305 Lecture 45 - Significance Testing

Brian Weatherson

July 29, 2020

Plan

- We’re going to end with a discussion of the role of significance testing in contemporary statistics.

Associated Reading

- Odds and Ends, Chapter 19 (and part of chapter 20)

Significance Testing

- The subjective approach is very popular within philosophy, but not within statistics or a lot of social sciences.
- This is in part because of its subjectivity.
- A lot of sciences want methods that are more objective.

Significance Testing

What is known as ‘classical statistics’ is based around the idea of significance testing.

- The intuitive idea is that we say that a correlation reflects a real pattern in the underlying data if (but only if) it would be really improbable if we got the data we did by chance.
- Intuitively, three heads in a row might be a coincidence, but ten heads in a row suggests something odd is going on.

Significance Testing

So say that we want to argue that two things are connected.

- To make this concrete, say that we want to argue that the survival rates for people who take our company's drug are higher than for people who do not.
- What we do is give a bunch of people the drug (after getting all the approvals!)
- We then look at their survival rates, and ask *How likely would this data be if our drug had no effect at all*?
- If that number is low enough, we conclude that our drug works. (Profit!)

Significance Testing

- That is, we work out something like $\Pr(E|H_0)$, where E is the data that we get, and H_0 is a **null hypothesis** that says there is nothing of interest here.
- If $\Pr(E|H_0)$ is low enough, we conclude that H_0 is false, and hopefully there is a natural alternative to H_0, such as that the drug works, that we infer.
- In those cases, we will say that the data shows a significant correlation between taking the drug and survival rates.
- This literally means that it would be really improbable to get this data by chance.

Significance Testing

How low is ‘really low’?

- As the book says, this is mostly a matter of convention.
- Which is ironic given the whole point was to get away from subjectivity.
- But a common idea is that it is less than 5%.
- So a correlation is significant if it is outside the central 95% of the distribution.

Inverting

- Isn't this all wrong though?
- Isn't it inferring from the fact that $\Pr(E|H_0)$ is low to the conclusion that $\Pr(H_0|E)$ is low, something that we've said over and over again not to do?

Inverting

- Isn't this all wrong though?
- Isn't it inferring from the fact that $\Pr(E|H_0)$ is low to the conclusion that $\Pr(H_0|E)$ is low, something that we've said over and over again not to do?
- Yes, but...

Inverting

- Isn't this all wrong though?
- Isn't it inferring from the fact that $\Pr(E|H_0)$ is low to the conclusion that $\Pr(H_0|E)$ is low, something that we've said over and over again not to do?
- Yes, but...
- In practice the method is supplemented by practical rules that avoid the worst consequences of allowing these inversions.

Stopping Rules

- As it stands, this method leads to all sorts of nonsense and, frankly, fraud.
- It needs to be supplemented with extra rules to avoid obvious mistakes.
- The first thing that is needed, as was realised fairly early on, was a commitment to external ‘stopping rules’.
- If you are allowed to keep collecting data until the probability of the evidence given the null is low, you will almost always get to reject the null.

An Experiment

- This is actually an experiment; the data are randomly generated anew every time I compile these slides.
- I'm going to compile these slides, and then present them whether the data comes up the way I hope it does or not.
- So it could go horribly wrong!

An Experiment

- I'm going to simulate tossing a coin 100,000 times.
- It uses the computer's random number generator, and it is set up so the probability of heads on each toss is 0.5, and the tosses are independent.
- But I'm going to measure the probability of the evidence given the null after each toss, not just at the end.
- And by 'the probability of the evidence', I mean the probability of getting at least that many heads.
- If that is outside $[0.025, 0.975]$ then we have, at this point, a rejection of the null.
- This is absurd of course; the null is programmed to be true.
- I'll run it five times to see how it goes.

Take One

t	result	heads	frequency	prob	distance
4685	1	2399	0.5120598	0.9521008	0.4521008
4687	1	2400	0.5120546	0.9520654	0.4520654
4684	1	2398	0.5119556	0.9506448	0.4506448
4686	0	2399	0.5119505	0.9506088	0.4506088
4688	0	2400	0.5119454	0.9505728	0.4505728

Here t is the trial number, result is how the coin landed on that trial, heads is how many heads to that point, frequency is frequency of heads to that point, prob is probability of getting at least that many heads, and distance is the distance of that number from 0.5. In this trial, we ended up with this many heads.

```
## [1] 50019
```

Take Two

t	result	heads	frequency	prob	distance
35413	0	17587	0.4966255	0.1029849	0.3970151
35417	0	17589	0.4966259	0.1029977	0.3970023
35433	0	17597	0.4966274	0.1030489	0.3969511
35412	0	17587	0.4966396	0.1039377	0.3960623
35414	1	17588	0.4966397	0.1039441	0.3960559

Here t is the trial number, result is how the coin landed on that trial, heads is how many heads to that point, frequency is frequency of heads to that point, prob is probability of getting at least that many heads, and distance is the distance of that number from 0.5. In this trial, we ended up with this many heads.

```
## [1] 50012
```

Take Three

t	result	heads	frequency	prob	distance
43629	1	21997	0.5041830	0.9601343	0.4601343
43482	1	21923	0.5041856	0.9599762	0.4599762
43628	1	21996	0.5041716	0.9597228	0.4597228
43630	0	21997	0.5041714	0.9597194	0.4597194
43632	1	21998	0.5041713	0.9597159	0.4597159

Here t is the trial number, result is how the coin landed on that trial, heads is how many heads to that point, frequency is frequency of heads to that point, prob is probability of getting at least that many heads, and distance is the distance of that number from 0.5. In this trial, we ended up with this many heads.

```
## [1] 50068
```

Take Four

t	result	heads	frequency	prob	distance
70523	0	34936	0.4953845	0.0071894	0.4928106
70623	0	34986	0.4953910	0.0072241	0.4927759
70522	0	34936	0.4953915	0.0072646	0.4927354
70524	1	34937	0.4953916	0.0072653	0.4927347
70395	0	34873	0.4953903	0.0072961	0.4927039

Here t is the trial number, result is how the coin landed on that trial, heads is how many heads to that point, frequency is frequency of heads to that point, prob is probability of getting at least that many heads, and distance is the distance of that number from 0.5. In this trial, we ended up with this many heads.

```
## [1] 49860
```

Take Five

t	result	heads	frequency	prob	distance
61387	0	30456	0.4961311	0.0278665	0.4721335
61391	0	30458	0.4961314	0.0278705	0.4721295
61395	0	30460	0.4961316	0.0278745	0.4721255
61397	0	30461	0.4961317	0.0278765	0.4721235
61386	0	30456	0.4961392	0.0281248	0.4718752

Here t is the trial number, result is how the coin landed on that trial, heads is how many heads to that point, frequency is frequency of heads to that point, prob is probability of getting at least that many heads, and distance is the distance of that number from 0.5. In this trial, we ended up with this many heads.

```
## [1] 49868
```

Stopping Rules

- As I said, this is a known bug in significance testing, and every responsible scientist who uses it knows that you aren't meant to stop just when you get the data you want.
- But there is another kind of problem that we've recently discovered the importance of.
- Here I was testing just one hypothesis.
- What if we try to test many more hypotheses at once?

P-Hacking

- This practice is known as **p-hacking**.
- It is the tactic of looking at results within all sorts of sub-groups within the data in the hope of finding a significant correlation somewhere.
- And with enough subgroups, the odds are pretty good that you'll find one.

Another Experiment

- Here I'm doing 32000 coin flips, but each flip is randomly assigned to either having or not having three different characteristics: C1, C2 and C3.
- In medical contexts, think of these as being like sex, age, race, etc.
- Again, I'm just doing coin flips here.
- And I'm going to run the trials to completion.
- But we're going to look for significant correlations among each group.
- I'll walk through three attempts to see how likely it is we get one.
- I haven't seen the data, so there isn't commentary on the slides, but it is quite unlikely that we'll get a significant correlation on the whole set. On the sub-samples though...

Experiment One

Characteristic	Heads	Trials	Probability	Distance
All	15906	32000	0.1479278	0.3520722

Experiment One - With One Characteristic

Characteristic	Heads	Trials	Probability	Distance
Yes C3	7885	16024	0.0228219	0.4771781
No C2	7885	15996	0.0376179	0.4623821
No C3	7885	15976	0.0524126	0.4475874
Yes C1	7923	15981	0.1445738	0.3554262
No C1	7983	16019	0.3405911	0.1594089
Yes C2	8021	16004	0.6210662	0.1210662

Experiment One - With Two Characteristics

Characteristic	Heads	Trials	Probability	Distance
+C3-C2	3835	7866	0.0139480	0.4860520
-C1-C2	3933	8024	0.0398247	0.4601753
+C3+C2	4151	8158	0.9457971	0.4457971
+C1-C3	3964	8036	0.1163135	0.3836865
-C1+C2	4050	7995	0.8820884	0.3820884
-C3+C2	3870	7846	0.1179285	0.3820715
+C1-C2	3952	7972	0.2265088	0.2734912
+C1+C2	3971	8009	0.2304140	0.2695860
-C3-C2	4050	8130	0.3738687	0.1261313
-C1-C3	3956	7940	0.3809431	0.1190569
+C1+C3	3959	7945	0.3852617	0.1147383
-C1+C3	4027	8079	0.3947306	0.1052694

Experiment One - With All Three Characteristics

Characteristic	Heads	Trials	Probability	Distance
-C1-C2+C3	1904	3967	0.0060563	0.4939437
-C1+C2+C3	2123	4112	0.9823723	0.4823723
+C1+C2-C3	1943	3963	0.1136634	0.3863366
+C1-C2+C3	1931	3899	0.2821295	0.2178705
+C1-C2-C3	2021	4073	0.3191546	0.1808454
-C1+C2-C3	1927	3883	0.3265963	0.1734037
+C1+C2+C3	2028	4046	0.5686469	0.0686469
-C1-C2-C3	2029	4057	0.5125244	0.0125244

Experiment Two

Characteristic	Heads	Trials	Probability	Distance
All	15996	32000	0.4843929	0.0156071

Experiment Two - With One Characteristic

Characteristic	Heads	Trials	Probability	Distance
Yes C1	7974	16014	0.3037506	0.1962494
No C1	8022	15986	0.6796206	0.1796206
Yes C3	8008	15990	0.5845391	0.0845391
No C2	8008	16044	0.4156012	0.0843988
Yes C2	7988	15956	0.5660189	0.0660189
No C3	8008	16010	0.5220591	0.0220591

Experiment Two - With Two Characteristics

Characteristic	Heads	Trials	Probability	Distance
-C1+C3	4055	8007	0.8774333	0.3774333
+C1+C3	3956	7983	0.2166804	0.2833196
+C1-C2	4002	8064	0.2555862	0.2444138
-C3-C2	3990	8027	0.3038267	0.1961733
-C1-C3	3967	7979	0.3111554	0.1888446
-C1-C2	4006	7980	0.6440889	0.1440889
-C1+C2	4016	8006	0.6185799	0.1185799
+C3-C2	4018	8017	0.5883752	0.0883752
+C3+C2	3993	7973	0.5622940	0.0622940
-C3+C2	3995	7983	0.5356724	0.0356724
+C1-C3	4018	8031	0.5266899	0.0266899
+C1+C2	3972	7950	0.4776404	0.0223596

Experiment Two - With All Three Characteristics

Characteristic	Heads	Trials	Probability	Distance
-C1-C2+C3	2033	4002	0.8479038	0.3479038
+C1-C2+C3	1985	4015	0.2437179	0.2562821
-C1+C2+C3	2022	4005	0.7363238	0.2363238
-C1-C2-C3	1973	3978	0.3115369	0.1884631
+C1+C2+C3	1971	3968	0.3457323	0.1542677
+C1+C2-C3	2001	3982	0.6303508	0.1303508
+C1-C2-C3	2017	4049	0.4129310	0.0870690
-C1+C2-C3	1994	4001	0.4247685	0.0752315

Experiment Three

Characteristic	Heads	Trials	Probability	Distance
All	15898	32000	0.1282284	0.3717716

Experiment Three - With One Characteristic

Characteristic	Heads	Trials	Probability	Distance
No C2	7977	16086	0.1508314	0.3491686
Yes C1	7920	15942	0.2118772	0.2881228
No C1	7978	16058	0.2127167	0.2872833
Yes C3	7977	16052	0.2219549	0.2780451
Yes C2	7921	15914	0.2867798	0.2132202
No C3	7977	15948	0.5221019	0.0221019

Experiment Three - With Two Characteristics

Characteristic	Heads	Trials	Probability	Distance
-C1-C2	4011	8142	0.0936151	0.4063849
-C3-C2	3954	7999	0.1571372	0.3428628
+C1+C2	3954	7998	0.1598257	0.3401743
+C1-C3	3918	7907	0.2155796	0.2844204
-C1+C3	3980	8017	0.2658441	0.2341559
+C3+C2	3959	7965	0.3031296	0.1968704
-C1-C3	3998	8041	0.3118280	0.1881720
+C3-C2	4023	8087	0.3282329	0.1717671
+C1+C3	4002	8035	0.3689348	0.1310652
-C3+C2	3962	7949	0.3938937	0.1061063
-C1+C2	3967	7916	0.5845504	0.0845504
+C1-C2	3966	7944	0.4508892	0.0491108

Experiment Three - With All Three Characteristics

Characteristic	Heads	Trials	Probability	Distance
-C1-C2-C3	1985	4040	0.1388334	0.3611666
+C1+C2-C3	1949	3948	0.2177434	0.2822566
-C1-C2+C3	2026	4102	0.2221184	0.2778816
+C1+C2+C3	2005	4050	0.2699983	0.2300017
-C1+C2-C3	2013	4001	0.6594775	0.1594775
+C1-C2-C3	1969	3959	0.3752975	0.1247025
+C1-C2+C3	1997	3985	0.5629323	0.0629323
-C1+C2+C3	1954	3915	0.4618037	0.0381963

Pre-Registration

- The solution to this problem (which I hope the numbers came out right on!) is to require that scientists **pre-register** their hypotheses.
- So you can't collect data then see what it supports, but have to say that one particular thing is what you're testing.
- This is still in the process of becoming a universal requirement in respectable science; it was very much not part of standard scientific practice 10-20 years ago.

Philosophical Question

- Should we trust a method if it requires these ad hoc rules like announced stopping rules and pre-registration?

Philosophical Question

- Should we trust a method if it requires these ad hoc rules like announced stopping rules and pre-registration?
- Maybe! It depends on the alternatives.
- But when you see a report on a statistical finding, you should really check if it satisfies these conditions.
- And if it comes from a for-profit entity, you should be really sceptical that it does unless they are super transparent.

For Next Time

We'll start looking at modal logic, the logic of 'must' and 'might'.