Lecture 2: Statistical Evidence

Bryan S. Graham, UC - Berkeley & NBER

January 29, 2019

Games of chance played an important role in the development of probability theory, deeply influencing, for example, the thinking of great 17th century scholars such as Blaise Pascal and Pierre de Fermat. Games of chance also play an important role in teaching probability and statistics. The randomness inherent in flipping a coin, or rolling a die, is well-understood by most people. Chance is also central to how we interpret *statistical evidence*: under what conditions are we justified in saying that the data in hand support one hypothesis versus another?

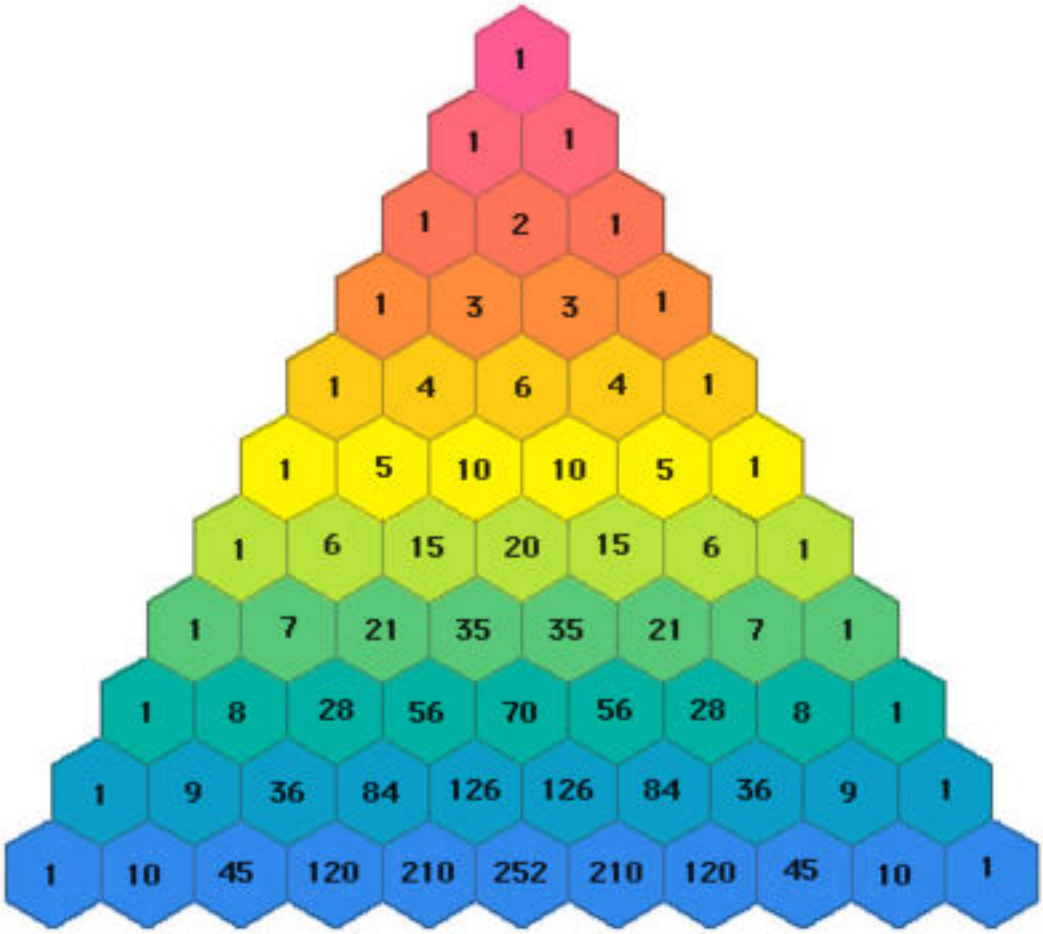
Consider the act of flipping a coin. If the coin is "fair", then the *ex ante* probability of the coin landing heads is one-half. Now say you flip the coin 10 times and the coin lands heads all ten times. This strikes you as unusual. There are two possibilities (i) the coin is "biased" (i.e., weighted such that it lands heads with probability greater than one half) or (ii) the coin is indeed fair and you just got lucky.

Each coin flip is independent. Whether a coin lands heads on the 3rd flip has no bearing on how it lands on, say, the 7th flip. The *statistical evidence* is that 10 independent coin flips have all landed heads. What conclusions should (may?) we draw from this evidence?

Since each coin flip has two outcomes, and you flip the coin ten times, there are $2^{10} = 1024$ possible sequences of heads and tails of length 10. While all of these sequences are equally likely, in only one of them do all ten flips land heads. If the coin is fair, the event "ten coin flips, all land heads" is indeed a lucky outcome. It has an ex ante chance of occurring equal to 1/1024. Say, instead, your coin flipping resulted in five heads and five tails. There are $\binom{10}{5} = 252$ flip sequences consisting of five heads and five tails. The event "ten coin flips, half land heads, half lands tails" is not especially rare. About one quarter of the time you will observe such a sequence if the coin is fair.

Lets introduce a bit more structure to our thinking. Our *null hypothesis* (or maintained null hypothesis) is that the coin is fair. Under the null we have just observed a 1 out of 1000 event (10 heads in a row). Since we have observed something that occurs very infrequently

Figure 1: Pascal's Triangle



Source: http://mathforum.org/mathimages/index.php/Pascal's_triangle.

Table 1: The logic of statistical testing

	Actual State of the World	
	Null is true (H_0)	Null in false (H_1)
Fail to Reject	$1-\alpha$	Type II Error ($1-\beta$)
Reject	Type I Error (α)	Power (β)

Notes: Size of test equals α. Power of test equals β.

under the null, we conclude that our null must be wrong. If the statistical evidence in hand would arise only very rarely under our maintained null beliefs about the state of the world, then we conclude that our beliefs are likely to be incorrect. We reject of null hypothesis in favor of the alternative that the coin is biased toward heads.

It is important to understand that we don't know for certain that our null hypothesis is wrong. What we do know is that the evidence in hand would not occur very often if it were true. This is the basis for our rejection; nothing more.

Now imagine we've been assigned a job as a coin tester at the Philadelphia Mint. Our job is to make sure coins produced at the mint are fair. Fair coins are allowed to enter into circulation, while coins deemed biased are melted back down to be recast. You develop the following procedure. If ten flips come up all heads or all tails you send the coin back for recasting. Otherwise the coin is certified "fair" and sent out into circulation.

This procedure will, of course, mistakenly send some fair coins back for recasting. This is unfortunate, but also unavoidable. The frequency with which fair coins are falsely deemed biased, a so called Type I error (see Table 1), is called the *size* of the procedure. Since all heads or all tails occurs with an *ex ante* chance of 2 out of 1024, the size of your procedure is $\frac{2}{1024}$. You end up remelting about 2 out of every 1000 fair coins. We generally use the greek symbol α to represent size.

Now say you test a biased coin. This biased coin lands heads with a probability of $\frac{3}{4}$. The chance of observing an all heads or all tails sequence for this coin is

$$\left(\frac{1}{4}\right)^{10} + \left(\frac{3}{4}\right)^{10} \approx 0.056.$$

So even if the coin is heavily biased, your procedure only successfully detects, and sends back for recasting, about 6 out of 100 times. The frequency with which your test correctly rejects the null is called *power* (see Table 1). We denote power by β.

The standard approach to testing in statistics is to fix the size of the test in advance. Typical values for α in the social sciences are 0.05 or 0.01, meaning that the test falsely rejects the null hypothesis about 1-out-of-20 or 1-out-of-100 times respectively. Holding size fixed, the

goal is then to construct procedures with good power. Researchers generally face a *size-power trade-off*. Reducing size (and the rate of Type I) errors, generally means reducing power (and increasing the rate of Type II errors).

Say you switched your procedure to reject anytime you observed at least nine heads or tails. There are 10 sequences each with nine heads or nine tails. So under this new procedure you will commit Type I errors at a rate of 22/1014, or about 2 percent of the time. What happens to power? The probability of a 9 heads sequence is $\left(\frac{3}{4}\right)^9 \frac{1}{4}$, of which their are 10 such sequences. The probability of a 9 tails sequence is $\left(\frac{1}{4}\right)^9 \frac{3}{4}$, of which there are also 10. So across biased coins that land heads with probability 0.75 you correctly reject, and send back for recasting,

$$\left(\frac{1}{4}\right)^{10} + 10\left(\frac{1}{4}\right)^{9}\frac{3}{4} + 10\left(\frac{3}{4}\right)^{9}\frac{1}{4}\left(\frac{3}{4}\right)^{10} \approx 0.24$$

of all coins with heads probability 0.75.

The size-power paradigm is the dominant one in modern statistics, but there are other principled approaches. In applying these methods in practice one should be mindful of the possibility of both Type I and Type II errors. There are also subtle issues of interpretation.

Incorporating prior knowledge when interpreting statistical evidence

Consider the classic example of testing for a rare disease. Suppose there is some disease in the population, very expensive to treat, that affects about 1 out of 1000 individuals. As part of a routine medical screening you are tested for the disease. Your test comes back positive. The doctor tells you that the power of the test is 0.95, while its size is 0.05. Before the test you had no reason to believe that you were especially likely to have the disease. How should you interpret the the positive test result? From Bayes' Law we have that

$$\begin{aligned}
\Pr\left(\text{Have Disease}|\text{Test Positive}\right) &= \frac{\Pr\left(\text{Test Postitive}|\text{Have Disease}\right)\Pr\left(\text{Have Disease}\right)}{\Pr\left(\text{Test Positive}\right)} \\
&= \frac{(0.95)\times\left(\frac{1}{1000}\right)}{(0.95)\times\left(\frac{1}{1000}\right)+(0.05)\times\left(\frac{999}{1000}\right)} \\
&\approx 0.02.
\end{aligned}$$

So, unless you have strong reason to believe you were sick prior to testing, chances are the test result is a false positive (i.e., Type I error). The optimal action here is to request a

second test. If you test positive again, then the chance that you have the disease is about $\frac{1}{4}$. After three positive tests almost 0.9. When confronted with evidence for a rare outcome seek out more evidence. This example is cautionary.

Now consider an impact evaluation example. By studying an online registry of impact evaluations with an international development focus (e.g., http://www.ridie.org/) you discover that roughly one half of evaluations result in a positive finding (i.e., conclude that the program is "effective" for the target outcome). You assume that the typical study has a power of $\beta = 0.80$ and size $\alpha = 0.05$. Let p be the fraction of evaluated studies that *actually* are effective. The fraction which are *found* to be effective can be expressed as a weighted average of correct rejections of the null and Type I errors:

$$\frac{1}{2} = 0.80p + 0.05\,(1-p)\,.$$

Solving for p yields $p = \frac{3}{5}$. So roughly 60 percent of all evaluated programs are *actually* effective for their target outcome (at this point I should emphasize that all these numbers are fictitious and used solely as a vehicle to make an expository point).

Now an independent consultant comes along and evaluates a program in which you have some supervisory role. The consult designs a study with power of $\beta = 0.80$ and size $\alpha = 0.05$. She finds no evidence against the null of no effect. For the sake of argument, lets assume that you don't have any particular views about the effectiveness of your program prior to the evaluation. How should you incorporate the consultant's findings into your beliefs? Again using Bayes' Law we have

$$\begin{aligned}
\Pr\left(\text{No Actual Effect}|\text{No Effect Found}\right) &= \frac{\Pr\left(\text{No Effect Found}|\text{No Actual Effect}\right)}{\Pr\left(\text{No Effect Found}\right)} \\
&\quad \times \Pr\left(\text{No Actual Effect}\right) \\
&= \frac{(1-\alpha)\,(1-p)}{(1-\beta)\,p + (1-\alpha)\,(1-p)} \\
&= \frac{0.95 \times \frac{2}{5}}{0.20 \times \frac{3}{5} + 0.95 \times \frac{2}{5}} \\
&= 0.76.
\end{aligned}$$

The evaluators negative finding should give you serious pause about the efficacy of your program. At the same time, across many similar situations, the evaluator will be wrong about one out of four times. Falsely concluding that an effective program is ineffective in such cases.

Now assume that prior to the evaluation you were rather confident in the efficacy of your program, believing that there was a 90 percent chance it was effective. Setting $p = 0.9$ we have

$$\Pr\left(\text{No Actual Effect}|\text{No Effect Found}\right) = \frac{(1-\alpha)(1-p)}{(1-\beta)p + (1-\alpha)(1-p)} \approx 0.35.$$

So, in this case, after studying the (negative) evaluation evidence you believe there is about a one-third chance your program is ineffective. This is much more than the 10 percent chance your gave to such an outcome prior to the consultant's report, but still leaves you believing the program is probably effective.

What if, instead, the consultant had designed a study with power $\beta = 0.95$ and size $\alpha = 0.01$. This is a considerably more rigorous evaluation design. Now if the consultant reports a no effect finding we get

$$\Pr\left(\text{No Actual Effect}|\text{No Effect Found}\right) = \frac{(1-\alpha)(1-p)}{(1-\beta)p + (1-\alpha)(1-p)} \approx 0.91$$

Faced with this evidence you update your beliefs rather dramatically. Although prior to the evaluation you were quite high on the program, giving it about a 90 percent chance of being effective, after the evaluation you think there is less than a 10 percent chance it is effective.

The point of these examples is to remind you that the interpretation of *statistical evidence* does not occur in a vacuum. In general statistical evidence comes in the form of "a $\alpha = 0.05$ test for the null hypothesis that the coin is fair rejected". If the test has been properly calibrated and implemented (which can be big "ifs" in practice) we know that rejections occur relatively infrequently under the null (no more than 5 percent of the time). By convention we often take such evidence as demonstrating that the null is false (i.e., the coin is biased).

However when the decisions that need to be made are very consequential, as in deciding whether to begin a costly medical treatment regimen, or whether to discontinue a long-standing program, it is perfectly reasonable to combine the information provided by the test alone with additional outside information as we have done in the examples above.

