PROBABILITY THEORY

CMPUT 296

Martha White

Winter, 2020

Quick Check on Pre-req Knowledge

- You should know how to take derivatives
  - Will rarely, if ever integrate
  - I will teach you about Partial Derivatives
- I assume you know about vectors and dot products, hopefully also about matrices
- Need to have learned about probability, I will cover
  - Expected Value (Mean)
  - Variance
  - Random Variables
  - Probability Density Functions…

WHY DO WE NEED PROBABILITIES?

- We could just assume a deterministic world
- I see an input x, I can produce the output y = f(x)
  - Example: in a game (e.g., chess), you take an action, and the outcome is deterministic
- But, even in a deterministic world, we have a problem: partial observability
- Outcomes look random because we don't have enough information
  - Example: Imagine a (high-tech) gumball machine, where f(x = has candy, battery charged) = output candy
  - You can only see if it has candy

Why do we need probabilities?

- But, even in a deterministic world, we have a problem: partial observability
- Outcomes look random because we don't have enough information
  - Example: Imagine a (high-tech) gumball machine, where f(x = has candy, battery charged) = output candy
  - You can only see if it has candy
  - From your perspective, when x = has candy, sometimes candy is outputted and sometimes not
  - Looks stochastic (dependent on hidden battery charge)

Space of outcomes and events

Ω = sample space, all outcomes of the experiment

$\mathcal{E}$ = event space, set of subsets of Ω

Ω and $\mathcal{E}$ must be non-empty

Sample Spaces

$$\Omega$$

discrete (countable)

e.g., Outcome of Dice Roll

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

$$\Omega = \mathbb{N}$$

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

continuous (uncountable)

$$\Omega = [0, 1]$$

$$\Omega = \mathbb{R}$$

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

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 I use (though its somewhat loose)
  - discrete and countable
  - continuous and uncountable

(Measurable) Space of outcomes and events

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

Ω and $\mathcal{E}$ must be non-empty

If the following conditions hold:

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

$\mathcal{E}$ is an event space

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

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
$\mathcal{E}$ = event space, set of subsets of Ω

If the following conditions hold:

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

Quick Check on background

- The complement of a set
- The union of sets
- A set of sets
- Any other confusing notation?

Sample Spaces

$$\Omega$$

discrete (countable)

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

$\Omega = \mathbb{N}$

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

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

Power set

continuous (uncountable)

$\Omega = [0, 1]$

$\Omega = \mathbb{R}$

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

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

Borel field

Axioms of Probability

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

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

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

is called a probability measure (probability distribution)

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

FINDING PROBABILITY DISTRIBUTIONS

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

Do you recognize this distribution?

Example: $\Omega = \{0, 1\}$

$\mathcal{E} = \{\emptyset, \{0\}, \{1\}, \Omega\}$

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

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
$\mathcal{E} = \mathcal{P}(\Omega)$

Probability mass function:

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

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

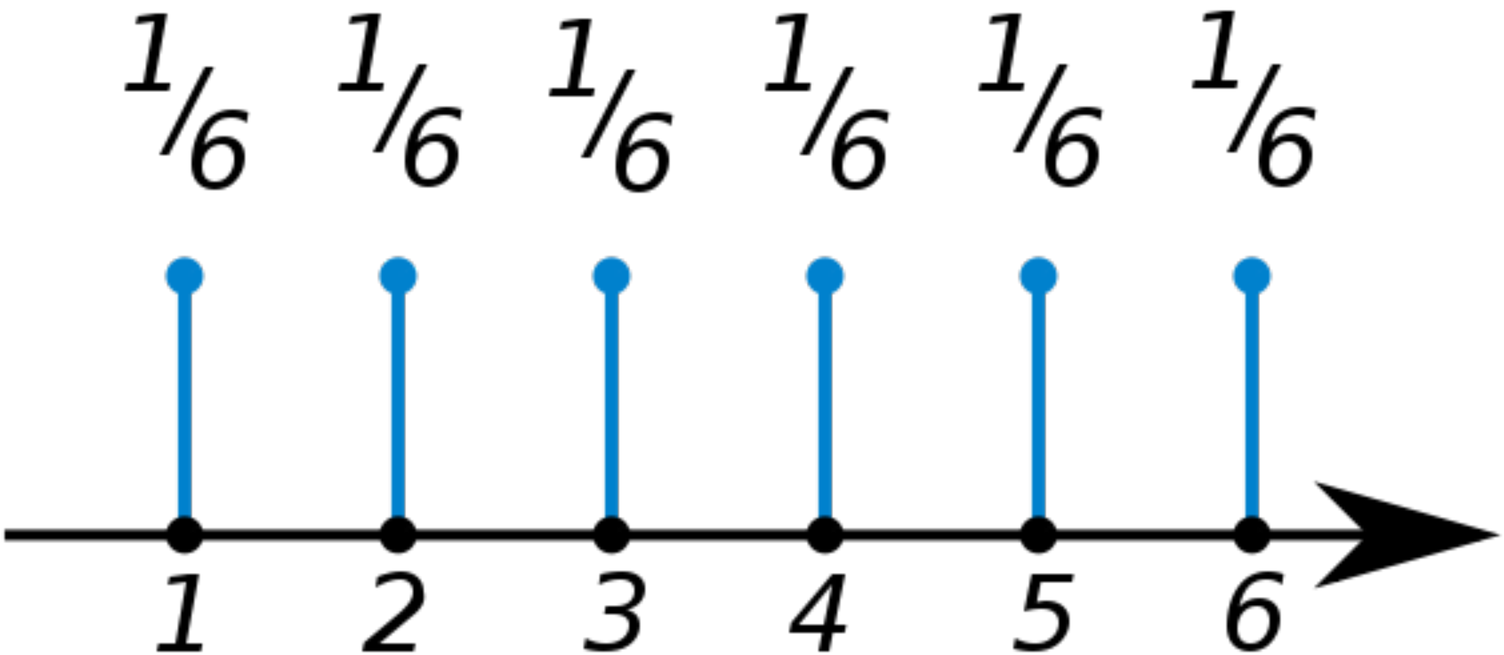
$$P(A) = \sum_{\omega \in A} p(\omega)$$

Arbitrary PMFs

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

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

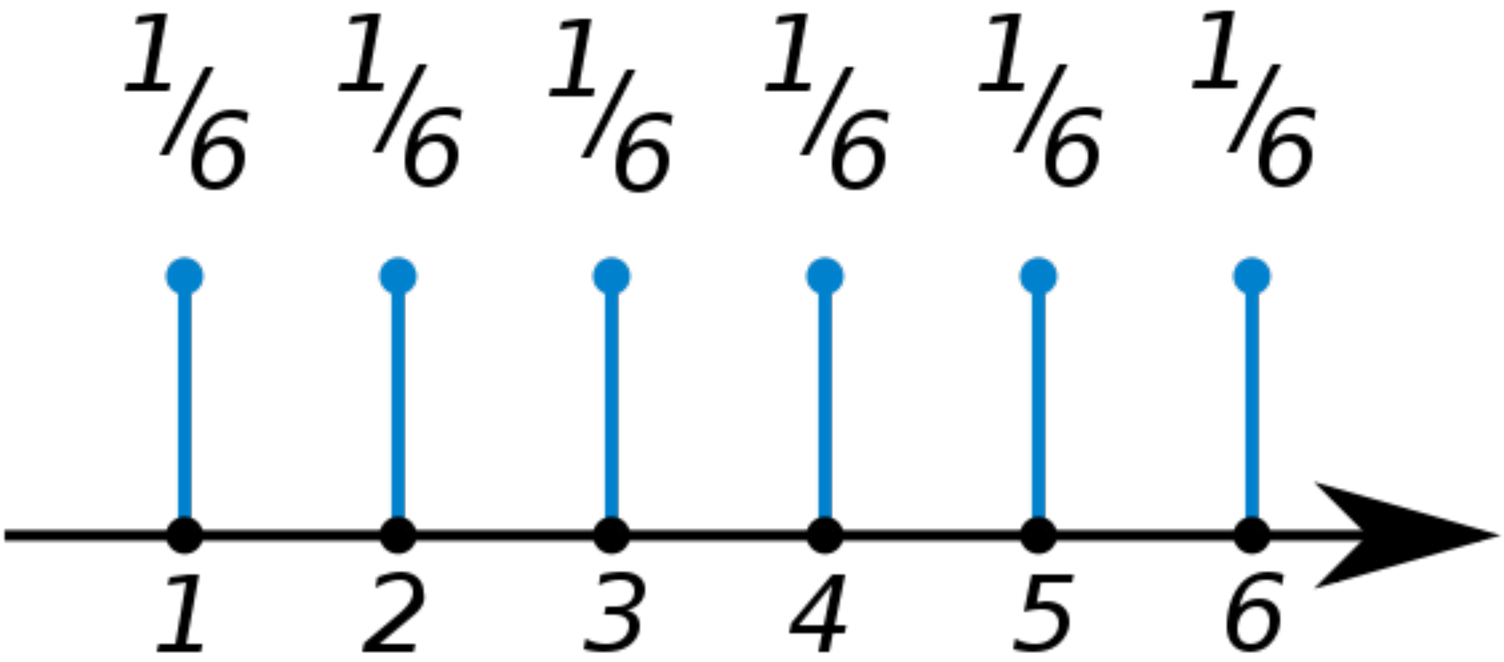
$$p(\omega) = 1/6 \quad \forall \omega \in \Omega$$



Exercise: Examples of events

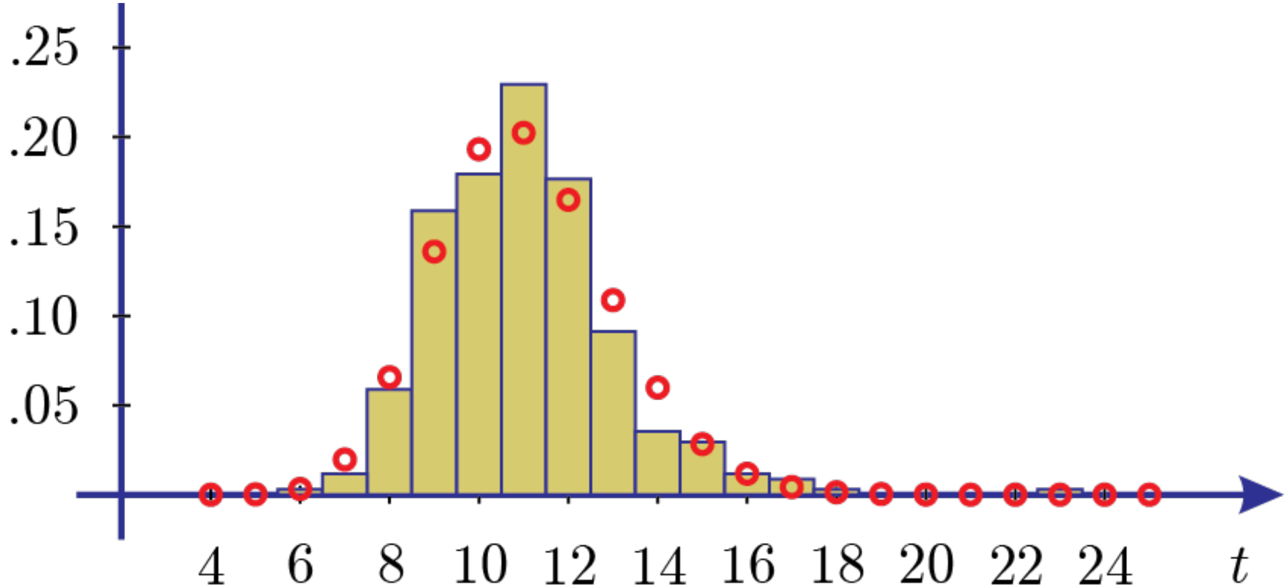
- What is a possible event? What is its probability
- What is the event space?

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



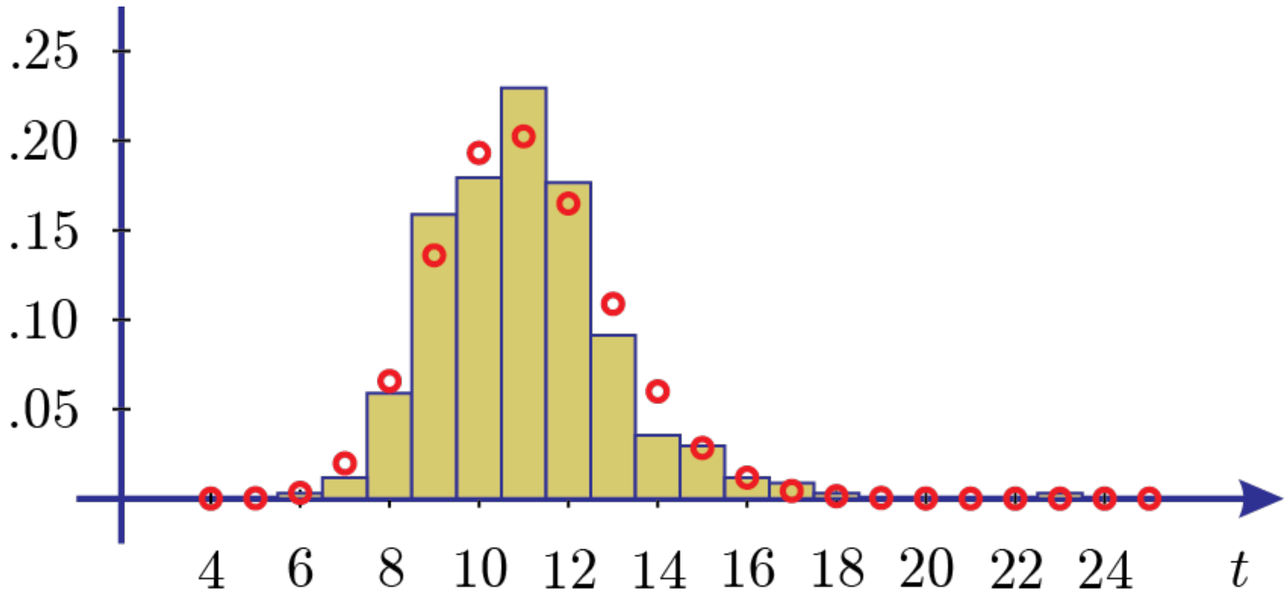
Exercise: How are PMFs useful as a model?

- Histograms!
- Imagine you recorded your commute times for a year, in minutes (365 recorded times)
- How do you get p(t)?
- How is p(t) useful?



Exercise: How are PMFs useful as a model?

- Histograms!
- Imagine you recorded your commute times for a year, in minutes
- Get p(t): count number of times t = 1, 2, 3, … occurs and then normalize probabilities by # samples



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

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

Reminders: January 9, 2020

- Assignment 1 is available on the website
  - https://marthawhite.github.io/mlbasics/
- You should start reading the notes
  - Chapters 1, 2 and 3 (about 30 pages)
- The notes are in-progress
  - Some sections still say "Coming soon…"
  - Avoid printing the full set of notes (if at all). At most, print what you are reading

A few other notes on Policy

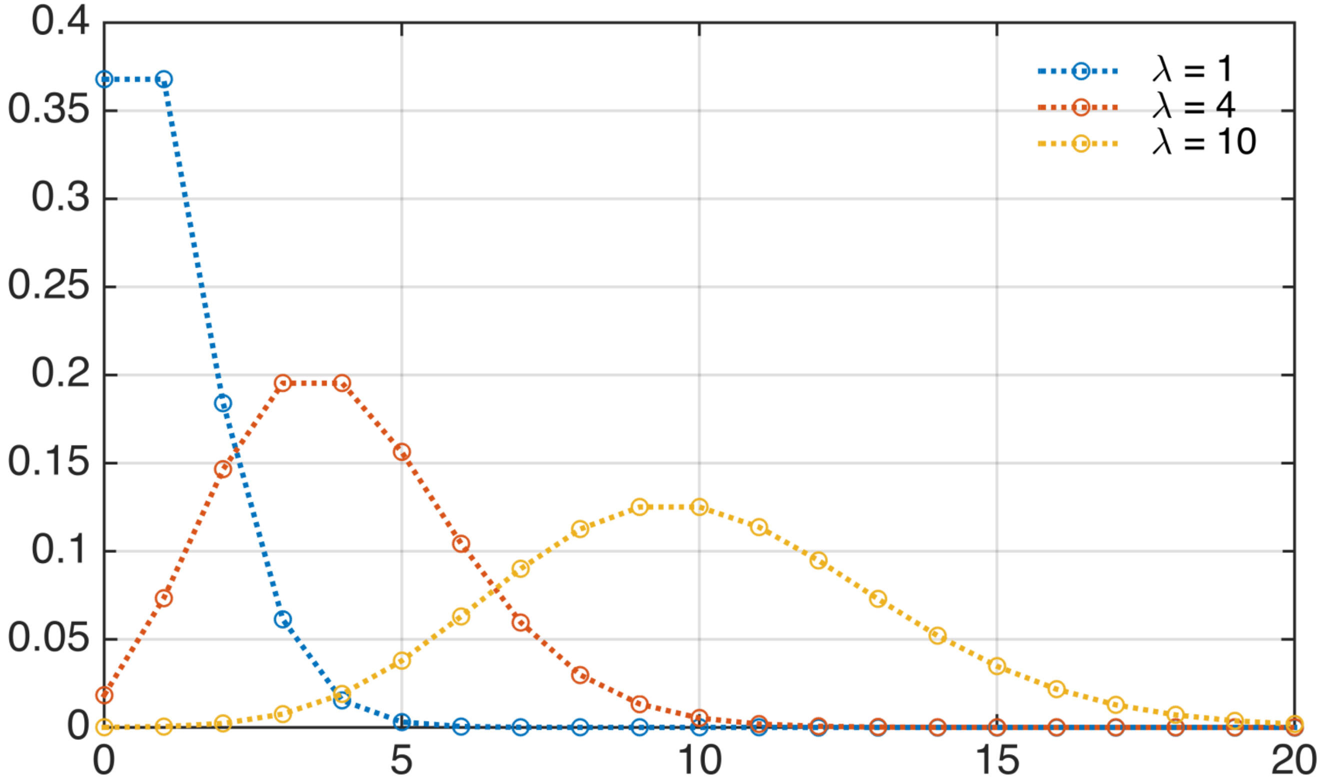
- If you have an issue (sickness, injury, family problem, etc.), and need an extension on an assignment, contact me about it before the assignment deadline.
    - I cannot give extensions after.
- You cannot submit Thought Questions late
    - only Assignments have a late policy that allows for late submission
    - if you submit 1 hour late, we won’t penalize you

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

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



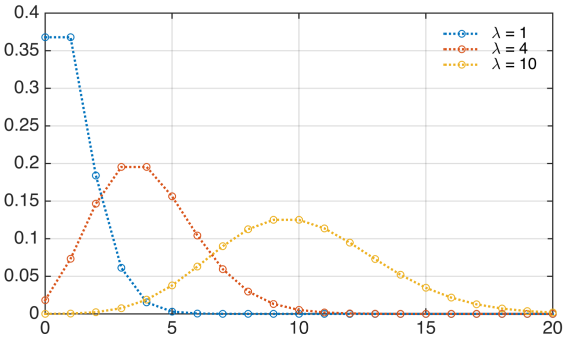
USEFUL PMFs

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

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

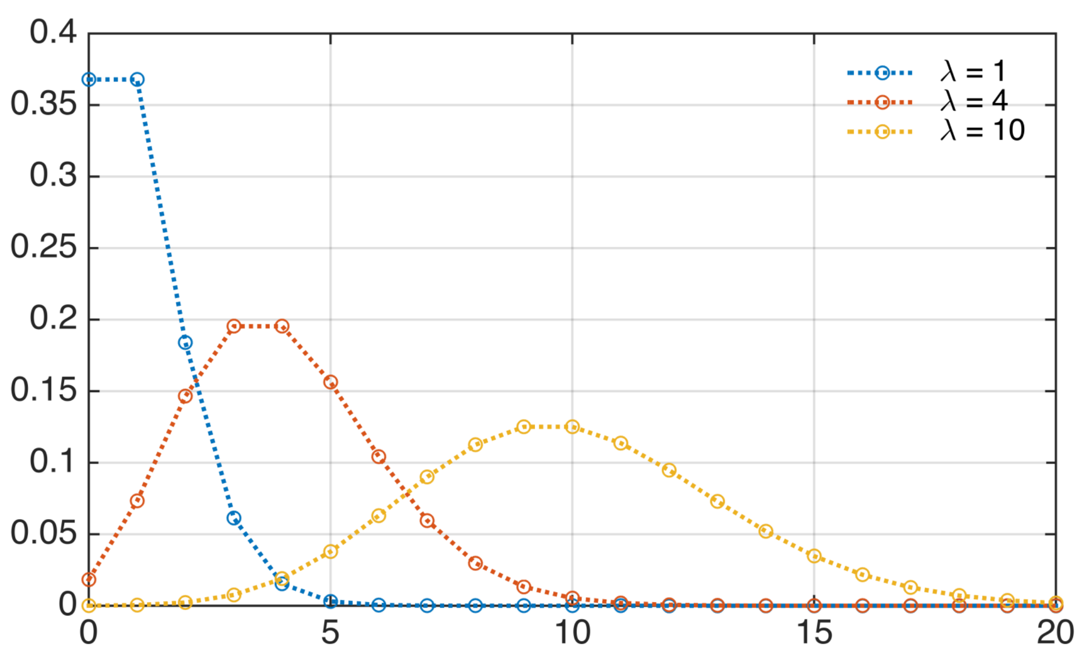
1. Can we use a table for this?
2. How do we know this is a valid pmf? How can you check?



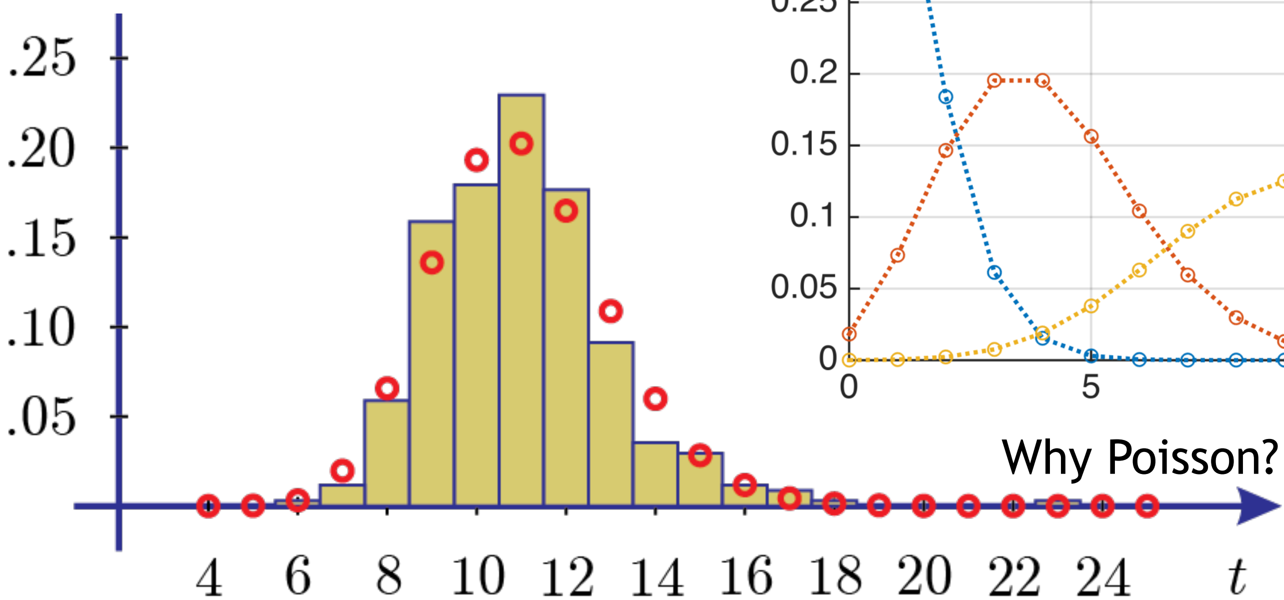
Exercise: Can we use a Poisson for commute times?

- 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?

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



Why Poisson? Why not just a histogram?



Probability Density Functions

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

Probability density function:

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

Who has never seen an integral?

Probability Density Functions

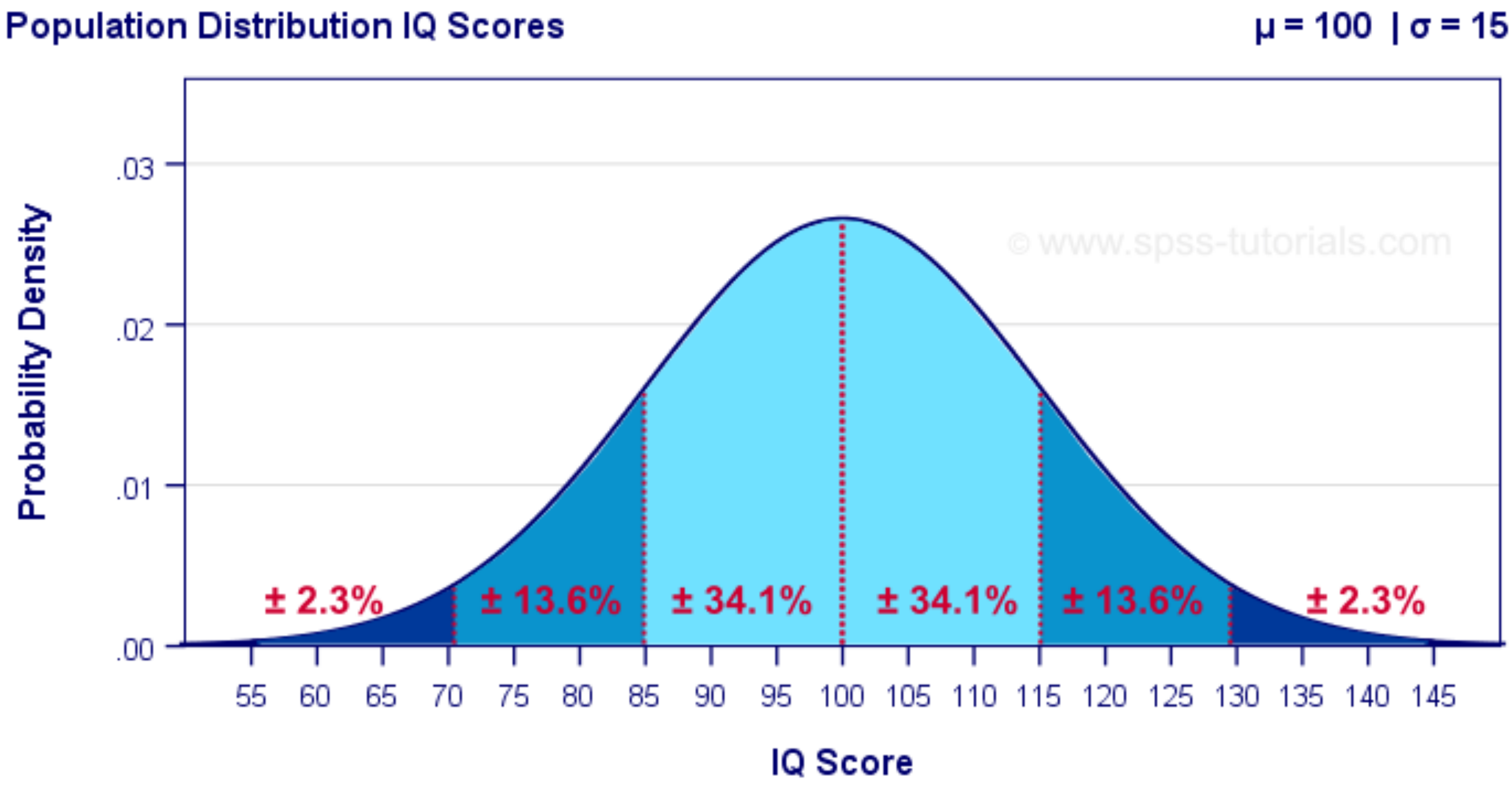
Ω = continuous sample space

$\mathcal{E} = \mathcal{B}(\Omega)$

Probability density function:

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

e.g. normal distribution (Gaussian)



Probability Density Functions

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

Probability density function:

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

Who has never seen an integral?

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

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

PMFs vs. PDFs

Ω = discrete sample space

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

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

Ω = continuous sample space

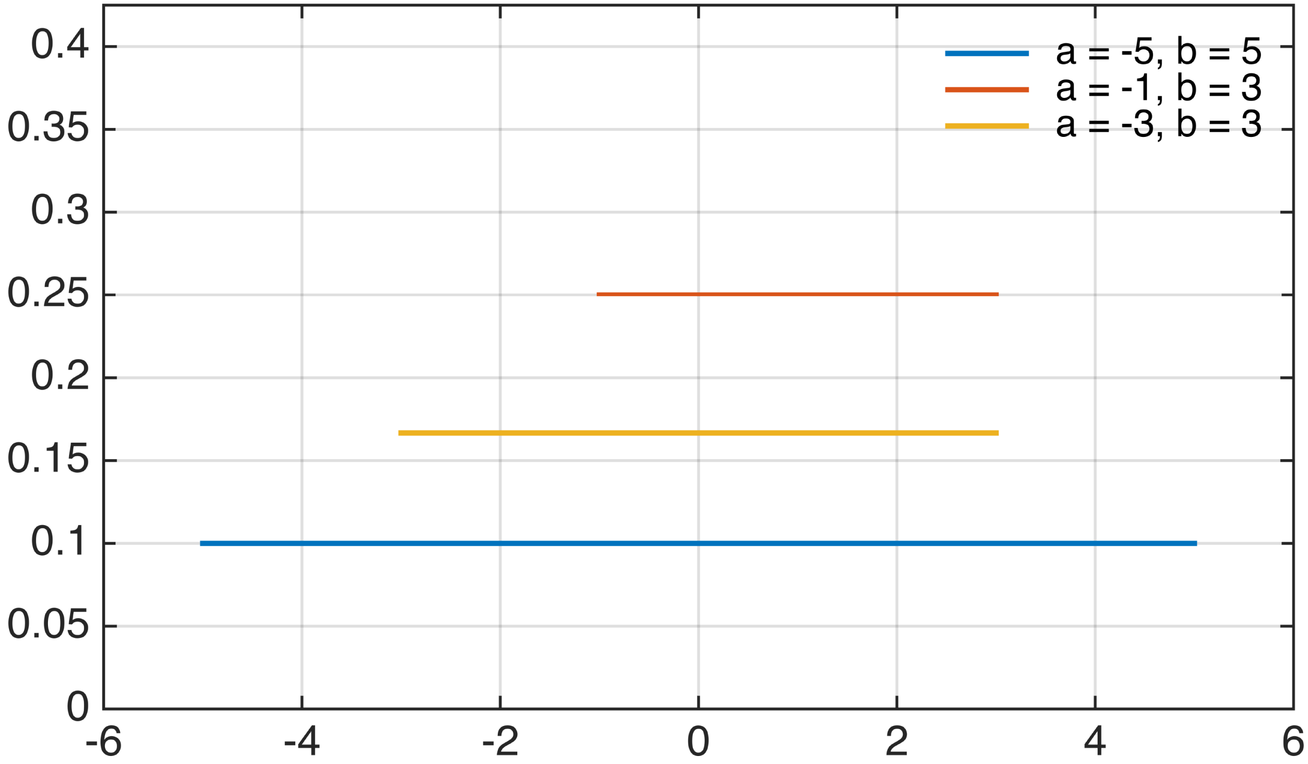
Example:

- 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 or space does it take up in [3,15]?)
- More reasonable to ask the probability of stopping between 3 to 3.5 seconds. How do we get that probability?

Useful PDFs

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

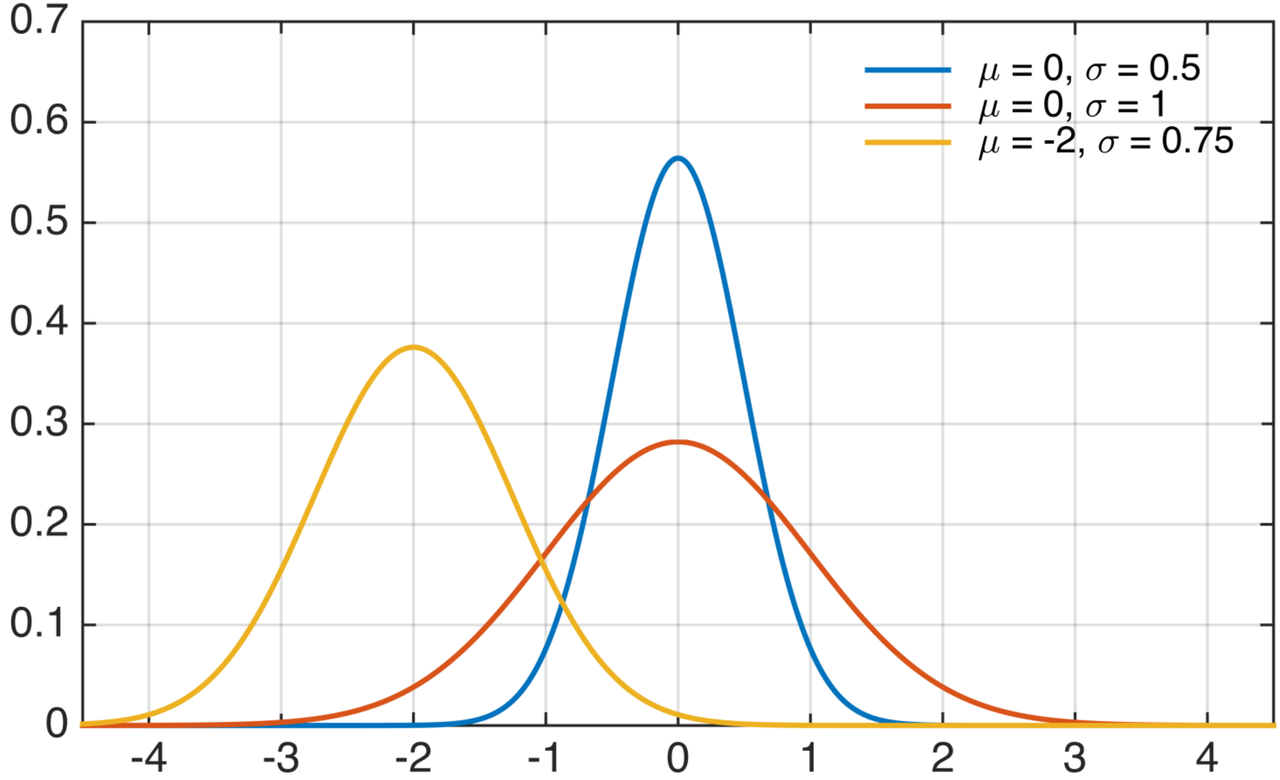
$$p(\omega) = \frac{1}{b-a} \qquad \forall \omega \in [a, b]$$



USEFUL PDFS

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

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

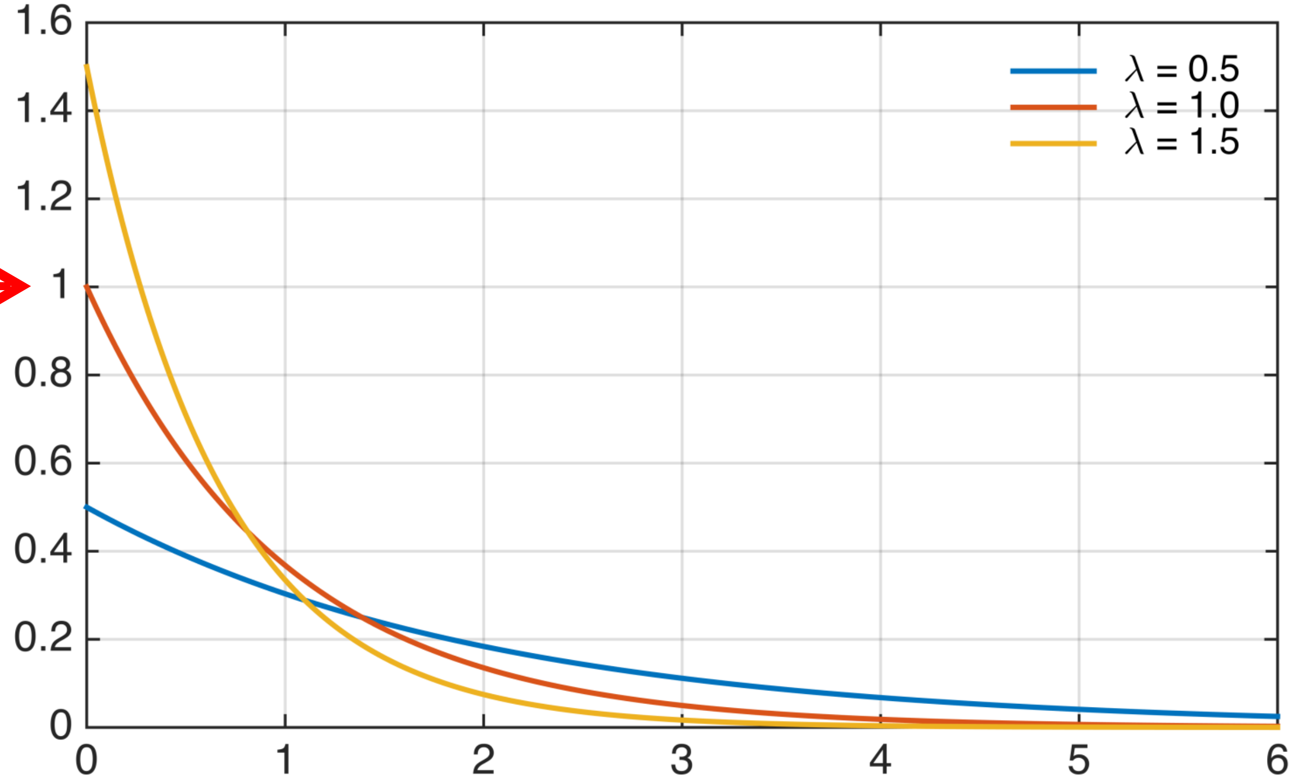


USEFUL PDFS

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

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

1 is here!



Why can the density be above 1?

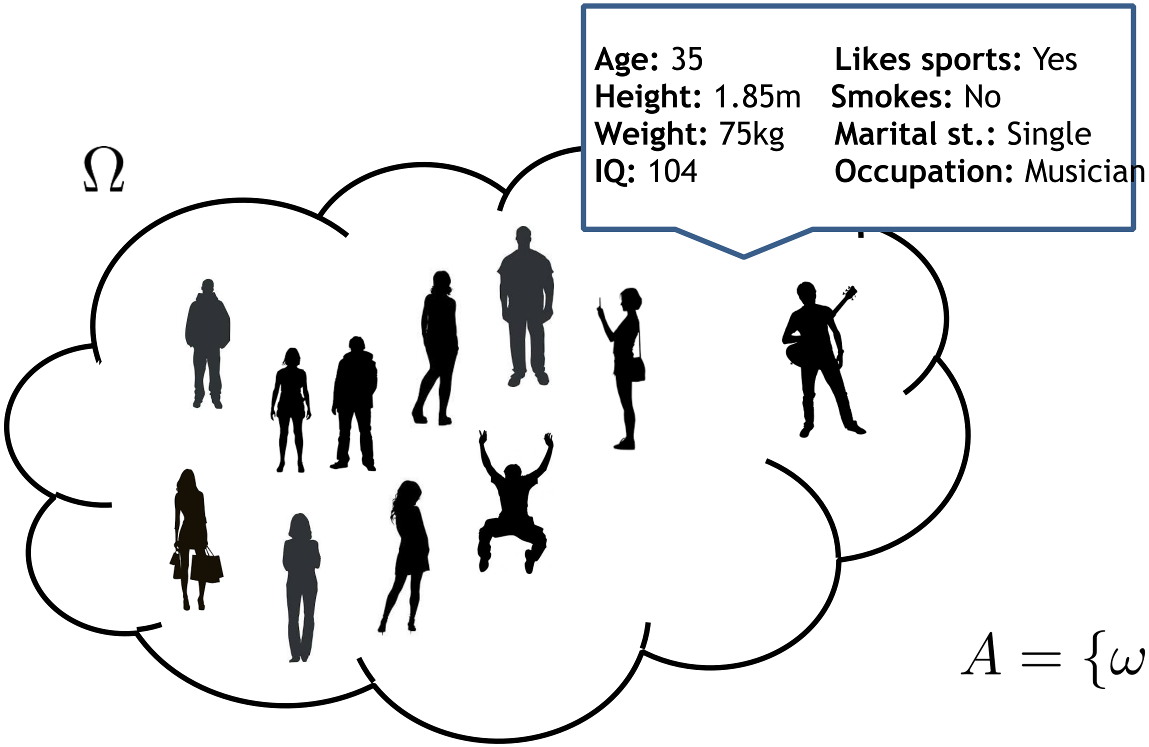
Consider an interval event $A = [x, x + \Delta x]$, where Δ is small

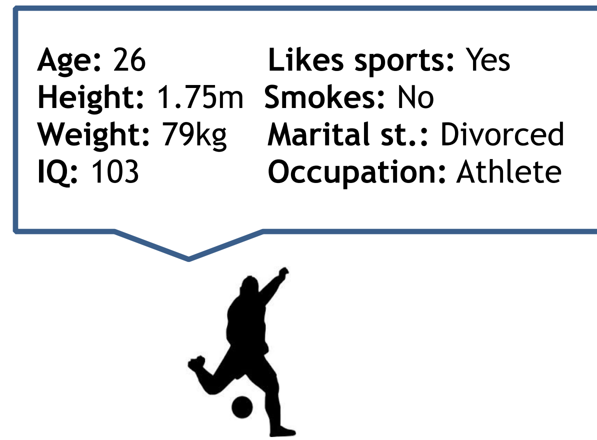
$$P(A) = \int_x^{x+\Delta x} p(\omega) d\omega$$
$$\approx p(x)\Delta x$$

p(x) can be big, because delta x is small
P(A) MUST be less than or equal to 1
p(x) can be bigger than 1

Random Variables

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





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

Musician is a random variable (a function)
A is a new event, let’s call it 1; Not-A is 0
Omega is {0, 1}
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

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

Random variable:

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

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

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

3 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

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

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

$$\Omega = \mathbb{R}^2, \text{e.g., } \omega = [-0.5, 10]$$
$$\Omega = [0, 1] \times [2, 5], \text{e.g., } \omega = [0.2, 3.5]$$

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	1
X	0	$P(X{=}0, Y{=}0) = 1/2$	$P(X{=}0, Y{=}1) = 1/100$
	1	$P(X{=}1, Y{=}0) = 1/10$	$P(X{=}1, Y{=}1) = 39/100$

Table 1.1: A joint probability table for random variables X and Y.

* these numbers are completely made up

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 know 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. So what is P(X = 0)?

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

No

Conditional Distributions

Conditional probability distribution:

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

If p(x,y) is small, does this imply that p(y|x) is small?

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 $p(x_i)$ is defined as

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

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 $p(x_i)$ is defined as

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

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

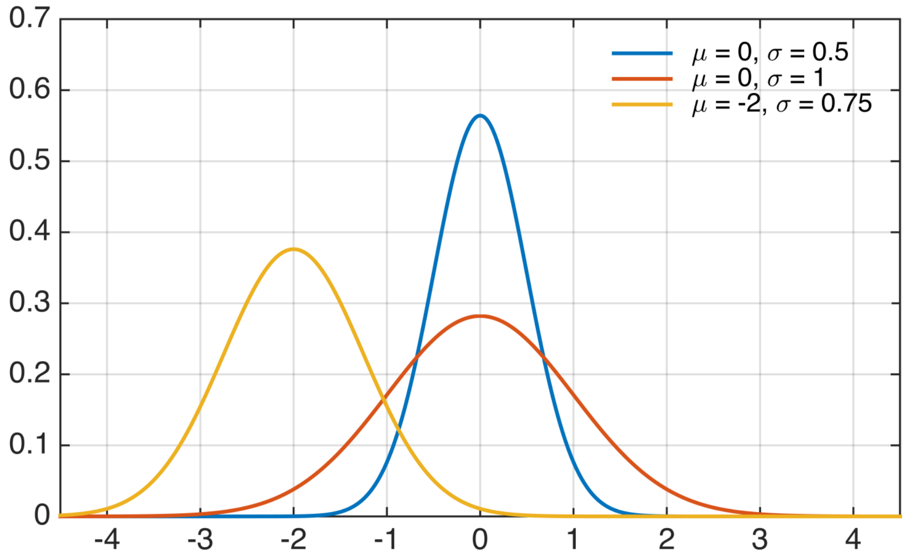
- Let **X** be a Bernoulli random variable (i.e., 0 or 1 with probability alpha)
- Let **Y** be a random variable in {10, 11, …, 1000}
- p(y | X = 0) and p(y | X = 1) are different distributions
- Two **types of books**: fiction (X=0) and non-fiction (X=1)
- Let Y correspond to **number of pages**
- 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
- 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
  - $p(y \mid X = 0) = p(X = 0, y)/p(X = 0)$
  - $p(X = 0 , y)$ = probability that a book is fiction and has y pages (imagine randomly sampling a book)
  - $p(X = 0)$ = probability that a book is fiction

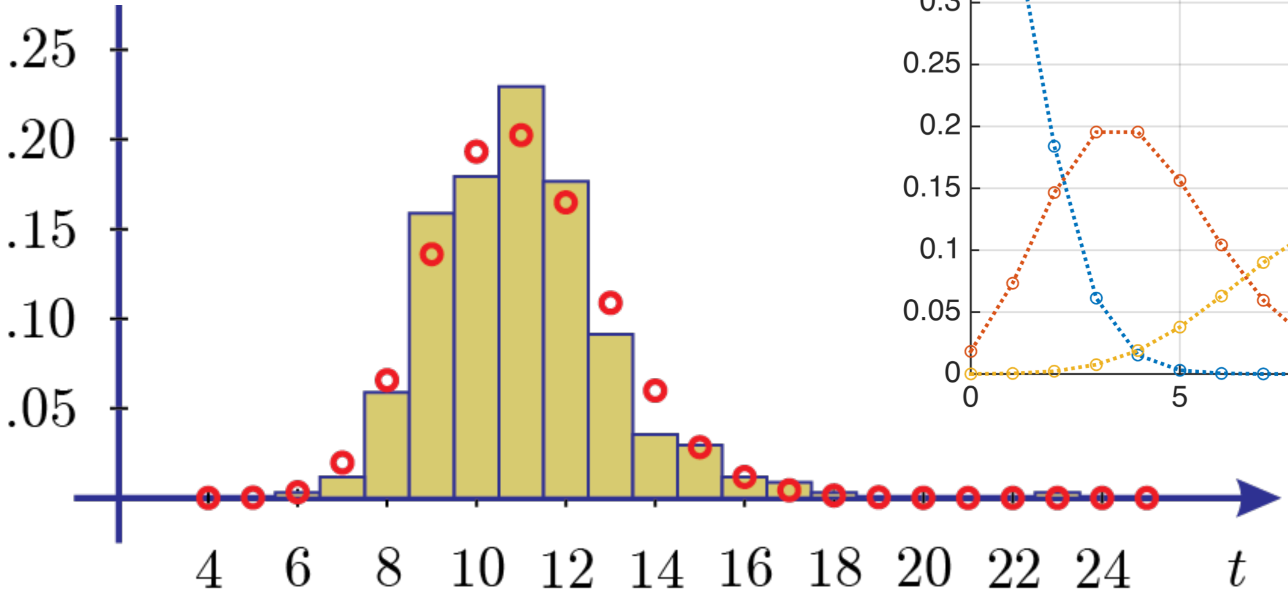
Another Example

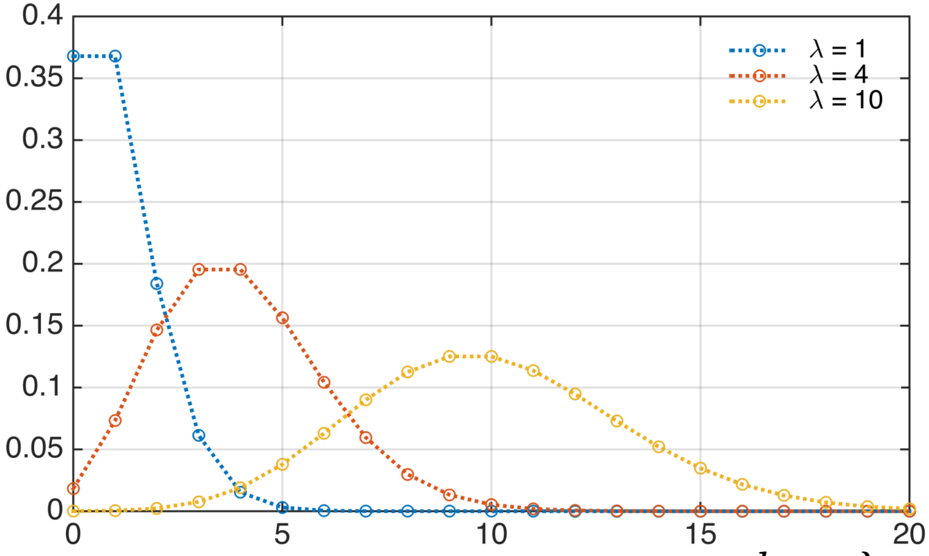
- Two types of books: fiction (X=0) and non-fiction (X=1)
- Let Y be a random variable over the reals, which corresponds to amount of money made
- p(y | X = 0) and p(y | X = 1) are different distributions
- e.g., even if both p(y | X = 0) and p(y | X = 1) are Gaussian, they likely have different means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



THINK-PAIR-SHARE (5 MINUTES)

- How might you use a given Poisson distribution, that models commute times? (Hint: recall modes)
- How might you pick lambda for a Poisson distribution, to model commute times? (Hint: the mean of a Poisson is lambda)





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

Review so far

- PMFs (discrete) and PDFs (continuous)
- Joint probabilities
- Marginals
- Conditional probabilities

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:

$$\begin{aligned} p(x_1,\ldots,x_k) &= p(x_k)\prod_{i=1}^{k-1} p(x_i|x_{i+1},\ldots,x_k) \\ &= p(x_1)\prod_{i=2}^{k} p(x_i|x_1,\ldots,x_{i-1}) \end{aligned}$$

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

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:

$$p(x_1,\ldots,x_k) = p(x_k)\prod_{i=1}^{k-1} p(x_i|x_{i+1},\ldots,x_k)$$

Three variable example

$$\begin{aligned} p(x,y,z) = p(y|x,z)p(x,z) &= p(y|x,z)p(x|z)p(z) \\ &= p(x|y,z)p(y|z)p(z) \\ &= p(x|y,z)p(z|y)p(y) \\ &\vdots \end{aligned}$$

How do we get Bayes rule?

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

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

Example: Drug test

- Imagine p(DT = True | User = True) = 0.99
- Imagine p(User = True) = 0.005
- Imagine p(DT = True | User = False) = 0.01
- What is p(User = True | DT = True)?

Reminders: January 14, 2020

- Thought Questions 1 due on January 23
  - Chapters 1-3
  - You should be reading now
  - if you read the material before I lecture, it will (a) help you understand a lot better and (b) be easier to write coherent thought questions
- Assignment 1 due on January 30
- Any questions?

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?

**(Basic example about an important issue in ML: hidden variables)

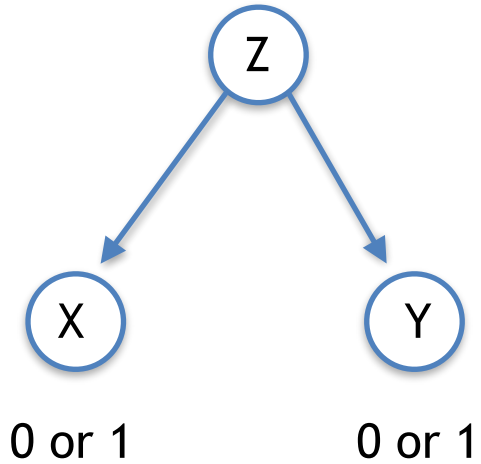
Conditional Independence is relative to distribution you pick, not objective for all distributions

- Explained on whiteboard
- What is $p(X)$ and $p(X \mid Z)$

Conditional Independence Examples
Example 7 in the notes

- Are X and Y independent? (don't know Z) $p(X, Y) = p(X)p(Y)?$
- Are X and Y conditionally independent, given Z?

$p(X, Y|Z) = p(X|Z)p(Y|Z)?$



bias

z	*0.3*	*0.5*	*0.8*
p(z)	*0.7*	*0.2*	*0.1*

probability of that bias

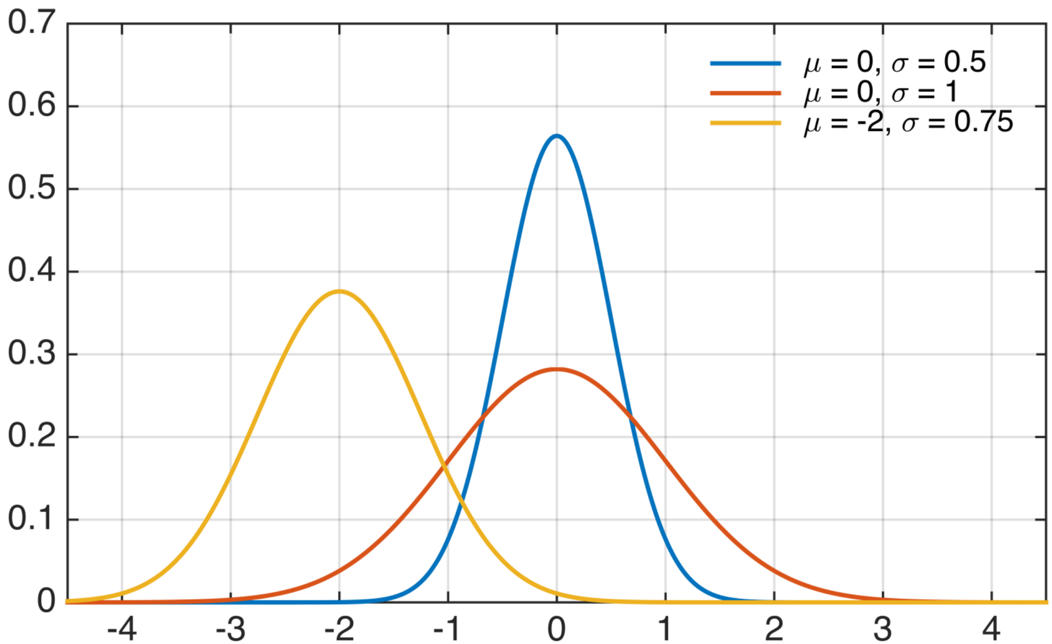
Imagine don't know Z and flip two 0s. Does that tell you anything about Z?

**(Basic example about an important issue in ML: hidden variables)

Expected value (Mean, Average)

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

Exercise: Biased coin with p(Heads) = 0.8



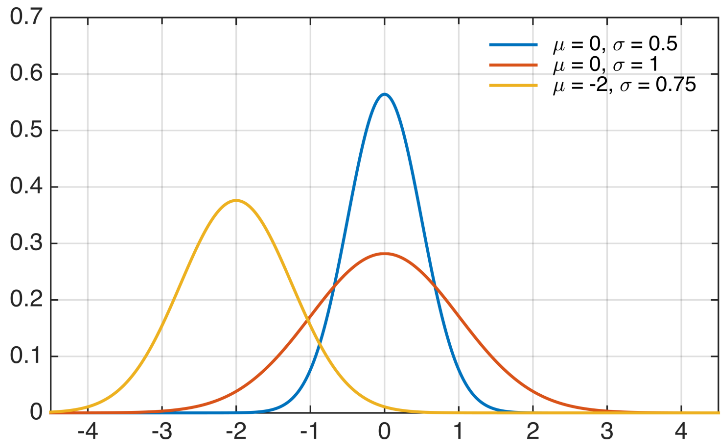
Expectations with functions

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

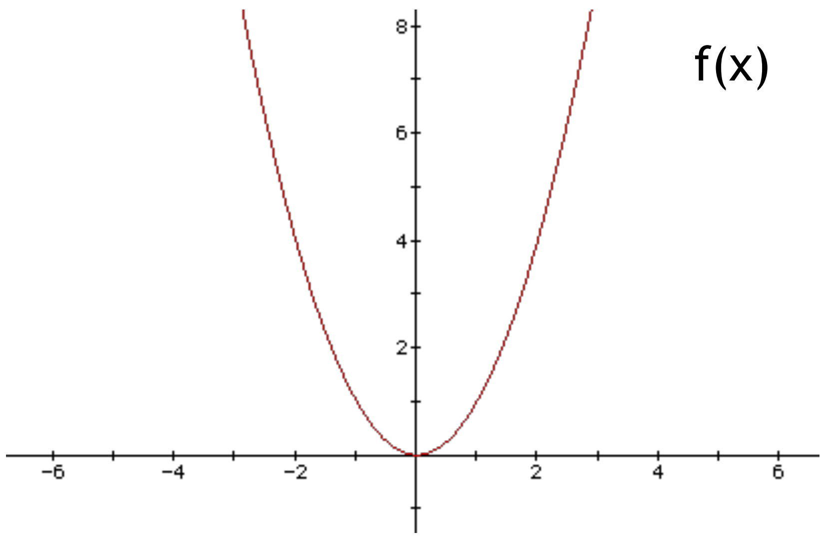
Exercise: Imagine you get 10 dollars if a Heads is flipped, lose 3 if Tails.

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

p(x)



f(x)



Conditional Expectations

$$\mathbb{E}\left[Y|X=x\right] = \begin{cases} \sum_{y\in\mathcal{Y}} yp(y|x) & Y : \text{discrete} \\ \\ \int_{\mathcal{Y}} yp(y|x)dy & Y : \text{continuous} \end{cases}$$

Different expected value, depending on which x is observed

3 Minute Break

Turn to your neighbour and discuss a question that you have
If you have nothing, consider the questions below

$$\mathbb{E}\left[Y|X=x\right] = \begin{cases} \sum_{y\in\mathcal{Y}} yp(y|x) & Y: \text{discrete} \\ \\ \int_{\mathcal{Y}} yp(y|x)dy & Y: \text{continuous} \end{cases}$$

What is the E[Y | x] for the following?
- p(y | x) = Gaussian with N(mu = x, sigma^2 = 10)
- p(y | x) = Gaussian with N(mu = f(x), sigma^2 = 0.1)
- p(y | x) = Bernoulli with alpha = x

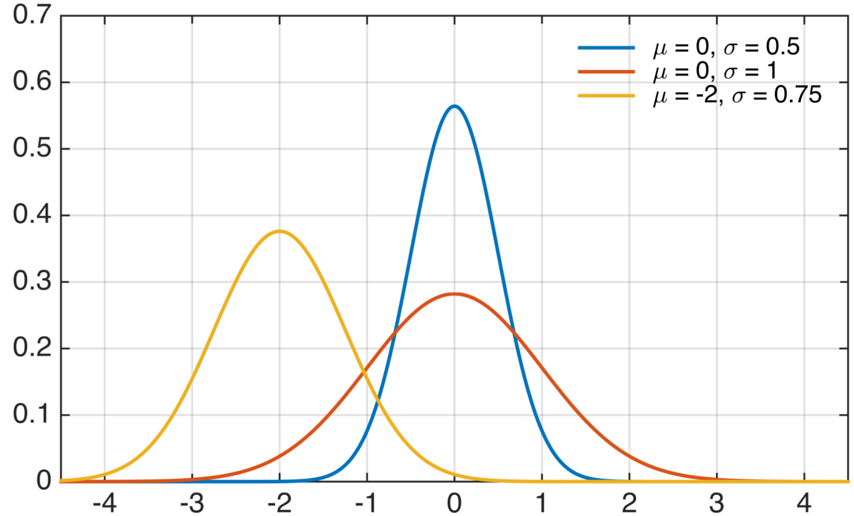
Properties of Expectations

- $E[cX] = c\,E[X]$, for a constant c
- $E[X + Y] = E[X] + E[Y]$ (linearity of expectation)
- If X and Y independent, then $E[XY] = E[X]\,E[Y]$
- $E[Y] = E[E[Y \mid X]]$, where outer expectation over X
  - called Law of Total Expectation
- (prove these on the whiteboard)

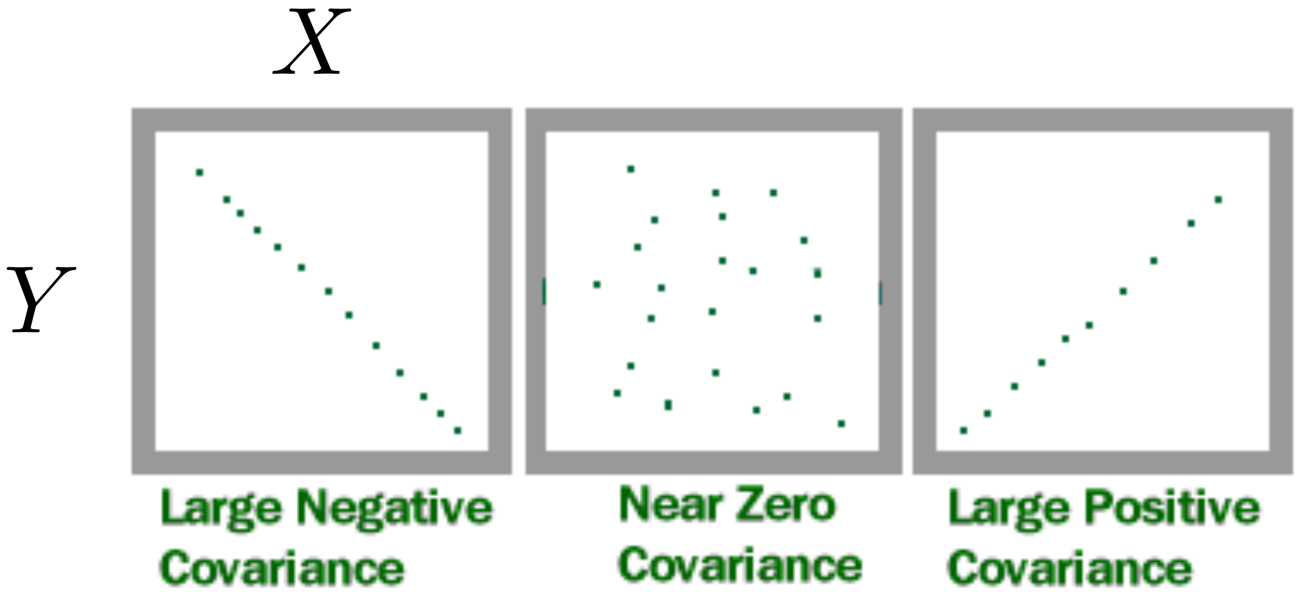
Variance

$$\text{Variance}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$$
$$= \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

Why? See if you can get this formula



COVARIANCE



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

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

Properties of Variances

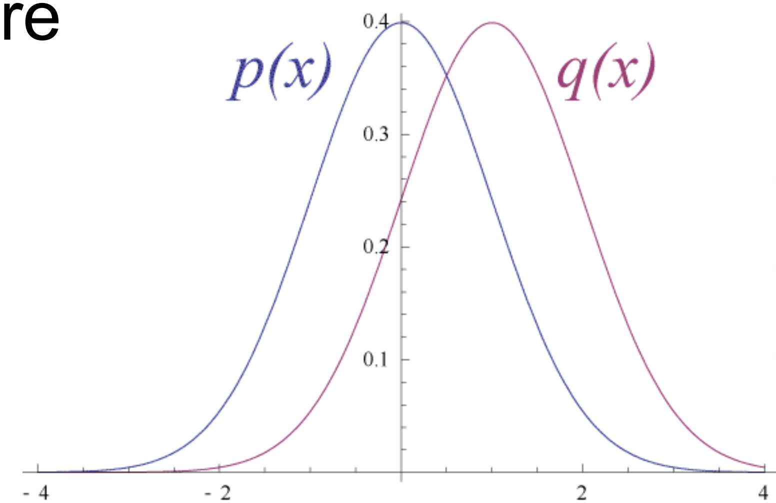
- V[c] = 0 for a constant c
- V[c X] = c^2 V[X]
- V[X + Y] = V[X] + V[Y] + 2 Cov[X,Y]
- If X and Y are independent, V[X + Y] = V[X] + V[Y]
  - i.e., Cov[X,Y] = 0

Independence and Decorrelation

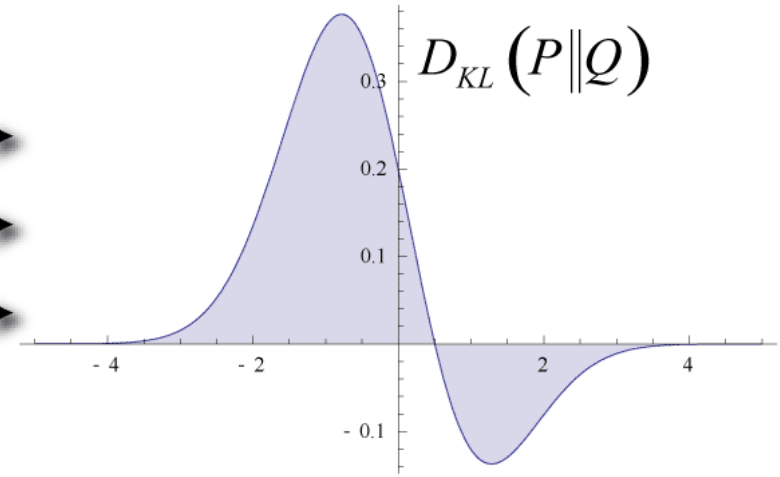
- Independent RVs have zero correlation
  - How can we tell?
  - Hint: use Cov[X,Y] = E[XY] - E[X]E[Y]
- Uncorrelated RVs (zero correlation) might be dependent
  - Correlation (Pearson's correlation) shows linear relationships; can miss nonlinear ones
  - Example: X normal RV, Y = X^2 (whiteboard)

ALTERNATIVES: MUTUAL INFORMATION (USING KL)

- KL-divergence measures how different two distributions are



Original Gaussian PDF's

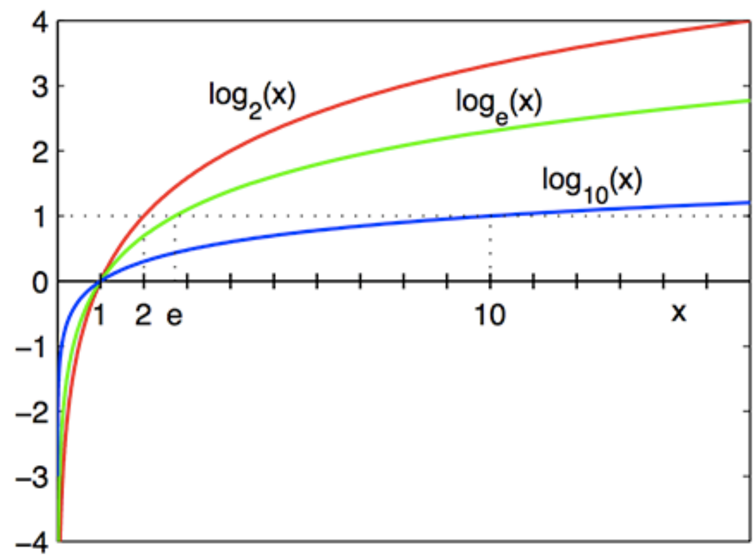


KL Area to be Integrated

$$\mathrm{KL}(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

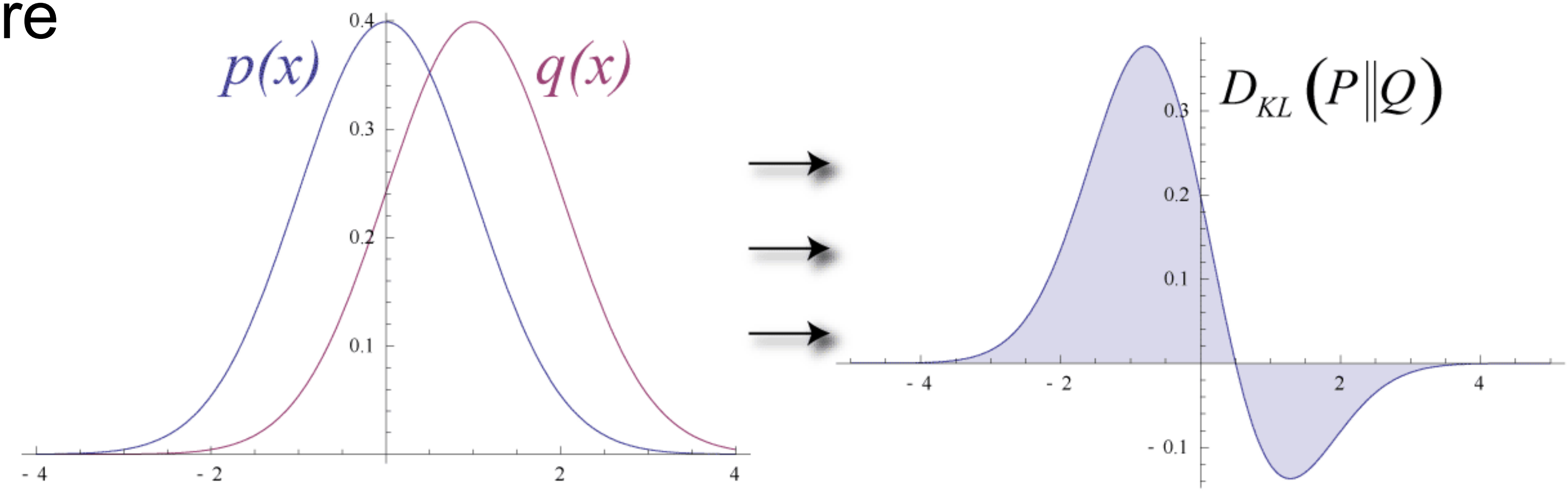
or

$$\mathrm{KL}(p||q) = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} dx$$



Alternatives: Mutual Information

- KL-divergence measures how different two distributions are



Original Gaussian PDF's

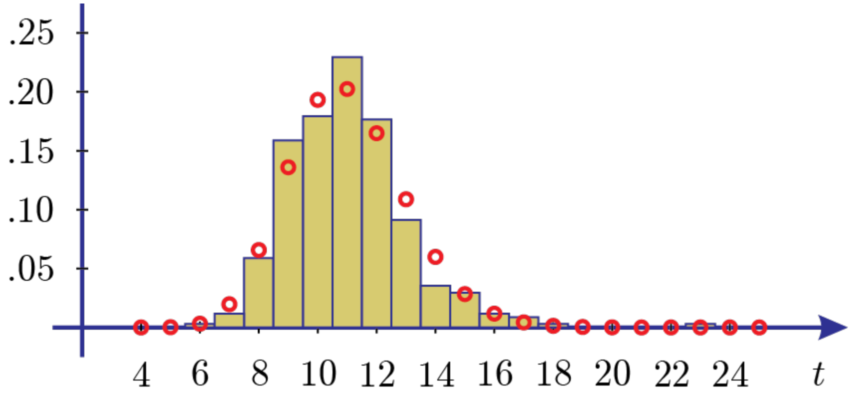
KL Area to be Integrated

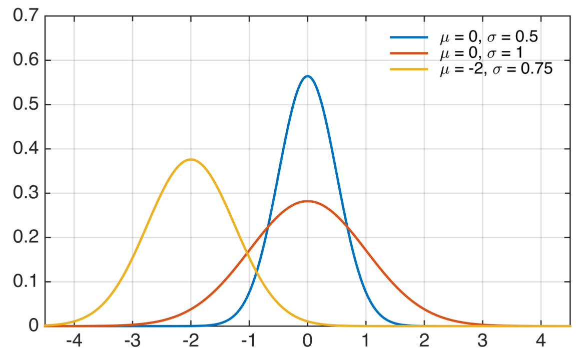
- Mutual Information between X and Y:
  - I(X; Y) = KL(p_{xy} || p_x p_y)
  - only zero when X and Y are independent
  - measure of price for encoding (X,Y) as independent RVs, even when they are not

Exercise: Modelling Commute Times

- Let's imagine we have 365 samples of commute times
- Say you wanted to model commute time C as a continuous RV (takes 7.6 minutes to get to work)
- This means we have to specify (or learn) the pdf; how?
- One option: pick distribution type (e.g., Gaussian), and find the "best" parameters that match the data
- What are the parameters to learn?
- Is a Gaussian a good choice?

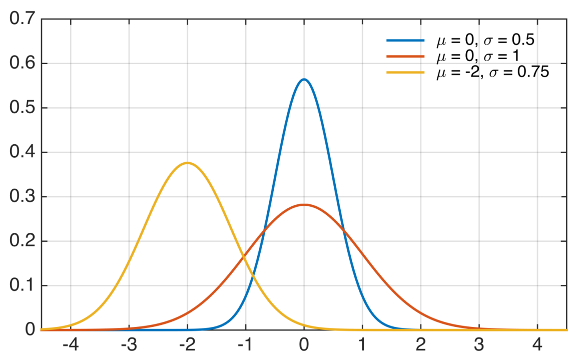
$$p(\omega) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\omega-\mu)^2}$$

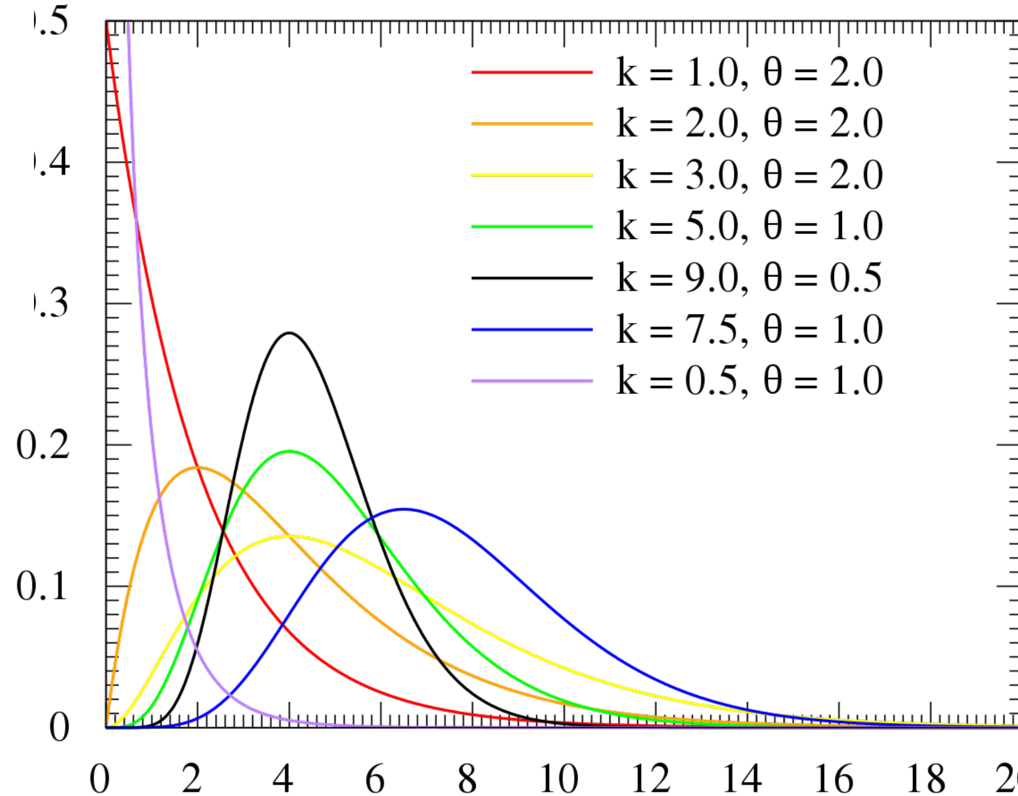


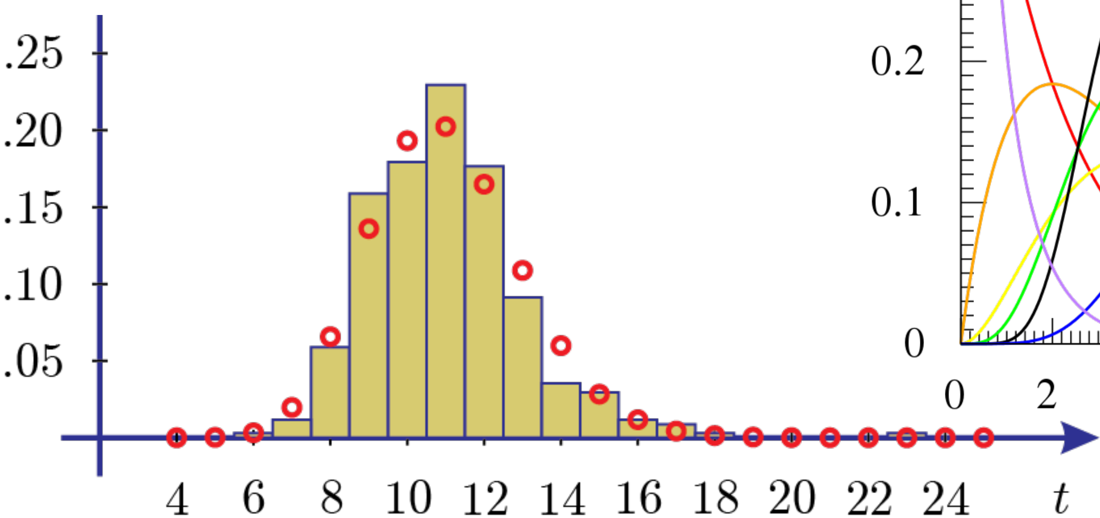


EXERCISE: MODELLING COMMUTE TIMES (CONT.)

- Note a better choice is actually a Gamma dist.
- Gaussian distribution (or gamma) for commute times extrapolates between recorded time in minutes

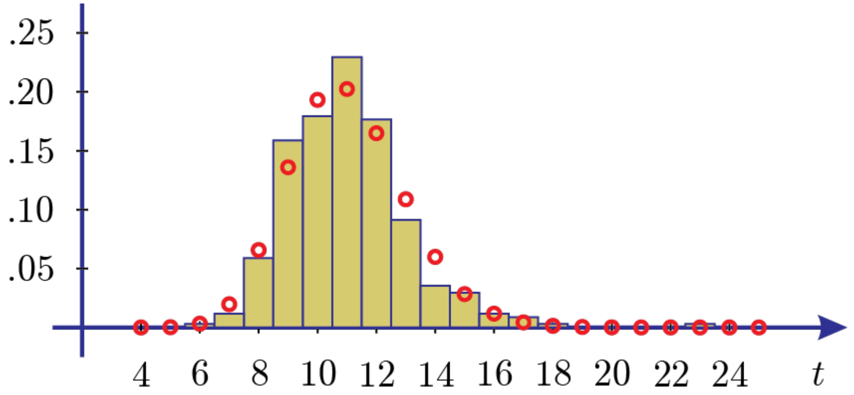






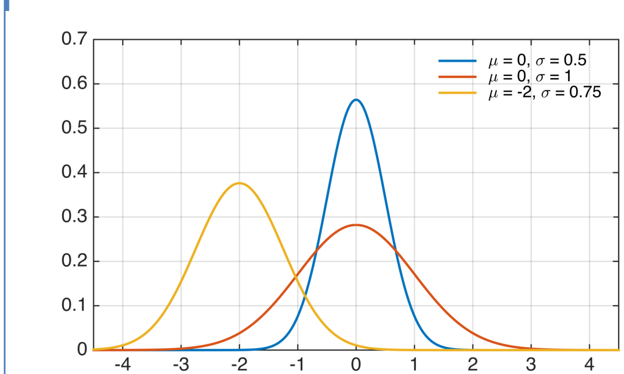
EXERCISE: CONDITIONAL PROBABILITIES

- 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 other x could we use?



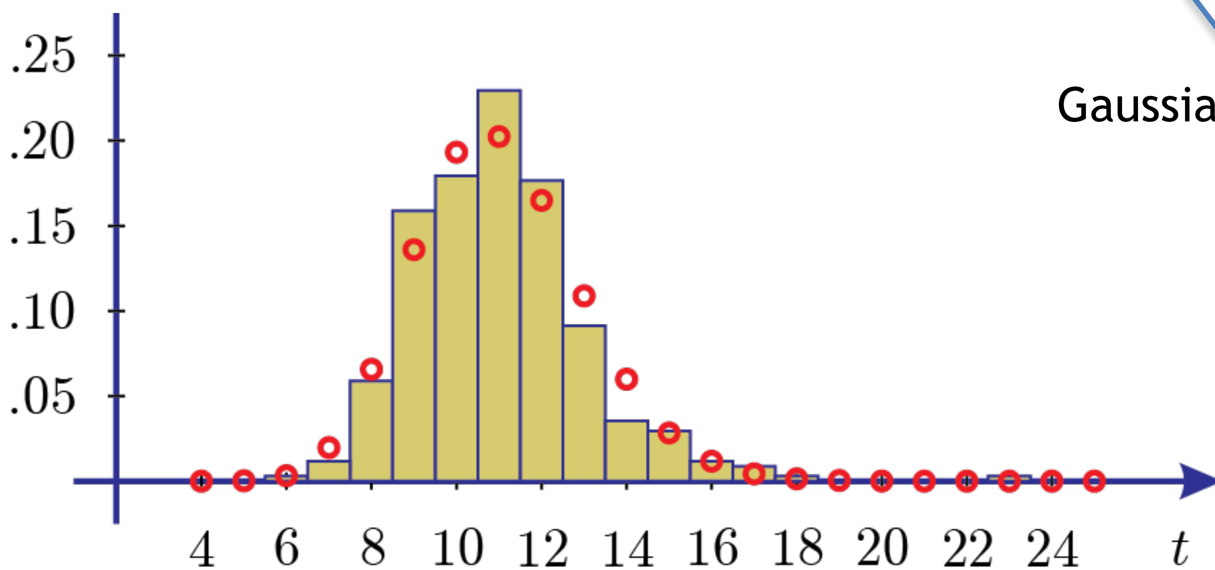
EXERCISE: ADDING IN AUXILIARY INFORMATION (1)

- Mean, variance for p(y | x) could depend on value of x
- Example: X = 1 if slippery out, and X = 0 else



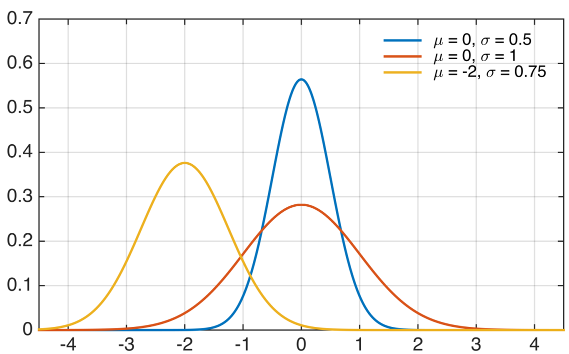
$$p(y|X=0) = \mathcal{N}\left(\mu_0, \sigma_0^2\right)$$
$$p(y|X=1) = \mathcal{N}\left(\mu_1, \sigma_1^2\right)$$

Gaussian denoted by N



Exercise: Adding in auxiliary information (2)

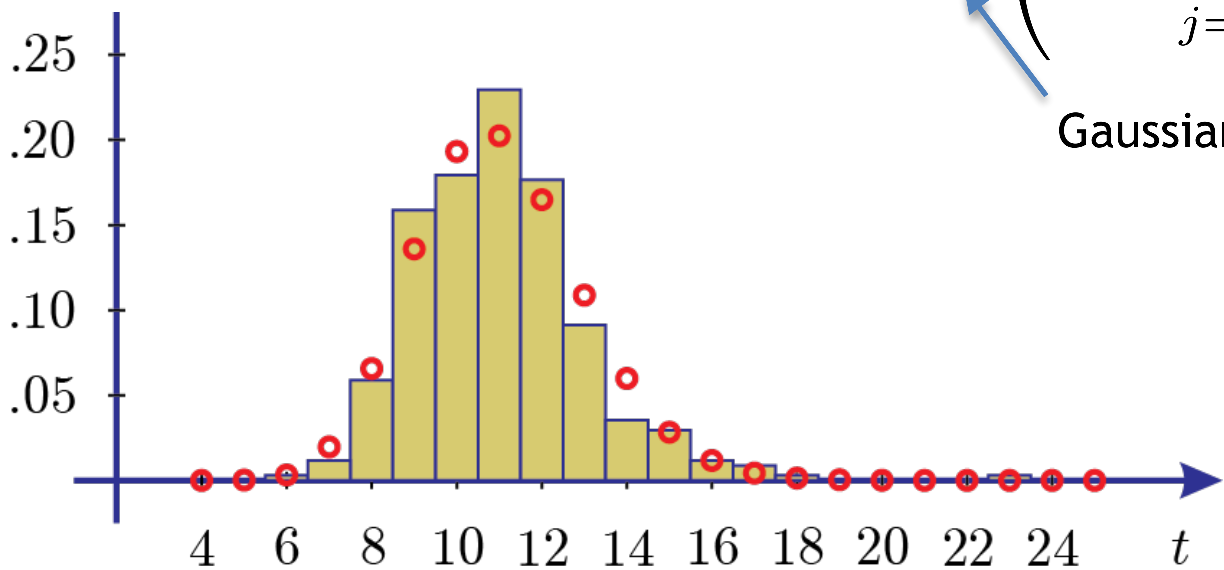
- Can incorporate external information (features) by modeling parameter = function(features)



$$\mu = \sum_{j=1}^{d} w_i x_i$$

$$p(y|\mathbf{x}) = \mathcal{N}\left(\mu = \sum_{j=1}^{d} w_i x_i, \sigma^2\right)$$

Gaussian denoted by N



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

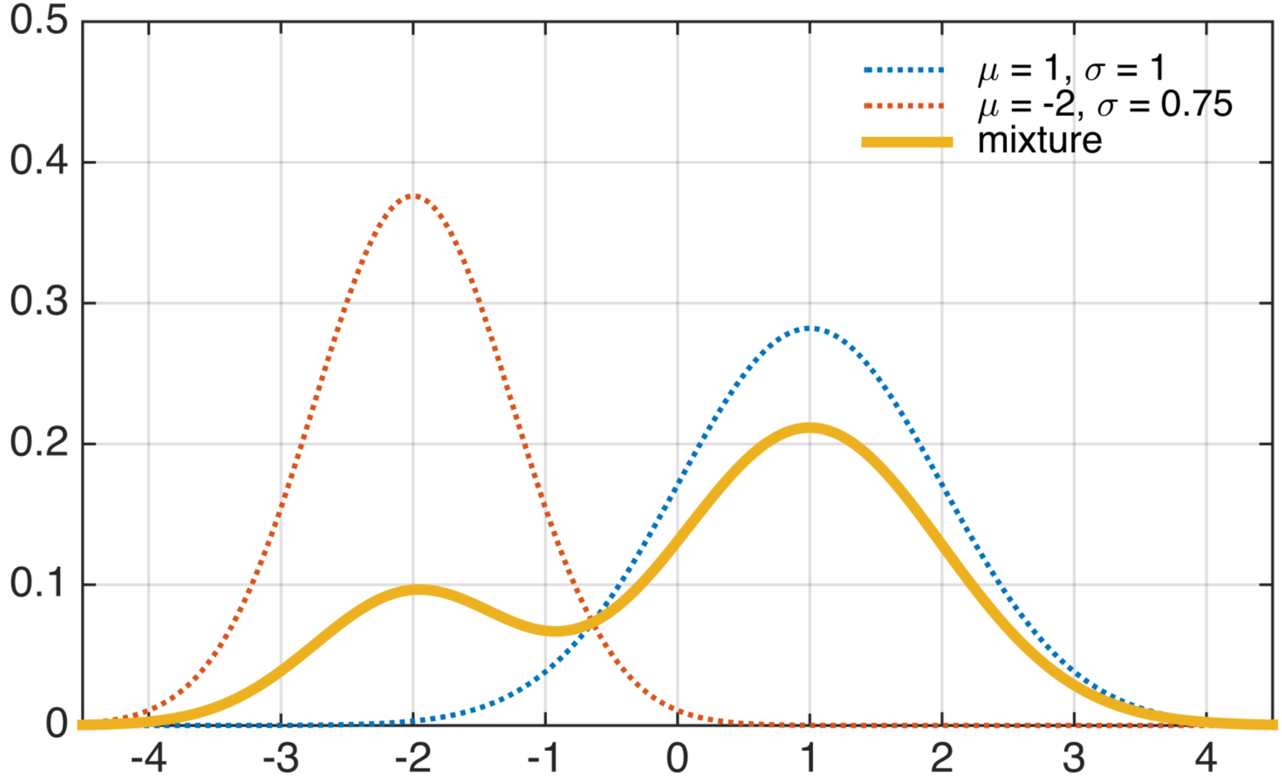
$$\sum_{i=1}^{m} w_i = 1$$

Mixtures of Gaussians

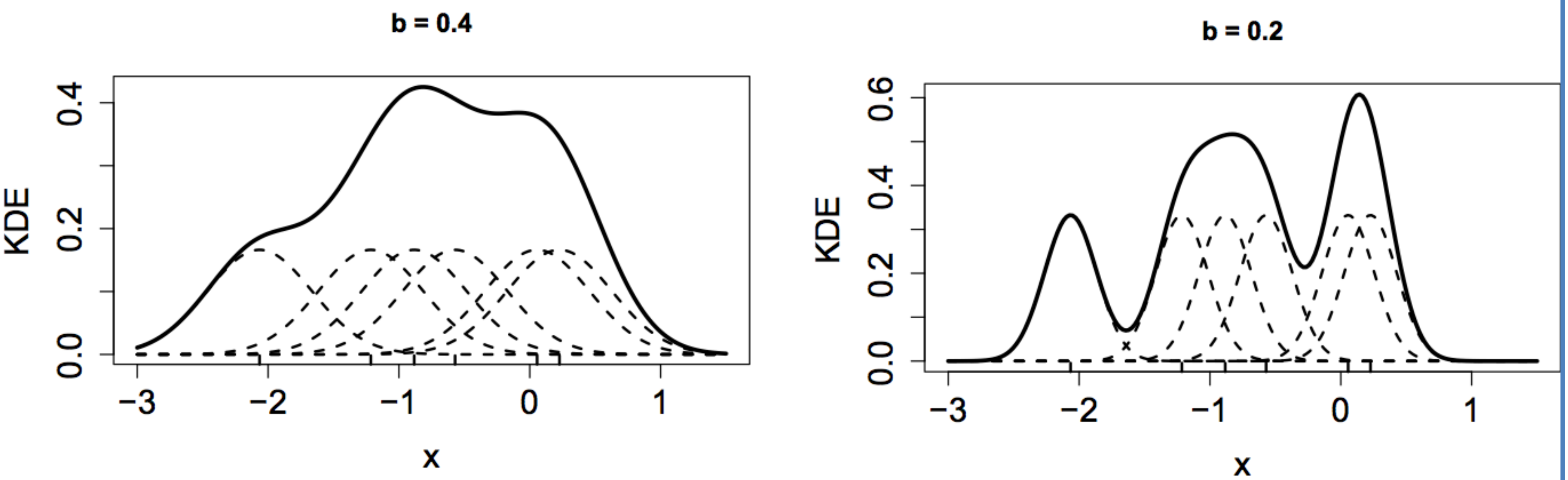
Mixture of $m = 2$ Gaussian distributions:

$w_1 = 0.75,\ w_2 = 0.25$

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



Mixtures can produce complex distributions



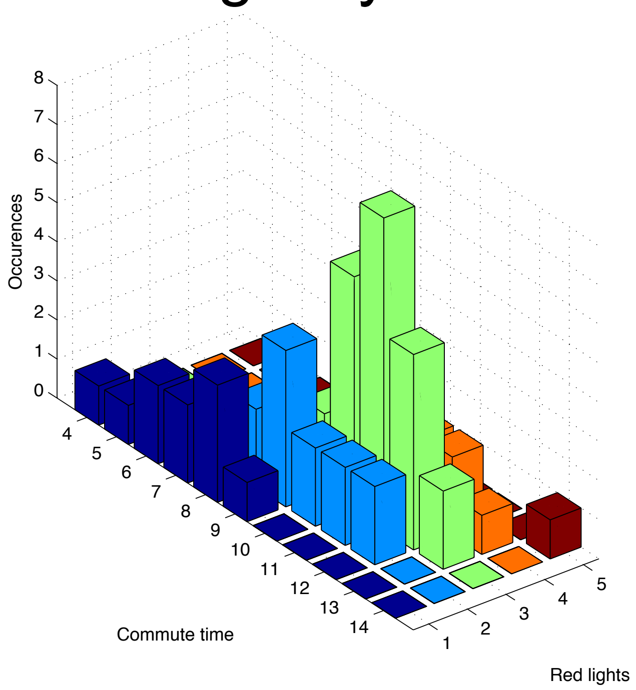
* Image from https://people.ucsc.edu/~ealdrich/Teaching/Econ114/LectureNotes/kde.html

Think-Pair-Share (5 minutes)

- Notice that moving to continuous RVs puts more restrictions on the distributions we can define
- For discrete RVs, distributions are tables of probabilities and so are highly flexible
  - we can define any possible distribution
- So, why not just discretize our variables?
- **Example**: imagine you have an RV in range [-10,10]
- You decide to discretize into chunks of size 0.01
- How many variables do you need to define the PMF?
- What if had instead modelled it as a Gaussian? Or a mixture of two Gaussians?

Think-Pair-Share

- **Example**: imagine you have an RV in range [-10,10]
- You decide to discretize into chunks of size 0.01
- How many variables do you need to define the PMF?
- What if had instead modelled it as a Gaussian? Or mixture of two Gaussians?
- **Additional question if you have time**: imagine you have a 2-dim. RV (in [-10, 10] x [-10,10]). Now imagine you discretize to the same level. How many variables in this PMF? (i.e., this multidimensional array?)



Exercise: Sample average is an 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 $\frac{1}{n}\sum_{i=1}^{n} x_i = \mu$