BIOMED SCI 552:

STATISTICAL THINKING

LECTURE 3: PROBABILITY PART 1

QUESTIONS FROM THURSDAY?

PROBLEM SET DISCUSSION

- GitHub Organizations and Permissions
- Conflicts

WHAT ARE YOUR INTERESTS?

- Answers for this ranged from “Hiking” to “Antibody dependent effector functions”
    - And *neither* is wrong
- In the real world, data collection is messy, questions are ambiguous, or answers can come from a different cultural context to your own
- This will almost *always* happen

SOME REAL EXAMPLES

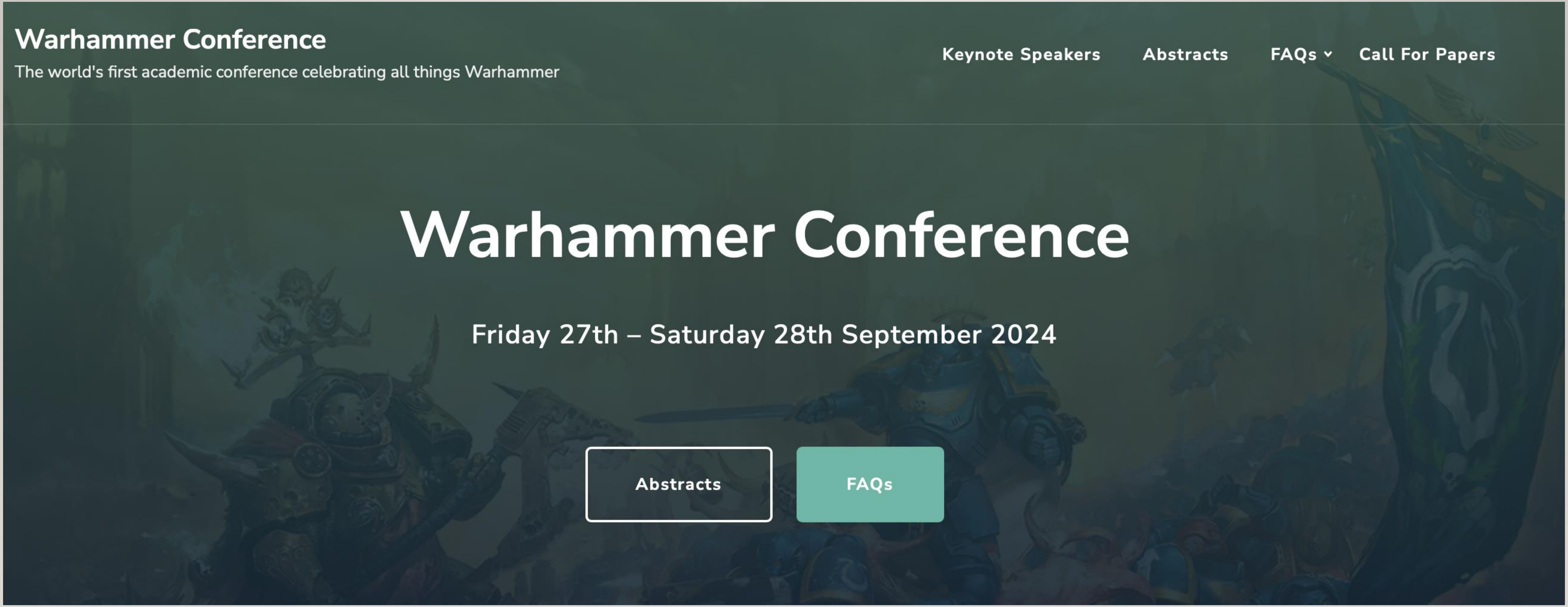
- How many hurricanes have you chosen not to evacuate for?
- How many people live in your household?
- Are you a commercial sex worker?
- Do you eat meat?

- It is also often hard to clean data with these answers
- One can *sometimes* use validation questions
  - Ask the same question in a different way
  - Ask questions which can be mathematically checked
  - Ask questions of someone else

DISCUSSION

- Question: "What are your interests?"
- Answer: "Warhammer."

- Is this a "Personal Hobby" type question, or an "Academic Interest" type question?



Warhammer Conference

The world's first academic conference celebrating all things Warhammer

Warhammer Conference

Friday 27th – Saturday 28th September 2024

PSYCHOLOGY STUDENTS AS A CONVENIENCE SAMPLE

- One example:
- Psychology students at Washington State University are taken as a sample of all college students in the United States
  - This is far from the most egregious case
- How does WSU differ?

	WSU	National
Percentage Women	53%	58%
Percentage White	60%	49%
Percentage Black/African American	3%	15%
Percentage NI/PI	2%	0%
Percentage bachelor’s degrees in Psychology	8.7%	6%
Percentage bachelor’s degrees in Mechanical Engineering	4.9%	1.5%
Percentage First Generation	35.5%	56%

*These are all approximate

BANFIELD QUESTION

- A researcher comes to you, wanting to do a study in dogs in Banfield pet clinics. They’re planning on using a consolidated electronic health record (EHR) from close to 1,000 clinics, looking at the risk factors for the development of parathyroid conditions in elderly dogs, by using clinical diagnoses and the dogs clinical records.
- Is this data prospective or retrospective, and why? Is this data passive or actively collected, and why? Is this data “Big Data”, and why?
- What follows are *my* answers, and may not be the only right answers

PROSPECTIVE OR RETROSPECTIVE

- Arguably, I didn't give you enough information to answer this question.
- Will they use the EHR system to enroll elderly dogs, and follow them over time to see if they develop parathyroid conditions?
    - That's prospective
- *Most* EHR studies are retrospective however, so it's not unreasonable to assume that what they're doing is looking for dogs in the system with and without parathyroid conditions

ACTIVE VS. PASSIVE

- I'd argue that EHR data is passively collected
    - There's no involvement of the participant
    - It's being collected anyway via clinical care, which would continue whether this study continues or not
    - Some of the data may truly be passively collected, obtained from pharmacy records, automatic monitoring, etc.

BIG DATA?

- EHR data exists in a fuzzy grey area between "Big Data" and "Data That Which Is Large", in my opinion
  - It tends to be somewhat more purposeful than a lot of "Big Data" examples
  - But the researcher who assumes that an EHR will be easy pickings for their research project does so at their peril
- It certainly meets many of the qualitative definitions of big data
  - You're not going to be able to just open it up on your laptop and have a look
  - Potentially lots of rows and columns, not all of which pertain to your question

QUESTIONS?

WHAT IS PROBABILITY?

- As a field: Probability is the branch of mathematics that concerns an event, and the chance it occurs
- The probability of an event is bounded by 0 (impossible) and 1 (certain)
- This means when we think about probability, we have to decide how to describe our event. Multiple event descriptions can have very different probabilities
    - The probability you pass this class
    - The probability you pass this class with an B
    - The probability you pass this class with an 87.5% grade or higher
    - The probability you pass this class with an 87.5% grade

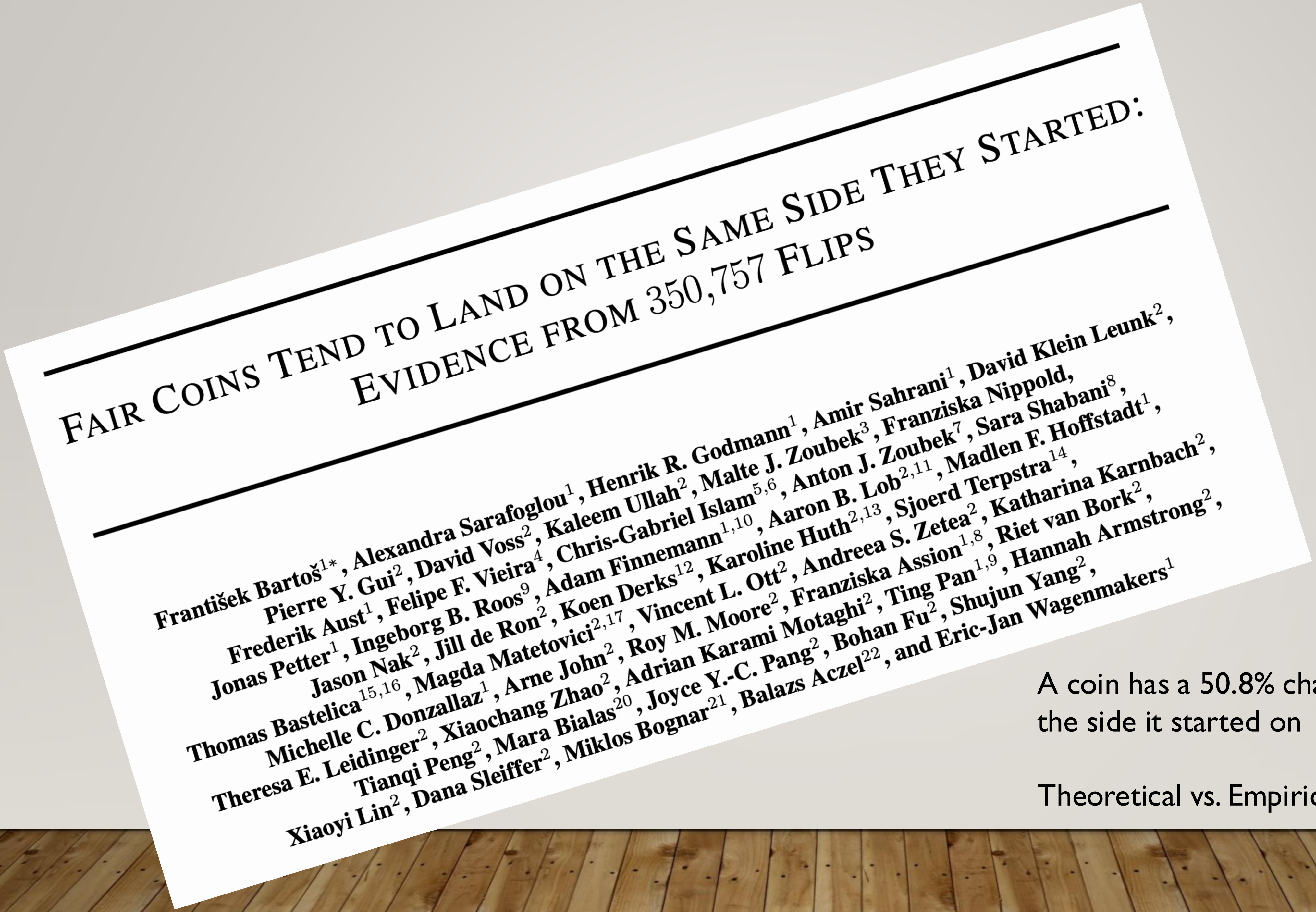
VERY BASIC PROBABILISTIC CALCULATIONS

- We can calculate the probability of an event occurring out of a number of discrete, equally probable events, by adding up the total number of events we're interested in (in many examples, these are called "successes") by the total number of events
- For example, in a deck of 52 cards, there are four queens, so the probability of drawing a queen is 4/52 = 0.0769

PARTICIPATION TIME!

- What's the probability of rolling a 1 on a six-sided die?
- What's the probability of rolling a 20 on a twenty-sided die?
- What's the probability of getting heads on a coin toss?

WHAT HAVE YOU BEEN ASSUMING?

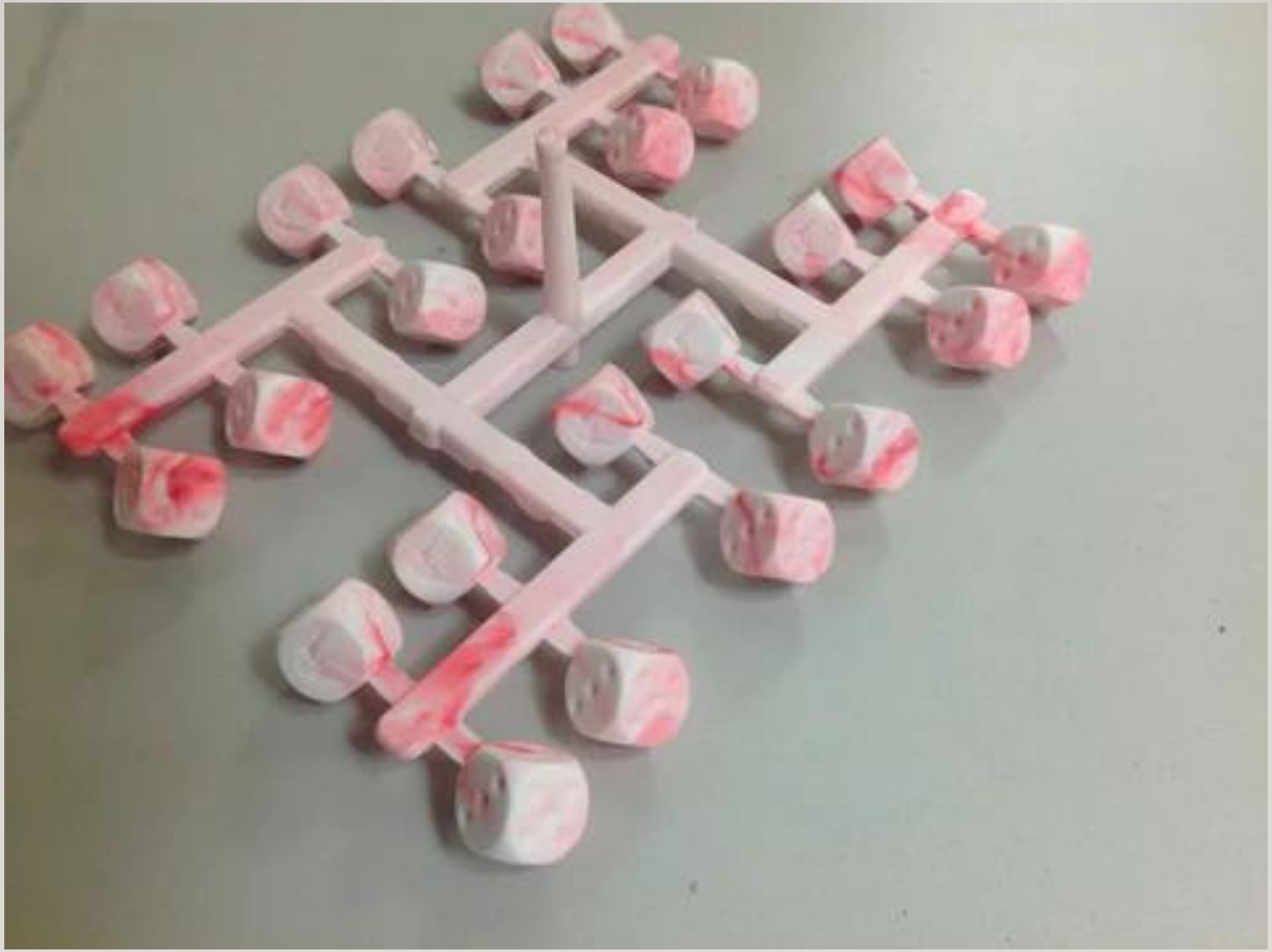


Fair Coins Tend to Land on the Same Side They Started: Evidence from 350,757 Flips

František Bartoš[1*], Alexandra Sarafoglou[1], Henrik R. Godmann[1], Amir Sahrani[1], David Klein Leunk[2], Pierre Y. Gui[2], David Voss[2], Kaleem Ullah[2], Malte J. Zoubek[3], Franziska Nippold, Frederik Aust[1], Felipe F. Vieira[4], Chris-Gabriel Islam[5,6], Anton J. Zoubek[7], Sara Shabani[8], Jonas Petter[1], Ingeborg B. Roos[9], Adam Finnemann[1,10], Aaron B. Lob[2,11], Madlen F. Hoffstadt[1], Jason Nak[2], Jill de Ron[2], Koen Derks[12], Karoline Huth[2,13], Sjoerd Terpstra[14], Thomas Bastelica[15,16], Magda Matetovici[2,17], Vincent L. Ott[2], Andreea S. Zetea[2], Katharina Karnbach[2], Michelle C. Donzallaz[1], Arne John[2], Roy M. Moore[2], Franziska Assion[1,8], Riet van Bork[2], Theresa E. Leidinger[2], Xiaochang Zhao[2], Adrian Karami Motaghi[2], Ting Pan[1,9], Hannah Armstrong[2], Tianqi Peng[2], Mara Bialas[20], Joyce Y.-C. Pang[2], Bohan Fu[2], Shujun Yang[2], Xiaoyi Lin[2], Dana Sleiffer[2], Miklos Bognar[21], Balazs Aczel[22], and Eric-Jan Wagenmakers[1]

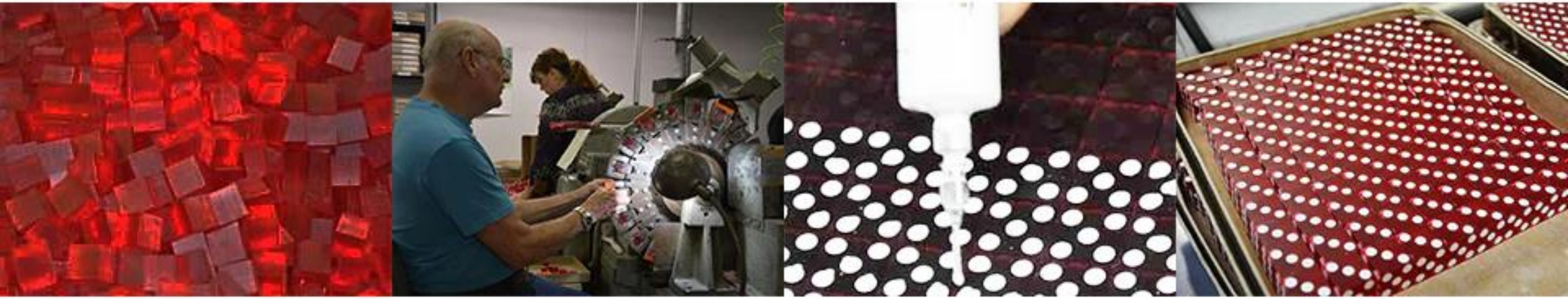
A coin has a 50.8% chance of landing on the side it started on

Theoretical vs. Empirical Probability



How Precision Casino Dice Are Made
by GMDICE.COM

Images from Midwest Game Supply Co.,
Kearney, MO



Step 1: Plastic is rough cut into cubes.

Step 2: Dice are machined to precise specifications.

Step 3: Epoxy is added with the same density as the dice for pips.

Step 4: Dice are packaged and shipped from the warehouse.

WEIGHTED PROBABILITY

- Previously, we’ve assumed all events are equally probable
- But what if they’re not?
- We can calculate what’s called a weighted probability, where the difference in the events is taken into account

EXAMPLE

- Randomly drawing a blue M&M out of a bag
- Red: 7
- Orange: 4
- Yellow: 9
- Green: 6
- Blue: 7
- Brown: 5
- Total: 38
- It's not 1/6th

- Instead, we calculate the *ways* you can draw a Blue M&M (there are 7 such ways) by the total number of possible events (38)

- 7/38 = 0.184

PATIENT Z (FOR ZEBRA)

- Pretend for a moment you’re a clinician, and you have a patient who has tested positive for a 1-in-10,000 condition. You also know the test has a 5% false positive rate. The test does not produce false negatives.

- What are the chances your patient has the condition?

- In 10,000 people:
  - There is 1 person with the disease
  - There have been 9,999*0.05 ≈ 500 people with false positives
- 1/501 = 0.001996008
- The probability your patient has the condition is actually quite low, even with a positive test
- This is called the "Base Rate Fallacy"

THINKING STOCHASTICALLY

- ”Stochastic” – just a fancy word for random
- Our world is just a single stochastic realization of staggering array of probabilistic outcomes
- It is often helpful (at least to me) to think about your studies this way – are there alternate realities where things turned out differently, and what does that mean?
- This occurs *a lot* in simulation studies, modeling, etc. but also happens in the real world
- If I say each of you has a 10% chance of walking out of this room, that *does not mean* that 1/10th of you will necessarily leave

ANOTHER EXAMPLE

- In the 2016 US presidential election, Hilary Clinton was forecast to have a 66% chance to win the election
  - Obviously, she lost
  - *Was the forecast wrong?*

STATISTICAL INDEPENDENCE

- Right now, we've been assuming that all of our examples are *independent*
- The probability of one event occurring is unrelated the the probability of another occurring
- Two six-sided dice don't talk to each other and decide to roll high
- There are *many* ways to violate statistical independence, some fairly extreme, some perhaps mild enough to be ignorable

EXTREME EXAMPLES

- If you draw two samples from the same…
  - Person (i.e. longitudinal measures of something, or measurements from both eyes, lungs, etc.)
  - Twins
  - Family
  - Herd
  - Household
  - Hospital Ward
  - And so forth
- …you are probably violating statistical independence.

INDEPENDENT UNITS OF REPLICATION

- Observational units are the independent unit to which measurements have been made, or experimental treatments assigned.
- In a perfect world, this would be individuals, but it often isn’t
- Within “units” that are treated the same way in some form, measurements *within* that unit are not independent
- For example, microbes on a petri dish are all on the same medium, got exposed to the same temperatures, etc. Cows within the same herd have been cared for by the same people, fed the same food, etc.

WHY THIS MATTERS

- You don’t have the information you think you do
  - You likely have *lower* uncertainty than you should
- An analysis that assumes that your samples are independent when they aren’t is likely to overestimate the precision of whatever you’re trying to estimate, and lead you astray
- In ecology, this is often referred to as pseudoreplication
  - i.e. you think you’re replicating your result, but what you’re actually doing is effectively taking multiple measurements of the same (or a very similar) sample

MORE MILD EXAMPLES

- The same *hospital system*
- The same major city
- Much of what we think about in statistics is not actually black-and-white, but is trying to understand the degree to which something is a problem, and how much our answer might be influenced by violations of our assumptions
- It's possible that we can make an assumption, be wrong, and it still won't matter
- There is a whole branch of analysis called sensitivity analysis that deals with this

DISCUSSION

- This lecture was written on a Boeing 737-800 while en route from Washington DC to Seattle, WA
- A very strange researcher wants to take two samples from this flight: One from myself, and one from the woman sitting next to me
- Are these samples independent?
  - And why?
- Are these samples biased?
  - And why?

CONDITIONAL PROBABILITY

- We tend to write probability as P(Thing) = 0.283
  - This is called the marginal probability – the probability of an event without reference to other variables.
- But what if we have more information?
- Or want to restrict probability in some way?
- This is called *conditional* probability, and is written as P(Thing | Other Thing)
- We've already used conditional probability once…
  - P (Disease | Test = Positive)

CONDITIONAL PROBABILITY

- Conditional probability is *all over* biomedical research
  - Marginal probabilities are still useful – they're used a lot in modeling
- Some very common ones:
  - P(Disease | Treatment = Yes)
  - P(Outcome | Demographics)

WHY IS CONDITIONAL PROBABILITY USEFUL

- You can look at specific strata (i.e. P(Disease | Sex = Female))
- You can *compare* specific strata
  - P (Disease | Exposed = 1) vs. P(Disease | Exposed = 0)
  - If you divide these, you have a relative risk
  - If you subtract them, you have a risk difference
- This is (usually) what is meant when someone says they control for other variables
  - P(Disease | Exposed =1, Age, Race, Sex, SES, etc.)
  - A comparison between strata here is a comparison holding everything else constant

WHAT IS THE CONDITIONAL PROBABILITY FOR AN RCT?

IT…DEPENDS

- “Intent to Treat” analyzes all individuals as allocated after randomization
  - P(Outcome | **Randomization**)
- ”Per Protocol” analyzes only those individuals who adhered to the study protocol
  - P(Outcome | Treatment)

- Which would you prefer, and why?

MORE PARTICIPATION

- Pregnancies resulting in an infant with Down Syndrome are *much* more common in older mothers
- $\frac{P(Down\ Syndrome\ |Mother's\ Age<30)}{P(Down\ Syndrome\ |Mother's\ Age>40} = 17$
- Are most infants with Down Syndrome born to older mothers?

THE PROSECUTOR'S FALLACY

- 51% of babies with Down Syndrome are born to young mothers, because *most mothers are young*
- This is called the Prosecutor's Fallacy, and is mistaking the probability of A | B with the probability of B | A
  - "The probability of someone owning the type of gun used in this murder is 0.33%. The suspect owned this type of gun. Therefore, the suspect is very likely guilty!"
    - In Pullman, there are 107 people who own that gun
    - In New York, there are 27,500 people
- We already saw an example of this earlier, without using conditional probabilities, in our rare disease example

BAYES' THEOREM

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

*maybe

BAYES' THEOREM

- P(A|B): The "Posterior" probability – usually what we're interested in
- P(B|A): The "Likelihood" – this is often rather hard to calculate
- P(A): The "Prior" – what we already know about A
- P(B): The "Marginalization" – the probability of B being true

$$P(Disease|Test = Positive) = \frac{P(Test = Positive|Disease)P(Disease)}{P(Test = Positive)}$$

$$P(Disease|Test = Positive) = \frac{1 * 0.0001}{\frac{(9999 * 0.05) + 1}{10{,}000}} = 0.001996207$$

THE PRIOR

- The prior reflects our understanding of the probability of an event occurring *before* we analyze our data
- If we say we know nothing, this is called an "uninformative" prior
- Priors *need not be uninformative*
- There is often a feeling that an uninformative prior is more rigorous, but it is often a very strong assumption – namely, that we know nothing

EXAMPLE

- Returning to our example, lets assume now you're a specialist in the field who is a known expert in this disease, the patient has been referred to you by a colleague, and the patient's paternal grandmother also had the condition
- Given all of this – but considering the rarity of the disease, you think that the probability the patient has the disease before you give them a test is around 12%
- You then give the patient the test, and it's positive
- What's the posterior probability that the patient has the disease?

$$P(Disease|Test = Positive) = \frac{P(Test = Positive|Disease)P(Disease)}{P(Test = Positive)}$$

$$P(Disease|Test = Positive) = \frac{1 * \mathbf{0.12}}{\frac{(9999 * 0.05) + 1}{10{,}000}} = 2.395449$$

Why Didn’t This Work?

$$P(Disease|Test = Positive) = \frac{P(Test = Positive|Disease)P(Disease)}{P(Test = Positive)}$$

$$P(Disease|Test = Positive) = \frac{1 * \mathbf{0.12}}{\frac{\mathbf{(8800 * 0.05) + 1200}}{10{,}000}} = 0.732$$