PROBABILITY THEORY REVIEW

CMPUT 466/551

Martha White

Fall, 2017

Reminders

- Assignment 1 is due on September 28
- Thought questions 1 are due on September 21
  - Chapters 1-4, about 40 pages
  - If you are printing, don't print all the notes yet
- Office hours
  - Martha: 3-5 p.m. on Tuesday (ATH 3-05)
  - Labs: W 5-8 p.m. and F 2-5 p.m.
- Start thinking about datasets for your mini-project
- I do not expect you to know formulas, like pdfs
- I will use a combination of slides and writing on the board (where you should take notes)

Background for course

- Need to know calculus, mostly derivatives
  - Will almost never integrate
  - I will teach you multivariate calculus
- Need to know linear algebra
  - I assume you know about vector, matrices and dot products
  - I will teach you more advanced topics, like singular value decompositions
- Need to have learned a bit about probability
- Concerns for final: will be testing only fundamentals, not intended to be difficult to finely separate students

PROBABILITY THEORY IS THE SCIENCE OF PREDICTIONS*

- The **goal of science** is to discover theories that can be used to predict how natural processes evolve or explain natural phenomenon, based on observed phenomenon.

- The **goal of probability theory** is to provide the foundation to build theories (= models) that can be used to reason about the outcomes of events, future or past, based on observations.
  - prediction of the unknown which may depend on what is observed and whose nature is probabilistic

*Quote from Csaba Szepesvari, https://eclass.srv.ualberta.ca/pluginfile.php/1136251/mod_resource/content/1/LectureNotes_Probabilities.pdf

(Measurable) Space of outcomes and events

Ω = sample space, all outcomes of the experiment
$\mathcal{F}$ = event space, set of subsets of Ω

Ω and $\mathcal{F}$ must be non-empty
$\mathcal{F}$ has been changed to $\mathcal{E}$ in the notes

If the following conditions hold:

1. $A \in \mathcal{F} \quad \Rightarrow \quad A^c \in \mathcal{F}$
2. $A_1, A_2, \ldots \in \mathcal{F} \quad \Rightarrow \quad \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

$\mathcal{F}$ is an event space

Note: terminology sigma field sounds technical, but it just means this event space

$$(\Omega, \mathcal{F}) = \text{a measurable space}$$

Why is this the Definition?

Intuitively,

1. A collection of outcomes is an event (e.g., either a 1 or 6 was rolled)
2. If we can measure two events separately, then their union should also be a measurable event
3. If we can measure an event, then we should be able to measure that that event did not occur (the complement)

Ω = sample space, all outcomes of the experiment
$\mathcal{F}$ = event space, set of subsets of Ω

If the following conditions hold:

1. $A \in \mathcal{F} \quad \Rightarrow \quad A^c \in \mathcal{F}$
2. $A_1, A_2, \ldots \in \mathcal{F} \quad \Rightarrow \quad \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

Axioms of Probability

$(\Omega, \mathcal{F})$ = a measurable space

Any function $P : \mathcal{F} \to [0, 1]$ such that

1. (unit measure) $P(\Omega) = 1$
2. (σ-additivity) Any countable sequence of disjoint events $A_1, A_2, \ldots \in \mathcal{F}$ satisfies $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

is called a probability measure (probability distribution)

$(\Omega, \mathcal{F}, P)$ = a probability space

A FEW COMMENTS ON TERMINOLOGY

- A few new terms, including countable, closure
    - only a small amount of terminology used, can google these terms and learn on your own
    - notation sheet in notes
- Countable: integers, {0.1,2.0,3.6},…
- Uncountable: real numbers, intervals, …
- Interchangeably use (though its somewhat loose)
    - discrete and countable
    - continuous and uncountable

Sample Spaces

$$\Omega$$

discrete (countable)

$\Omega = \{1, 2, 3, 4, 5, 6\}$

$\Omega = \mathbb{N}$

$e.g., \mathcal{F} = \{\emptyset, \{1,2\}, \{3,4,5,6\}, \{1,2,3,4,5,6\}\}$

Typically: $\mathcal{F} = \mathcal{P}(\Omega)$

Power set

continuous (uncountable)

$\Omega = [0, 1]$

$\Omega = \mathbb{R}$

$e.g., \mathcal{F} = \{\emptyset, [0, 0.5], (0.5, 1.0], [0, 1]\}$

Typically: $\mathcal{F} = \mathcal{B}(\Omega)$

Borel field

$\Omega = [0, 1] \cup \{2\} = \text{mixed space}$

Finding Probability Distributions

$(\Omega, \mathcal{F})$ = a measurable space

Example: $\Omega = \{0, 1\}$
$\mathcal{F} = \{\emptyset, \{0\}, \{1\}, \Omega\}$

$$P(A) = \begin{cases} 1-\alpha & A = \{0\} \\ \alpha & A = \{1\} \\ 0 & A = \emptyset \\ 1 & A = \Omega \end{cases} \qquad \alpha \in [0, 1]$$

How can we choose P in practice?

Clearly, we cannot do it arbitrarily.

How can we satisfy all constraints?

Probability Mass Functions

Ω = discrete sample space
$\mathcal{F} = \mathcal{P}(\Omega)$

Probability mass function:

1. $p : \Omega \to [0, 1]$
2. $\sum_{\omega \in \Omega} p(\omega) = 1$

The probability of any event $A \in \mathcal{F}$ is defined as

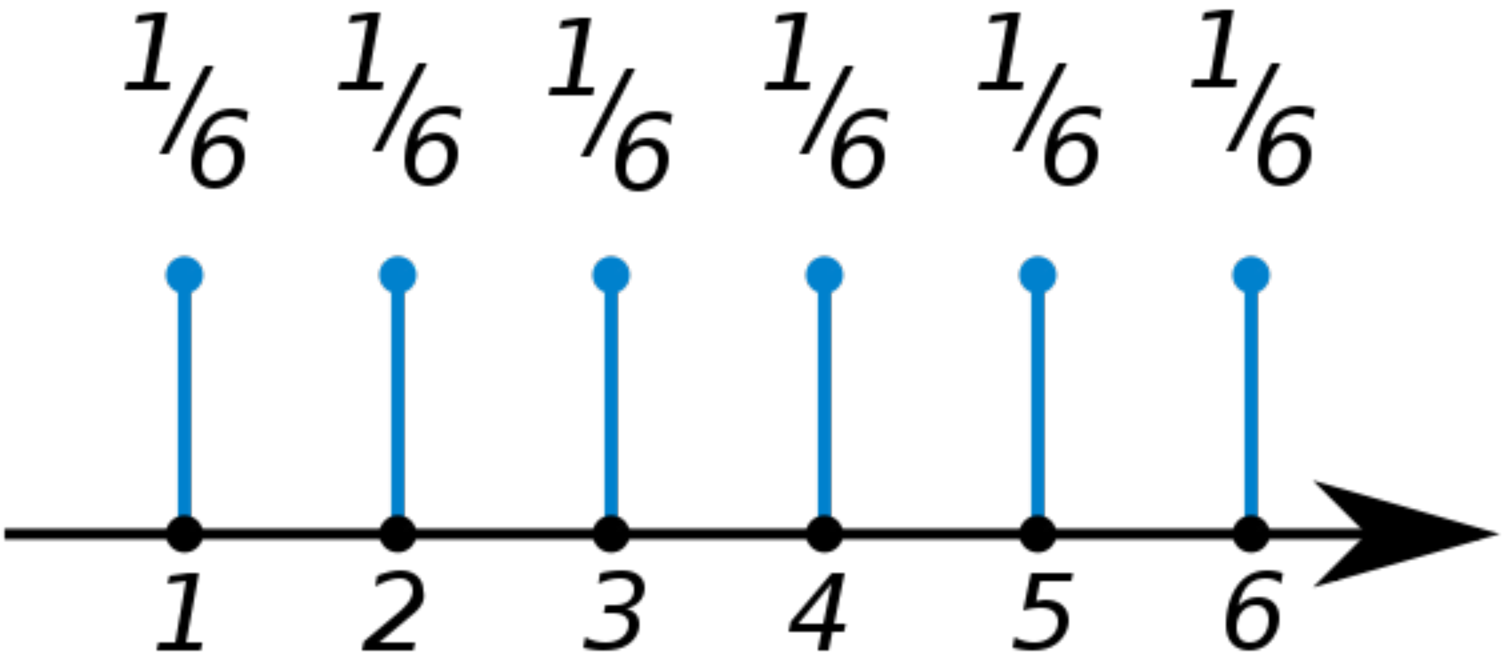
$$P(A) = \sum_{\omega \in A} p(\omega)$$

ARBITRARY PMFS

e.g. PMF for a fair die (table of values)

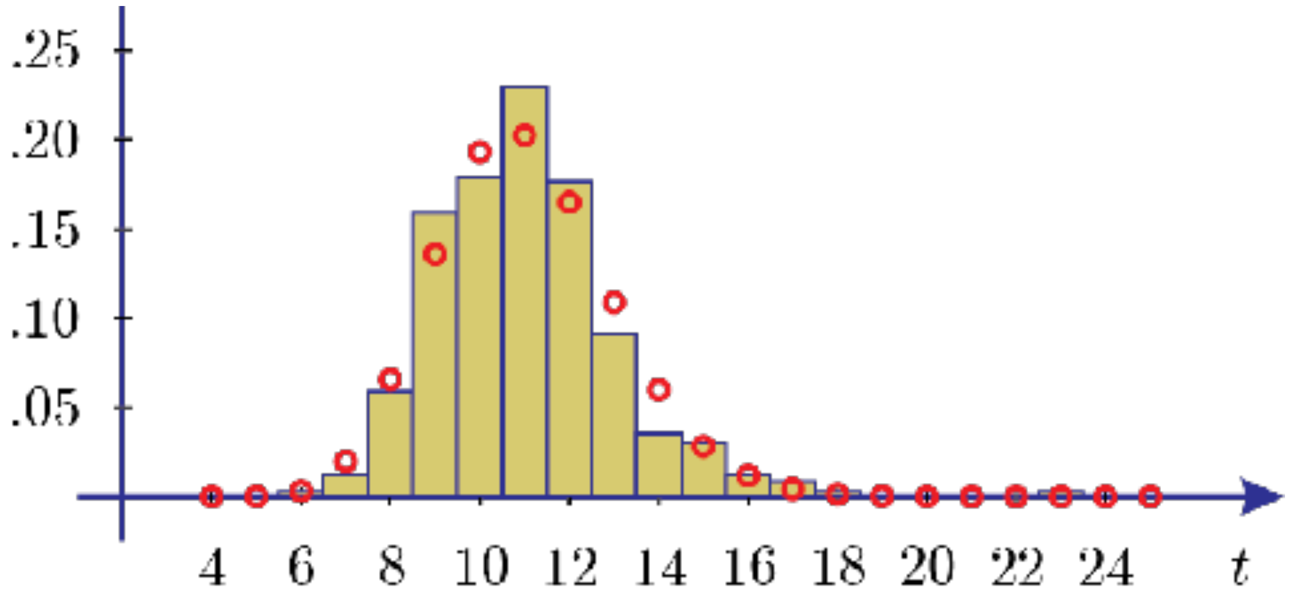
$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

$$p(\omega) = 1/6 \quad \forall \omega \in \Omega$$



EXERCISE: HOW ARE PMFS USEFUL AS A MODEL?

- Recall we wanted to model commute times
- We could use a probability table for minutes: count number of times t = 1, 2, 3, … occurs and then normalize probabilities by # samples
- Pick t with the largest p(t)



Useful PMFs

Bernoulli distribution: $\Omega = \{S, F\} \quad \alpha \in (0, 1)$

$$p(\omega) = \begin{cases} \alpha & \omega = S \\ 1 - \alpha & \omega = F \end{cases}$$

Alternatively, $\Omega = \{0, 1\}$

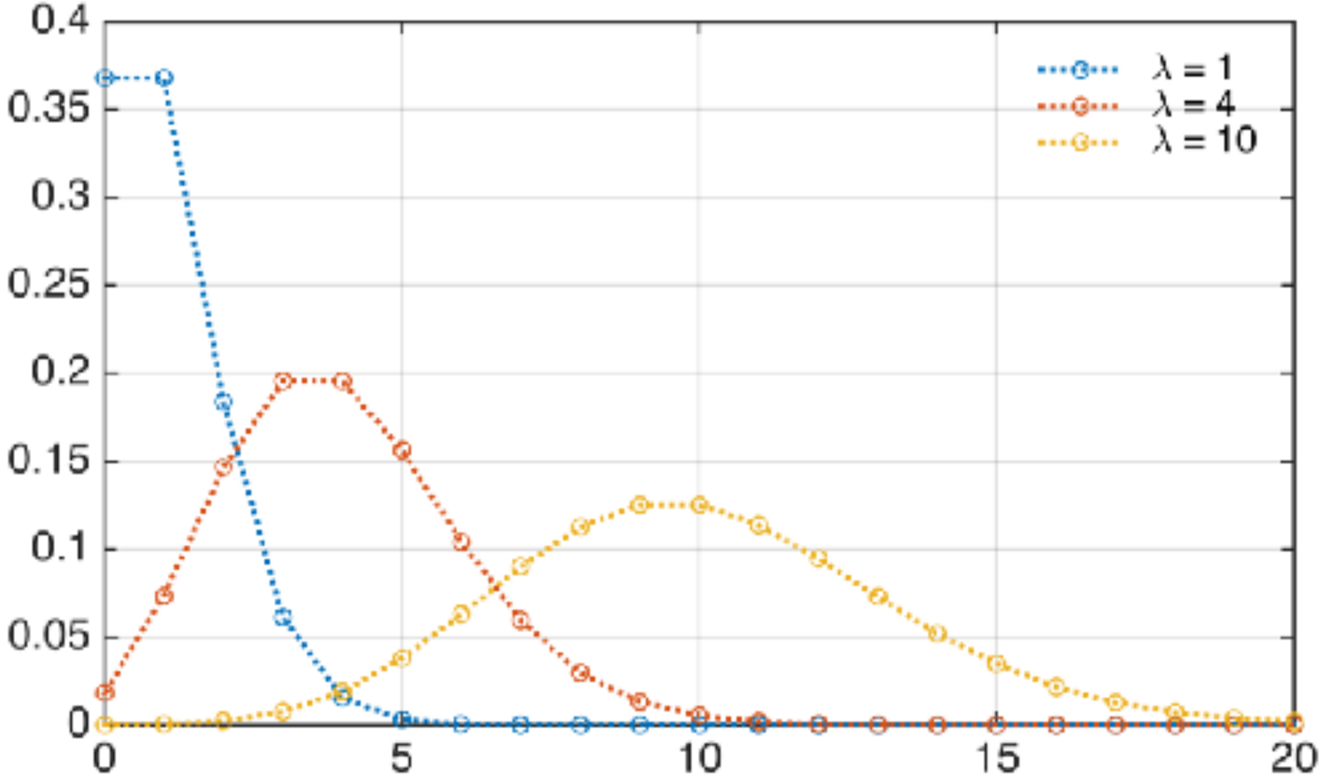
$$p(k) = \alpha^k \cdot (1 - \alpha)^{1-k} \qquad \forall k \in \Omega$$

USEFUL PMFs

Poisson distribution: $\Omega = \{0, 1, \ldots\}$ $\lambda \in (0, \infty)$

e.g., amount of mail received in a day
number of calls received by call center in an hour

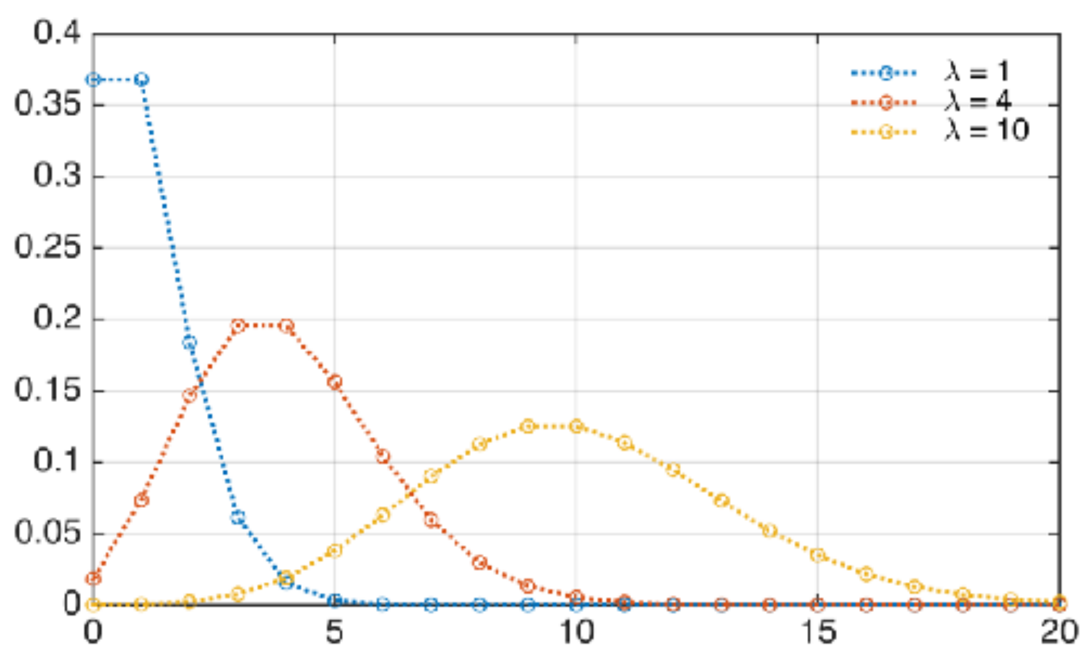
$$p(k) = \frac{\lambda^k e^{-\lambda}}{k!} \qquad \forall k \in \Omega$$

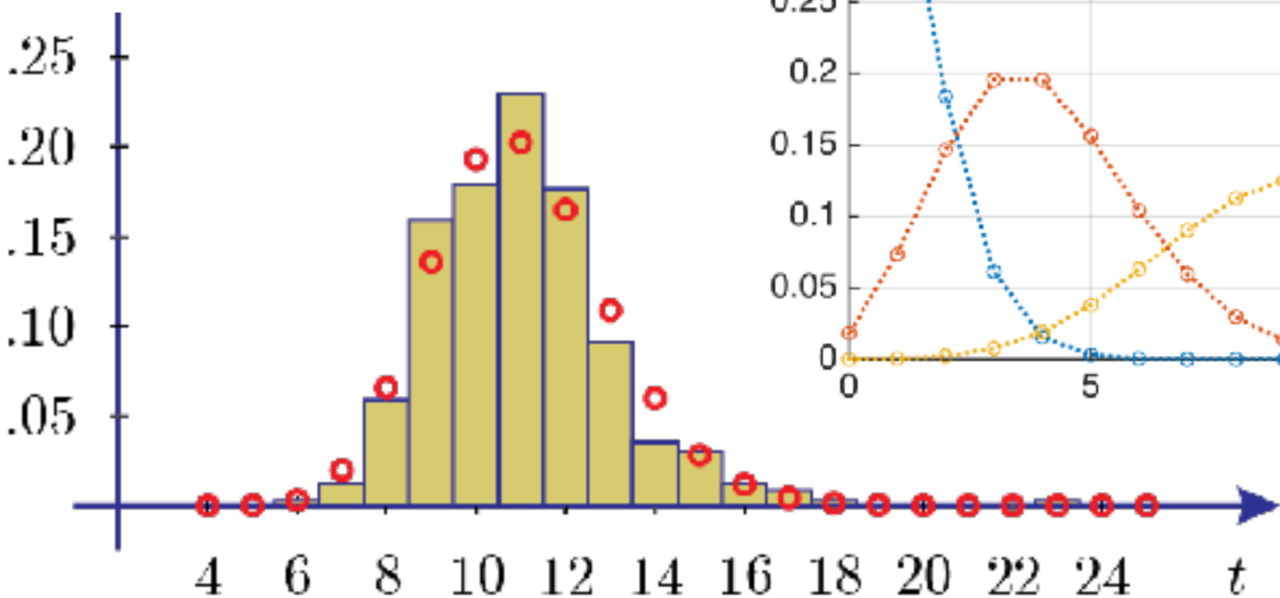


EXERCISE: CAN WE USE A POISSON FOR COMMUTE TIMES?

- Used a probability table (histogram) for minutes: count number of times t = 1, 2, 3, … occurs and then normalize probabilities by # samples
- Can we use a Poisson?

$$p(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$





Probability Density Functions

Ω = continuous sample space
$\mathcal{F} = \mathcal{B}(\Omega)$

Probability density function:

1. $p : \Omega \to [0, \infty)$
2. $\int_\Omega p(\omega) d\omega = 1$

The probability of any event $A \in \mathcal{F}$ is defined as

$$P(A) = \int_A p(\omega) d\omega.$$

PMFs vs. PDFs

Ω = discrete sample space

Consider a singleton event $\{\omega\} \in \mathcal{F}$, where $\omega \in \Omega$

$$P(\{\omega\}) = p(\omega)$$

Ω = continuous sample space

Example:

- Stopping time of a car, in interval [3,15]. What is the probability of seeing a stopping time of exactly 3.141596? (How much mass in [3,15]?)
- More reasonable to ask the probability of stopping between 3 to 3.5 seconds.

PMFs vs. PDFs

$\Omega =$ discrete sample space

Consider a singleton event $\{\omega\} \in \mathcal{F}$, where $\omega \in \Omega$

$$P(\{\omega\}) = p(\omega)$$

$\Omega =$ continuous sample space

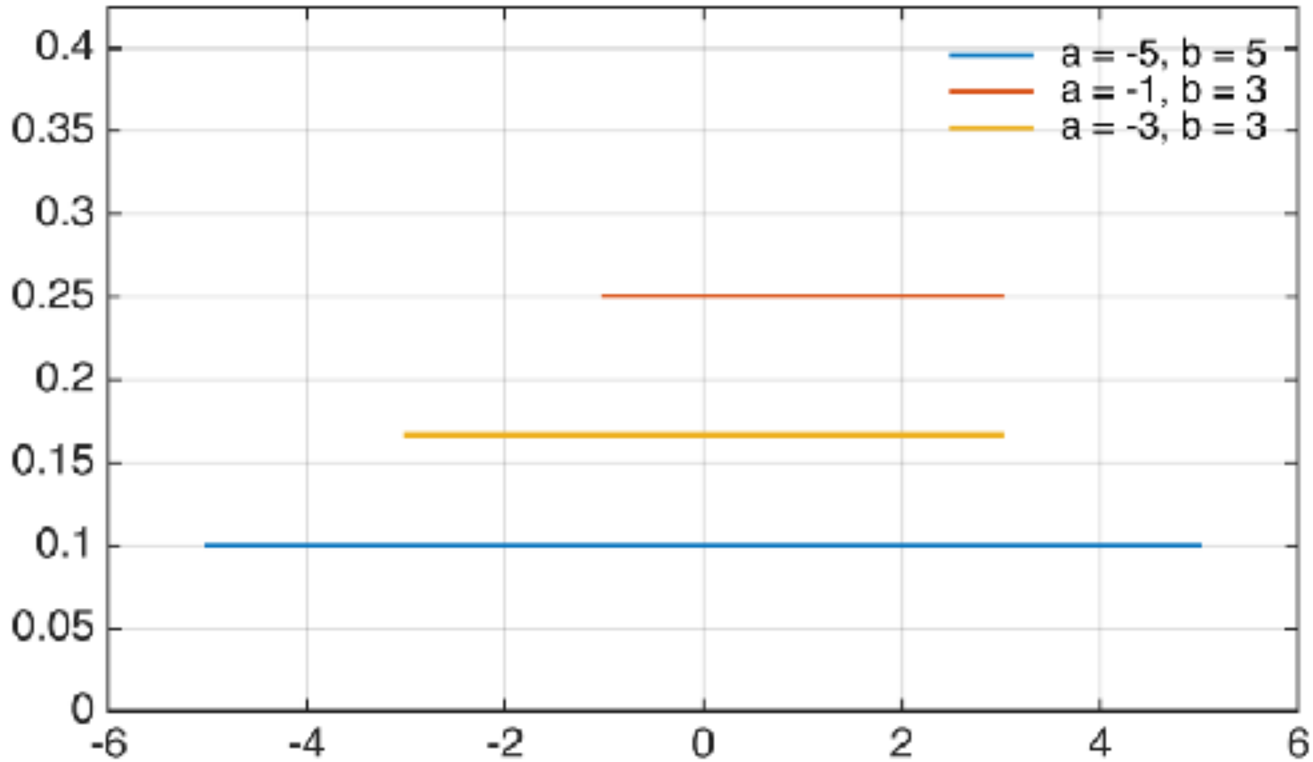
Consider an interval event $A = [x, x + \Delta x]$, where Δ is small

$$P(A) = \int_x^{x+\Delta x} p(\omega) d\omega$$
$$\approx p(x)\Delta x$$

Useful PDFs

Uniform distribution: $\Omega = [a, b]$

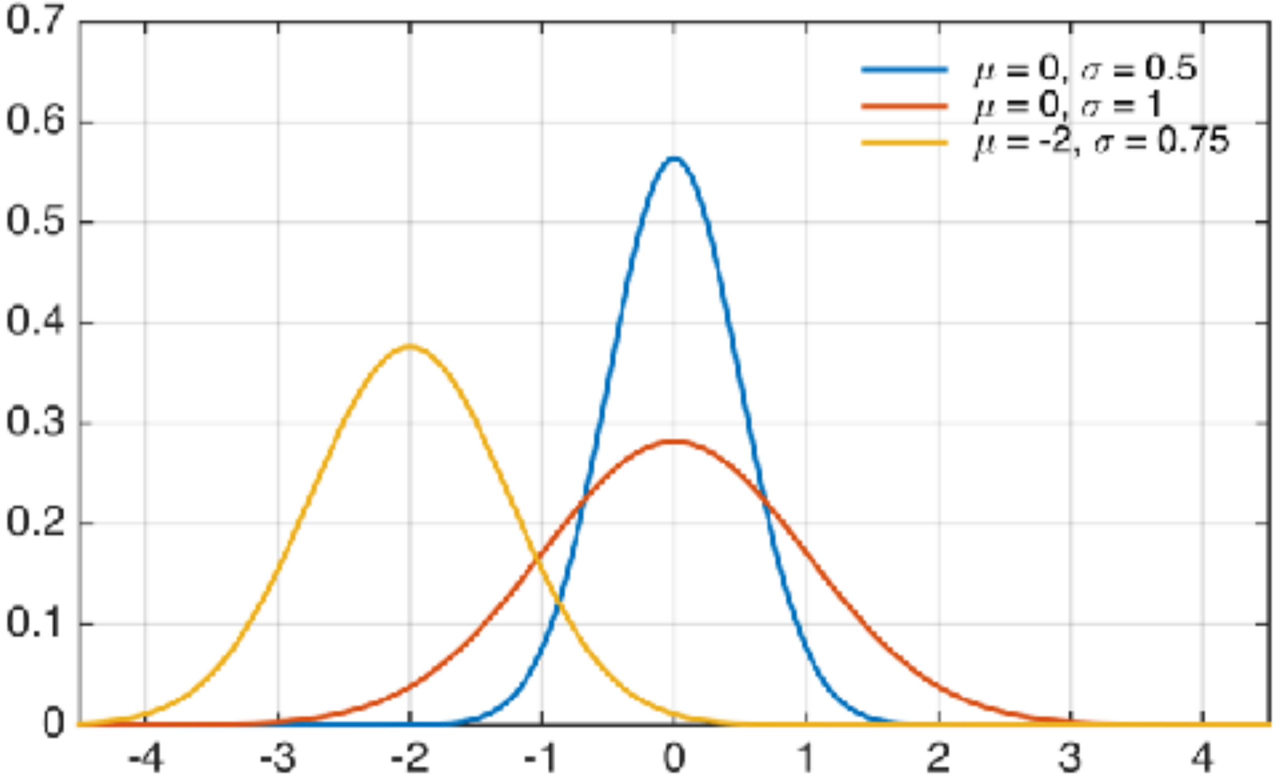
$$p(\omega) = \frac{1}{b-a} \qquad \forall \omega \in [a, b]$$



Useful PDFs

Gaussian distribution: $\Omega = \mathbb{R}$ $\mu \in \mathbb{R},\ \sigma \in \mathbb{R}^+$

$$p(\omega) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\omega-\mu)^2} \qquad \forall \omega \in \mathbb{R}$$

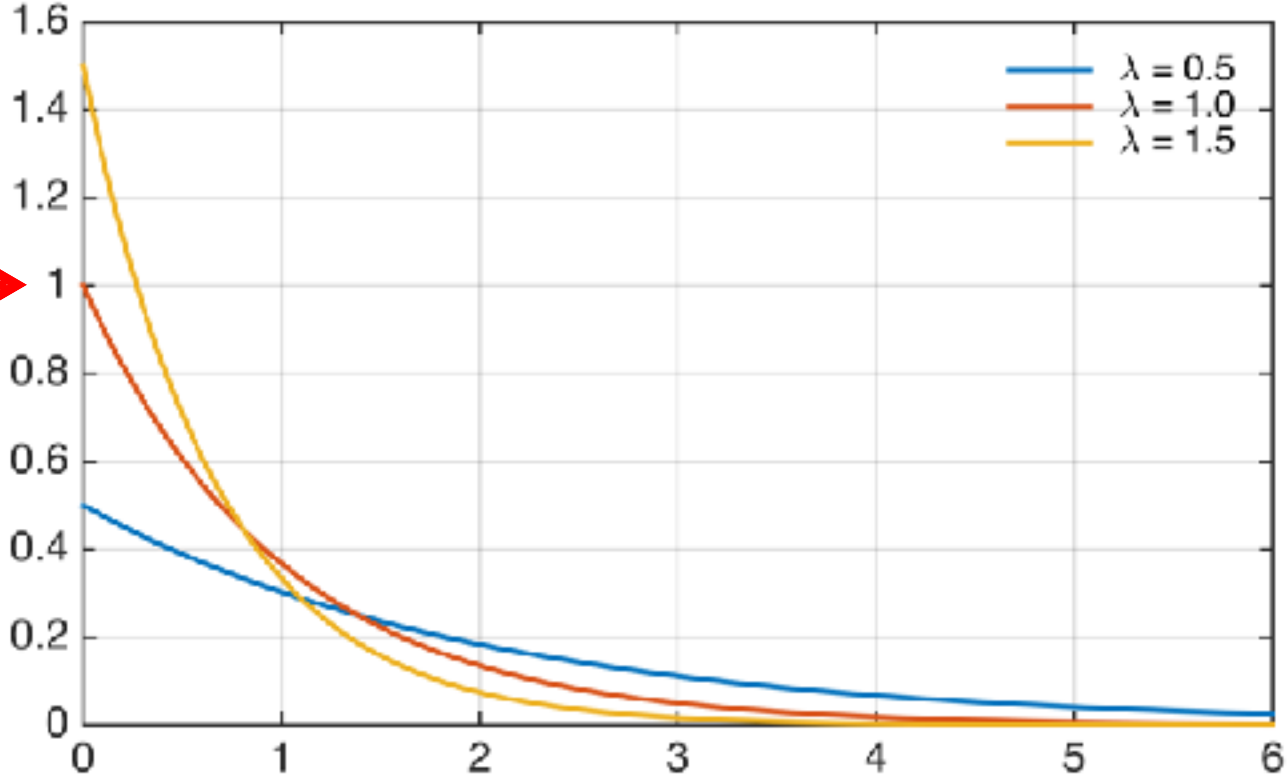


USEFUL PDFS

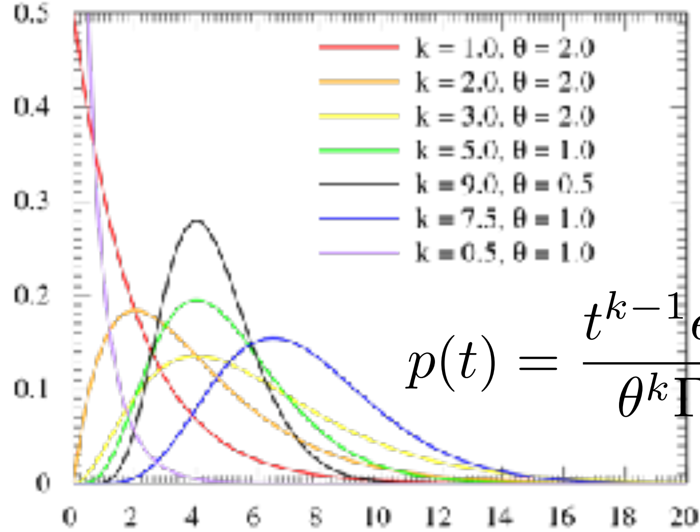
Exponential distribution: $\Omega = [0, \infty) \quad \lambda > 0$

$$p(\omega) = \lambda e^{-\lambda\omega} \qquad \forall \omega \geq 0$$

1 is here!

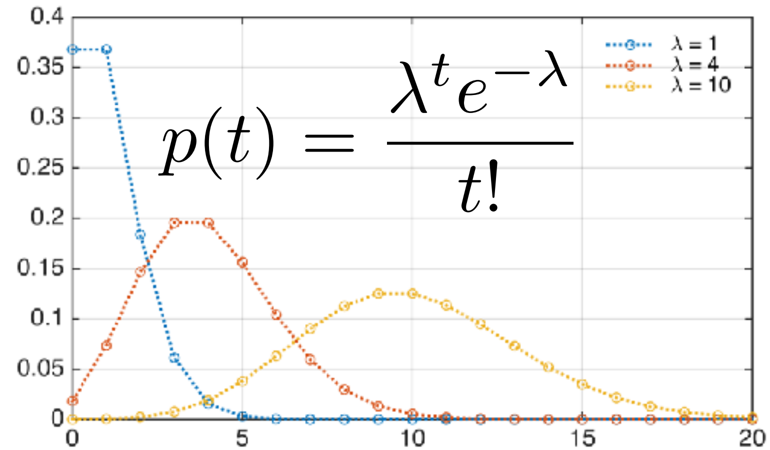


EXERCISE: MODELING COMMUTE TIMES



$$p(t) = \frac{t^{k-1} e^{-t/\theta}}{\theta^k \Gamma(k)}$$

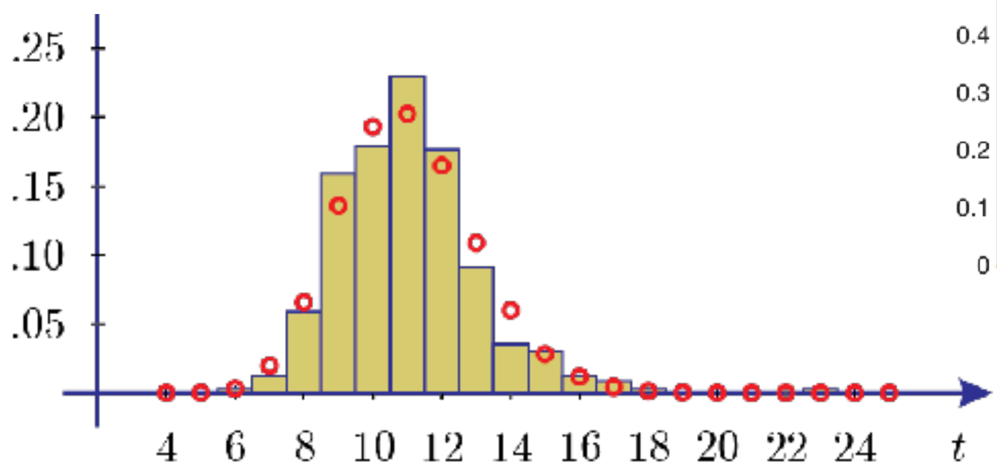
Gamma

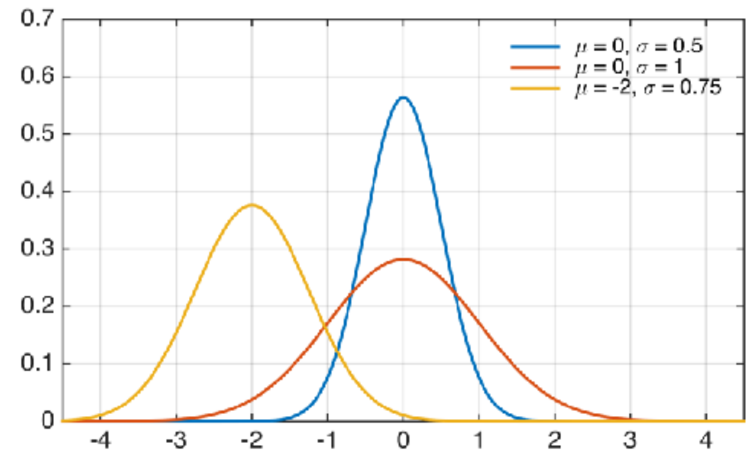


$$p(t) = \frac{\lambda^t e^{-\lambda}}{t!}$$

Poisson

Which might you choose?





Gaussian

$$p(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}$$

Random Variables

$(\Omega, \mathcal{F}, P)$

Age: 35 **Likes sports:** Yes
Height: 1.85m **Smokes:** No
Weight: 75kg **Marital st.:** Single
IQ: 104 **Occupation:** Musician

Ω

Age: 26 **Likes sports:** Yes
Height: 1.75m **Smokes:** No
Weight: 79kg **Marital st.:** Divorced
IQ: 103 **Occupation:** Athlete

$$A = \{\omega \in \Omega : Musician(\omega) = \text{yes}\}$$

Musician is a random variable (a function)
A is the new event space
Can ask P(M = 0) and P(M = 1)

We instinctively create this transformation

Assume Ω is a set of people.

Compute the probability that a randomly selected person $\omega \in \Omega$ has a cold.

Define event $A = \{\omega \in \Omega : \text{Disease}(\omega) = \text{cold}\}$.

Disease is our new random variable, $P(Disease = cold)$

Disease is a function that maps outcome space to new outcome space $\{\text{cold}, \text{not cold}\}$

Disease is a function, which is neither a variable nor random
BUT, this term is still a good one since we treat Disease as a variable
And assume it can take on different values
(randomly according to some distribution)

Random Variable: Formal Definition

$(\Omega, \mathcal{F}, P)$ = a probability space

Random variable:

1. $X : \Omega \to \Omega_X$
2. $\forall A \in \mathcal{B}(\Omega_X)$ it holds that $\{\omega : X(\omega) \in A\} \in \mathcal{F}$

It follows that: $P_X(A) = P(\{\omega : X(\omega) \in A\})$

Example $X : \Omega \to [0, \infty)$

Ω is set of (measured) people in population

with associated measurements such as height and weight

$X(\omega) =$ height

$A =$ interval $= [5'1'', 5'2'']$

$P(X \in A) = P(5'1'' \leq X \leq 5'2'') = P(\{\omega : X(\omega) \in A\})$

5 Minute Break and Exercise

- Let X be a random variable that corresponds to the ratio of hard-to-easy problems on an assignment. Assume it takes values in {0.1, 0.25, 0.7}. Is this discrete or continuous? Does it have a PMF or PDF? Further, where could the variability come from? i.e., why is this a random variable?
- Let X be the stopping time of a car, taking values in [3,5] union [7,9]. Is this discrete or continuous?
- Think of an example of a discrete random variable (RV) and a continuous RV
- We provided several named PMFs. Why do we use these explicit functional forms? Why not just tables of values, which is more flexible?

What if we have more than two variables…

- So far, we have considered scalar random variables
- Axioms of probability defined abstractly, apply to vector random variables

$$\Omega = \text{sample space, all outcomes of the experiment}$$
$$\mathcal{F} = \text{event space, set of subsets of } \Omega$$

$$\Omega = \mathbb{R}^2, \text{e.g.}, \ \omega = [-0.5, 10]$$

$$\Omega = [0, 1] \times [2, 5], \text{e.g.}, \ \omega = [0.2, 3.5]$$

But, when defining probabilities, we will want to consider how the variables interact

Two discrete random variables

Random variables X and Y
Outcome spaces $\mathcal{X}$ and $\mathcal{Y}$

$$p(x,y) = P(X = x, Y = y)$$

$$\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) = 1.$$

$$\mathcal{X} = \{\text{young}, \text{old}\} \text{ and } \mathcal{Y} = \{\text{no arthritis}, \text{arthritis}\}$$

		Y = 0	Y = 1
X	0	$1/2$	$1/100$
	1	$1/10$	$39/100$

Some questions we might ask now that we have two random variables

$$\mathcal{X} = \{\text{young}, \text{old}\} \text{ and } \mathcal{Y} = \{\text{no arthritis}, \text{arthritis}\}$$

		Y = 0	Y = 1
X	0	$1/2$	$1/100$
	1	$1/10$	$39/100$

Are these two variables related?
Or do they change completely independently of each other?

Given this joint distribution, can we determine just the distribution over arthritis? i.e., P(Y = 1)? (Marginal distribution)

If we knew something about one of the variables, say that the person Is young, do we now the distribution over Y? (Conditional distribution)

Example: Marginal and Conditional Distribution

$$\mathcal{X} = \{\text{young}, \text{old}\} \text{ and } \mathcal{Y} = \{\text{no arthritis}, \text{arthritis}\}$$

		Y = 0	Y = 1
X	0	$1/2$	$1/100$
	1	$1/10$	$39/100$

P(Y = 1) = P(Y = 1, X = 0) + P(Y = 1, X = 1) = 40/100
What is P(Y = 0)?

P(X = 1) = 49/100

P(Y = 1 | X = 0) = ?
Is it 1/100, where the table tells us P(Y = 1, X=0)?

CONDITIONAL DISTRIBUTIONS

Conditional probability distribution:

$$p(y|x) = \frac{p(x, y)}{p(x)}$$

The probability of an event A, given that $X = x$, is:

$$P(Y \in A | X = x) = \begin{cases} \sum_{y \in A} p(y|x) & Y : \text{discrete} \\ \int_A p(y|x) dy & Y : \text{continuous} \end{cases}$$

Exercise: Conditional Distribution

$$\mathcal{X} = \{\text{young}, \text{old}\} \text{ and } \mathcal{Y} = \{\text{no arthritis}, \text{arthritis}\}$$

		$Y=0$	$Y=1$
X	0	$1/2$	$1/100$
	1	$1/10$	$39/100$

$$p(y|x) = \frac{p(x,y)}{p(x)}$$

P(Y = 1 | X = 0) = ?

What is P(Y = 0 | X = 0)?
Should P(Y = 1 | X = 0) + P(Y = 0 | X = 0) = 1?

JOINT DISTRIBUTIONS FOR MANY VARIABLES

In general, we can consider d-dimensional random variable $\boldsymbol{X} = (X_1, X_2, \ldots, X_d)$ with vector-valued outcomes $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)$, such that each x_i is chosen from some $\mathcal{X}_i$. Then, for the discrete case, any function $p : \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_d \to [0, 1]$ is called a multidimensional probability mass function if

$$\sum_{x_1 \in \mathcal{X}_1} \sum_{x_2 \in \mathcal{X}_2} \cdots \sum_{x_d \in \mathcal{X}d} p\left(x_1, x_2, \ldots, x_d\right) = 1.$$

or, for the continuous case, $p : \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_d \to [0, \infty]$ is a multidimensional probability density function if

$$\int_{\mathcal{X}_1} \int_{\mathcal{X}_2} \cdots \int_{\mathcal{X}d} p\left(x_1, x_2, \ldots, x_d\right) dx_1 dx_2 \ldots dx_d = 1.$$

MARGINAL DISTRIBUTIONS

A *marginal distribution* is defined for a subset of $\boldsymbol{X} = (X_1, X_2, \ldots, X_d)$ by summing or integrating over the remaining variables. For the discrete case, the marginal distribution $p\left(x_i\right)$ is defined as

$$p\left(x_i\right) = \sum_{x_1 \in \mathcal{X}_1} \cdots \sum_{x_{i-1} \in \mathcal{X}_{i-1}} \sum_{x_{i+1} \in \mathcal{X}_{i+1}} \cdots \sum_{x_d \in \mathcal{X}_d} p\left(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_d\right),$$

where the variable x_i is fixed to some value and we sum over all possible values of the other variables. Similarly, for the continuous case, the marginal distribution $p\left(x_i\right)$ is defined as

$$p\left(x_i\right) = \int_{\mathcal{X}_1} \cdots \int_{\mathcal{X}_{i-1}} \int_{\mathcal{X}_{i+1}} \cdots \int_{\mathcal{X}_d} p\left(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_d\right) dx_1 \ldots dx_{i-1} dx_{i+1} \ldots dx_d.$$

Natural question: Why do you use p for p(xi) and for p(x1, …., xd)?
They have different domains, they can’t be the same function!

Dropping subscripts

Instead of:

$$p_{Y|X}(y|x) = \frac{p_{XY}(x,y)}{p_X(x)}$$

We will write:

$$p(y|x) = \frac{p(x,y)}{p(x)}$$

Another example for conditional distributions

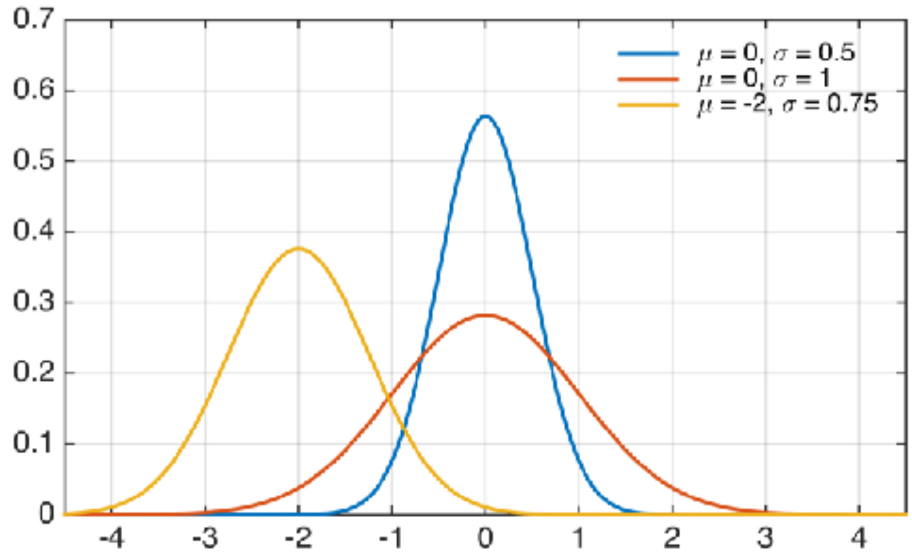
- Let X be a Bernoulli random variable (i.e., 0 or 1 with probability alpha)
- Let Y be a random variable in {10, 11, …, 1000}
- $p(y \mid X = 0)$ and $p(y \mid X = 1)$ are different distributions
- Two types of books: fiction (X=0) and non-fiction (X=1)
- Let Y corresponds to number of pages
- Distribution over number of pages different for fiction and non-fiction books (e.g., average different)

Example continued

- Two types of books: fiction ($X=0$) and non-fiction ($X=1$)
- Y corresponds to number of pages
- $p(y \mid X = 0) = p(X = 0, y)/p(X = 0)$
- $p(X = 0 , y)$ = probability that a book is fiction and has y pages (imagine randomly sampling a book)
- $p(X = 0)$ = probability that a book is fiction
- If most books are non-fiction, $p(X = 0, y)$ is small even if y is a likely number of pages for a fiction book
- $p(X = 0)$ accounts for the fact that joint probability small if $p(X = 0)$ is small

Another Example

- Two types of books: fiction (X=0) and non-fiction (X=1)
- Let Y be a random variable over the reals, which corresponds to amount of money made
- p(y | X = 0) and p(y | X = 1) are different distributions
- e.g., even if both p(y | X = 0) and p(y | X = 1) are Gaussian, they likely have different means and variances

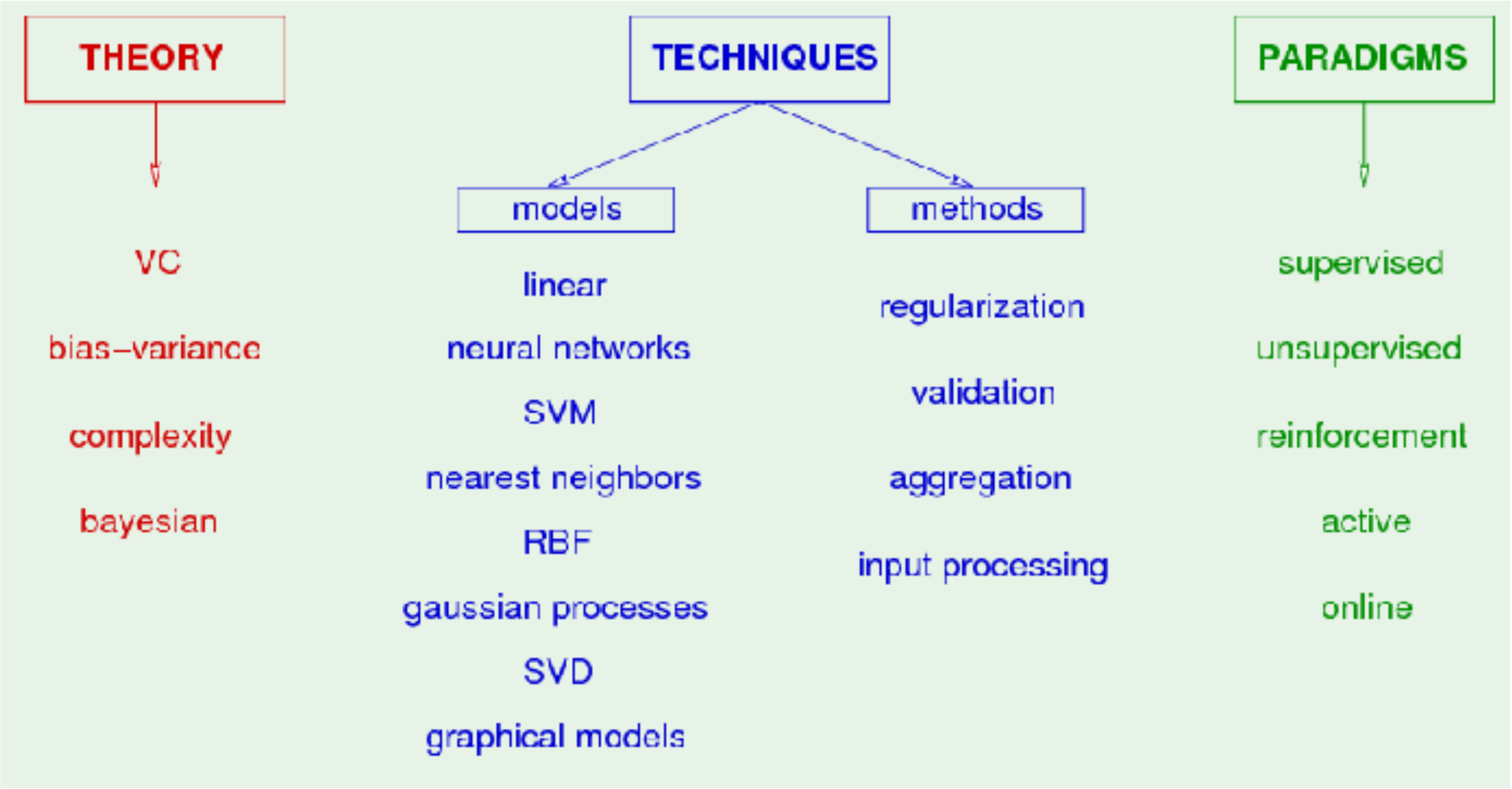


What do we know about p(y)?

- We know p(y | x)
- We know marginal p(x)
- Correspondingly we know p(x, y) = p(y | x) p(x)
  - from conditional probability definition that p(y | x) = p(x, y) / p(x)
- What is the marginal p(y)?

$$
\begin{aligned}
p(y) &= \sum_x p(x, y) \\
&= \sum_x p(y|x)p(x) \\
&= p(y|X=0)p(X=0) + p(y|X=1)p(X=1)
\end{aligned}
$$

Sept 12: Probability review continued



Machine learning topic overview
* from Yaser Abu-Mostafa, https://work.caltech.edu/library/

REMINDERS

- Assignment 1 (September 28),
  - small typo fixes (bold X, and range of lambda to [0, infty) rather than (0, infty))
  - “express in terms of givens a, b, c” does not mean you have to use all a, b and c, but that the final expression should include (some subset) of these given values
- Office hours and labs start this week
  - Martha: 3-5 p.m. on Tuesday (ATH 3-05)
  - Labs: W 5-8 p.m. and F 2-5 p.m.

Chain Rule

Conditional probability distribution:

$$p(x_k|x_1,\ldots,x_{k-1}) = \frac{p(x_1,\ldots,x_k)}{p(x_1,\ldots,x_{k-1})}$$

This leads to:

$$p(x_1,\ldots,x_k) = p(x_1)\prod_{l=2}^{k} p(x_l|x_1,\ldots,x_{l-1})$$

Two variable example $p(x,y) = p(x|y)p(y) = p(y|x)p(x)$

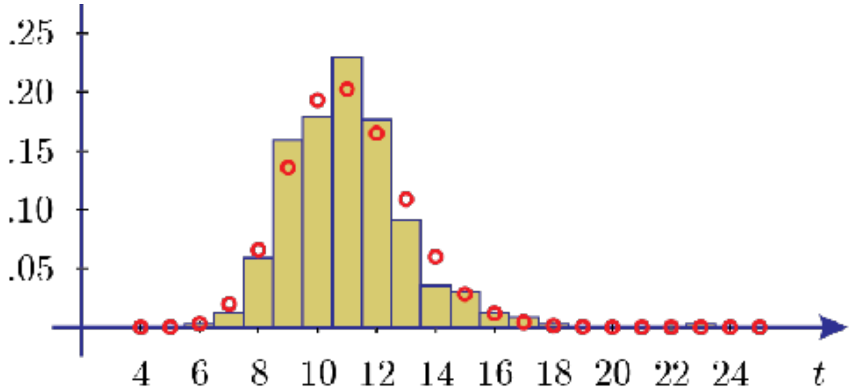
How do we get Bayes rule?

Recall chain rule: $p(x, y) = p(x|y)p(y) = p(y|x)p(x)$

Bayes rule: $$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

EXERCISE: CONDITIONAL PROBABILITIES

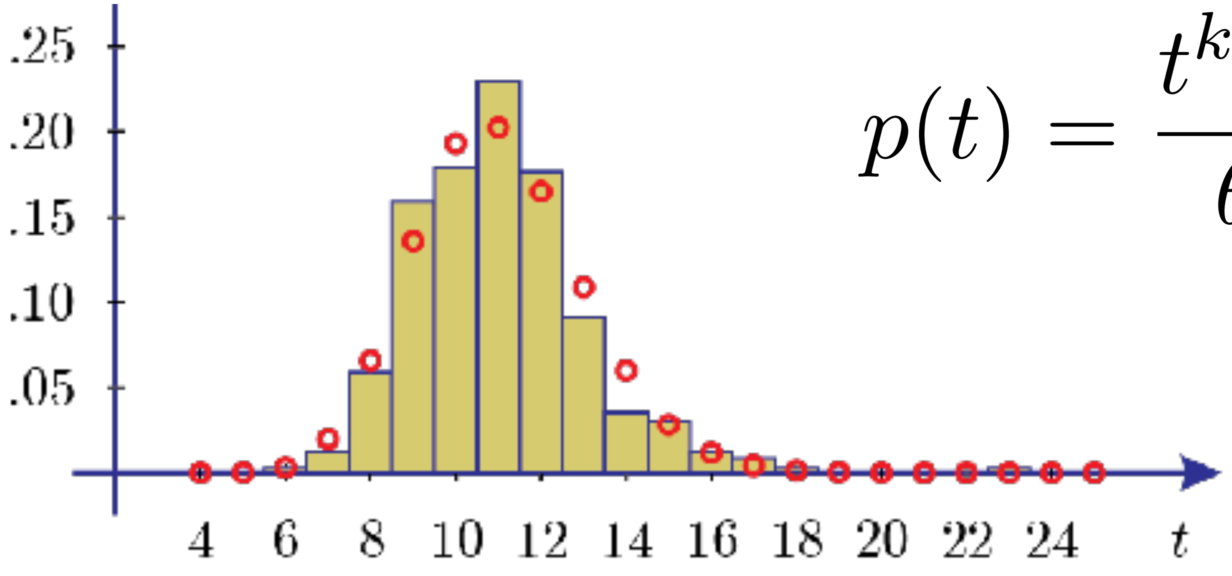
- Using conditional probabilities, we can incorporate other external information (features)
- Let y be the commute time, x the day of the year
- Array of conditional probability values —> p(y | x)
    - y = 1, 2, … and x = 1, 2, …, 365
- What are some issues with this choice for x?
- What other x could we use feasibly?



EXERCISE: ADDING IN AUXILIARY INFORMATION

- Gamma distribution for commute times extrapolates between recorded time in minutes
- Can incorporate external information (features) by modeling theta = function(features)

$$\theta = \sum_{i=1}^{d} w_i x_i$$



$$p(t) = \frac{t^{k-1} e^{-t/\theta}}{\theta^k \Gamma(k)}$$

Independence of Random Variables

X and Y are **independent** if:

$$p(x, y) = p(x)p(y)$$

X and Y are **conditionally independent** given Z if:

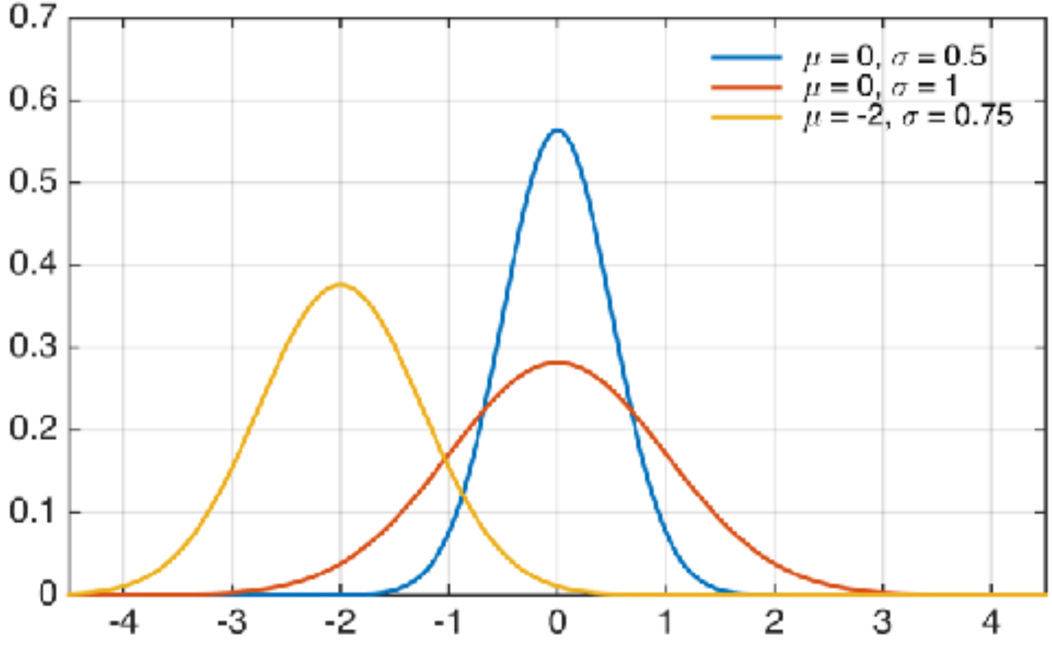
$$p(x, y|z) = p(x|z)p(y|z)$$

Conditional Independence Examples
Example 7 in the notes

- Imagine you have a biased coin (does not flip 50% heads and 50% tails, but skewed towards one)
- Let Z = bias of a coin (say outcomes are 0.3, 0.5, 0.8 with associated probabilities 0.7, 0.2, 0.1)
    - what other outcome space could we consider?
    - what kinds of distributions?
- Let X and Y be consecutive flips of the coin
- Are X and Y independent?
- Are X and Y conditionally independent, given Z?

Expected value (Mean)

$$\mathbb{E}[X] = \begin{cases} \sum_{x \in \mathcal{X}} x p(x) & X : \text{discrete} \\ \\ \int_{\mathcal{X}} x p(x) dx & X : \text{continuous} \end{cases}$$



Expectations with functions

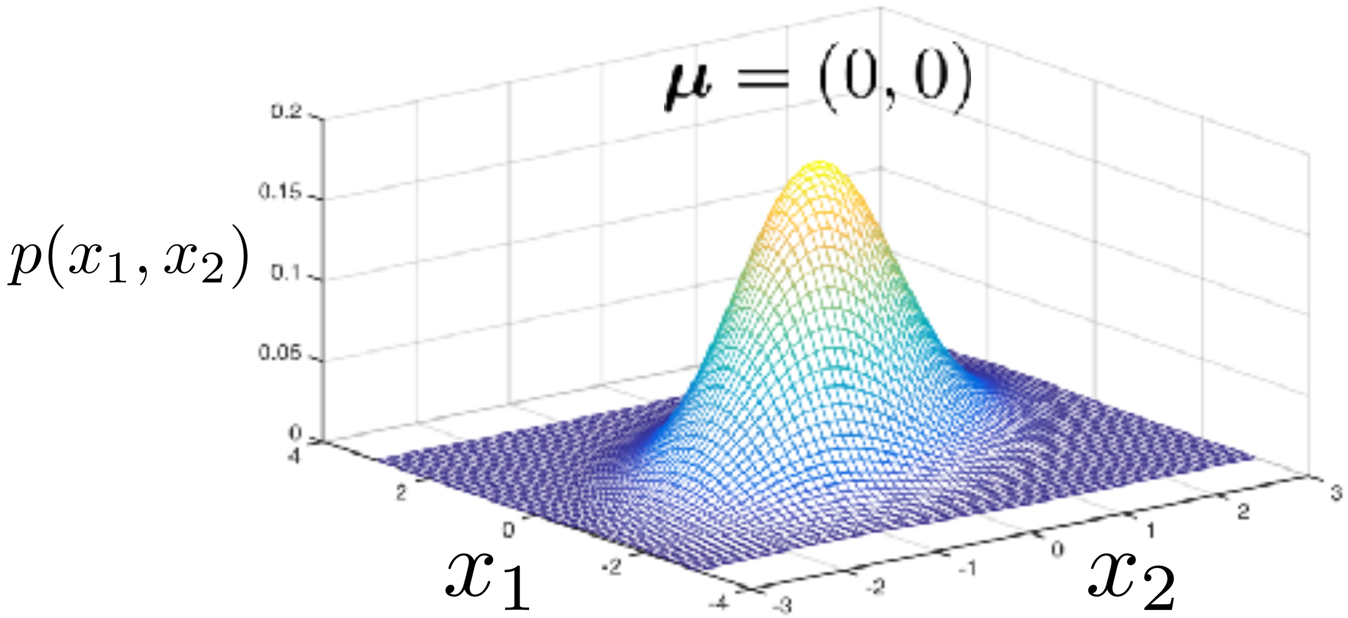
$$f : \mathcal{X} \to \mathbb{R}$$

$$\mathbb{E}\left[f(X)\right] = \begin{cases} \sum_{x \in \mathcal{X}} f(x)p(x) & X : \text{discrete} \\ \\ \int_{\mathcal{X}} f(x)p(x)dx & X : \text{continuous} \end{cases}$$

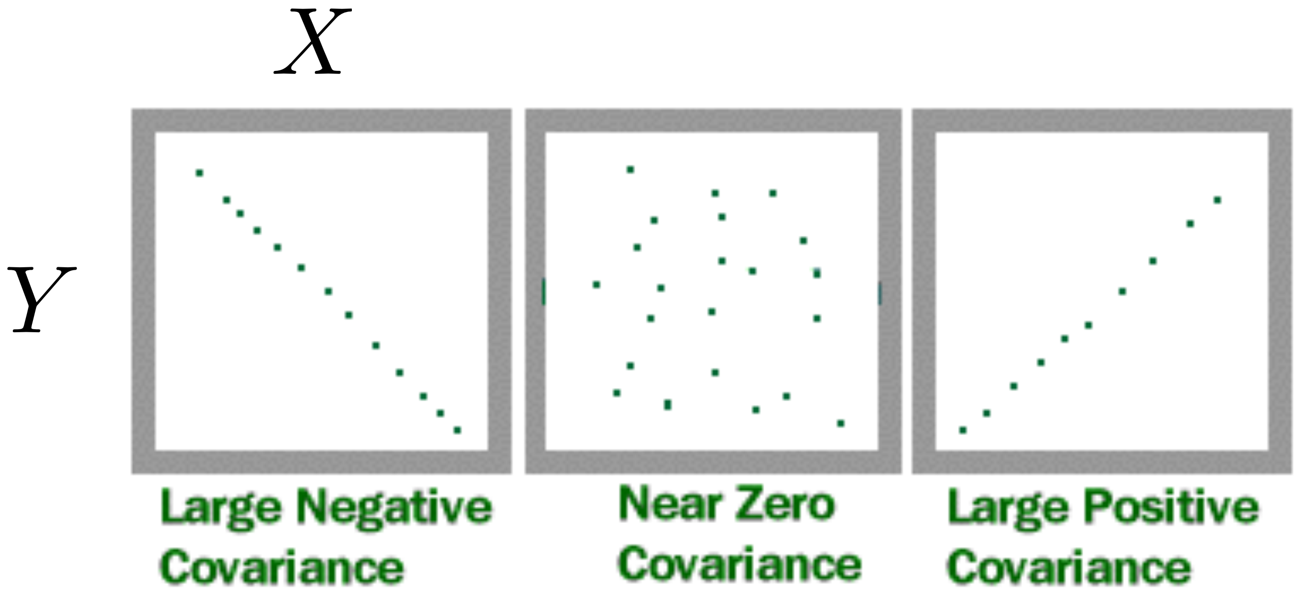
Expected value for multivariate

$$\mathbb{E}\left[\boldsymbol{X}\right] = \begin{cases} \sum_{\boldsymbol{x}\in\mathcal{X}} \boldsymbol{x}p(\boldsymbol{x}) & \boldsymbol{X}: \text{discrete} \\ \\ \int_{\mathcal{X}} \boldsymbol{x}p(\boldsymbol{x})d\boldsymbol{x} & \boldsymbol{X}: \text{continuous} \end{cases}$$

Each instance x is a vector, p is a function on these vectors



Covariance



$$\begin{aligned}\mathrm{Cov}[X, Y] &= \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y],\end{aligned}$$

$$\mathrm{Corr}[X, Y] = \frac{\mathrm{Cov}[X, Y]}{\sqrt{V[X] \cdot V[Y]}},$$

Covariance for more than two dimensions

$$\boldsymbol{X} = [X_1, \ldots, X_d]$$

$$\begin{aligned}\Sigma_{ij} &= \mathrm{Cov}[X_i, X_j] \\ &= \mathbb{E}\left[\left(X_i - \mathbb{E}\left[X_i\right]\right)\left(X_j - \mathbb{E}\left[X_j\right]\right)\right]\end{aligned}$$

$$\begin{aligned}\boldsymbol{\Sigma} &= \mathrm{Cov}[\boldsymbol{X}, \boldsymbol{X}] \quad \in \mathbb{R}^{d \times d} \\ &= \mathbb{E}[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathbb{E}(\boldsymbol{X})^\top] \\ &= \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top] - \mathbb{E}[\boldsymbol{X}]\mathbb{E}[\boldsymbol{X}]^\top.\end{aligned}$$

Covariance for more than two dimensions

$$\boldsymbol{X} = [X_1, \ldots, X_d]$$

$$\begin{aligned} \boldsymbol{\Sigma} = \mathrm{Cov}[\boldsymbol{X}, \boldsymbol{X}] &\in \mathbb{R}^{d \times d} \\ &= \mathbb{E}[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathbb{E}(\boldsymbol{X})^\top] \\ &= \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top] - \mathbb{E}[\boldsymbol{X}]\mathbb{E}[\boldsymbol{X}]^\top. \end{aligned}$$

$$\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$$

Dot product

$$\mathbf{x}^\top \mathbf{y} = \sum_{i=1}^{d} x_i y_i$$

Outer product

$$\mathbf{x}\mathbf{y}^\top = \begin{bmatrix} x_1 y_1 & x_1 y_2 & \cdots & x_1 y_d \\ x_2 y_1 & x_2 y_2 & \cdots & x_2 y_d \\ \vdots & \vdots & & \vdots \\ x_d y_1 & x_d y_2 & \cdots & x_d y_d \end{bmatrix}$$

Some useful properties

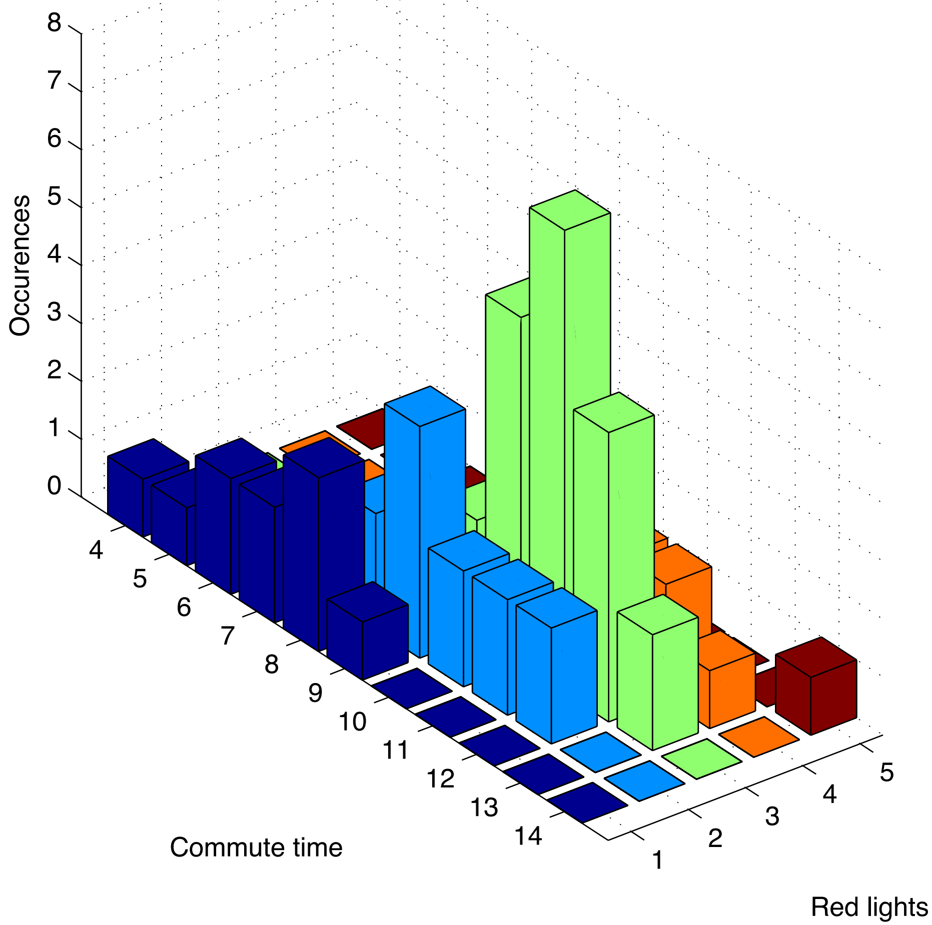
1. $\mathbb{E}\left[c\boldsymbol{X}\right] = c\mathbb{E}\left[\boldsymbol{X}\right]$
2. $\mathbb{E}\left[\boldsymbol{X} + \boldsymbol{Y}\right] = \mathbb{E}\left[\boldsymbol{X}\right] + \mathbb{E}\left[\boldsymbol{Y}\right]$
3. $V\left[c\right] = 0$ $\triangleright$ the variance of a constant is zero
4. $\mathrm{V}[\boldsymbol{X}] \succeq 0$ (i.e., is positive semi-definite), where for $d = 1$, $\mathrm{V}[\boldsymbol{X}] \geq 0$
   $\mathrm{V}[\boldsymbol{X}]$ is shorthand for $\mathrm{Cov}[\boldsymbol{X}, \boldsymbol{X}]$.
5. $\mathrm{V}[c\boldsymbol{X}] = c^2\mathrm{V}[\boldsymbol{X}]$.
6. $\mathrm{Cov}[\boldsymbol{X}, \boldsymbol{Y}] = \mathbb{E}[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{Y} - \mathbb{E}(\boldsymbol{Y})^\top] = \mathbb{E}[\boldsymbol{X}\boldsymbol{Y}^\top] - \mathbb{E}[\boldsymbol{X}]\mathbb{E}[\boldsymbol{Y}]^\top$
7. $\mathrm{Cov}[\boldsymbol{X} + \boldsymbol{Y}] = \mathrm{V}[\boldsymbol{X}] + \mathrm{V}[\boldsymbol{Y}] + 2\mathrm{Cov}[\boldsymbol{X}, \boldsymbol{Y}]$

Multidimensional PMF

Now record both commute time and number red lights

$\Omega = \{4, \ldots, 14\} \times \{1, 2, 3, 4, 5\}$

PMF is normalized 2-d table (histogram) of occurrences

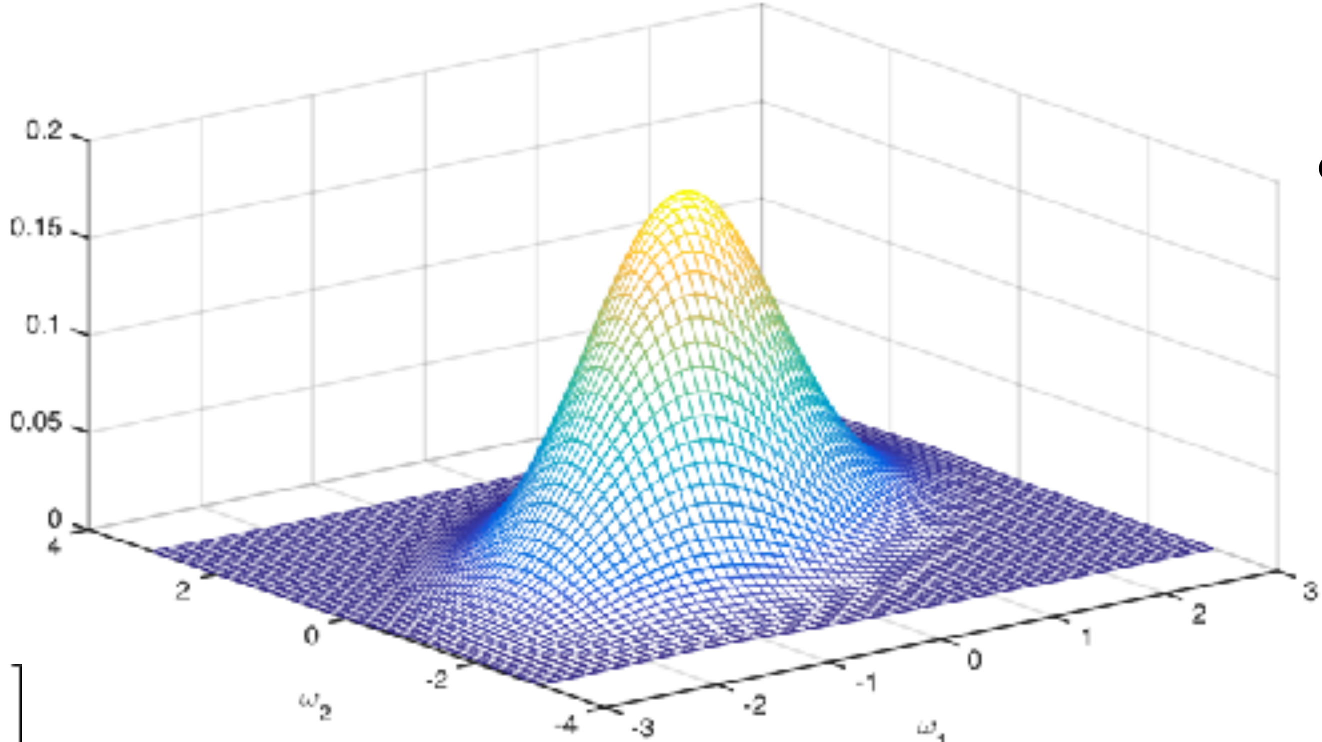


MULTIDIMENSIONAL GAUSSIAN

$\Omega = \mathbb{R}^d$
$\mathcal{F} = \mathcal{B}(\mathbb{R})^d$

$\boldsymbol{\mu} \in \mathbb{R}^d$
$\boldsymbol{\Sigma}$ = positive definite $d \times d$ matrix
$|\boldsymbol{\Sigma}|$ = determinant of $\boldsymbol{\Sigma}$

$$p(\omega) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\omega - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\omega - \boldsymbol{\mu})\right)$$



d = 2

$\boldsymbol{\mu} = (0, 0)$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & .75 \\ .75 & 1 \end{bmatrix}$$

Mixtures of Distributions

Mixture model:

A set of m probability distributions, $\{p_i(x)\}_{i=1}^{m}$

$$p(x) = \sum_{i=1}^{m} w_i p_i(x)$$

where $\boldsymbol{w} = (w_1, w_2, \ldots, w_m)$ and non-negative and

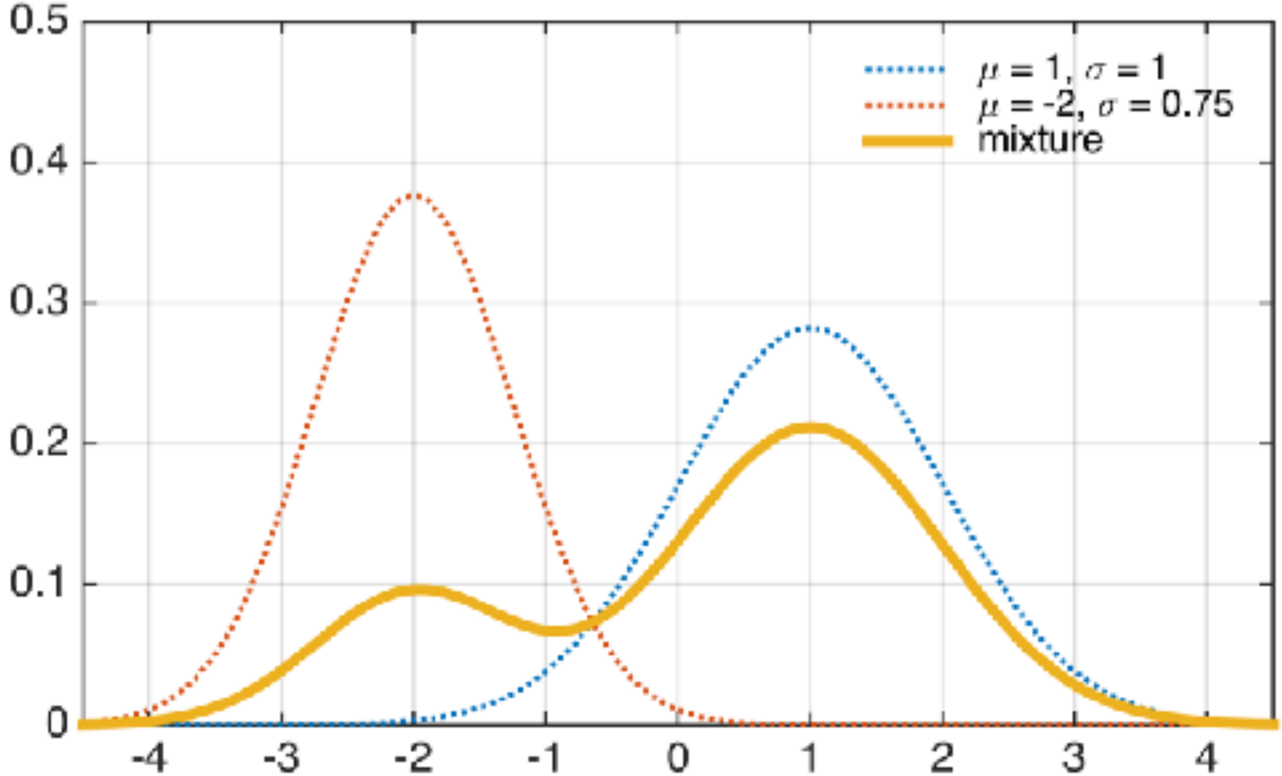
$$\sum_{i=1}^{m} w_i = 1$$

Mixtures of Gaussians

Mixture of $m = 2$ Gaussian distributions:

$w_1 = 0.75$, $w_2 = 0.25$

$$p(x) = \sum_{i=1}^{m} w_i p_i(x)$$



Example: Sample average is unbiased estimator

Obtain instances $x_1, \ldots, x_n$

What can we say about the sample average?

This sample is random, so we consider i.i.d. random variables $X_1, \ldots, X_n$

Reflects that we could have seen a different set of instances x_i

$$\begin{aligned}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mu \\ &= \mu\end{aligned}$$

For any one sample $x_1, \ldots, x_n$, unlikely that $\frac{1}{n}\sum_{i=1}^{n} x_i = \mu$