University of Nevada, Reno

Lecture 5: Rules of Probability

Perry Williams, PhD

NRES 779

Bayesian Hierarchical Modeling in Natural Resources

Road-map for next three lectures

- The rules of probability supporting Bayes' Theorem
  - Conditional probability
  - The law of total probability

  Some other useful rules
  - Independence
  - Probability of disjoint events
  - Probability of alternative events
  - The chain rule of probability
- Factoring joint probabilities using the chain rule
- Factoring using Bayesian networks
- Probability distributions
- Marginal distributions
- Moment matching

Why do we need to know this?

Concept to be taught	Why do you need to understand this concept?
Conditional probability	The foundation for Bayes' Theorem
The law of total probability	Basis for the denominator of Bayes' Theorem $[y]$
Independence	Allows us to simplify fully factored joint distributions.
Factoring	The procedure for building hierarchical models
Chain rule of probability	The algebra of hierarchical factoring joint distributions
Probability distributions	Our toolbox for applying the Bayesian approach
Moments	Allow us to summarize properties of probability distributions. Basis for inference from MCMC.
Marginal distributions	Bayesian inference is based on marginal distributions of unobserved quantities.
Moment matching	Allows us to embed deterministic models into any probability distribution.

Why Rules of Probability?

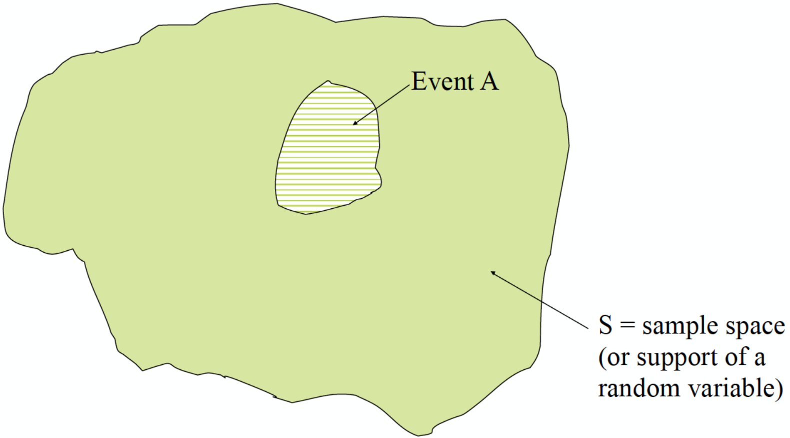
- Bayesian inference treats all unobserved quantities as random variables.
- We can think of the rules of probability as the "algebra of random variables."
- This turns out to be pretty useful because it allows us to use the rules of probability to build and interpret Bayesian models. This is not true for other branches of statistics.

Random Variables and "Events"

- A random variable is a quantity whose values are governed by chance.
- "Governed by chance" means these values arise from a probability distribution.
- Bayesian inference treats *all* unobserved quantities as random variables.
- We seek to understand those distributions.
  - We use a "$\sim$" to denote this in our models.
  - We often use "$=$" for deterministic functions (but there are variants in notation.
- An event is a specific outcome of an experiment or survey, a specific value of a *random variable*.

Concept of Probability

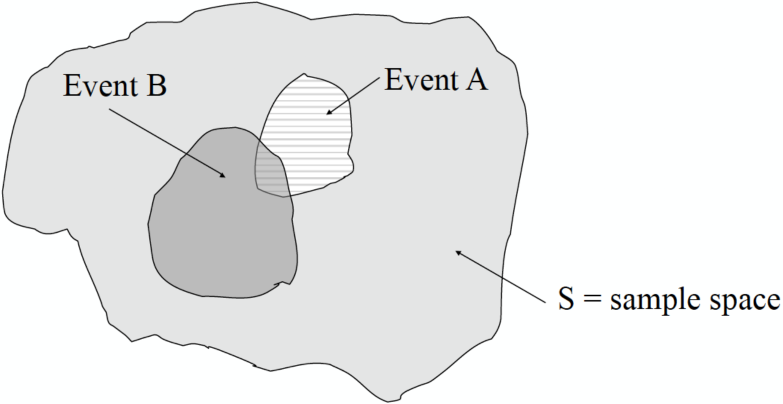
$P(A)$=probability that event A occurs $= \frac{\text{area of } A}{\text{area of } S}$



Conditional Probability

Probability that event A occurs, given that we know B occurred:

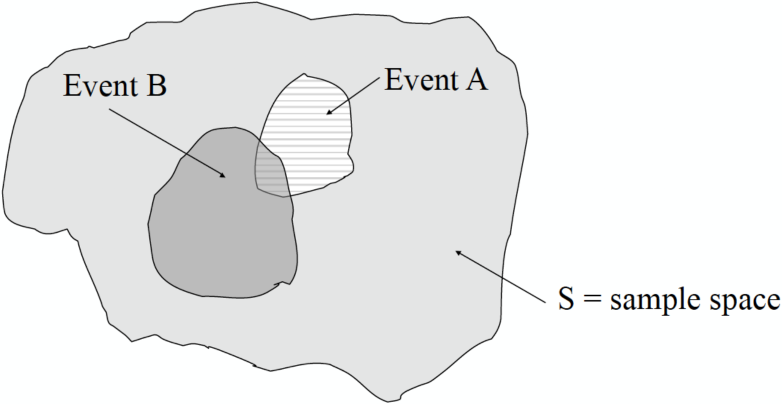
$$P(A|B) = \frac{\text{area of } A \text{ intersecting with } B}{\text{area of } B} = \frac{P(A,B)}{P(B)}$$



Conditional Probability

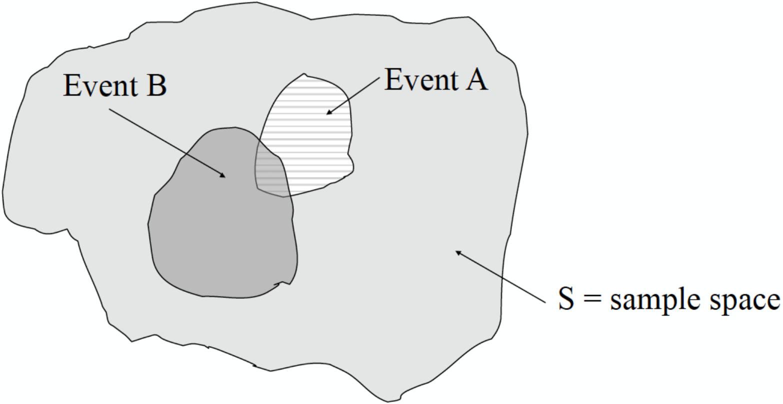
Probability that event B occurs, given that we know A occurred:

$$P(B|A) = ?$$



Conditional Probability

Is $P(A|B) = P(B|A)$?



Conditional Probability

$$P(A|B) = \frac{\text{area of } A \text{ intersecting with } B}{\text{Area of } B} = \frac{P(A,B)}{P(B)}$$

We will make use of the rearrangement of this equation, i.e.,

$$P(A, B) = P(A|B)P(B)$$

and equivalently,

$$P(A, B) = P(B|A)P(A)$$

The right hand side is the *factored* version of the joint distribution of A and B.

Independence of A and B

Knowledge that B occurred provides no information about the probability of A if events A and B are independent.
If A and B are independent, then:

$$P(A|B) = P(A)$$

and ,

$$P(B|A) = P(B)$$

What this means is that the area of A intersecting with the area of B divided by the area of B is exactly the same as the area of A divided by the sample space. (Drawn approximately here to allow use of blobs.)

Derivation Exercise

Use the definition of independence

$$\begin{aligned} P(A|B) &= P(A) \\ P(B|A) &= P(B) \end{aligned} \tag{1}$$

and the definition of conditional probability

$$\begin{aligned} P(A|B) &= \frac{P(A,B)}{P(B)} \\ P(B|A) &= \frac{P(A,B)}{P(A)} \end{aligned} \tag{2}$$

to show that

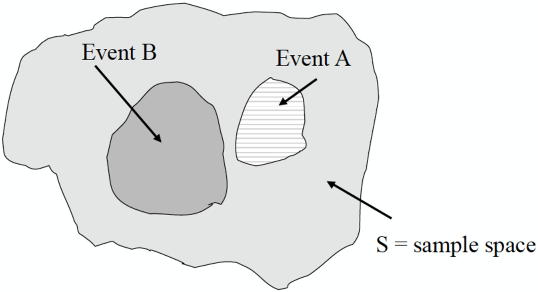
$$P(A,B) = P(A)P(B)$$

when event A and B are independent.

A and *B* are *disjoint*

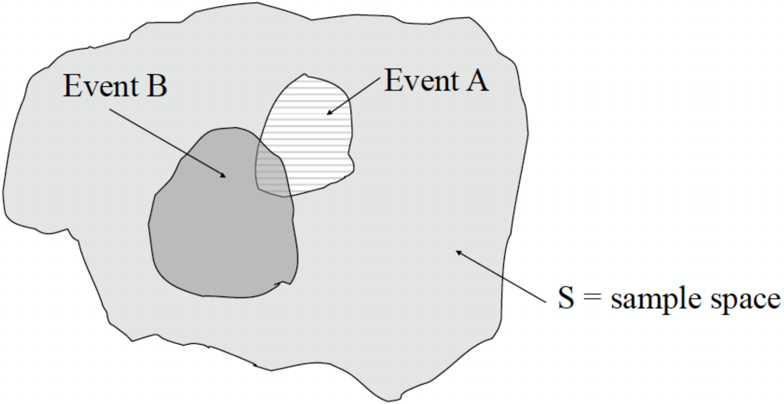
If A and B are disjoint events (or mutually exclusive), then the knowledge that event B has occurred means that we know that A has not occurred.

$$P(A \mid B) = 0$$



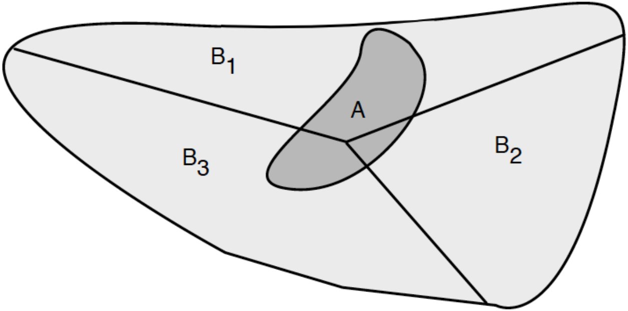
A or B

Probability of A or B occurring:
$P(B \cup A) = P(A) + P(B) - P(A, B))$



Why do we subtract the joint probability of A and B?

Law of Total Probability



We define a set of events $\{B_n : n = 1, 2, \ldots\}$, which taken together, cover the entire sample space, $\sum_n B_n = S$.
The probability of A is:

$$Pr(A) = \sum_n Pr(A|B_n)Pr(B_n).$$

When the area of events becomes “small”

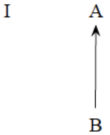
$$Pr(A) = \int_B Pr(A|B)Pr(B_n)dB.$$

Generalizing: The chain rule of probability

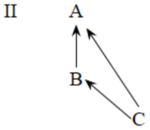
$$Pr(z_1, z_2, \ldots, z_n) = Pr(z_n | z_{n-1}, \ldots, z_1) \times \ldots \times Pr(z_3 | z_2, z_1) Pr(z_2 | z_1) Pr(z_1).$$

- Notice the pattern here.
- z_n can be scalars or vectors.
- When we build models, we choose a sequence that makes sense (Berliner 1996).

Diagramming joint and conditional probabilities



$$\Pr(A,B) = \Pr(A|B)\Pr(B)$$



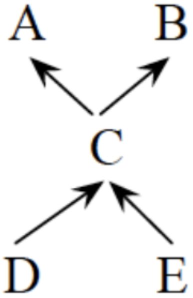
$$\Pr(A,B,C) = \Pr(A|B,C) \times \Pr(B|C)\Pr(C)$$

- Letters are nodes
- Heads of arrows are "children"
- Tails of arrows are "parents"

Factoring Joint Distributions

RULES

- All nodes at head of arrows must be on left hand side of conditioning symbol "|"
- All nodes at tails of arrows must be on right hand side of conditioning symbol.
- Any node at the tail of an arrow without an arrow leading into it must be expressed unconditionally.



$$\Pr(A, B, C, D, E) = \Pr(A|C) \times \Pr(B|C)\Pr(C|D, E) \times \Pr(D)\Pr(E)$$

Generalizing

$Pr(z_1, \ldots, z_n) = \prod_{i=1}^{n} Pr(z_i | \{p_i\})$
$\{p_i\}$ is the set of parents of node z_i