ISTA410/INFO510 Bayesian Modelling and Inference

Lecture 4 – Independence, Distributions, Continuity

Dr. Kunal Arekar
College of Information
University of Arizona, Tucson

September 8, 2025

Basic rules (so far)

Marginalization

$$P(A) = \sum_{b \in B} P(A, B)$$ *

Normalization

$$P(X = x) = \frac{p(x)}{\sum_x p(x)}$$ *

Conditional probability (definition)

$$P\left(A \middle| B\right) \equiv \frac{P(A \cap B)}{P(B)}$$ *

Chain (Product) Rule

$$P\left(A_1 \cap A_2\right) = P(A_1)P(A_2 | A_1)$$ *

$$P\left(A_1 \cap A_2 \cap \ \ .... \ \ A_N\right) = P(A_1)P(A_2 | A_1)P(A_3 | A_1 \cap A_2) \ \ .... \ \ P(A_N | A_1 \cap A_2 \cap \ \ .... \ \ A_{N-1})$$

Bayes Rule

$$P\left(A \middle| B\right) = \frac{P(B|A)P(A)}{P(B)}$$ *

Structure in relations between RVs:

(Absolute, Marginal) **Independence and Conditional Independence**

In probability and statistics, understanding the structure in relationships between random variables (RVs) is crucial

These relationships determine how information flows between variables, how dependencies arise, and how we can simplify complex probability models

(Absolute, Marginal) **Independence**

First, recall the product rule (which holds generally):
$P(X, Y) = P(X)P(Y|X)$

$$P(A_1 \cap A_2) = P(A_1)P(A_2|A_1)$$

$$P(A_1 \cap A_2 \cap \;\; \dots \;\; A_N) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \;\; \dots \;\; P(A_N|A_1 \cap A_2 \cap \;\; \dots \;\; A_{N-1})$$ *

expresses the probability of a sequence of dependent events in terms of conditional probabilities

(Absolute, Marginal) **Independence**

First, recall the product rule
(which holds generally):

$$P(X, Y) = P(X)P(Y|X)$$

"X is independent of Y"

$$X \perp Y \quad \Leftrightarrow \quad P(X,Y) = P(X)P(Y) \qquad *$$

$$X \perp Y \quad \Leftrightarrow \quad P(X|Y) = P(X) \quad \text{or } \ \mathrm{P(Y)=0} \qquad *$$

Some authors use ⫫ instead of ⊥

(Absolute, Marginal) **Independence**

First, recall the product rule
(which holds generally):

$$P(X, Y) = P(X)P(Y|X)$$

This can cause confusion. If P(Y) is zero, then P(X|Y) is ill-defined (divide by zero). Since if P(Y)=0, Y never happens, we have a choice to declare whether (a) X is independent from Y or (b) not. Under scrutiny, (a) make sense, and allows consistency with the first definition. The first formula works in this (weird) case because if P(Y)=0, then P(X,Y) is also 0.

"X is independent of Y"

$$X \perp Y \quad \Leftrightarrow \quad P(X,Y) = P(X)P(Y) \qquad *$$

$$X \perp Y \quad \Leftrightarrow \quad P(X|Y) = P(X) \quad \text{or} \;\; \mathrm{P(Y)=0} \qquad *$$

Some authors use ⫫ instead of ⊥

(Absolute, Marginal) **Independence**

"X is independent of Y"

$$X \perp Y \quad \Leftrightarrow \quad P(X,Y) = P(X)P(Y) \qquad *$$

$$X \perp Y \quad \Leftrightarrow \quad P(X|Y) = P(X) \quad \text{or } \mathrm{P(Y)=0} \qquad *$$

$P(T)$

T	P
warm	0.5
cold	0.5

$P_1(T,W)$

T	W	P
warm	sun	0.4
warm	rain	0.1
cold	sun	0.2
cold	rain	0.3

$P(W)$

W	P
sun	0.6
rain	0.4

(Absolute, Marginal) **Independence**

"X is independent of Y"

$$X \perp Y \quad \Leftrightarrow \quad P(X,Y) = P(X)P(Y) \quad *$$

$$X \perp Y \quad \Leftrightarrow \quad P(X|Y) = P(X) \quad \text{or} \;\; \mathrm{P(Y)=0} \quad *$$

$P(T)$

T	P
warm	0.5
cold	0.5

$P_1(T, W)$

T	W	P
warm	sun	0.4
warm	rain	0.1
cold	sun	0.2
cold	rain	0.3

$P_2(T, W)$

T	W	P
warm	sun	0.3
warm	rain	0.2
cold	sun	0.3
cold	rain	0.2

$P(W)$

W	P
sun	0.6
rain	0.4

Conditional Independence

$$X \perp Y \mid Z \quad \Leftrightarrow \quad P(X \mid Y, Z) = P(X \mid Z) \quad \text{or} \quad \mathrm{P(Y,Z)=0} \quad *$$

$$P\,(Y \mid X, Z) = P\,(Y \mid Z)$$

Conditional Independence

$$X \perp Y \mid Z \quad \Leftrightarrow \quad P(X \mid Y, Z) = P(X \mid Z) \quad \text{or} \quad \mathrm{P(Y,Z)=0} \quad *$$

$$P(Y \mid X, Z) = P(Y \mid Z)$$

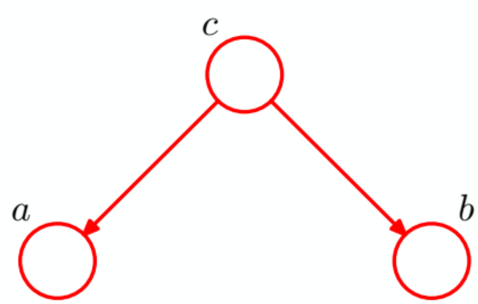
Equivalent, sometimes more convenient definition

$$X \perp Y \mid Z \quad \Leftrightarrow \quad P(X, Y \mid Z) = P(X \mid Z) P(Y \mid Z) \quad *$$

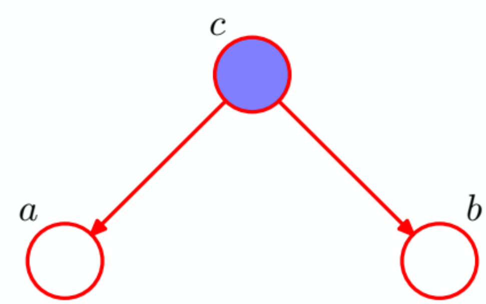
Conditional Independence

Equivalent, sometimes more convenient definition

$$X \perp Y \mid Z \quad \Leftrightarrow \quad P(X,Y \mid Z) = P(X \mid Z)P(Y \mid Z)$$ *



$a \not\perp\!\!\!\perp b$



$a \perp\!\!\!\perp b \mid c$

$$p(a, b, c) = p(c)p(a|c)p(b|c)$$ (what the graph represents in general)

$$p(a, b|c) = p(a|c)p(b|c)$$

(right side: with c observed, p(c) = 1.0;
left side: we're *conditioning* on c)

This is the definition of $a \perp\!\!\!\perp b \mid c$

Discrete Distributions:

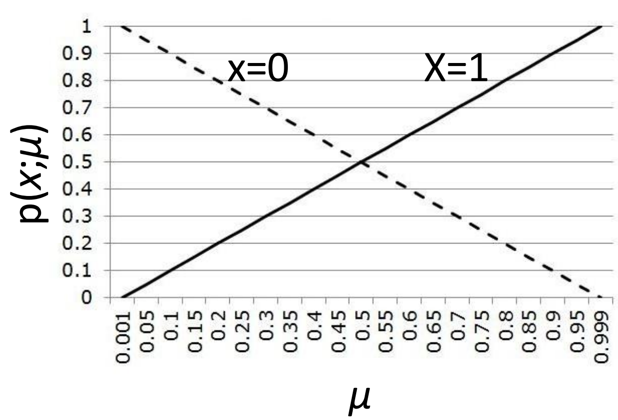
Probability *Mass* Functions (pmf)

Bernoulli Distribution

$x \in \{0,1\}$ (e.g., 1 is "heads" and 0 is "tails")

$\mathrm{p}(x=1|\mu)=\mu$ and $\mathrm{p}(x=0|\mu)=1-\mu$

$$\mathrm{p}(x|u)=\begin{cases} u & \text{if x=1} \\ 1-u & \text{if x=0} \end{cases}$$

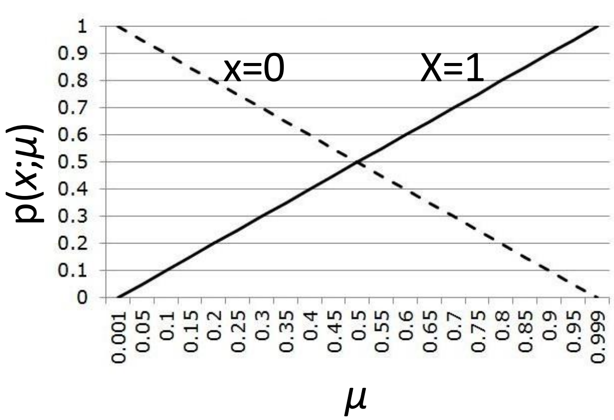


Bernoulli Distribution

$x \in \{0,1\}$ (e.g., 1 is "heads" and 0 is "tails")

$\mathrm{p}(x=1|\mu)=\mu$ and $\mathrm{p}(x=0|\mu)=1-\mu$

$Bern(x|\mu) = \mu^x(1-\mu)^{(1-x)}$



Binomial Distribution

Probability distribution for getting *m* "heads" out of *N* tosses.

$$Bin(m|N,\mu) = \underbrace{\binom{N}{m}}_{\substack{\text{Number of}\\ \text{ways to get}\\ m \text{ heads in}\\ N \text{ tosses}}} \cdot \underbrace{\mu^m(1-\mu)^{(N-m)}}_{\substack{\text{Probability of}\\ \textbf{\textit{each}} \text{ way to}\\ \text{get } m \text{ heads in}\\ N \text{ tosses}}}$$

Example event

N=3, *m*=2

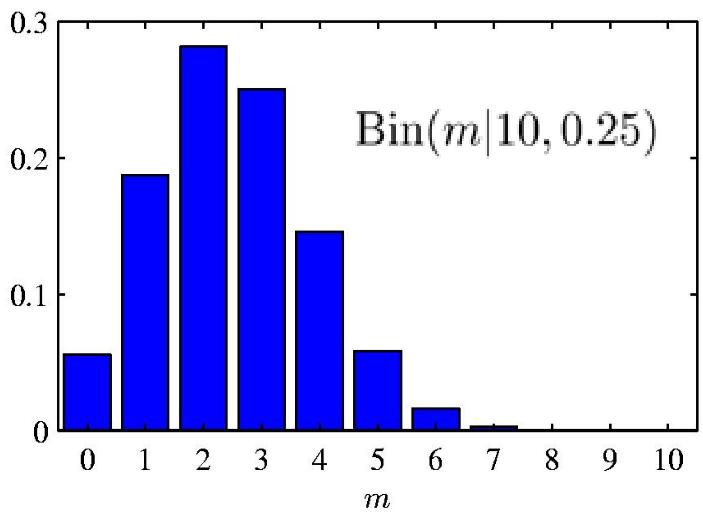
- HHT
- HTH
- THH

where $\binom{N}{m} = \frac{N!}{(N-m)!m!}$

Binomial Distribution

Probability distribution for getting *m* “heads” out of *N* tosses.

$$Bin(m|N,\mu) = \underbrace{\binom{N}{m}}_{\substack{\text{Number of} \\ \text{ways to get} \\ m \text{ heads in} \\ \text{N tosses}}} \cdot \underbrace{\mu^m (1-\mu)^{(N-m)}}_{\substack{\text{Probability of} \\ \textbf{\textit{each}} \text{ way to} \\ \text{get } m \text{ heads in} \\ N \text{ tosses}}}$$



where $\binom{N}{m} = \frac{N!}{(N-m)!m!}$

Categorical Distribution

- Simple extensions to Bernoulli to multiple outcomes (e.g., a six sided die)
- Let K be the number of possible outcomes (e.g., for six sided die, K=6)
- Now we use vectors for μ and x, i.e., $\boldsymbol{\mu}$ and **x**.
- **x** is a vector of 0's and exactly one 1 for observed outcomes (e.g., rolling 3 with a six sided die is (0,0,1,0,0,0)).
- $\boldsymbol{\mu}$ is a vector of real values representing the distribution over which element of **x** will be value 1.

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \boldsymbol{\mu}_k^{x_k} \qquad \text{(note that } \sum_{k=1}^{K} \boldsymbol{\mu}_k = 1)$$

Categorical Distribution

- Simple extensions to Bernoulli to multiple outcomes (e.g., a six sided die)
- Let K be the number of possible outcomes (e.g., for six sided die, K=6)
- Now we use vectors for μ and x, i.e., $\boldsymbol{\mu}$ and $\mathbf{x}$.
- $\mathbf{x}$ is a vector of 0's and exactly one 1 for observed outcomes (e.g., rolling 3 with a six sided die is (0,0,1,0,0,0)).
- $\boldsymbol{\mu}$ is a vector of real values representing the distribution over which element of $\mathbf{x}$ will be value 1.

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \boldsymbol{\mu}_k^{x_k} \qquad (\text{note that } \sum_{k=1}^{K} \boldsymbol{\mu}_k = 1)$$

Multinomial Distribution

Extension of binomial to multiple outcomes.

Let K be the number of outcomes, with probabilities $\boldsymbol{\mu}$.

$Mult(m_1, m_2, ..., m_K \; \boldsymbol{\mu}, | N)$ denotes the probaility of seeing m_1 of outcome 1, m_2 of outcome 2, and so on.

Multinomial Distribution

Extension of binomial to multiple outcomes.

Let K be the number of outcomes, with probabilities μ.

$$Mult(m_1, m_2, \dots, m_K \mid \boldsymbol{\mu}, N) = \binom{N}{m_1 \quad m_2 \quad \dots \quad m_K} \prod_{k-1}^{K} \mu_k^{m_k}$$

where $\binom{N}{m_1 \quad m_2 \quad \dots \quad m_K} = \left(\frac{N!}{m_1! \quad m_2! \quad \dots \quad m_K!} \right)$

and $\sum_{k=1}^{K} m_k = N$

Continuous Distributions:

Probability *Density* Functions (pdf)

Continuous Spaces

- Outcome space is observation of a real value
  - E.g., person's height, temperature outside, etc…
- Example: a random variable X that can take on any value in [0,1] with equal probability
  - We say that X is **uniformly** distributed.

- Challenge: $P(X = x) = 0$
- Instead of assigning probabilities to points, we instead assign probabilities to regions (within some range or interval):

$P(x_1 \leq X \leq x_2)$ but not $P(X = x)$

Continuous Spaces

For a Uniform Distribution on [0,1]

The **PDF** is $f(x)=1$ for $0 \leq x \leq 1$

The probability of picking a number in some subrange [x1,x2] is given by

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} p(x)\, dx$$

Here,
$p(x)=1$
$dx = x_2 - x_1$

$P(0.2 \leq x \leq 0.5) = 0.5 - 0.2 = 0.3$

Probability Density Functions

To compute the probability that X lies in some range, we compute the definite integral of the density function:

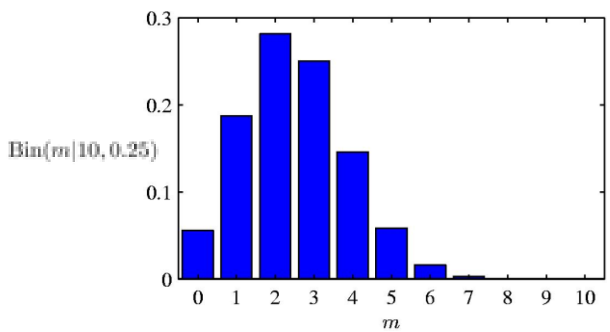
$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} p(x)\, dx$$

Probability Density Functions

To compute the probability that X lies in some range, we compute the definite integral of the density function:

$$P(x_1 \le X \le x_2) = \int_{x_1}^{x_2} p(x)\,dx$$

With discrete outcome spaces…
Probability **Mass** Function (**pmf**)

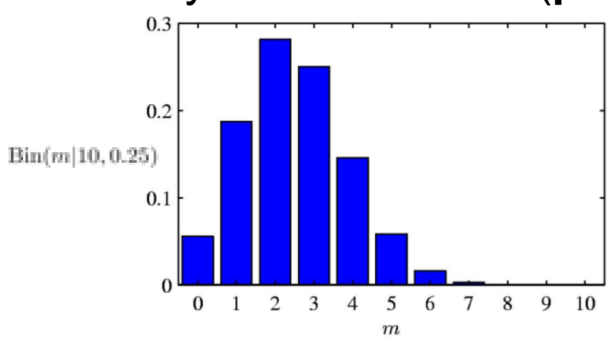


Probability Density Functions

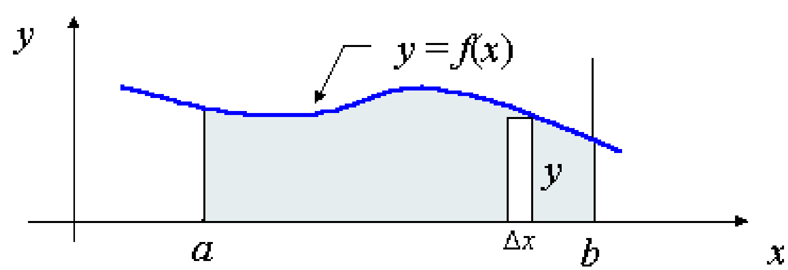
To compute the probability that X lies in some range, we compute the definite integral of the density function:

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} p(x)\, dx$$

With discrete outcome spaces…
Probability **Mass** Function (**pmf**)



Probability **Density** Function (**pdf**)



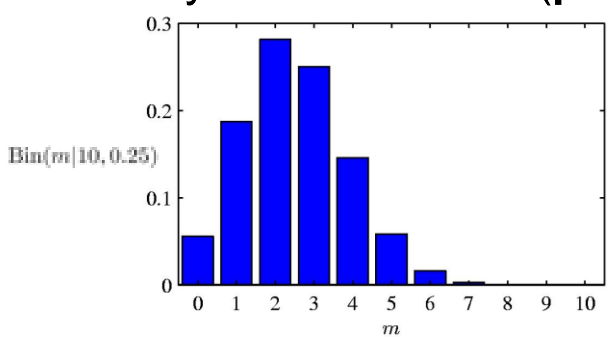
Probability Density Functions

To compute the probability that X lies in some range, we compute the definite integral of the density function:

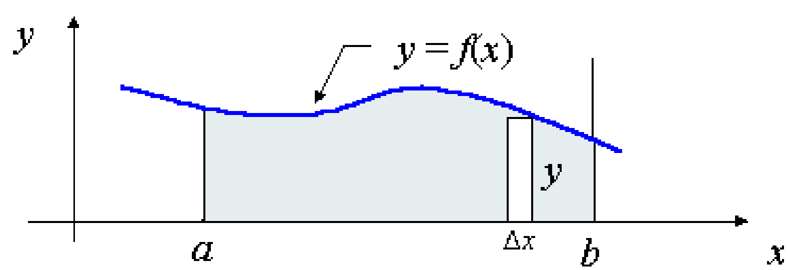
$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} p(x)\, dx$$

$$\underset{\text{density}}{\rho} = \frac{\overset{\text{mass}}{m}}{\underset{\text{volume}}{V}}$$

With discrete outcome spaces…
Probability **Mass** Function (**pmf**)



Probability **Density** Function (**pdf**)



Probability Density Functions

To compute the probability that X lies in some range, we compute the definite integral of the density function:

$$P(x_1 \le X \le x_2) = \int_{x_1}^{x_2} p(x)\,dx$$

mass

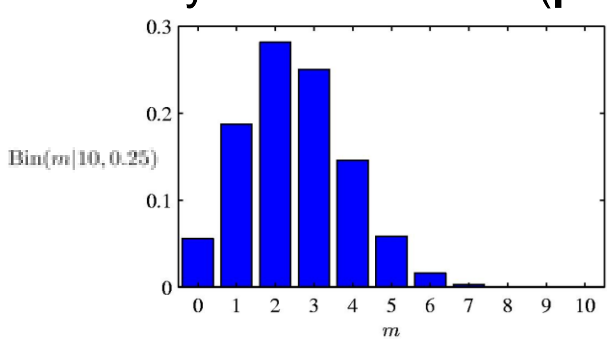
density **volume**

mass

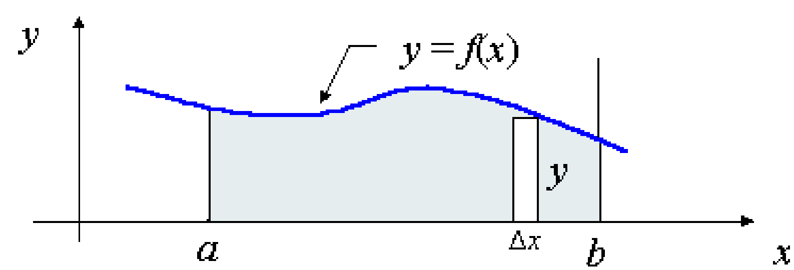
$$\underset{\text{density}}{\rho} = \frac{m}{V}$$

volume

With discrete outcome spaces…
Probability **Mass** Function (**pmf**)



Probability **Density** Function (**pdf**)



Probability Density Functions

To compute the probability that X lies in some range, we compute the definite integral of the density function:

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} p(x)\, dx$$

mass

density volume

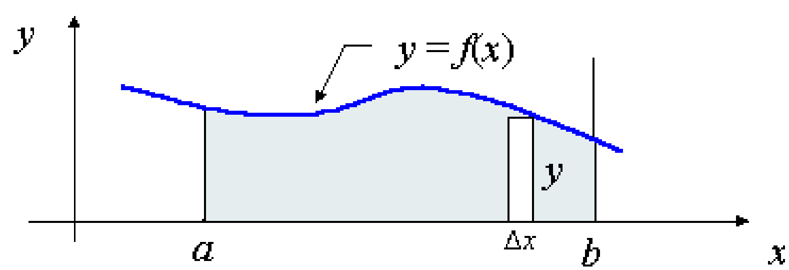
$$p : \mathbb{R} \mapsto \mathbb{R}$$

$$p(x) \geq 0$$

$$\int_{-\infty}^{\infty} p(x)\, dx = 1$$

Note: p(x) can be > 1 !

Probability **Density** Function (**pdf**)



The actual probability of an event is obtained by integrating the PDF over a specific interval

Probability Density Functions

To compute the probability that X lies in some range, we compute the definite integral of the density function:

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} p(x)\, dx$$

mass

density volume

$p(\boldsymbol{x}) = p(x_0, x_1, ..., x_n)$ Probability vector is just a joint probability!

Joint $P(x_1 \leq X \leq x_2, y_1 \leq Y \leq y_x) = \int_{x=x_1}^{x_2} \int_{y=y_1}^{y_2} p(x, y)\, dx\, dy$

Conditional $P(x_1 \leq X \leq x_2, Y = y) = \int_{x=x_1}^{x_2} p(x|Y = y)\, dx$

Marginalization $P(y) = \int_{x=x_1}^{x_2} p(x, y)\, dx$ (where $x_1 \leq X \leq x_2$ describes the sample space of *X*)