Bayesian statistics – Nanjing Forestry University

Lecture 1

Wei Wu, The University of Southern Mississippi

December, 2016

Acknowledgement: SESYNC Bayesian modeling for socio-environmental data workshop by Hobbs, Collins, and Che-Castaldo

THE UNIVERSITY OF SOUTHERN MISSISSIPPI

GULF COAST RESEARCH LABORATORY

Bayesian inference

- Bayes' theorem

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

$$P(\theta|data) = \frac{P(data|\theta)P(\theta)}{P(data)} \propto P(data|\theta)P(\theta)$$

Posterior = likelihood*prior/normalizing constant

Posterior $\propto$ likelihood*prior

- Advantage of Bayesian inference

-- simplicity of interpretation of the results

-- being consistent and ability of leaning from previous knowledge

-- estimating different sources of errors

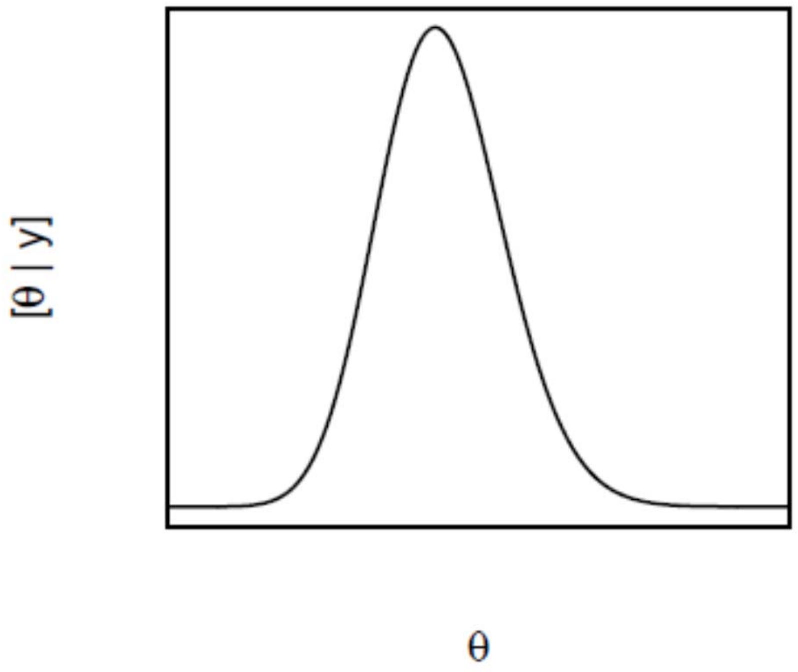
-- data assimilation is automatic and the data need not be at the same scale and time

Some notation

- y data
- Θ a parameter or other unknown quantity of interest
- $[y|\Theta]$ the probability distribution of y conditional on Θ
- $[\theta|y]$ the probability distribution of θ conditional on y
- $P(y|\theta)=p(y|\theta)=[y|\theta]=f(y|\theta)$, different notation that means the same thing.

What sets Bayesian apart

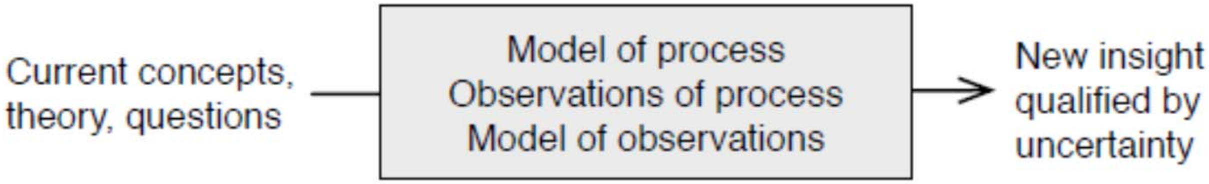
- Bayesian analysis is the only branch of statistics that treats all unobserved quantities as random variables (θ).
- The data are random variables before they are observed and fixed after they have been observed. We seek to understand the probability distribution of unobserved using fixed observations, i.e. [θ|y]
- Those distribution quantify our uncertainty about θ.



What sets Bayesian apart

- You can understand it.

- Rules of probability
    - Conditioning and independence
    - Law of total probability
    - Factoring joint probabilities
- Distribution theory
- Markov chain Monte Carlo



Current concepts, theory, questions

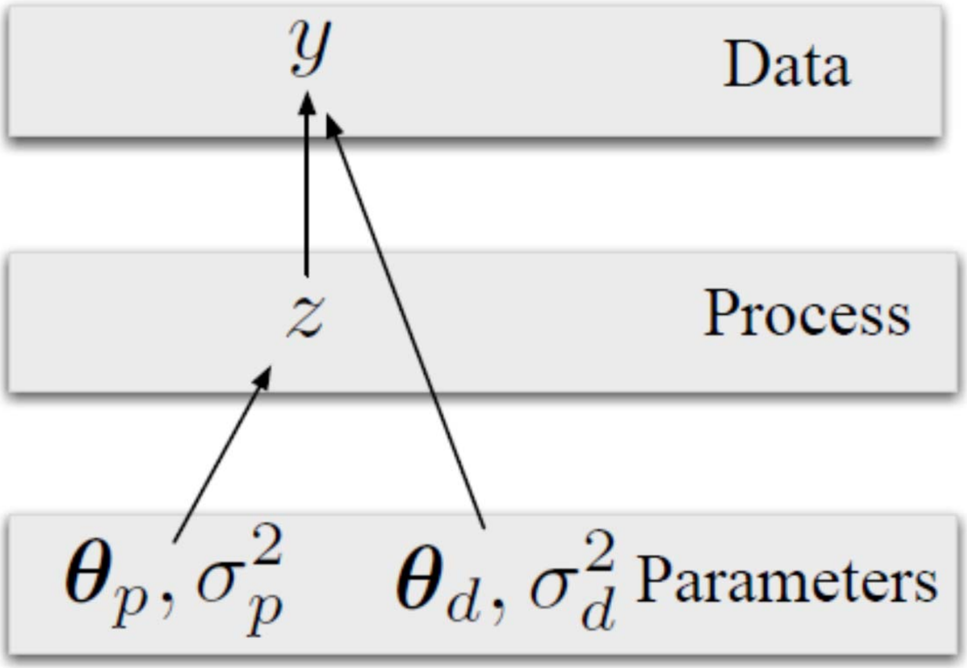
Model of process
Observations of process
Model of observations

New insight qualified by uncertainty

What sets Bayesian apart

One approach applies to many problems

- An unobservable state of interest, z
- A deterministic model of a process g(θ,x), controlling the state
- A model of the data
- Models of parameters


y
Data
z
Process
$\boldsymbol{\theta}_p, \sigma_p^2$
$\boldsymbol{\theta}_d, \sigma_d^2$
Parameters

Probability laws – Uncertainty

Bayesian methods are the only branch of statistics that treat all unobserved quantities as random.

A random variable is a quantity governed by chance (they arise from a probability distribution).

Probability laws – Source of Uncertainty

Process variance: error associated with the model itself

Observation variance: imperfect sampling, measurement

Parameter variance: errors associate with parameters

Variation among individuals: genetics, environmental variation

Model selection uncertainty: inference is conditional on a model

Rules of probability

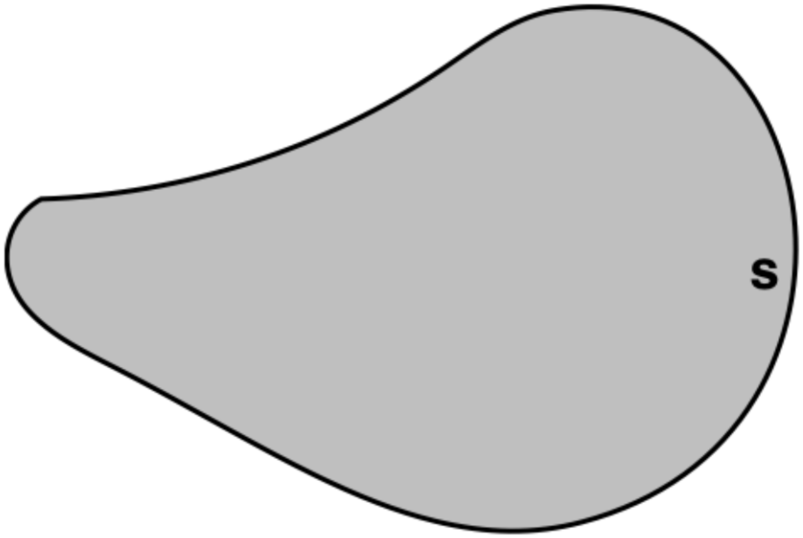
All random variables have probability distributions, which are governed by rules that determine how we gain insights.

A Bayesian framework treats all unobserved quantities as random variables.

Sample space (S)

The set of all possible outcomes of an experiment.

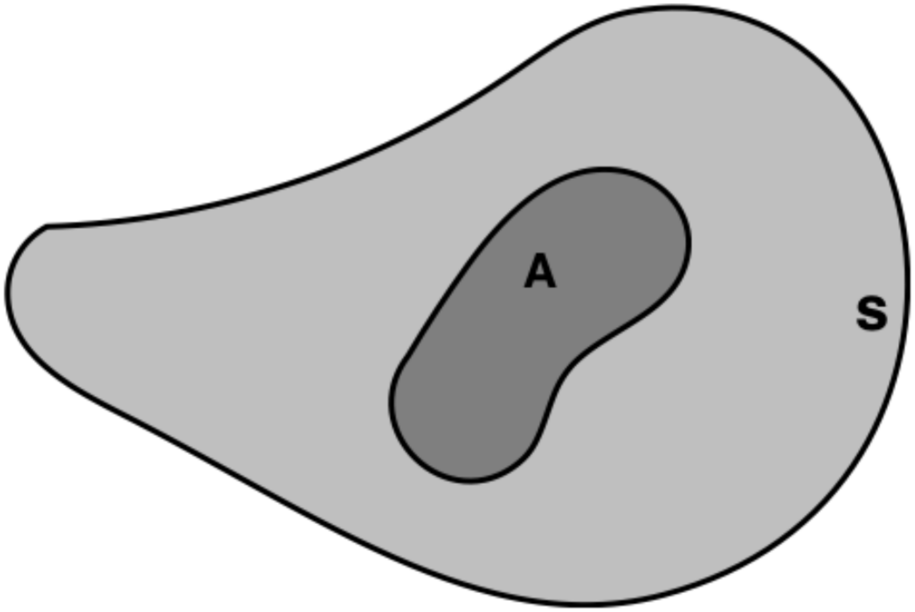
The sample space has a specific area.



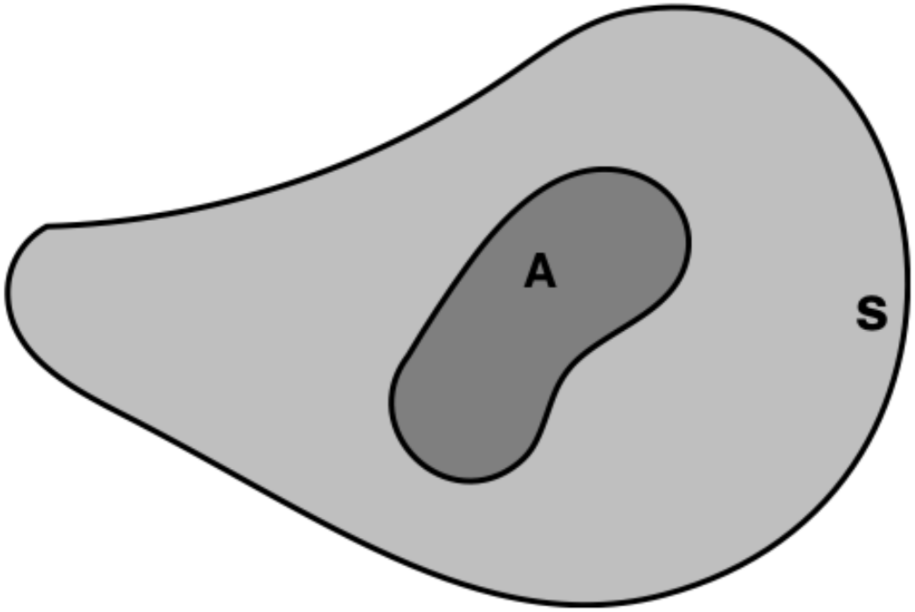
Events in S

Event A is a set of outcomes with a specific area.

The area of event A is less than the area of the sample space.



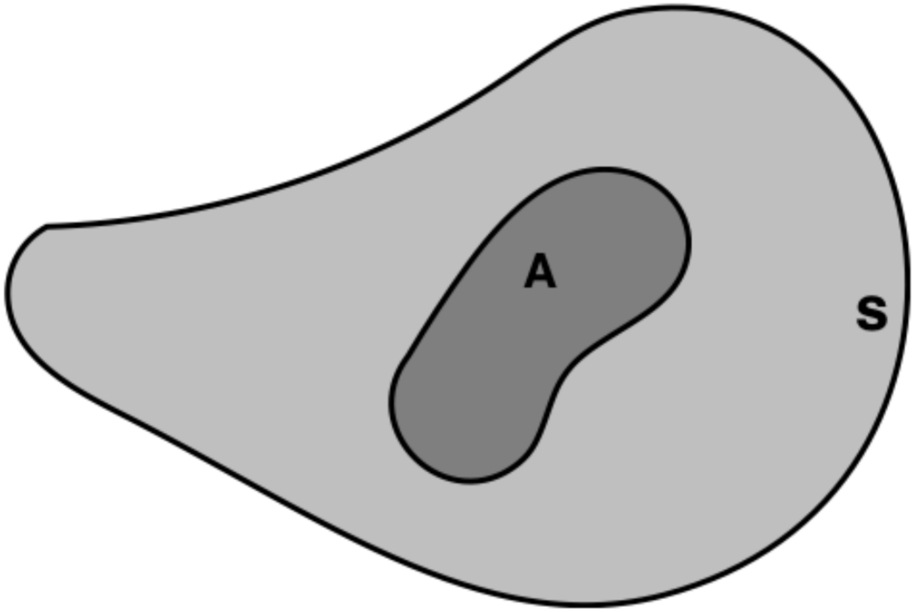
What is the probability of event A?



Pr(A) = Area of A / Area of S

What is the probability of event A?

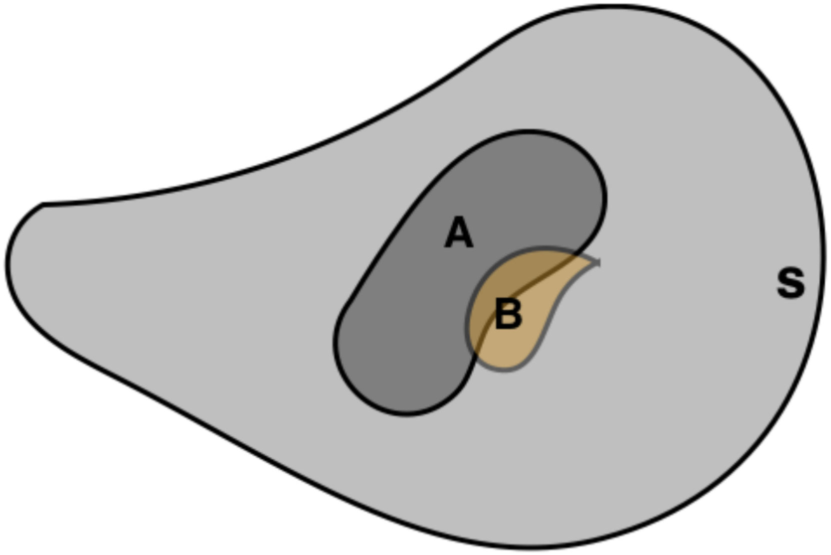
Area of A = 4
Area of S = 20



Pr(A) = Area of A / Area of S = 4/20 = .2

Conditional probability

What about the case where: When event A happens, we learn something about another event, B?



Pr(A) = Area of A / Area of S = 4/20 = .2

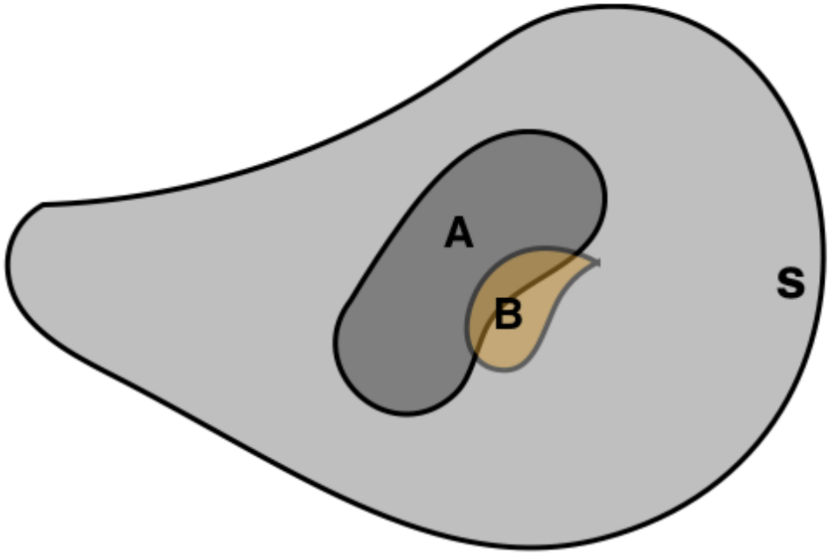
What is the probability of a new event B, given we know that event A has occurred?

Area A = 4

Area B = 2

Area $A \cap B$ = 1

Area S = 20



Pr(B|A) = Probability of B conditional on knowing A has occurred

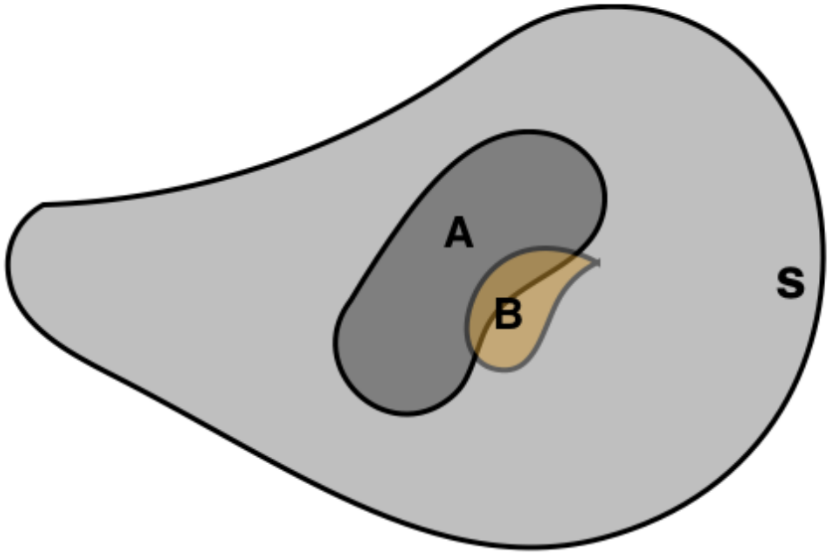
What is the probability of a new event B, given we know that event A has occurred?

Area A = 4

Area B = 2

Area A $\cap$ B = 1

Area S = 20



$$\Pr(B \mid A) = \frac{\textit{Jo}\,\text{int Prob}}{\text{Prob of A}} = \frac{\Pr(A \cap B)}{\Pr(A)} = \frac{\Pr(A, B)}{\Pr(A)}$$

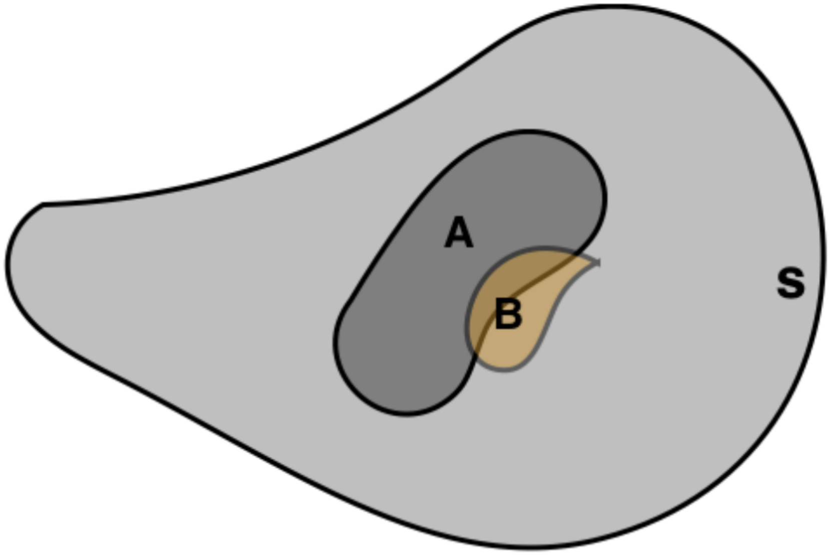
What is the probability of a new event B, given we know that event A has occurred?

Area A = 4

Area B = 2

Area $A \cap B$ = 1

Area S = 20



$$\Pr(B \mid A) = \frac{\Pr(A, B)}{\Pr(A)} = \frac{0.05}{0.2} = 0.25$$

We can rearrange the equation

$$\Pr(B \mid A) = \frac{\Pr(A,B)}{\Pr(A)}$$

$$\Pr(A,B) = \Pr(A \mid B)\Pr(B) \quad \text{and equivalently} \qquad (1)$$

$$\Pr(\mathrm{A},\mathrm{B}) = \Pr(\mathrm{B} \mid \mathrm{A})\Pr(\mathrm{A}) \qquad (2)$$

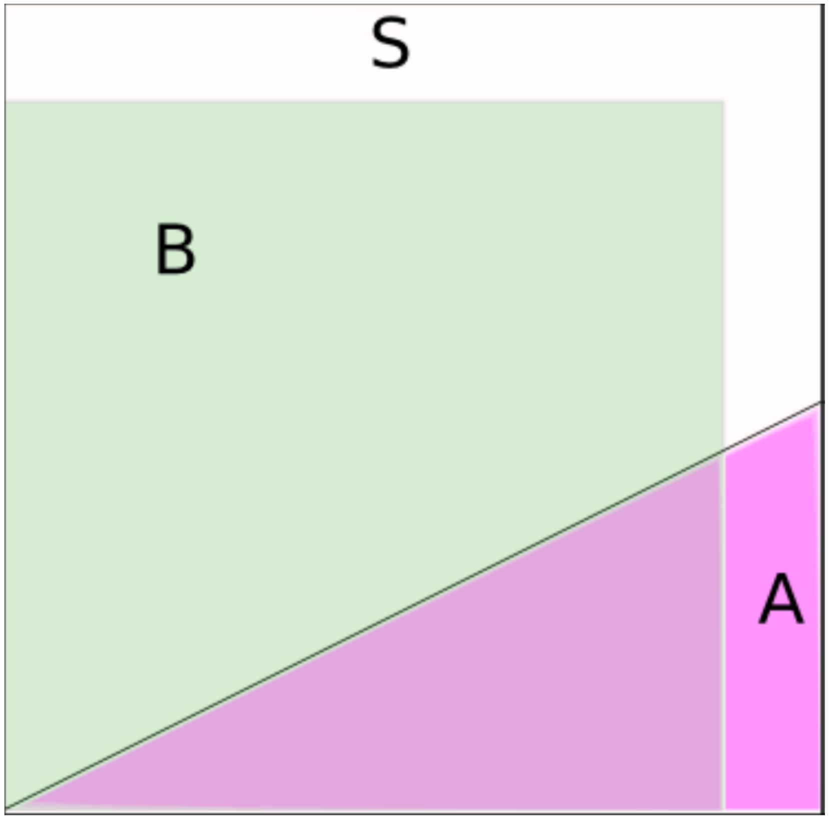
Conditional probability

True or false

$$\Pr(\mathrm{B} \mid \mathrm{A}) = \Pr(\mathrm{A} \mid \mathrm{B})$$

Conditional probability

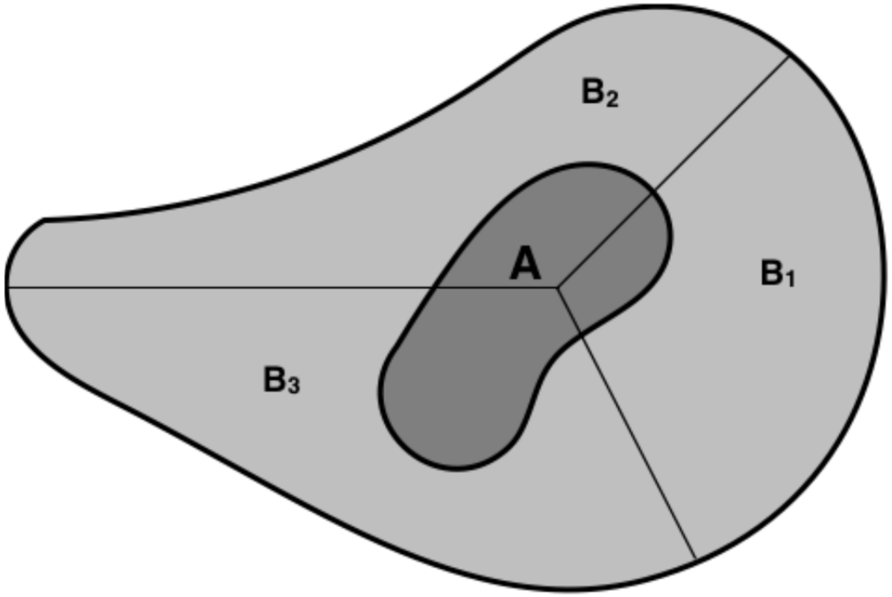
Events are independent if and only if Pr(A|B) = Pr(A)



The law of total probability

The sample space (S) is a non-overlapping group of events.

We can define a set of events {B_n: n=1,2,3,…},
Which taken together define the entire sample
Space, $\sum_n B_n = S$



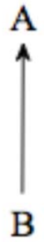
Factoring joint probabilities

The chain rule allows us to take complicated joint distributions of random variables and break them down into simpler conditional probabilities.

Conditional probabilities can be analyzed one at a time as if all other random variables are known and constant.

Factoring joint probabilities provide a usable graphical and mathematical foundation, which is critical for accomplishing the model specification step in the general modeling process.

Consider a Bayesian network (represented by a directed acyclic graph or DAG)



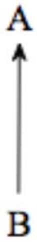
Bayesian networks specify how joint distributions are factored into conditional distributions using nodes to represent random variables (RV) and arrows to represent dependencies.

Notes at the heads of arrows must be on the left hand side of the conditioning symbols.

Nodes at the tails of arrows are on the right hand side of the conditioning symbols.

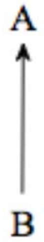
Any node at the tail of an arrow without an arrow leading into it must be expressed unconditionally.

Factoring with DAGs



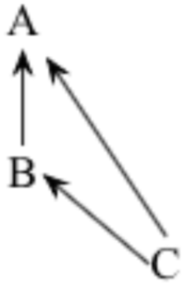
Pr(A,B) =

Factoring with DAGs



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

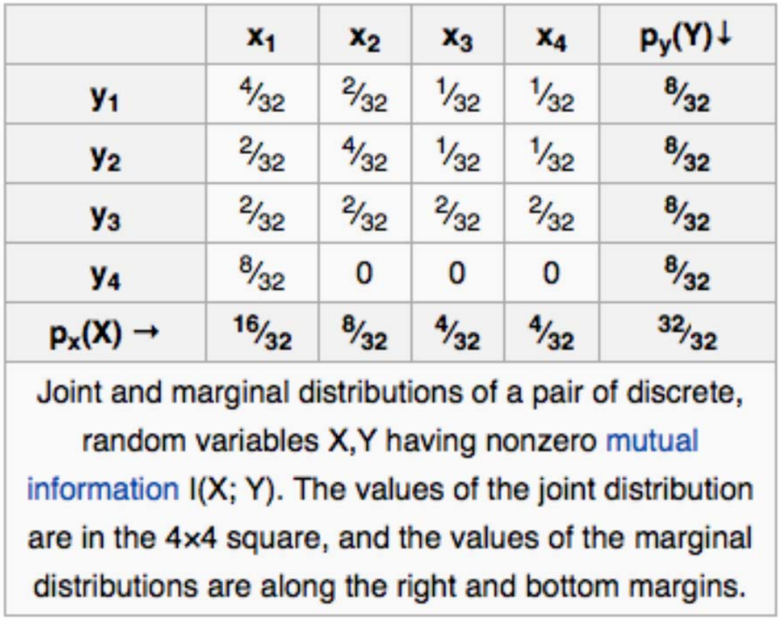
Factoring with DAGs



Pr(A,B,C) = Pr(A|B,C)Pr(B|C)Pr(C)

Marginal distributions

The marginal distribution of a subset of a collection of random variable is the probability distribution of the random variables contained in the subset.

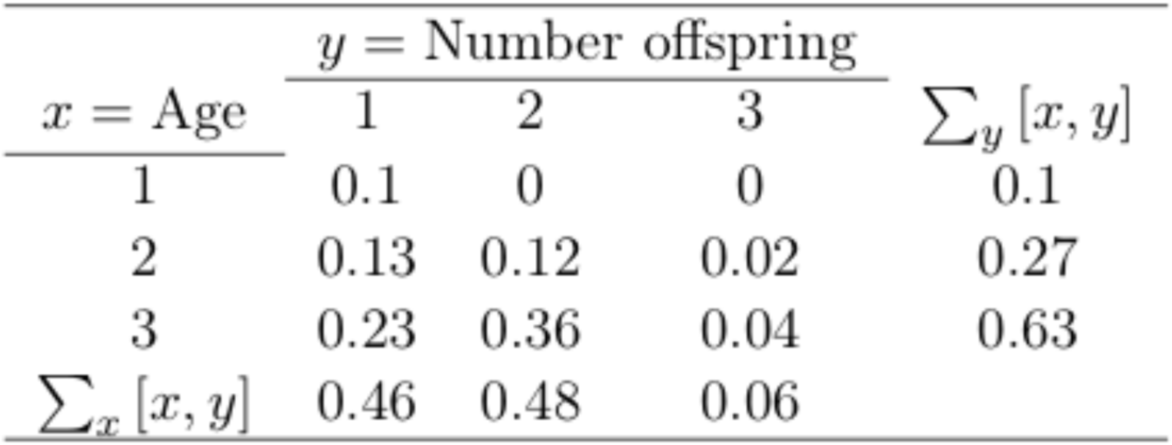


	x_1	x_2	x_3	x_4	$p_y(Y)$↓
y_1	4/32	2/32	1/32	1/32	8/32
y_2	2/32	4/32	1/32	1/32	8/32
y_3	2/32	2/32	2/32	2/32	8/32
y_4	8/32	0	0	0	8/32
$p_x(X)$ →	16/32	8/32	4/32	4/32	32/32

Joint and marginal distributions of a pair of discrete, random variables X,Y having nonzero mutual information I(X; Y). The values of the joint distribution are in the 4×4 square, and the values of the marginal distributions are along the right and bottom margins.

Marginal distributions example

Consider two discrete random variables that are jointly distributed such that [x,y]≡ Pr(x,y).
We are studying a species for which births occur in pulses. We observe 100 females and record the age of each animal and the number of offspring produced.



	$y =$ Number offspring			
$x =$ Age	1	2	3	$\sum_y [x,y]$
1	0.1	0	0	0.1
2	0.13	0.12	0.02	0.27
3	0.23	0.36	0.04	0.63
$\sum_x [x,y]$	0.46	0.48	0.06	

When we have a joint distribution of two random variables, we can focus on one by summing over the probabilities of the other.

Marginal distributions example

We care about marginal distributions because they allow us to represent the univariate distribution of unknown quantities that are parts of joint distributions.