

Statistics and Lies

As an example of a research question, we'll ask: Does a coin have a slight preference for flipping to "heads" or "tails" ? (In our final report we will use the more academic-sounding expressions "obverse" and "reverse".) Since we don't have a theory of why one side would be favored over the other, we will not hypothesize one or the other, but start with an exploratory 50 throws.

Our randomly selected *research subjects* are ten quarters, "A" through "J", identified by their dates:

A	1976 D "Centennial"
B	1979 D
C	1981 D
D	1985 D
E	1999 D "Delaware"
F	1999 D "Georgia"
G	2001 D "Kentucky"
H	2003 D "Illinois"
I	2004 D "Florida"
J	2005 D "California"

(We observe that they all are from the Denver mint, so our results will be restricted to quarters from Denver.) Our *research protocol* is to spin each quarter and record the outcome, and then repeat four more times.

The results from the exploratory tests (marked "1" for heads and "0" for tails) are:

Trial	A	B	C	D	E	F	G	H	I	J
1.	1	0	1	0	1	1	0	1	1	1
2.	0	0	1	0	0	1	0	1	0	0
3.	0	0	1	1	0	1	1	1	0	0
4.	1	0	1	1	1	0	1	0	1	0
5.	1	0	1	1	0	1	0	0	1	1

We see that there were 27 heads and 23 tails; we form the hypothesis that heads is favored. To test our hypothesis, we now run another 50 throws:

Trial	A	B	C	D	E	F	G	H	I	J
6.	0	0	1	0	1	1	1	1	0	0
7.	1	0	1	1	0	1	0	1	1	0
8.	0	0	1	0	0	0	1	0	0	1
9.	0	0	0	0	1	1	1	0	0	0
10.	0	0	0	1	1	1	0	0	0	0

In this second set, there were 20 heads and 30 tails. We do not need fancy statistics to reject our hypothesis. Indeed, it looks like we should have postulated the opposite hypothesis!

Since more data is always better, we decide to include *all* available data in our model:

Trial	A	B	C	D	E	F	G	H	I	J
1.	1	0	1	0	1	1	0	1	1	1
2.	0	0	1	0	0	1	0	1	0	0
3.	0	0	1	1	0	1	1	1	0	0
4.	1	0	1	1	1	0	1	0	1	0
5.	1	0	1	1	0	1	0	0	1	1
6.	0	0	1	0	1	1	1	1	0	0
7.	1	0	1	1	0	1	0	1	1	0
8.	0	0	1	0	0	0	1	0	0	1
9.	0	0	0	0	1	1	1	0	0	0
10.	0	0	0	1	1	1	0	0	0	0

We now have twice as many data points, with 47 heads and 53 tails. Our final hypothesis is that tails are favored slightly; we apply some criterion that compares the extra 3 tails with random fluctuations in the heads/tails balance (proportional to the square-root of 100, the number of trials) and conclude that our tests give only weak confirmation of the theory.

We don't notice that we have engaged in the Wharton fallacy: in economic forecasting, where you have to wait a year before you get new data, it is very tempting to include *all* your data in your model. As a result, the model is not tested until next year. Next year's data will show that the model should be rejected, but we will use the new data to fine-tune a new, better model, which "predicts" the past perfectly. The problem is that the current model is always untested. If you use all your data to construct your theory, you don't have any way to test it statistically!

Inspection of the results leads us to suspect that certain quarters favor one outcome of the other: coin "B" always came up tails in 10 tests, and coin "C" came up heads 8 times out of ten trials. Compared to the square-root of 10, those results appear to be highly significant. After some thought, we report only that we have identified a 1979 quarter (our "B"), which strongly favors tails. To avoid confusing the issue, we do not report that we also tested 9 other coins with inconsistent results.

We have now engaged in the correlation fallacy identified by Judith Rich Harris: in educational research, it is common to split a data set many different ways, until we find some correlation somewhere. We then apply chi-square analysis as if these were the only data we had taken, and as if we had taken data *after* our hypothesis was formed. That's why there are so many inconsistent studies reported: the results are not reproducible, and will not be consistent with additional data.

To deflect criticism and confirm our hypothesis that some coins favor tails, we decide to take coin “B” and throw it 90 more times; we get 44 heads and 46 additional tails. Summing it all up, that coin had 44 heads and 56 tails, a statistically significant result.

We have capped our efforts with the Duke/Rhine fallacy: in ESP research, it was common to combine the data by which the psychics were identified, with the data by which they were subsequently tested. The overall result was always weaker than the original identification data, but tended in the same direction because selection data were included.

Having obtained publishable results, we obtain further funding and have several graduate students expand the research to other coins. Most of our students turn out to have little talent for research and obtain only indifferent result, but one student is successful and reports on a dime which strongly favors heads.