Probability and Mathematical Statistics in Data Science

Lecture 04: Section 2.2: Symmetry in Sampling
Section 2.3: Baye’s Rule

Symmetry in Simple Random Sampling

- One of the topics we will revisit many times is ***simple random sampling***. Sampling without replacement, where each outcome is equally likely to occur

- Sampling with replacement: We keep putting the sampled outcomes back before sampling again.

- We need to count number of possible outcomes from repeating an action such as sampling.

- We will use the **product rule of counting.**

Product rule of counting

- If a set of actions (call them $A_1, A_2, \ldots, A_n$) can result, respectively, in $k_1, k_2, \ldots, k_n$ possible outcomes, then the entire set of actions can result in:

$$k_1 \times k_2 \times k_3 \times \cdots \times k_n \text{ possible outcomes}$$

- For example: toss a coin twice. Each toss can have 2 possible outcomes, therefore 2 tosses can have 4 possible outcomes.

Product Rule of Ordered Pairs

PROPOSITION

If the first element or object of an ordered pair can be selected in n_1 ways, and for each of these n_1 ways the second element of the pair can be selected in n_2 ways, then the number of pairs is $n_1 n_2$.

Q. Home modelling Job: If there are 12 plumbing contractors and 9 electrical contractors available in the area, in how many ways can the contractors be chosen?

A More General Product Rule

Product Rule for k-Tuples

Suppose a set consists of ordered collections of k elements (k-tuples) and that there are n_1 possible choices for the first element; for each choice of the first element, there are n_2 possible choices of the second element; . . . ; for each possible choice of the first $k - 1$ elements, there are n_k choices of the kth element. Then there are $n_1 n_2 \cdot \cdots \cdot n_k$ possible k-tuples.

- If you roll a die 3 times, there are $6^3 = 216$ sequences of faces.
- If you toss a coin 10 times, there are $2^{10} = 1024$ possible sequences of heads and tails.
- If you deal all 52 cards in a standard deck, there are $52!$ possible sequences or *permutations*. That'a lot of possibilities:

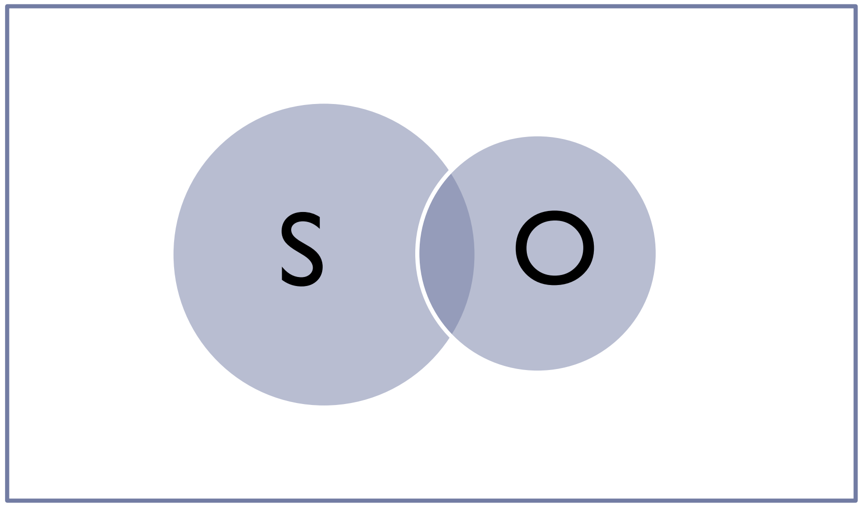
Conditional Probability

- **P(B|A)** is the **conditional probability of B given A**.
- *This describes the chance that event B happens, in the situation that we know event A happens.*
- We can reason about this the same way we did before, but restricting ourselves to the case that A happens.

Conditional Probability: Notation

- What do we mean by P(B|A)?
- We are asking: what is the probability of the event B occurring **given** the event A has occurred?
- It can be challenging (at first) to reason with the difference between the P(A and B) and the P(B|A)
- Thinking about the concepts with real world examples and finite populations can help.

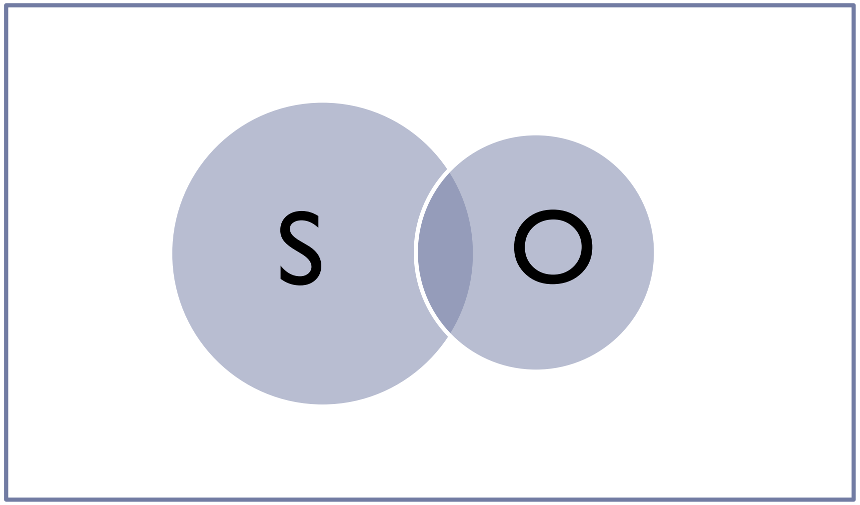
Conditional Probability: Hospital Example



- S="surgery"; O="obstetrics";
- P(S)=0.14; P(O)=0.08; P(S and O)=0.02

- What is P(S|O)? In other words, given that a patient is in hospital for obstetrics, what is the probability they are also there for surgery?

Conditional Probability: Hospital Example



- P(S)=0.14; P(O)=0.08; P(S and O)=0.02
- P(S|O) = *P*(S and O) / P (O) = 0.02 / 0.08 = 0.25 (or 25%)
- The probability a person is in the hospital for surgery given that we know they are there for obstetrics is 0.25

Conditional Probability

- When we want the probability of an event from a *conditional* distribution, we write $P(\mathbf{B}|\mathbf{A})$ and pronounce it "the probability of **B** *given* **A**."

- A probability that takes into account a given condition is called a conditional probability.

- To find the probability of the event **B** *given* the event **A**, we restrict our attention to the outcomes in **A**. We then find the fraction of *those* outcomes **B** that also occurred.

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

High Blood Pressure Fact Sheet

- Having high blood pressure puts you at risk for heart disease and stroke, which are leading causes of death in the United States.

- **About 75 million** American adults (32%) have high blood pressure—that's **1 in every 3 adults**.

- About **1 in 3 American adults** has **prehypertension**—blood pressure numbers that are higher than normal—but not yet in the high blood pressure range.

- https://www.cdc.gov

High Cholesterol in the United States

In 2015–2016, more than 12% of adults age 20 and older had total cholesterol higher than 240 mg/dL.

The chart below shows the prevalence of high total cholesterol (240 mg/dL or more) among adults age 20 and older in the United States from 2015 to 2016.

Racial or Ethnic Group	Men, %	Women, %
Non-Hispanic Blacks	10.6	10.3
Hispanics	13.1	9.0
Non-Hispanic Whites	10.9	14.8
Non-Hispanic Asians	11.3	10.3

https://www.cdc.gov

U.S. Adults: Cholesterol and Blood Pressure

U.S. Adult Population: 234 million

- P(High Cholesterol - HC) = 0.12 (28 million U.S. adults)
- P(High Blood Pressure - HBP) = 0.32 (75 Million U.S. adults)
- P(HC and HBP) = 0.07 (16.4 million U.S. adults)

Q. What is the P(HC|HBP)?

- P(HC|HBP) = P(HC and HBP) / P(HBP) = 0.07/0.32 = 0.22

- 22% of U.S. adults with high blood pressure also have high cholesterol or 16.5 million adults.

Q. What is P(HBP|HC)?

Example

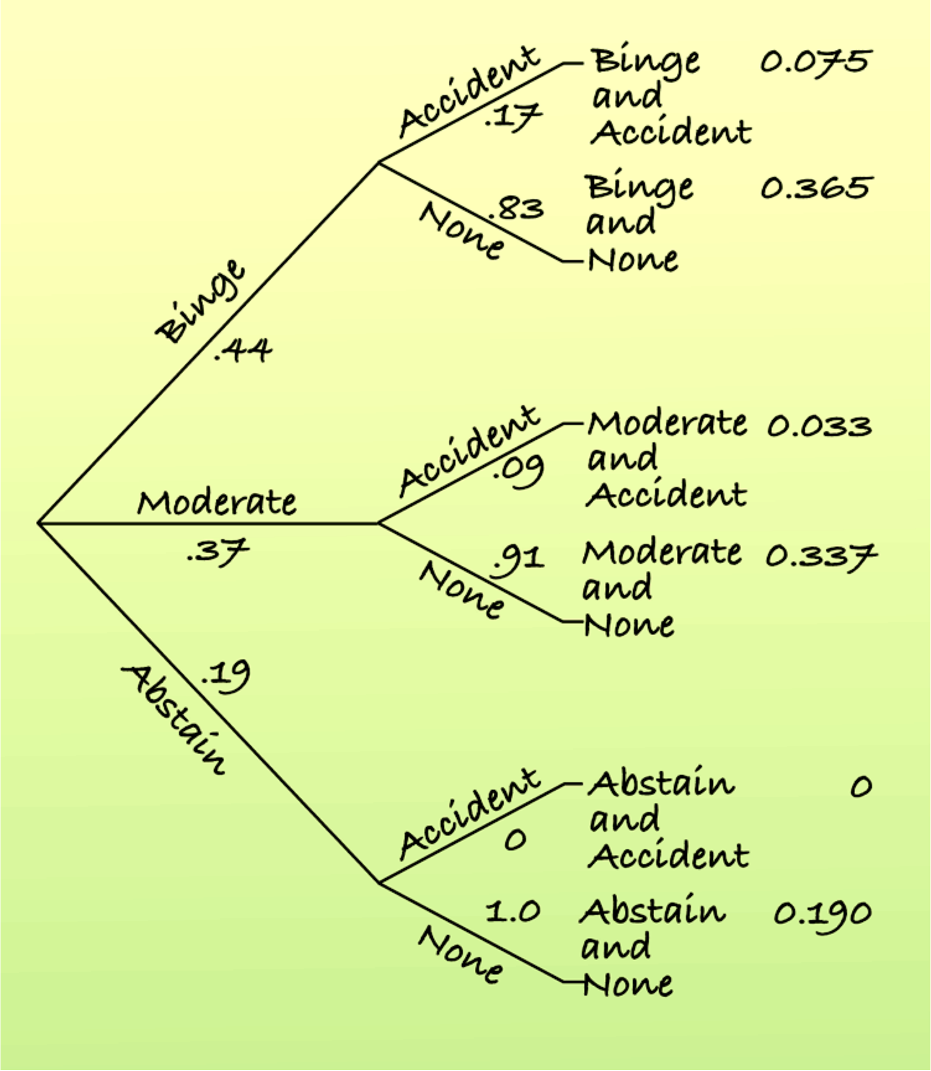
- Binge Drinking on Campus: Results of a National Study:

- 44% of college students engage in binge drinking, 37% drink moderately, and 19% abstain entirely.

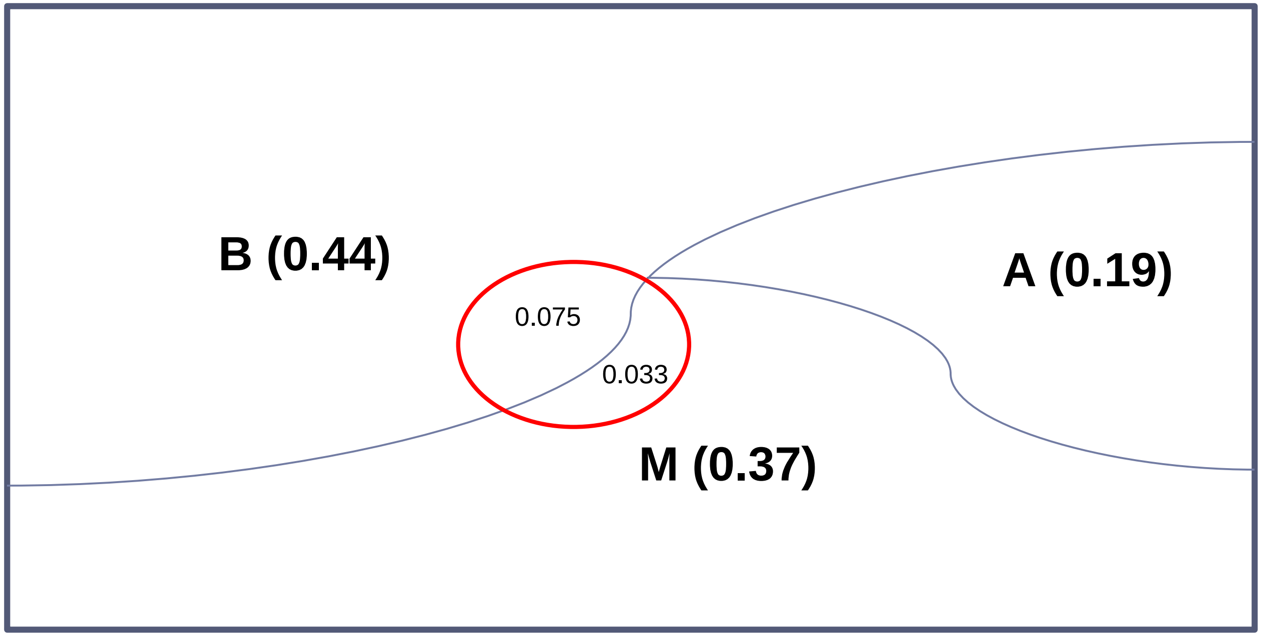
- Another study finds that among binge drinkers aged 21 to 34, 17% have been involved in an alcohol-related accident, and among nonbingers only 9% have been involved in such accidents

Example: Tree Diagrams

- Q. What is the probability of being a binge drinker and having an accident?
- Q. What is the probability that a student had an accident?
- Q. Are the type of drinker a student is and whether they have an accident independent events?



Example: Venn Diagram



Q1. What is the Pr(Accident)?

Q2. What is the Pr(Binge Drinker|Accident)?

Bayes's Rule - Reversing the Conditioning

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A \mid B)P(B) + P(A \mid B^C)P(B^C)}$$

- What is probability that a student is a binge drinker (B) given they had an accident (A)?