Lecture 9: Understanding randomness

Criminology 250

Prof Maria Cuellar

University of Pennslyvania

Understanding randomness

"The most decisive conceptual event of twentieth century physics has been the discovery that the world is not deterministic... A space was cleared for chance." - Ian Hocking, The Taming of Chance.

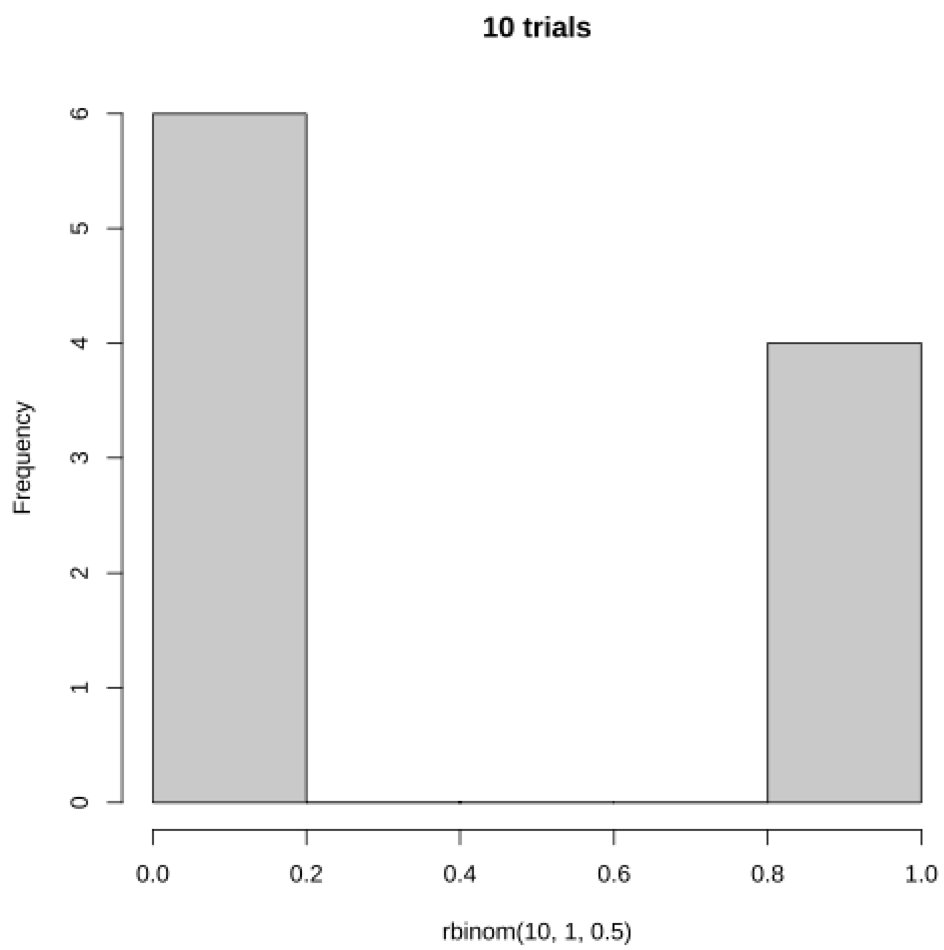
What is randomness?

- Nobody can guess the outcome before it happens.
- When we want things to be fair, usually some underlying set of outcomes will be equally likely.
- You can't predict how a fair coin will land on any single toss, but you're pretty confident that if you flipped it thousands of times you'd see about 50% heads.

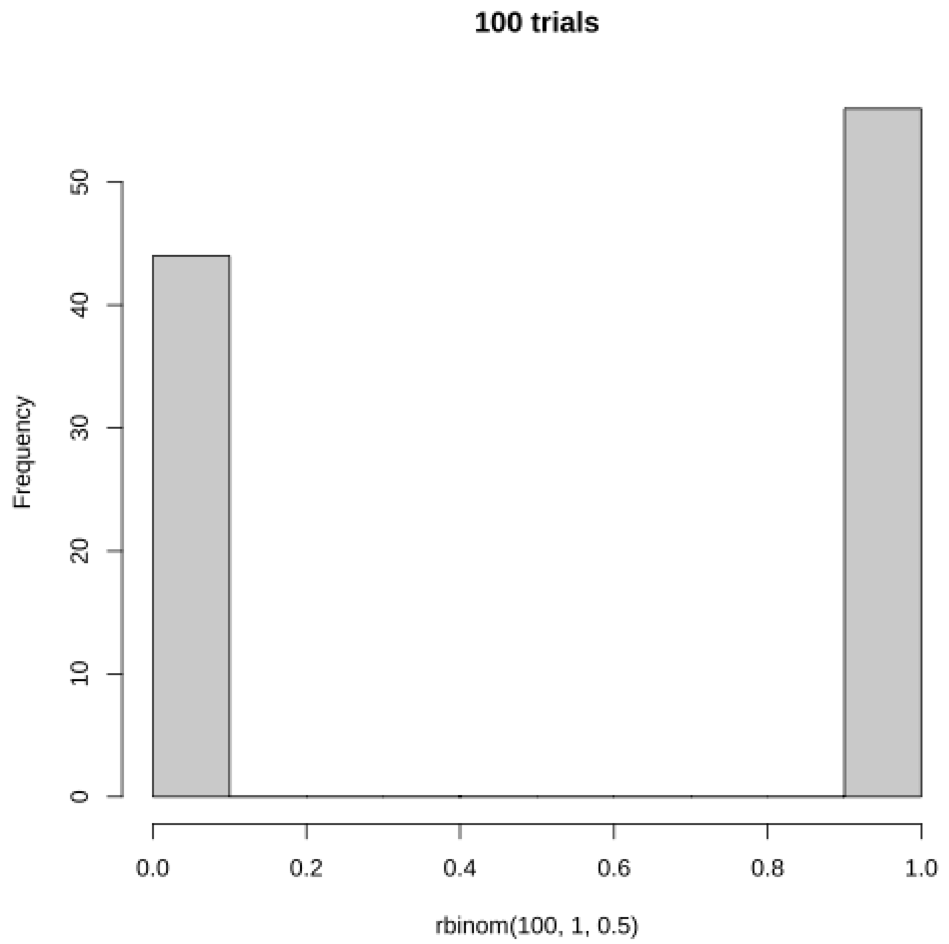
Randomly generate flips of a coin

- It is not feasible to generate truly random numbers using a computer, so R generates numbers that appear random (and have the qualities of randomness that we care about in statistics): pseudorandom.
- R use a variety of pseudorandom number generators.
- e.g., `rbinom(400, size = 10, prob = 0.2)` gives the results of 400 runs of 10 coin flips each, returning the number of successes in each run.
- You can also find a truly random number by going online: https://www.random.org/integers/ . This site generates numbers using methods from nature like listening to atmospheric noise or timing the decay of a radioactive element.

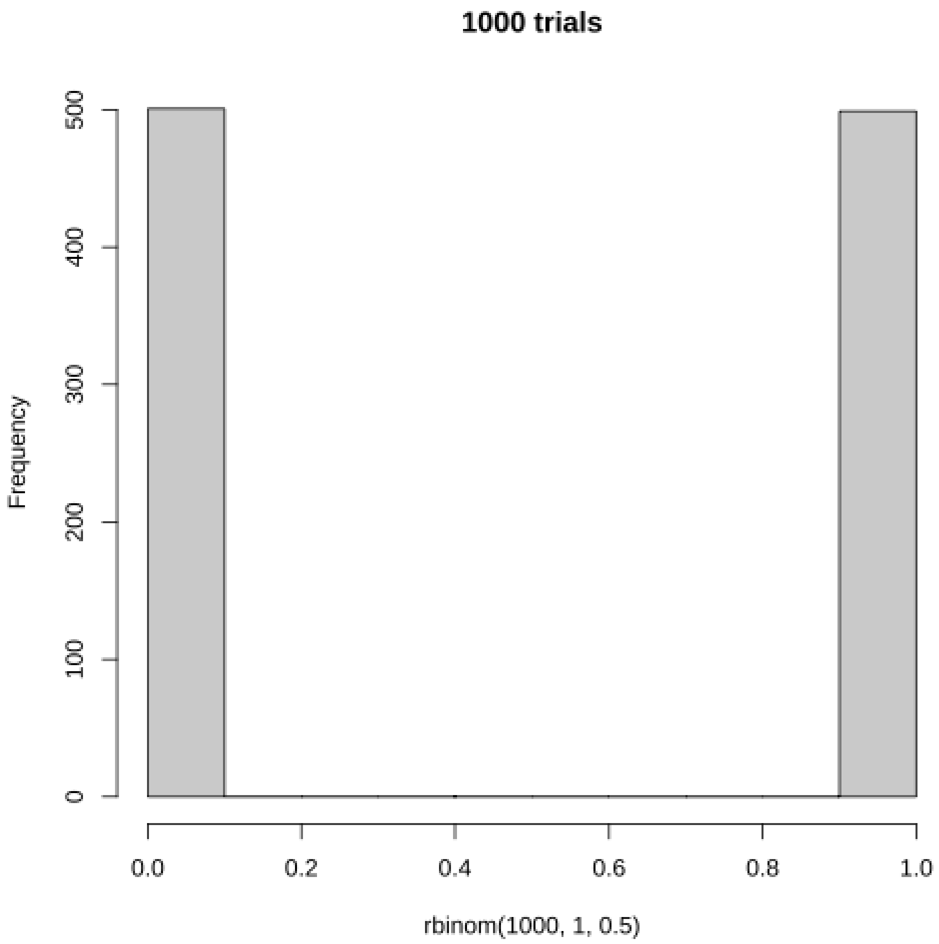
Histogram of R-generated (fair) coin flips



Histogram of R-generated (fair) coin flips



Histogram of R-generated (fair) coin flips



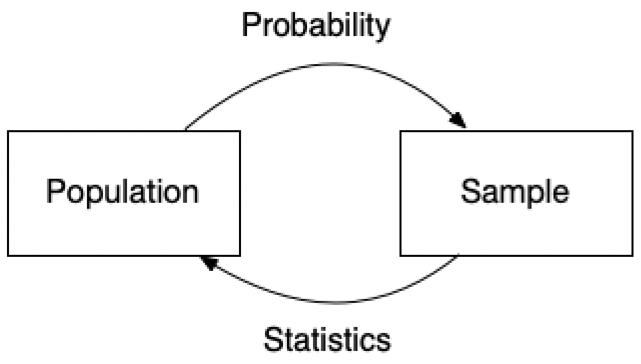
Probability

Probability allows us to describe how likely something is to happen.

How are statistics and probability related?

- Suppose I know the exact distribution of car purchases in Pennsylvania. Then I can find the probability that the first car I see on the road is a Ford. This is probabilistic reasoning since I know the population and predict a sample.

- Suppose I do not know the distribution, but I would like to estimate it. I can observe a random sample of cars in the street and use it to estimate the proportions of the population. This is statistical inference.



What is a random phenomenon?

- A random phenomenon is a situation in which we know what outcomes can occur, but we don't know which particular outcome we'll get.
- Each time we observe the random phenomenon is called a trial, and each trial has an outcome.
- Outcomes combine to make events. We call the collection of all possible outcomes the sample space.

Law of large numbers

A law that describes the result of performing the same experiment a large number of times.

Definition - Law of Large Numbers: The long-term relative frequency of an event's occurance gets closer and closer to the relative frequency as the number of trials increases.

This figure shows how the average dice roll changes as we have more rolls. As you roll the die more times, the average of the values of all results approaches 3.5 (Expected average for fair die: (1+2+3+$+5+6)/6 = 3.5).

Definition - Empirical probability: For any event A,

$$P(A) = \frac{\text{Number of times } A \text{ occurs}}{\text{Total number of trials}}.$$

1. Empirical probability example

This table shows the distribution of whether students voted in the last presidential election.

	No, eligible but didn't	No, not eligible	Yes	**Total**
First year	3	38	3	44
Sophomore	10	40	14	64
Junior	7	6	41	54
Senior	4	1	9	14
Total	24	85	67	176

- Exercise 1.1: What is the probability that a randomly chosen student has voted in the last presidential election? Want: P(voted).
- Answer: 67/176.
- Exercise 1.2: What is the probability that a randomly chosen student is a junior and voted? Want: P(junior and voted)
- Answer: 41/176

	No, eligible but didn't	No, not eligible	Yes	**Total**
First year	3	38	3	44
Sophomore	10	40	14	64
Junior	7	6	41	54
Senior	4	1	9	14
Total	24	85	67	176

Exercise 1.3: What is the probability that a randomly chosen student voted given that they are a junior?

- Answer:41/54
- Exercise 1.4: What is the probability that a randomly chosen student is a junior or has voted in the last presidential election?
- Answer: 80/176
- Exercise 1.5: Do these data suggest an association between class year and whether students have voted in the last presidential election?
- Answer: Yes. P(voted) $\neq$ P(voted | junior).

Formal probability theory

Axioms of probability

1. A probability is a number between 0 and 1: $0 \leq P(A) \leq 1$ for any event A.
2. The set of all possible outcomes must have probability 1: $P(S) = 1$ where S is the set of all possible outcomes.
3. The probability of an event occurring is 1 minus the probability that it doesn't occur: $P(A) = 1 - P(A^c)$, where C denotes the complement, or "not A".

Formal probability theory

Definition - Disjoint events: If A and B are disjoint events, they cannot both occur. They are mutually exclusive.

If A and B are disjoint events, then $P(A \text{ or } B) = P(A) + P(B)$.

Definition - Independent events: Two events are independent if learning that one event happened does not change the probability of the other event.

IF A and B are independent events, then $P(A \text{ and } B) = P(A) \times P(B)$.

Formal probability theory

Example: Let A = {You get an A in this class} and B = {You get a B in this class}. Are these events disjoint? Are they independent?

Yes, they are disjoint. They cannot both happen.

They are not independent. If we know that A is the case, then we know $P(B) = 0$.

Conditional probability

Sometimes we know something about what actually has happened, and we want to know the probability of a different event conditional on what we already know. This is called conditional probability.

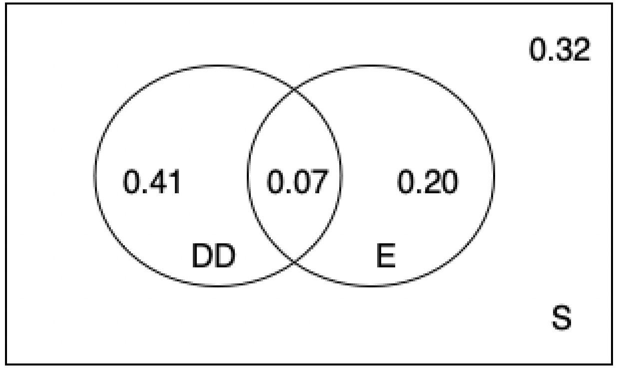
Example: In a sample of pages from a statistics textbook, 48% pof pages had some sort of data display, 27% of pages had an equation, and 7% of the pages had both a datat display and an equation.

(a) Display these results as a Venn diagram.

(b) What is the probability that a randomly selected sample page had neither a data display nor an equation?

(c) What is the probability that a randomly selected sample page had a data display but no equation?

Answers

(a) 

(b) P(DD or E) = 0.41 + 0.07 + 0.20 = 0.68.

P(neither DD or E) = 1-0.68 = 0.32.

(c) P(DD and E^c) = 0.41.

Given that a randomly selected page had data