class’s discoveries disappear, no matter how lucky anyone got?
3. Show that 26 flips is the smallest number that clears 5×10^{-8} , and say what “more flips” corresponds to in a real genetic study.
4. One of the two 2008 groups calculated their threshold and the other measured it. What is the advantage of each, and why is it reassuring that they nearly agreed?

To stretch

1. Bonferroni correction assumes the tests are independent. Positions near each other on a chromosome are inherited together and are not independent. Does that make the correction too strict or too lenient, and why did the 2008 group have to estimate the number of independent tests rather than just counting the markers?
2. A strict threshold reduces false discoveries and increases missed real ones. Who bears the cost of each kind of error? Is the answer the same for a study of a common disease and a study of a rare one?
3. If the correct burden is a million tests in Europeans and two million in Africans, and the field uses 5×10^{-8} for both, what happens? Now argue the other side: what happens to a field that already under-studies some populations if it also makes their results harder to publish?
4. Only 6 of 166 associations replicated in 2002. Suppose the threshold had been 5×10^{-8} all along. Would all 160 of the others have vanished, or would some real findings have gone with them?

Teacher notes — where each question is going

1. The expected count is $(\text{studies} \div 16)$. The observed count will not match it, and the gap is worth a minute on its own: this is sampling variation on top of a

false-discovery rate, and both are in the room at once.

2. A weighted coin would mean there was something real to find, and the discoveries would no longer all be false. The lesson depends on knowing the null is true — which is exactly what a real study never knows, and that is the point to leave hanging.

3. “How many things did they test?” Everything else follows.

4. Looking for a strong answer, not a technical one: “something that happens one time in twenty will definitely happen if you try a million times.”

5. Because the correction moved the bar below the best result the design can physically produce. This separates two ideas students routinely merge: strength of evidence is bounded by the study, and luck cannot exceed that bound.

6. $2/2^n \leq 5 \times 10^{-8}$ requires $2^n \geq 4 \times 10^7$, so $n = 26$ ($2^{26} = 67,108,864$, giving 2.98×10^{-8}). More flips means more participants: sample size is the only lever that moves the strongest available p-value.

7. Calculation generalises and is cheap; measurement captures the messiness of real data including the correlation structure. Agreement between two methods with different assumptions is worth more than either alone — the same logic as the vitiligo and blood pressure pages on this site.

8. Too strict: correlated tests mean fewer effective tests than markers, so dividing by the marker count over-corrects. Hence estimating independent tests. Strong students may notice this makes the threshold a property of the population, not of the technology — which is question 10.

9. No single right answer. Push for who is actually harmed: a false discovery wastes research money and can reach patients; a missed one may delay a treatment by a decade. For rare disease, missing is often the worse error, and some fields adjust for exactly that reason.

10. This is the hardest and the best. The statistics has a clean answer and the consequence does not. Do not resolve it for them. If nobody raises it, point out that the populations with the higher testing burden are the ones with the fewest studies to begin with.

11. Some real findings would certainly have been lost — several of the 6 that replicated would have failed a genome-wide bar in a study of that size. The honest answer is that the field traded discoveries for trustworthiness and decided the trade was worth it.

Take it further

- The alleles your parents kept — a study of three million people, and what a very large

sample buys you. Good follow-on reading for the stretch questions.

- Our data page — the live counts, including how many variants and papers are behind them.
- Where this data comes from — the full set of filters an association has to pass before it gets a page here, of which the p-value is only the first.
- C-reactive protein — a worked example of the next question after significance: the association is real, so is it a cause?

Use this freely. Print it, copy it, cut it up, put it on your own worksheet, translate it, change the questions. No permission needed and nothing to pay. If you credit it, mygenelog.com is enough.

If you teach with it and something in it does not work, tell us — that is worth more to us than a thank-you.