Probability 2: Conditional Probability

Howard Liu

2024-10-10

Roadmap

1. Conditional Probability
2. Bayes's Rule
3. Independence

1. Conditional Probability

Conditional probability

- **Conditional probability**: if we know that B has occurred, what is the probability of A?
- Conditioning our analysis on B having occurred.

Conditional probability

- **Conditional probability**: if we know that B has occurred, what is the probability of A?
- Conditioning our analysis on B having occurred.
- Examples:
  - What is probability of two states going to war if they are both democracies?
  - What is the probability of a judge ruling in a pro-choice direction **conditional** on having daughters?
  - What is the probability that there will be a coup in a country **conditional** on having a presidential system?
- Conditional probability is the cornerstone of quantitative social science.

Conditional Probability definition

- Definition: If $\mathbb{P}(B) > 0$ then the **conditional probability** of A given B, $\mathbb{P}(A|B)$ is ?

Conditional Probability definition

- Definition: If $\mathbb{P}(B) > 0$ then the **conditional probability** of A given B, $\mathbb{P}(A|B)$ is ?

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

- How often both A and B occur divided by how often B occurs.

Conditional Probability definition

- Definition: If $\mathbb{P}(B) > 0$ then the **conditional probability** of A given B, $\mathbb{P}(A|B)$ is ?

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

- How often both A and B occur divided by how often B occurs.
- **WARNING!** $\mathbb{P}(A|B)$ does not, in general, equal $\mathbb{P}(B|A)$.
  - $\mathbb{P}$(smart | in POLI 502) is high
  - $\mathbb{P}$(in POLI 502 | smart) is low

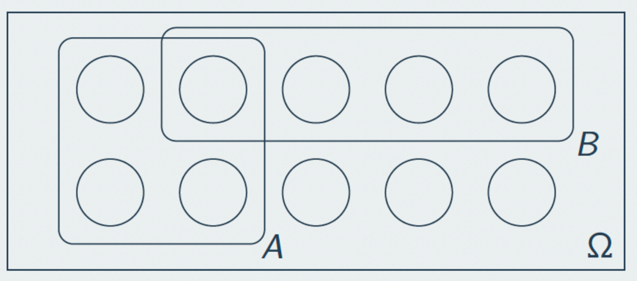
Conditional Probability definition

- Definition: If $\mathbb{P}(B) > 0$ then the **conditional probability** of A given B, $\mathbb{P}(A|B)$ is ?

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

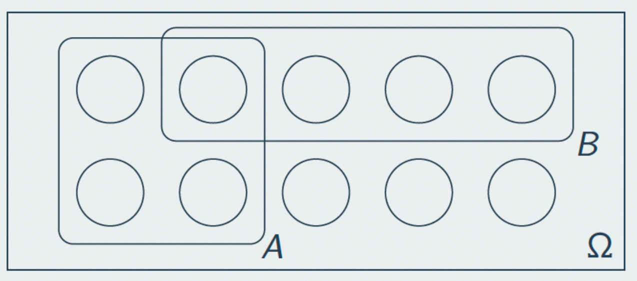
- How often both A and B occur divided by how often B occurs.
- **WARNING!** $\mathbb{P}(A|B)$ does not, in general, equal $\mathbb{P}(B|A)$.
  - $\mathbb{P}$(smart | in POLI 502) is high
  - $\mathbb{P}$(in POLI 502 | smart) is low
  - There are many many smart people out there who are not in this class (perhaps in other universities or other departments or on the street)!
  - Also known as the **prosecutor's fallacy**
  - $\mathbb{P}$(innocent | evidence) is not the same as $\mathbb{P}$(evidence | innocent)

Examples



- A $=$ {you get an A grade}, B $=$ {everyone gets an A grade}
- If B occurs then A must also occur, so $Pr(\text{A}|\text{B}) = 1$.
- Does this mean that $Pr(\text{B}|\text{A}) = 1$ as well?

Examples



- A $=$ {you get an A grade}, B $=$ {everyone gets an A grade}
- If B occurs then A must also occur, so $Pr(\text{A}|\text{B}) = 1$.
- Does this mean that $Pr(\text{B}|\text{A}) = 1$ as well?
- Now let A $=$ {you get a B grade}.
- The intersection $\text{A} \cap \text{B} = \emptyset$, so that $Pr(\text{A}|\text{B}) = 0$.
- Intuitively, it's because B occurring precludes A from occurring.

Conditional probabilities are probabilities

- Condition probabilities $\mathbb{P}(A|B)$ are valid probability functions:
  1. $\mathbb{P}(A|B) \geq 0$
  2. $\mathbb{P}(\Omega|B) = 1$
  3. Addition/Partition rule: If A_1 and A_2 are disjoint, then $\mathbb{P}(A_1 \cup A_2|B) = \mathbb{P}(A_1|B) + \mathbb{P}(A_2|B)$
- $\rightsquigarrow$ rules of probability apply to left-hand side of conditioning bar (A)
- All probabilities normalized to event B, $\mathbb{P}(B \mid B) = 1$.
- Not for right-hand side, so even if B and C are disjoint, $\mathbb{P}(A|B \cup C) \neq \mathbb{P}(A|B) + \mathbb{P}(A|C)$

Joint probabilities from conditional probabilities

- Joint probabilities: probability of two events occurring (intersections)
- Often replace $\cap$ with commas: $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A, B, C)$

Joint probabilities from conditional probabilities

- Joint probabilities: probability of two events occurring (intersections)
- Often replace $\cap$ with commas: $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A, B, C)$
- Definition of conditional prob. implies:
  $\mathbb{P}(A \cap B) \equiv \mathbb{P}(A, B) = \mathbb{P}(B)\mathbb{P}(A \mid B) = \mathbb{P}(A)\mathbb{P}(B \mid A)$
- Look at what we defined before for cond. prob., we're just "moving parts" in the same equation!

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

Joint probabilities from conditional probabilities

- Joint probabilities: probability of two events occurring (intersections)
- Often replace ∩ with commas: $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A, B, C)$
- Definition of conditional prob. implies:
  $\mathbb{P}(A \cap B) \equiv \mathbb{P}(A, B) = \mathbb{P}(B)\mathbb{P}(A \mid B) = \mathbb{P}(A)\mathbb{P}(B \mid A)$
- Look at what we defined before for cond. prob., we're just "moving parts" in the same equation!

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

- Let's draw a Venn diagram?
- What about three events? $\mathbb{P}(A, B, C) = \mathbb{P}(A)\mathbb{P}(B \mid A)\mathbb{P}(C \mid A, B)$
- Generalize to the intersection of N events: $\mathbb{P}(A_1, ..., A_N) = \mathbb{P}(A_1)\mathbb{P}(A_2 \mid A_1)\mathbb{P}(A_3 \mid A_1, A_2) \cdots \mathbb{P}(A_N \mid A_1, ..., A_{N-1})$

Joint probabilities from conditionals, example

- Draw three cards at random from a deck **without** replacement
- What's the probability that we draw three Aces (using cond. prob.)?

Joint probabilities from conditionals, example

- Draw three cards at random from a deck **without** replacement
- What's the probability that we draw three Aces (using cond. prob.)?

$$\mathbb{P}(Ace1 \cap Ace2 \cap Ace3) = \mathbb{P}(Ace1)\mathbb{P}(Ace2 \mid Ace1)\mathbb{P}(Ace3 \mid Ace1 \cap Ace2)$$

- What are these probabilities?
  - 4 Aces to pick out of 52 cards

Joint probabilities from conditionals, example

- Draw three cards at random from a deck **without** replacement
- What's the probability that we draw three Aces (using cond. prob.)?

$$\mathbb{P}(Ace1 \cap Ace2 \cap Ace3) = \mathbb{P}(Ace1)\mathbb{P}(Ace2 \mid Ace1)\mathbb{P}(Ace3 \mid Ace1 \cap Ace2)$$

- What are these probabilities?
  - 4 Aces to pick out of 52 cards $\rightsquigarrow \mathbb{P}(Ace_1) = \frac{4}{52}$
  - 3 Aces left in the 51 remaining cards

Joint probabilities from conditionals, example

- Draw three cards at random from a deck **without** replacement
- What's the probability that we draw three Aces (using cond. prob.)?

$$\mathbb{P}(Ace1 \cap Ace2 \cap Ace3) = \mathbb{P}(Ace1)\mathbb{P}(Ace2 \mid Ace1)\mathbb{P}(Ace3 \mid Ace1 \cap Ace2)$$

- What are these probabilities?
  - 4 Aces to pick out of 52 cards $\rightsquigarrow \mathbb{P}(Ace_1) = \frac{4}{52}$
  - 3 Aces left in the 51 remaining cards $\rightsquigarrow \mathbb{P}(Ace_2 | Ace_1) = \frac{3}{52}$
  - 2 Aces left in the 50 remaining cards

Joint probabilities from conditionals, example

- Draw three cards at random from a deck **without** replacement
- What's the probability that we draw three Aces (using cond. prob.)?

$$\mathbb{P}(Ace1 \cap Ace2 \cap Ace3) = \mathbb{P}(Ace1)\mathbb{P}(Ace2 \mid Ace1)\mathbb{P}(Ace3 \mid Ace1 \cap Ace2)$$

- What are these probabilities?
  - 4 Aces to pick out of 52 cards $\rightsquigarrow \mathbb{P}(Ace_1) = \frac{4}{52}$
  - 3 Aces left in the 51 remaining cards $\rightsquigarrow \mathbb{P}(Ace_2|Ace_1) = \frac{3}{52}$
  - 2 Aces left in the 50 remaining cards $\rightsquigarrow \mathbb{P}(Ace_3|Ace_2 \cap Ace_1) = \frac{2}{52}$
- Thus, $\mathbb{P}(Ace_1 \cap Ace_2 \cap Ace_3) = \frac{4}{52} \times \frac{3}{51} \times \frac{2}{50}$

2. Bayes' rule

Bayes' rule

- Reverend Thomas Bayes (1701–61): English minister and statistician

Bayes' rule

- Recall conditional probability

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

$$\mathbb{P}(\mathsf{A}\cap\mathsf{B}) = \mathbb{P}(\mathsf{A}|\mathsf{B})\cdot\mathbb{P}(\mathsf{B}) = \mathbb{P}(B|A)\cdot\mathbb{P}(A)$$

- **Bayes' rule**: if $\mathbb{P}(\mathsf{B}) > 0$, then:

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

- We often call $\mathbb{P}(B)$ as our prior, and $\mathbb{P}(\mathsf{A}|\mathsf{B})$ as our data/observations

Bayes' rule example

- Use the Covid test example
  - Want to know $\mathbb{P}$(covid | test positive) or $\mathbb{P}$(C | PT)
  - $\mathbb{P}(test\ positive|covid) = 0.8$ true positive rate
  - $\mathbb{P}(covid) = 0.007$ rough prevalance of active Covid cases
  - $\mathbb{P}(test\ positive) = 0.011$ rough prevalance of active Covid cases

$$\mathbb{P}(covid|test\ positive) =$$

Bayes' rule example

- Use the Covid test example
  - Want to know $\mathbb{P}$(covid | test positive) or $\mathbb{P}$(C | PT)
  - $\mathbb{P}(test\ positive|covid) = 0.8$ true positive rate
  - $\mathbb{P}(covid) = 0.007$ rough prevalance of active Covid cases
  - $\mathbb{P}(test\ positive) = 0.011$ rough prevalance of active Covid cases

$$\mathbb{P}(covid|test\ positive) = \frac{\mathbb{P}(test\ positive|covid) \cdot \mathbb{P}(covid)}{\mathbb{P}(test\ positive)}$$

$$\mathbb{P}(C|PT) = \frac{\mathbb{P}(PT|C) \cdot \mathbb{P}(C)}{\mathbb{P}(PT)} = \frac{0.8 \cdot 0.007}{0.011} = 0.53$$

More general version of Bayes Theorum

- Suppose X and θ are two continuous random variables.
  - θ corresponds to the unobserved effects $E's$
  - X corresponds to the observed outcome F (or data)
- We have avaiable two pieces of information:
  - The marginal p.d.f of θ, $\xi(\theta)$, corresponding to $\mathbb{P}(E_i)$
  - The conditional p.d.f of X given θ, $f(x|\theta)$, corresponding to $\mathbb{P}(F|E_i)$

More general version of Bayes Theorum

- The joint p.d.f. of X and θ

$$f(x, \theta) = \xi(\theta|x) f_x(x) \qquad (1)$$

- From this, we get that for x such that $f_x(x) > 0$

$$\xi(\theta|x) = \frac{f(x,\theta)}{f_x(x)} = \frac{f(x|\theta)\xi(\theta)}{f_x(x)} \qquad (2)$$

- Since this is a p.d.f., it must integrate to 1. That is,

$$\int_{-\infty}^{\infty} \frac{f(x|u)\xi(\theta)du}{f_x(x)} = 1 \qquad (3)$$

- Therefore, we have

$$f_x(x) = \int_{-\infty}^{\infty} f(x|u)\xi(\theta)du \qquad (4)$$

- Now, plug in what you have in the equation 2

$$\xi(\theta|x) = \frac{f(x|\theta)\xi(\theta)}{\int_{-\infty}^{\infty} f(x|u)\xi(\theta)du} \qquad (5)$$

More general version of Bayes Theorum

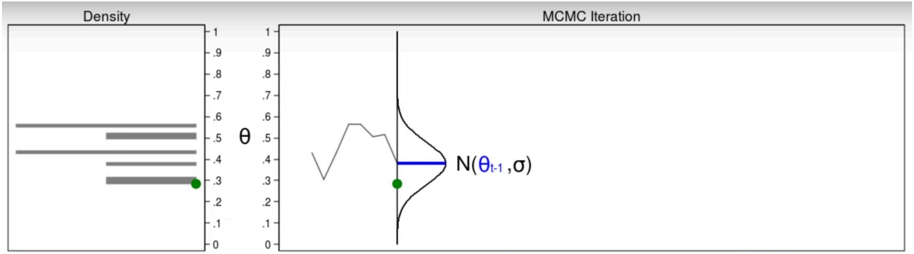
- The denominator is only a normalizing constant so that the density in θ integrates to 1.
- We can simply the equation (5) to this:

$$\underbrace{\xi(\theta|x)}_{\text{posterior dist.}} \propto \underbrace{f(x|\theta)}_{\text{data, given Likelihood}} \underbrace{\xi(\theta)}_{\text{prior dist.}} \tag{6}$$

$$\underbrace{\xi(\theta|data)}_{\text{posterior dist}} \propto \underbrace{f(data|\theta)}_{\text{data}} \underbrace{\xi(\theta)}_{\text{prior dist}} \tag{7}$$

More general version of Bayes Theorum

- Random draws of $\xi(\theta|x)$ are done by MCMC (Monte Carlo Markov Chain) iterations in Bayesian models



Step 1: $r(\theta_{new}, \theta_{t-1}) = \dfrac{\text{Posterior}(\theta_{new})}{\text{Posterior}(\theta_{t-1})} = \dfrac{\text{Beta}(1,1,0.286) \times \text{Binomial}(10,4,0.286)}{\text{Beta}(1,1,0.380) \times \text{Binomial}(10,4,0.380)} = 0.747$

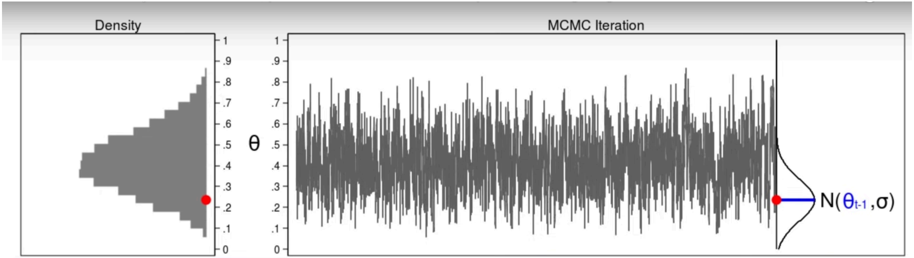
Step 2: Acceptance probability $\alpha(\theta_{new}, \theta_{t-1}) = \min\{r(\theta_{new}, \theta_{t-1}), 1\} = \min\{0.747, 1\} = 0.747$

Step 3: Draw $u \sim \text{Uniform}(0,1) = 0.094$

Step 4: If $u < \alpha(\theta_{new}, \theta_{t-1}) \rightarrow$ If $0.094 < 0.747$ Then $\theta_t = \theta_{new} = 0.286$

Otherwise $\theta_t = \theta_{t-1} = 0.380$

More general version of Bayes Theorum



Step 1: $r(\theta_{new}, \theta_{t-1}) = \dfrac{\text{Posterior}(\theta_{new})}{\text{Posterior}(\theta_{t-1})} = \dfrac{\text{Beta}(1,1,0.180) \times \text{Binomial}(10,4,0.180)}{\text{Beta}(1,1,0.235) \times \text{Binomial}(10,4,0.235)} = 0.523$

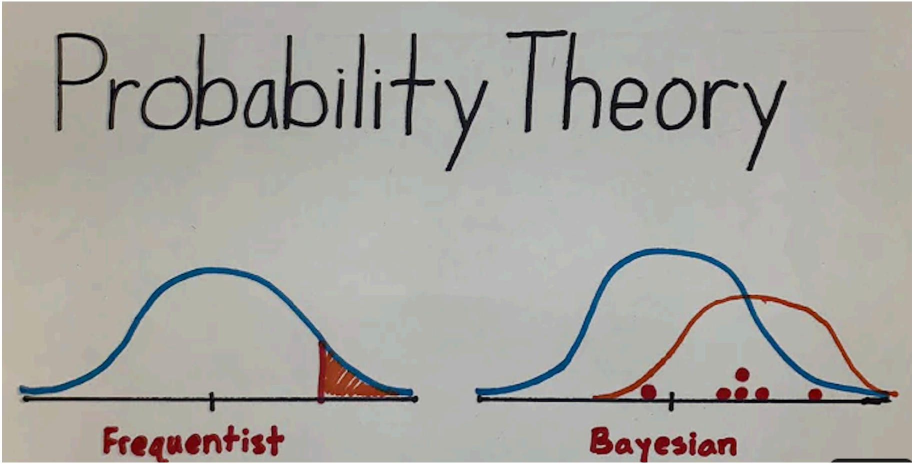
Step 2: Acceptance probability $\alpha(\theta_{new}, \theta_{t-1}) = \min\{r(\theta_{new}, \theta_{t-1}), 1\} = \min\{0.523, 1\} = 0.523$

Step 3: Draw $u \sim \text{Uniform}(0,1) = 0.867$

Step 4: If $u < \alpha(\theta_{new}, \theta_{t-1})$ → If $0.867 < 0.523$ Then $\theta_t = \theta_{new} = 0.180$
Otherwise $\theta_t = \theta_{t-1} = 0.235$

* If you're interested in this process, watch this intro to Bayes stats video from 2:40 → YouTube link

Frequentist vs. Bayesian Hypothesis Testing



Why Should I learn Bayesian statistics

- It allows us to acquire information on the strength of **evidence** for our results. We don't get this information with the p-value point estimate. Such information is highly valuable in research.
- Advantages
  - More **intuitive**
  - Gives you a range between which you can be certain for or against your hypotheses rather than a point-estimate
  - All information is contained within the data itself as opposed to unobserved frequencies
  - Calculates the probability distribution of the hypotheses
- Disadvantages
  - Setting the prior probabilities of the hypothesis can be different values because they're subjective, thus making them appear arbitrary
  - Bayesian analyses are complex and can require advanced statistical packages and software
  - Require more advanced statistical knowledge and computing power

3. Independence

Independence

- Heart of Bayes's rule: knowing B occurs often changes probability of A.
  - What if B provides no information? $\rightsquigarrow$ independence
- Two events A and B are independent if $\mathbb{P}(A \cap B) =$

Independence

- Heart of Bayes's rule: knowing B occurs often changes probability of A.
  - What if B provides no information? $\rightsquigarrow$ independence
- Two events A and B are independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$
- Sometimes written as $A \perp\!\!\!\perp B$
- Symmetric: $A \perp\!\!\!\perp B$ equivalent to $B \perp\!\!\!\perp A$
- Important consequence: if $A \perp\!\!\!\perp B$ and $\mathbb{P}(B) > 0$ then:

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

- Knowing B occurs has no impact on the probability of A.
- Works other way too: if $P(A) > 0$ and $A \perp\!\!\!\perp B \rightsquigarrow \mathbb{P}(B \mid A) = \mathbb{P}(B)$
- Common misunderstanding: **independent is different than disjoint**!
- Mutually exclusive events provide information

Independence and random sampling

- How we draw the random sample matters:
  - Sample n > 1 with replacement $\rightsquigarrow$ independent events
  - Sample n > 1 without replacement $\rightsquigarrow$ dependent events
- Sampling with replacement n for gathering:

$$\mathbb{P}(A_n) = \mathbb{P}(A_1 \cap \cdots \cap A_n) = \mathbb{P}(A_1) \cdot \mathbb{P}(A_2) \cdots \mathbb{P}(A_n)$$

Conditional independence

- A and B are *conditionally independent* given E if $\mathbb{P}(A \cap B \mid E) = \mathbb{P}(A \mid E)\mathbb{P}(B \mid E)$
- Massively important in statistics and causal inference.
- Warning: independence $\neq$ conditional independence.
  - Cond. ind. $\not\Rightarrow$ ind.: flipping a coin with unknown bias.
  - Ind. $\not\Rightarrow$ cond. ind.: test scores, athletics, and college admission.
- You will learn later on why the assumption of independence among observations matters and why it is **not** okay we constantly violate this assumption.