# 305 Lecture 10.7 - Significance Testing

Brian Weatherson

# 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.

# Tables

In the tables that follow

- 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.

# Take One

| t | result | heads | frequency | prob | distance |
|---|---|---|---|---|---|
| 2006 | 1 | 1089 | 0.5428714 | 0.9999444 | 0.4999444 |
| 1939 | 1 | 1054 | 0.5435792 | 0.9999440 | 0.4999440 |
| 2008 | 1 | 1090 | 0.5428287 | 0.9999440 | 0.4999440 |
| 2012 | 1 | 1092 | 0.5427435 | 0.9999431 | 0.4999431 |
| 1968 | 1 | 1069 | 0.5431911 | 0.9999426 | 0.4999426 |

Number of heads

```
## [1] 50121
```

# Take Two

| t | result | heads | frequency | prob | distance |
|---|---|---|---|---|---|
| 55157 | 1 | 27775 | 0.5035626 | 0.9532907 | 0.4532907 |
| 55211 | 1 | 27802 | 0.5035591 | 0.9532105 | 0.4532105 |
| 55213 | 1 | 27803 | 0.5035589 | 0.9532075 | 0.4532075 |
| 55114 | 1 | 27753 | 0.5035563 | 0.9529375 | 0.4529375 |
| 55156 | 1 | 27774 | 0.5035536 | 0.9528748 | 0.4528748 |

Number of heads

```
## [1] 50228
```

# Take Three

| t | result | heads | frequency | prob | distance |
|---|---|---|---|---|---|
| 6551 | 0 | 3183 | 0.4858800 | 0.0114993 | 0.4885007 |
| 6536 | 0 | 3176 | 0.4859241 | 0.0117964 | 0.4882036 |
| 6542 | 0 | 3179 | 0.4859370 | 0.0118284 | 0.4881716 |
| 6550 | 0 | 3183 | 0.4859542 | 0.0118712 | 0.4881288 |
| 6552 | 1 | 3184 | 0.4859585 | 0.0118818 | 0.4881182 |

Number of heads

```
## [1] 50107
```

# Take Four

| t | result | heads | frequency | prob | distance |
|---|---|---|---|---|---|
| 8821 | 1 | 4476 | 0.5074255 | 0.9200580 | 0.4200580 |
| 8823 | 1 | 4477 | 0.5074238 | 0.9200343 | 0.4200343 |
| 8827 | 1 | 4479 | 0.5074204 | 0.9199870 | 0.4199870 |
| 8818 | 1 | 4474 | 0.5073713 | 0.9184998 | 0.4184998 |
| 8820 | 1 | 4475 | 0.5073696 | 0.9184759 | 0.4184759 |

Number of heads

```
## [1] 50002
```

# Take Five

| t | result | heads | frequency | prob | distance |
|---|---|---|---|---|---|
| 10822 | 1 | 5526 | 0.5106265 | 0.9868111 | 0.4868111 |
| 10824 | 1 | 5527 | 0.5106245 | 0.9868042 | 0.4868042 |
| 10821 | 1 | 5525 | 0.5105813 | 0.9864853 | 0.4864853 |
| 10823 | 0 | 5526 | 0.5105793 | 0.9864783 | 0.4864783 |
| 10825 | 0 | 5527 | 0.5105774 | 0.9864712 | 0.4864712 |

Number of heads

```
## [1] 50313
```

# 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 | 15907 | 32000 | 0.1505256 | 0.3494744 |

# Experiment One - With One Characteristic

| Characteristic | Heads | Trials | Probability | Distance |
|---|---|---|---|---|
| No C1 | 7852 | 15960 | 0.0217691 | 0.4782309 |
| No C3 | 7916 | 16048 | 0.0448297 | 0.4551703 |
| Yes C3 | 7916 | 15952 | 0.1730473 | 0.3269527 |
| No C2 | 7916 | 15938 | 0.2027865 | 0.2972135 |
| Yes C2 | 7991 | 16062 | 0.2665301 | 0.2334699 |
| Yes C1 | 8055 | 16040 | 0.7124655 | 0.2124655 |

| Characteristic | Heads | Trials | Probability | Distance |
|---|---|---|---|---|
| -C1-C2 | 3875 | 7923 | 0.0266553 | 0.4733447 |
| -C1-C3 | 3914 | 7962 | 0.0680400 | 0.4319600 |
| -C1+C3 | 3938 | 7998 | 0.0880275 | 0.4119725 |
| +C1-C3 | 4085 | 8086 | 0.8277371 | 0.3277371 |
| -C1+C2 | 3977 | 8037 | 0.1801815 | 0.3198185 |
| +C3-C2 | 3919 | 7917 | 0.1903455 | 0.3096545 |
| +C1-C2 | 4041 | 8015 | 0.7762376 | 0.2762376 |
| +C3+C2 | 3989 | 8035 | 0.2660742 | 0.2339258 |
| -C3-C2 | 3997 | 8021 | 0.3857913 | 0.1142087 |
| -C3+C2 | 4002 | 8027 | 0.4030149 | 0.0969851 |
| +C1+C3 | 3970 | 7954 | 0.4420544 | 0.0579456 |
| +C1+C2 | 4014 | 8025 | 0.5178073 | 0.0178073 |

| Characteristic | Heads | Trials | Probability | Distance |
|---|---|---|---|---|
| -C1-C2+C3 | 1927 | 3987 | 0.0182797 | 0.4817203 |
| -C1+C2-C3 | 1966 | 4026 | 0.0713598 | 0.4286402 |
| +C1+C2-C3 | 2036 | 4001 | 0.8724992 | 0.3724992 |
| +C1+C2+C3 | 1978 | 4024 | 0.1454380 | 0.3545620 |
| +C1-C2+C3 | 1992 | 3930 | 0.8098470 | 0.3098470 |
| -C1-C2-C3 | 1948 | 3936 | 0.2670929 | 0.2329071 |
| +C1-C2-C3 | 2049 | 4085 | 0.5866906 | 0.0866906 |
| -C1+C2+C3 | 2011 | 4011 | 0.5751388 | 0.0751388 |

# Experiment Two

| Characteristic | Heads | Trials | Probability | Distance |
|---|---|---|---|---|
| All | 15928 | 32000 | 0.2120311 | 0.2879689 |

# Experiment Two - With One Characteristic

| Characteristic | Heads | Trials | Probability | Distance |
|---|---|---|---|---|
| Yes C3 | 7856 | 16042 | 0.0046933 | 0.4953067 |
| No C2 | 7856 | 15964 | 0.0234836 | 0.4765164 |
| No C3 | 7856 | 15958 | 0.0262221 | 0.4737779 |
| Yes C2 | 8072 | 16036 | 0.8053126 | 0.3053126 |
| No C1 | 7962 | 16014 | 0.2409352 | 0.2590648 |
| Yes C1 | 7966 | 15986 | 0.3375410 | 0.1624590 |

| Characteristic | Heads | Trials | Probability | Distance |
|---|---|---|---|---|
| -C3-C2 | 3868 | 7986 | 0.0026637 | 0.4973363 |
| -C3+C2 | 4105 | 7972 | 0.9962854 | 0.4962854 |
| -C1-C2 | 3896 | 7926 | 0.0675970 | 0.4324030 |
| +C3+C2 | 3967 | 8064 | 0.0754237 | 0.4245763 |
| +C1-C2 | 3960 | 8038 | 0.0959441 | 0.4040559 |
| +C1+C3 | 3939 | 7948 | 0.2194773 | 0.2805227 |
| +C1+C2 | 4006 | 7948 | 0.7670273 | 0.2670273 |
| -C1+C3 | 4016 | 8094 | 0.2488779 | 0.2511221 |
| -C1+C2 | 4066 | 8088 | 0.6915916 | 0.1915916 |
| -C1-C3 | 3946 | 7920 | 0.3807975 | 0.1192025 |
| +C1-C3 | 4027 | 8038 | 0.5751942 | 0.0751942 |
| +C3-C2 | 3988 | 7978 | 0.4955337 | 0.0044663 |

| Characteristic | Heads | Trials | Probability | Distance |
|---|---|---|---|---|
| +C1+C2-C3 | 2052 | 3969 | 0.9845709 | 0.4845709 |
| -C1-C2-C3 | 1893 | 3917 | 0.0188874 | 0.4811126 |
| +C1-C2-C3 | 1975 | 4069 | 0.0321605 | 0.4678395 |
| -C1+C2-C3 | 2053 | 4003 | 0.9498940 | 0.4498940 |
| +C1+C2+C3 | 1954 | 3979 | 0.1335600 | 0.3664400 |
| -C1+C2+C3 | 2013 | 4085 | 0.1820802 | 0.3179198 |
| +C1-C2+C3 | 1985 | 3969 | 0.5126624 | 0.0126624 |
| -C1-C2+C3 | 2003 | 4009 | 0.4874009 | 0.0125991 |

# Experiment Three

| Characteristic | Heads | Trials | Probability | Distance |
|---|---|---|---|---|
| All | 15968 | 32000 | 0.3623517 | 0.1376483 |

# Experiment Three - With One Characteristic

| Characteristic | Heads | Trials | Probability | Distance |
|---|---|---|---|---|
| Yes C3 | 8110 | 15857 | 0.9980783 | 0.4980783 |
| Yes C2 | 7858 | 15923 | 0.0512850 | 0.4487150 |
| No C2 | 8110 | 16077 | 0.8719573 | 0.3719573 |
| Yes C1 | 7929 | 15995 | 0.1411105 | 0.3588895 |
| No C3 | 8110 | 16143 | 0.7303609 | 0.2303609 |
| No C1 | 8039 | 16005 | 0.7207019 | 0.2207019 |

| Characteristic | Heads | Trials | Probability | Distance |
|---|---|---|---|---|
| +C1+C2 | 3854 | 7857 | 0.0474882 | 0.4525118 |
| -C1-C2 | 4035 | 7939 | 0.9307610 | 0.4307610 |
| +C3+C2 | 3906 | 7940 | 0.0770391 | 0.4229609 |
| +C1+C3 | 3914 | 7918 | 0.1586094 | 0.3413906 |
| -C3-C2 | 4120 | 8160 | 0.8150560 | 0.3150560 |
| -C3+C2 | 3952 | 7983 | 0.1913336 | 0.3086664 |
| +C3-C2 | 3990 | 7917 | 0.7640155 | 0.2640155 |
| -C1+C2 | 4004 | 8066 | 0.2628242 | 0.2371758 |
| -C1-C3 | 4057 | 8066 | 0.7073250 | 0.2073250 |
| +C1-C3 | 4015 | 8077 | 0.3043835 | 0.1956165 |
| -C1+C3 | 3982 | 7939 | 0.6147805 | 0.1147805 |
| +C1-C2 | 4075 | 8138 | 0.5572914 | 0.0572914 |

| Characteristic | Heads | Trials | Probability | Distance |
|---|---|---|---|---|
| +C1+C2+C3 | 1903 | 3907 | 0.0548104 | 0.4451896 |
| -C1-C2-C3 | 2056 | 4033 | 0.8961185 | 0.3961185 |
| -C1-C2+C3 | 1979 | 3906 | 0.8017872 | 0.3017872 |
| +C1+C2-C3 | 1951 | 3950 | 0.2272853 | 0.2727147 |
| -C1+C2-C3 | 2001 | 4033 | 0.3183244 | 0.1816756 |
| -C1+C2+C3 | 2003 | 4033 | 0.3411217 | 0.1588783 |
| +C1-C2+C3 | 2011 | 4011 | 0.5751388 | 0.0751388 |
| +C1-C2-C3 | 2064 | 4127 | 0.5124178 | 0.0124178 |

# 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'.