Probability Theory Review

CSCI-B455

Martha White
Department of Computer Science and Informatics
Indiana University, Bloomington

Spring, 2017

Reminders

- Assignment 1 is due on February 1
  - requires reading Chapters 1 and 2
- Thought questions 1 are due on January 25
  - Chapters 1, 2 and 3 (and read the preface)
- Recommendation: do not print out all the notes just yet; later parts will be slightly modified and improved
- See appendix for some background material
  - e.g. a notation sheet
- No class on Monday, January 16 (MLK Jr. day)

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

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 called a sigma field (sigma algebra)

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

Why not the simpler definition of finite unions?

In most cases, additivity is enough

2. $\forall A, B \in \mathcal{F}$ and $A \cap B = \emptyset \;\Rightarrow\; P(A \cup B) = P(A) + P(B)$

Why these seemingly arbitrary rules?

- These rules ensure nice properties of measures
- Other possibilities, these ones chosen

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

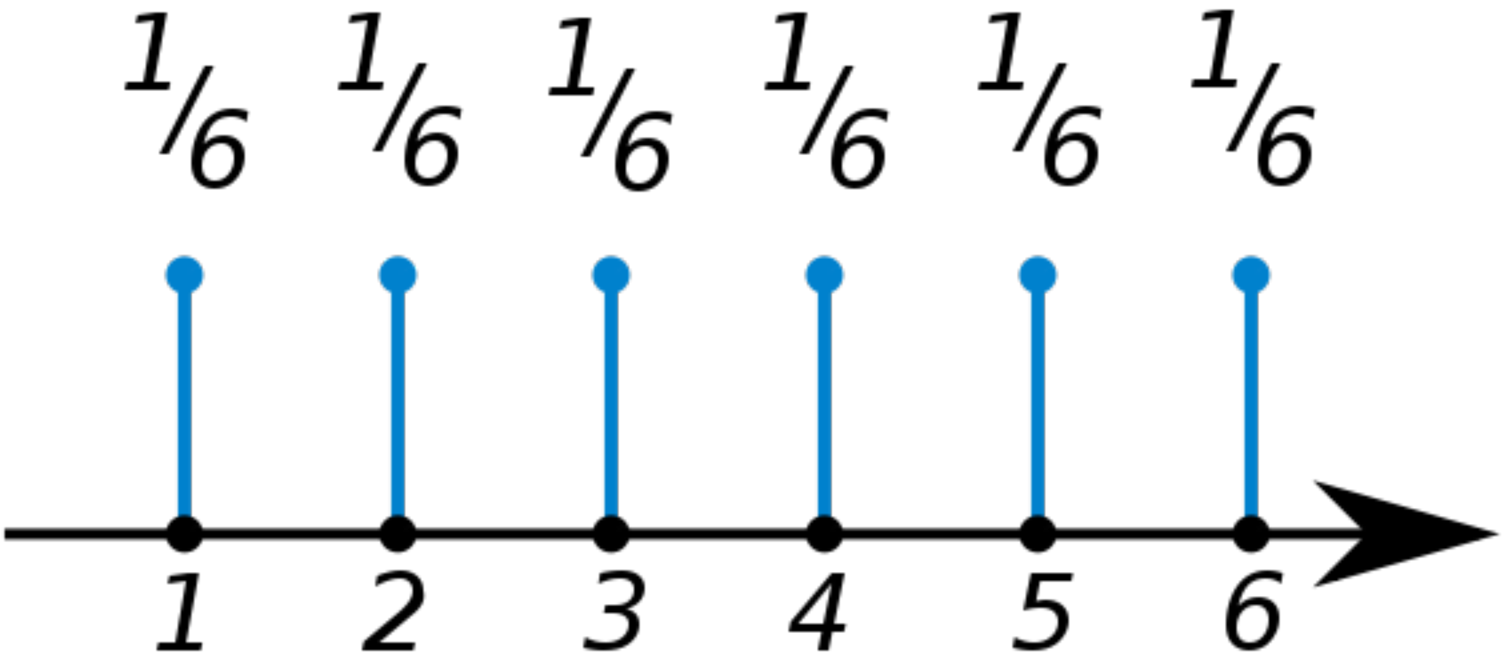
$$P(A) = \sum_{\omega \in A} p(\omega)$$

Arbitrary PMFs

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

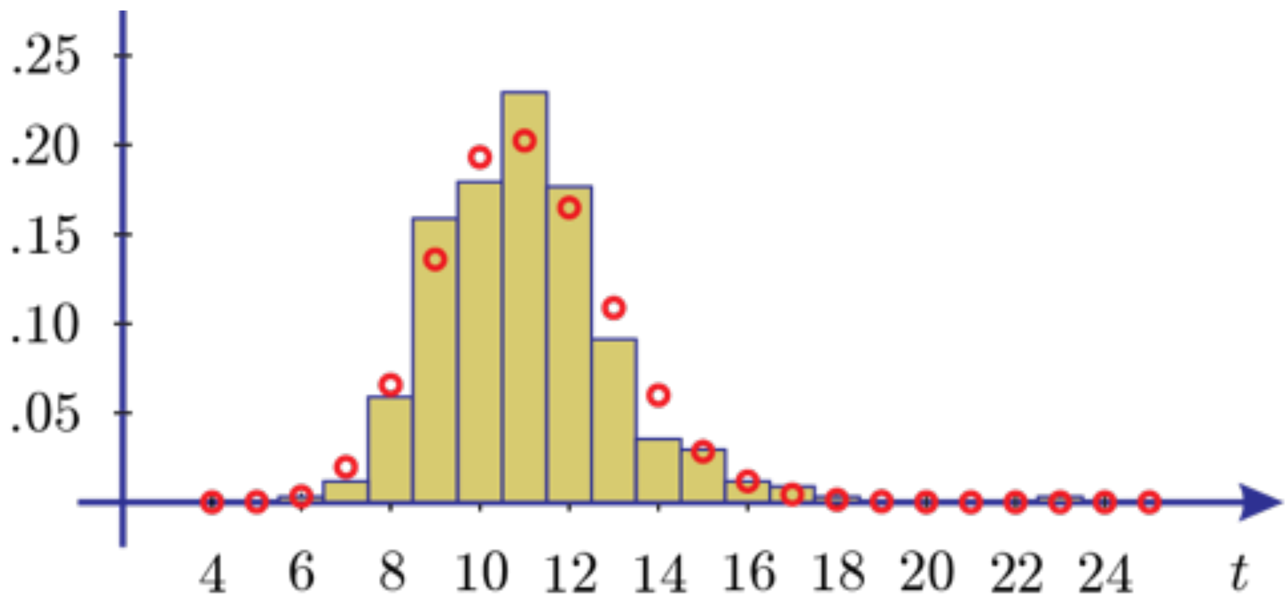
$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

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



Exercise: How are PMFs useful as a model?

- Recall we modeled commute times using a gamma distribution (continuous time t)
- Instead could use a probability table for minutes: count number of times t = 1, 2, 3, … occurs and then normalize probabilities by # samples
    - why normalize by number of 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

Binomial distribution: $\Omega = \{0, 1, \dots, n\} \quad \alpha \in (0,1)$

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

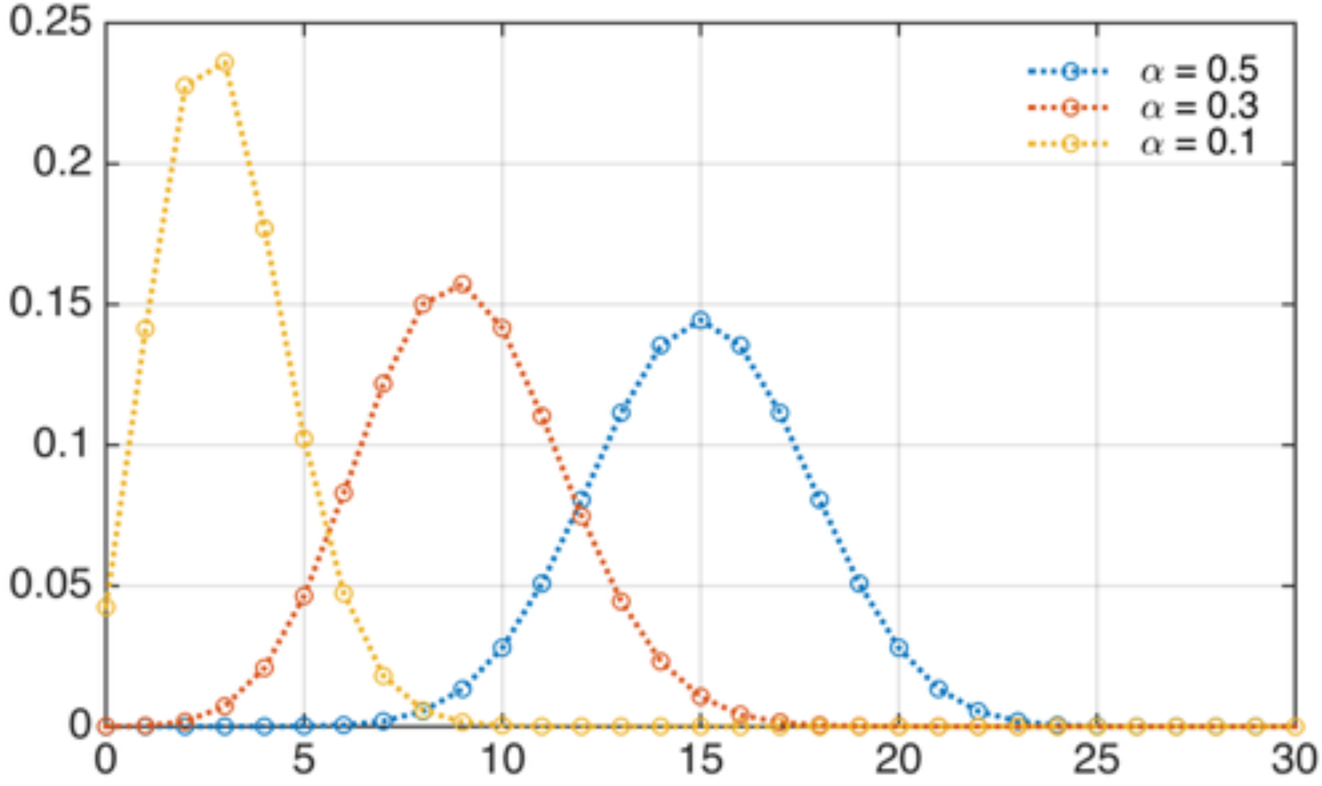
The values are k: the number of successes in a sequence of n independent 0/1 Bernoulli(α) experiments

http://www.math.uah.edu/stat/apps/BinomialCoinExperiment.html

USEFUL PMFs

Binomial distribution: $\Omega = \{0, 1, \ldots, n\} \quad \alpha \in (0, 1)$

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



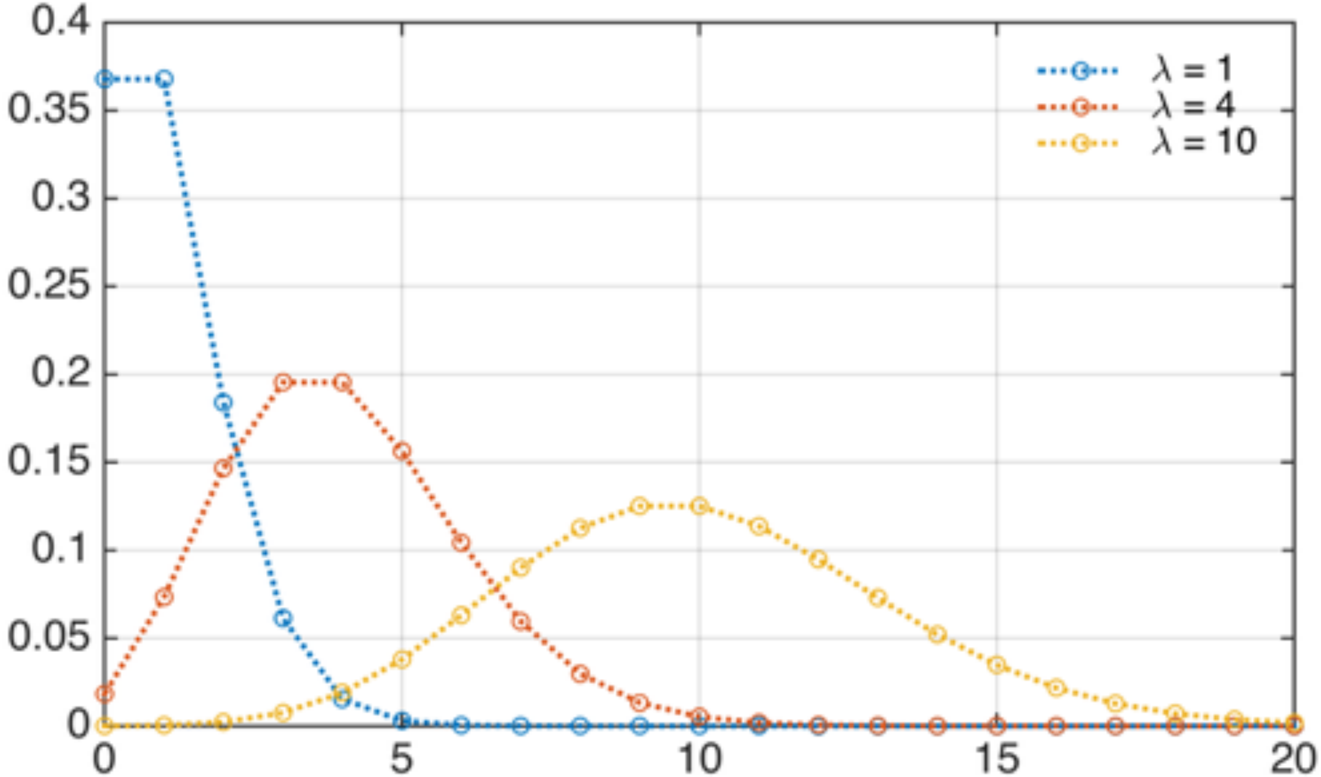
$n = 30$

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

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



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

$\Omega =$ 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

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

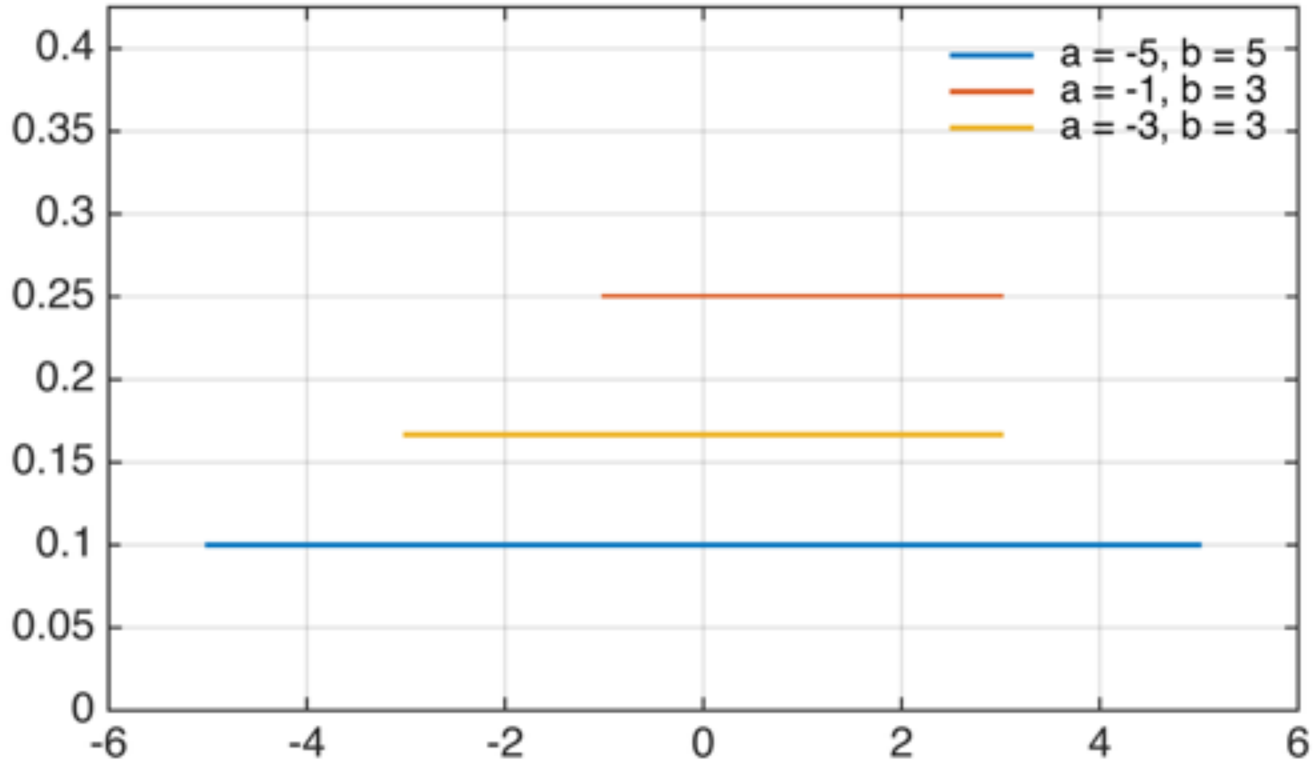
A FEW COMMENTS ON TERMINOLOGY

- A few new terms, including countable, closure
  - only a small amount of terminology used, can google these terms and learn on your own
  - notations sheet in Appendix of notes
- Countable: integers, rational numbers, …
- Uncountable: real numbers, intervals, …
- Why this matters: measures (probability) on these sets is different
- Example: for discrete uniform distribution on {0.1,2.0,3.6}, what is the probability of seeing 3.6?
- Example: for uniform distribution on [0,1], what is the probability of seeing 0.1?

USEFUL PDFs

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

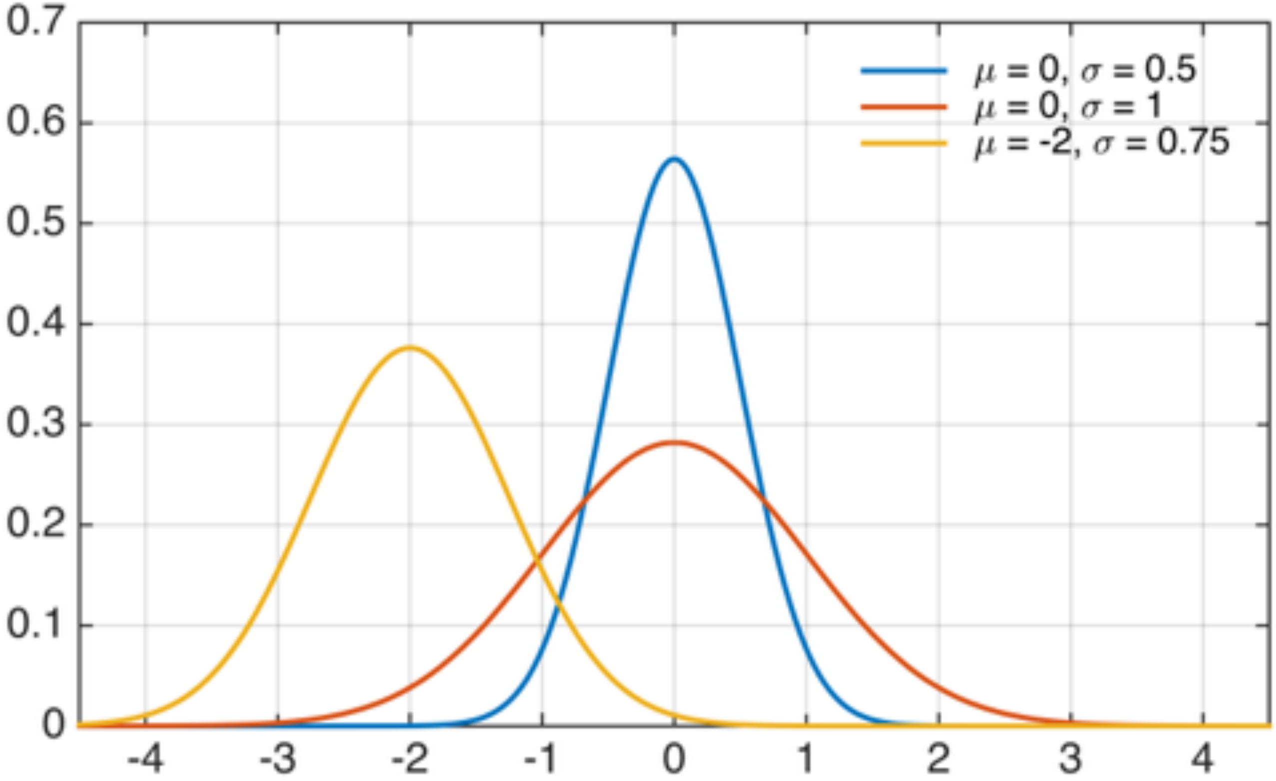
$$p(\omega) = \frac{1}{b-a} \qquad \forall \omega \in [a, b]$$



USEFUL PDFS

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

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

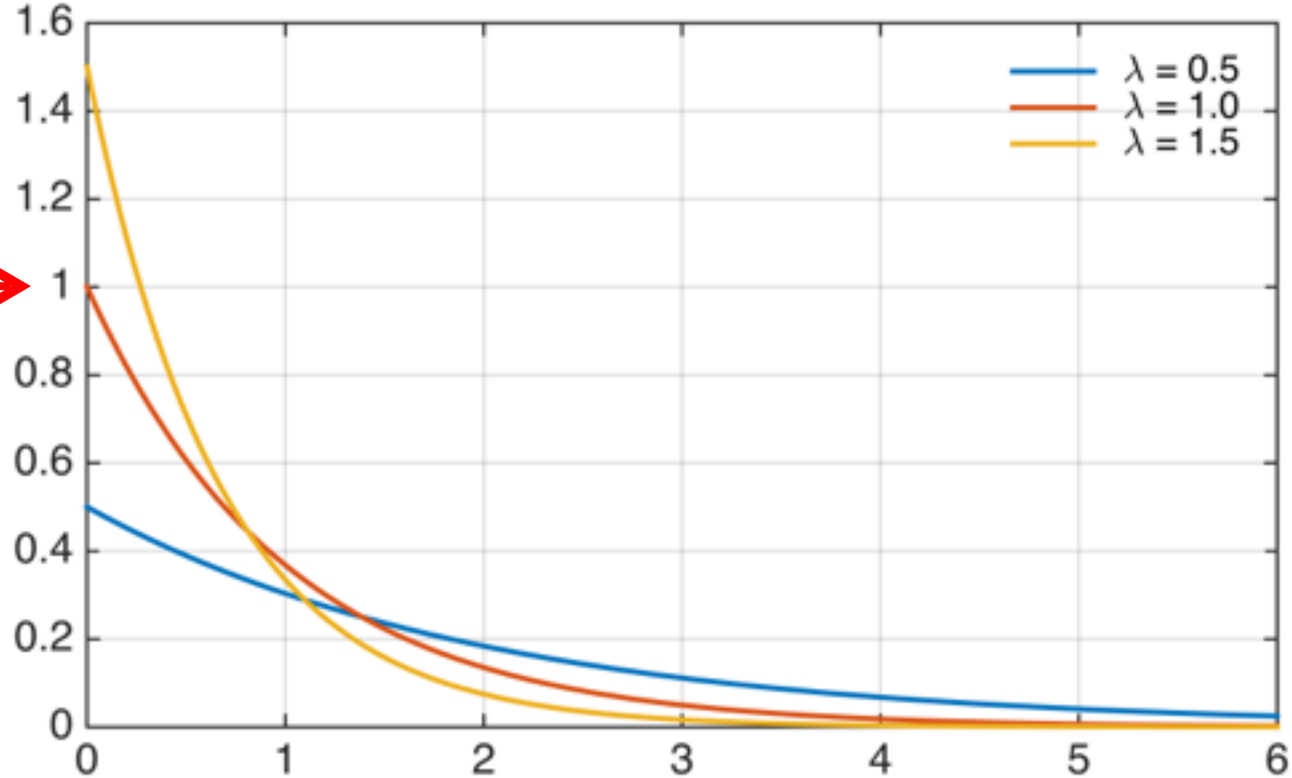


Useful PDFs

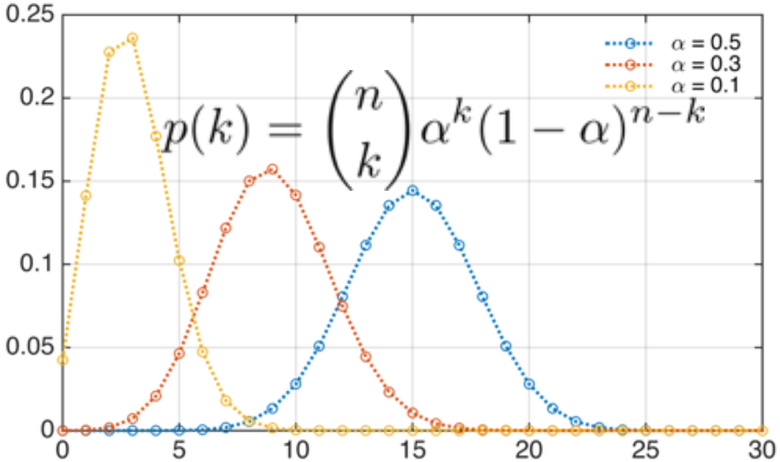
Exponential distribution: $\Omega = [0, \infty) \quad \lambda > 0$

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

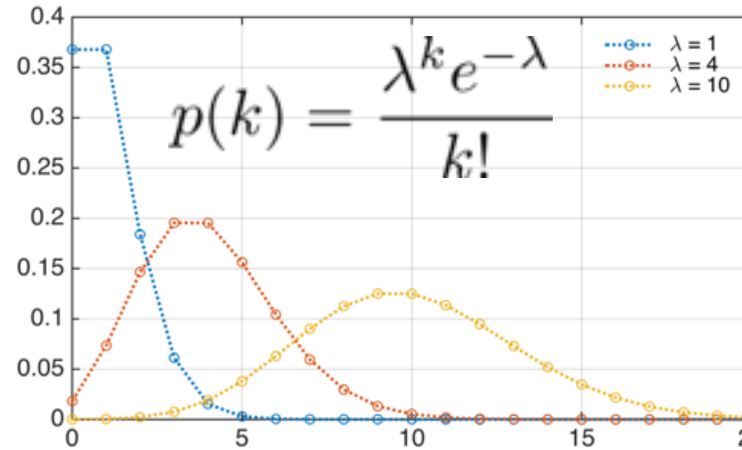
1 is here!



Exercise: modeling commute times

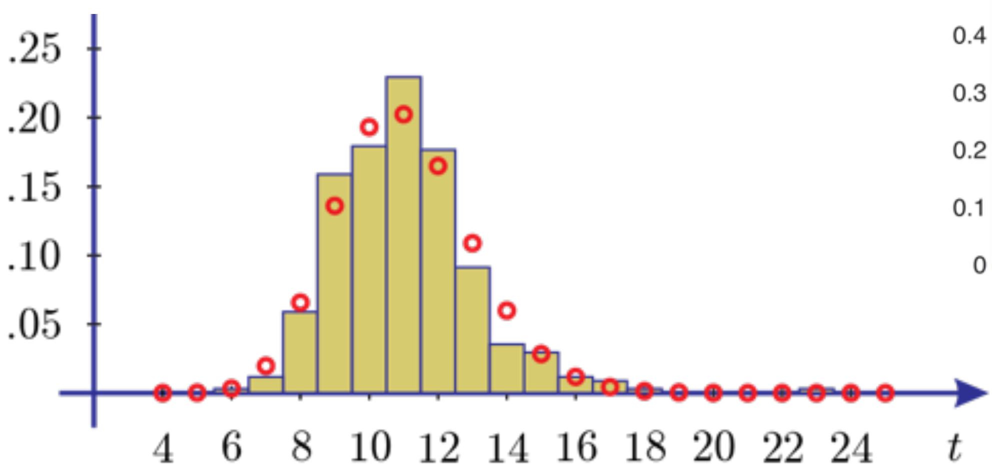


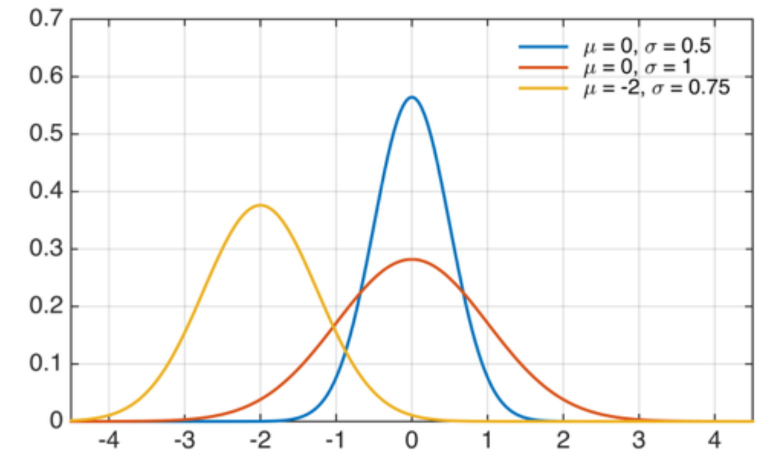
Binomial



Poisson

Which might you choose?





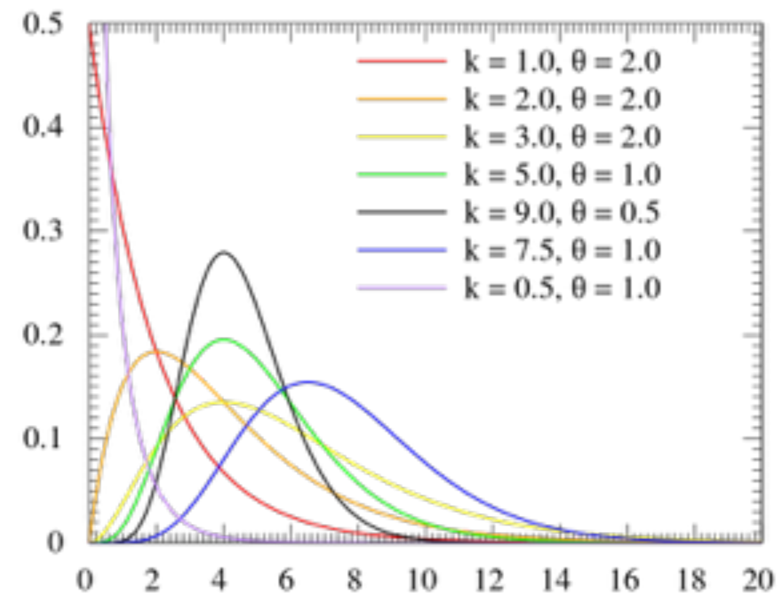
Gaussian

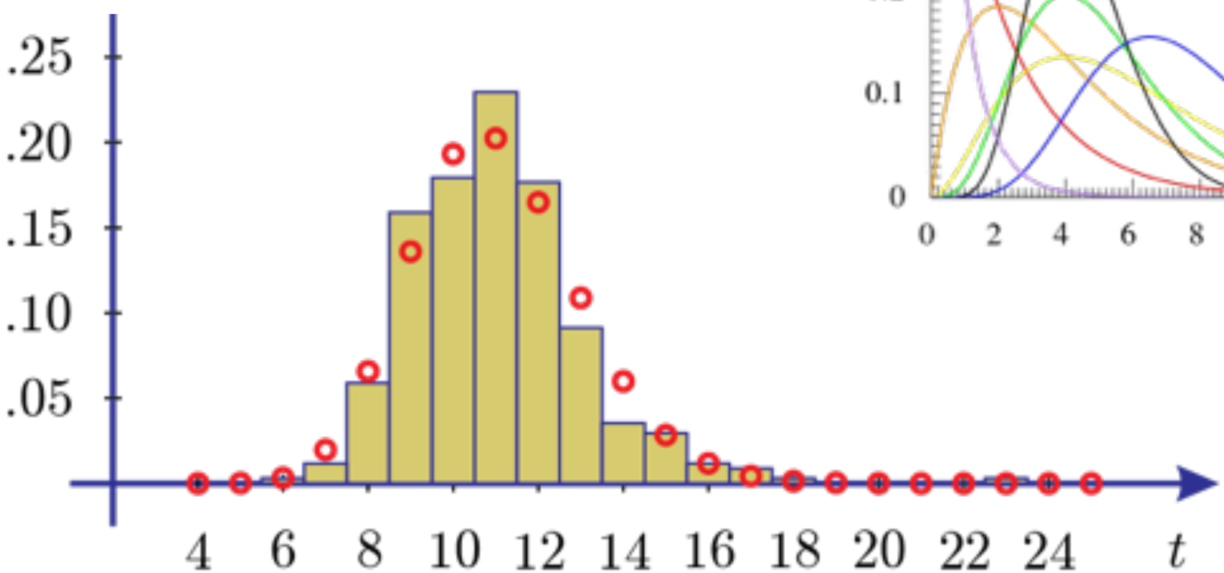
$$p(\omega) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\omega-\mu)^2}$$

EXERCISE: UTILITY OF PDFS AS A MODEL

- Gamma distribution for commute times extrapolates between recorded time in minutes

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





Multidimensional PMFs

$\Omega = \Omega_1 \times \Omega_2 \times \ldots \times \Omega_k$
$\mathcal{F} = \mathcal{P}(\Omega)$

Probability mass function:

1. $p : \Omega_1 \times \Omega_2 \times \ldots \times \Omega_k \to [0, 1]$
2. $\sum_{\omega_1 \in \Omega_1} \cdots \sum_{\omega_k \in \Omega_k} p\left(\omega_1, \omega_2, \ldots, \omega_k\right) = 1$

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

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

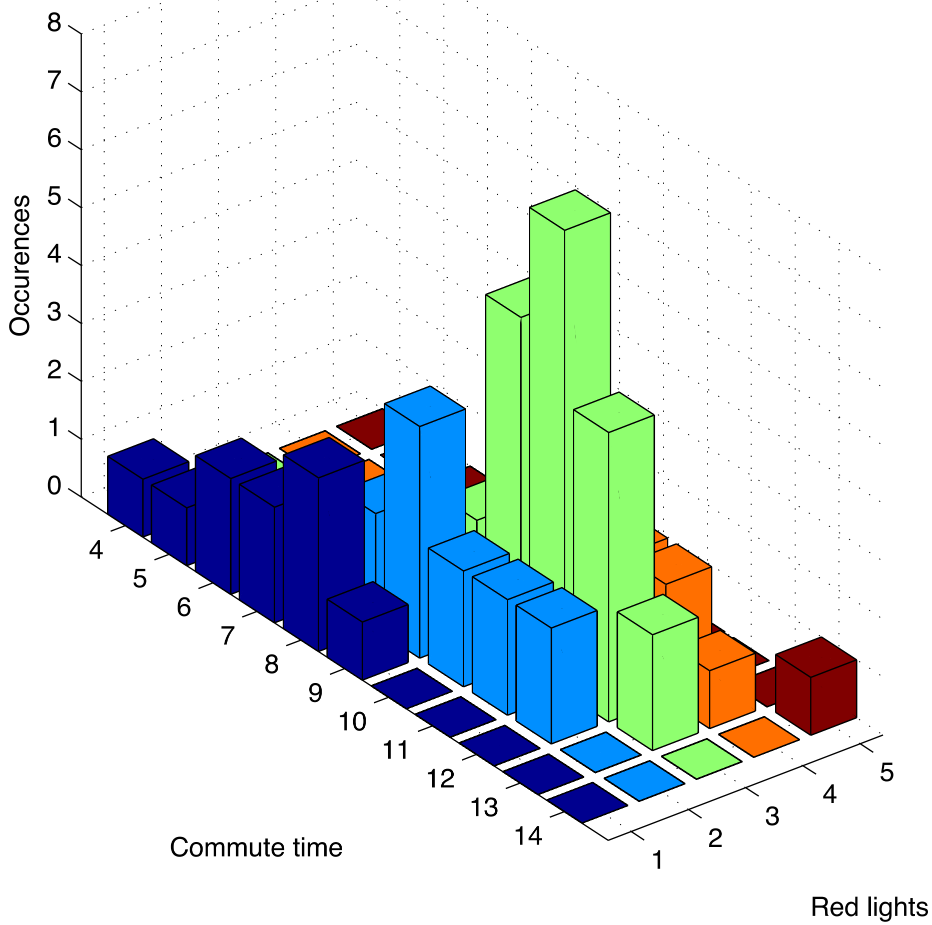
$\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_k)$

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 PDFs

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

Probability density function:

1. $p : \mathbb{R}^k \to [0, \infty)$
2. $\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p(\omega_1, \omega_2, \ldots, \omega_k) d\omega_1 \cdots d\omega_k = 1$

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

$$P(A) = \int_{\boldsymbol{\omega} \in A} p(\boldsymbol{\omega}) d\boldsymbol{\omega}.$$

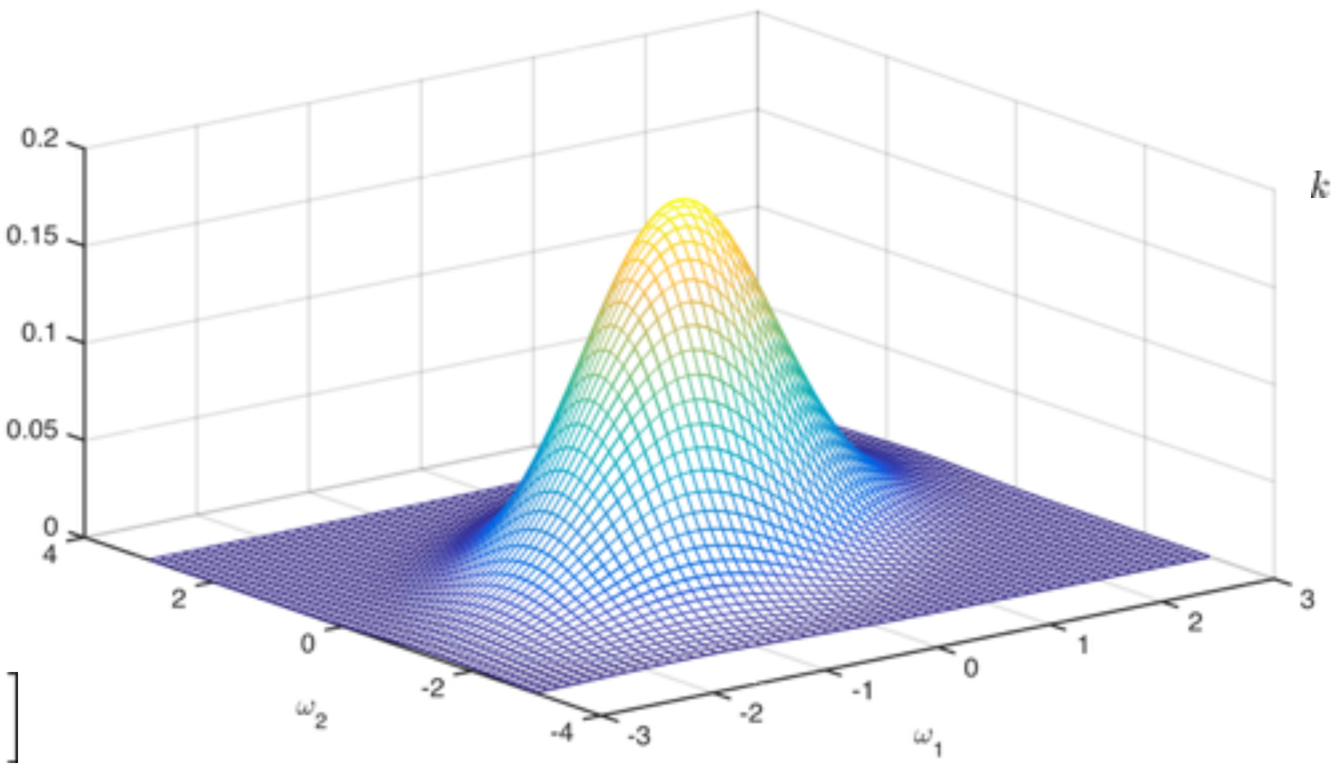
$\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_k)$

Multidimensional Gaussian

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

$\boldsymbol{\mu} \in \mathbb{R}^k$
$\boldsymbol{\Sigma}$ = positive definite k-by-k matrix
$|\boldsymbol{\Sigma}|$ = determinant of $\boldsymbol{\Sigma}$

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



$k = 2$

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

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

Quick survey

- Who has heard of vectors?
- Who has heard of dot products?
- Who has heard of matrices?

Multiple variables

- A vector can be thought of as a 1-dimensional array of length d
- A matrix can be thought of as a 2-dimensional array, of dimension n x d

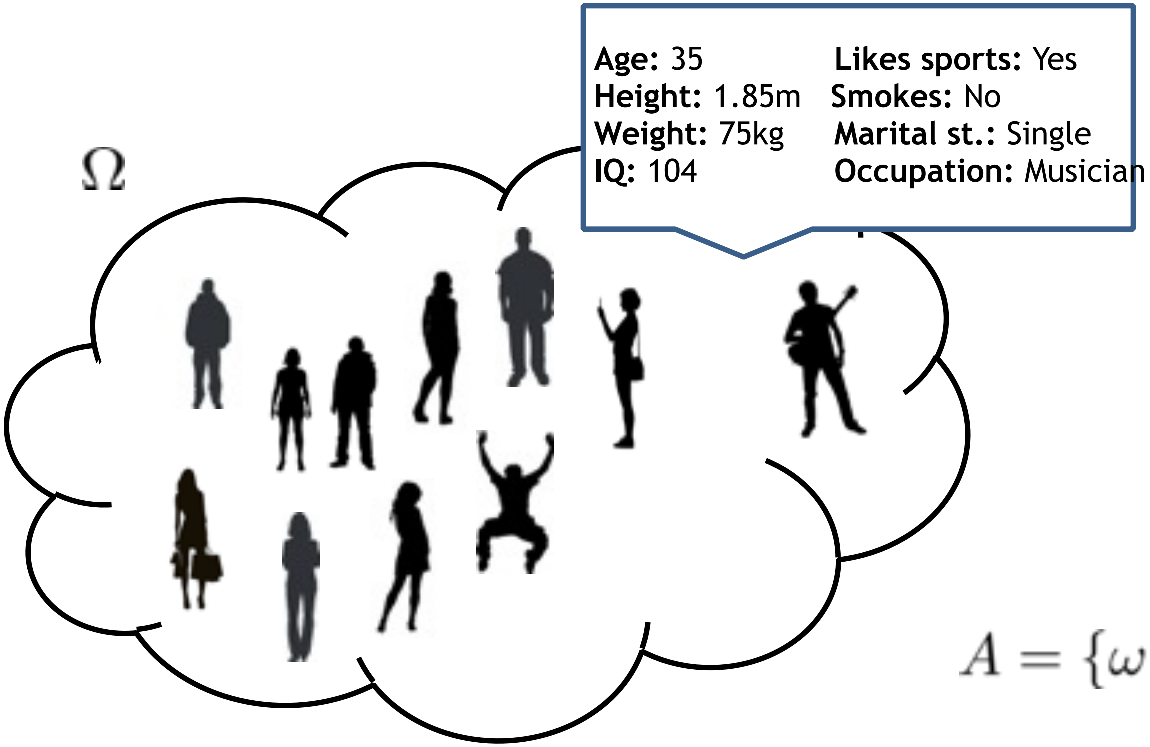
Two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$

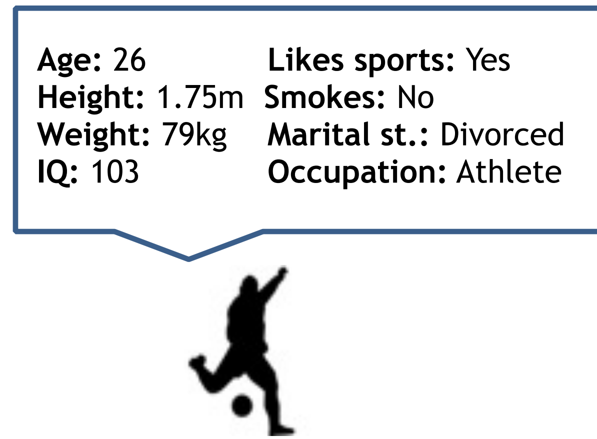
Dot product $\mathbf{x}^\top \mathbf{y} = \sum_{i=1}^{d} x_i y_i$

Random Variables

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

Ω





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

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 $\{\text{cold}, \text{not cold}\}$

Random Variables

Example: three consecutive (fair) coin tosses
X = the number of heads in the first toss
Y = the number of heads in all three tosses
Find the probability spaces after the transformations.

Where is the probability space $(\Omega, \mathcal{F}, P)$?

Where is the randomness?

$$\Omega = \{\text{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}\}$$

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

$$P = \,?$$

$$P(\Omega) = 1$$

$$P(\{\text{HHH}, \text{TTT}\}) = \tfrac{2}{8}$$

$$\vdots$$

Random Variables

$$X : \Omega \to \{0, 1\}$$
$$Y : \Omega \to \{0, 1, 2, 3\}$$

ω	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT
$X(\omega)$	1	1	1	1	0	0	0	0
$Y(\omega)$	3	2	2	1	2	1	1	0

What are the probability spaces $(\Omega_X, \mathcal{F}_X, P_X)$ and $(\Omega_Y, \mathcal{F}_Y, P_Y)$?

Where does the randomness come from?

Once we have these new spaces, same pdf and pdf definitions apply

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

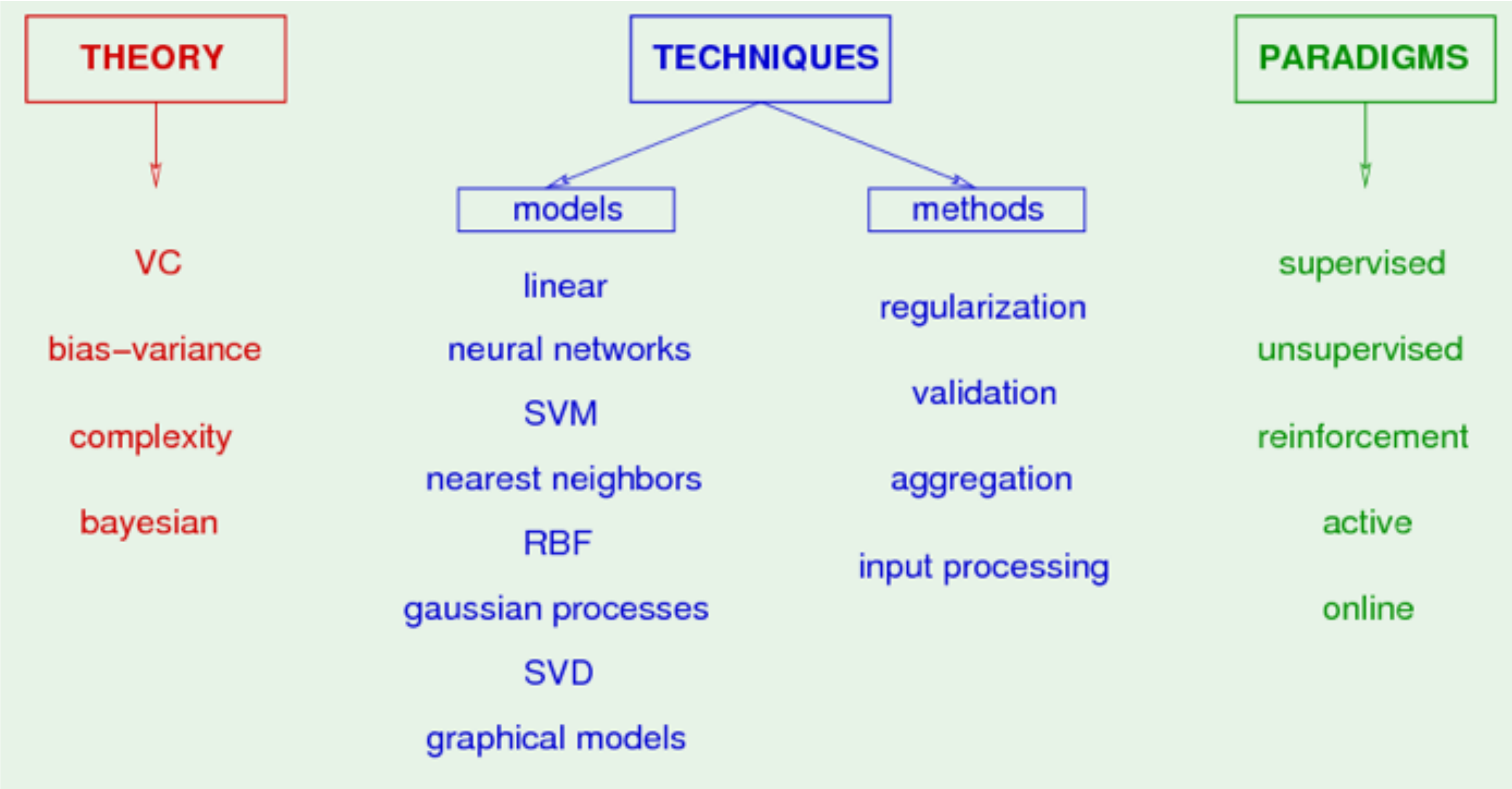
with associated measurements such as height and weight

$X(\omega) =$ height

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

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

JAN. 18: PROBABILITY REVIEW CONTINUED



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

Reminders

- Assignment 1 is due on February 1
- Thought questions 1 are due on January 25
- Office hours
  - Martha: 3-5 p.m. on Tuesday (LH 401E)
  - Inhak: 3:30 - 5:30 p.m. on Tuesday (LH 325)
  - Andrew: 12:00 - 2:00 p.m. on Thur (LH 215D)
- Up-front background material
  - Immersion style: understanding more in-depth as we use the ideas from probability
- Lecture notes posted before class
- I do not expect you to know formulas, like pdfs

Key points so far

- Many of our variables will be random
- These random variables can be discrete or continuous
  - discrete e.g. {0, 1, 2}
  - continuous, e.g. [-100, 100]
- Several named PMFs and PDFs to provide distributions over the possible values
  - why? explicit functional form will be useful later
  - e.g., p(x) = lambda exp(-lambda x)
- Multi-variate distributions natural extensions of scalar distributions; probabilities over vector instances, e.g., x in [-10,10]^2

Conditional Distributions

Conditional probability distribution:

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

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

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

Dropping subscripts

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

We will write: $$p(y|x) = \frac{p(x,y)}{p(x)}$$

EXAMPLE

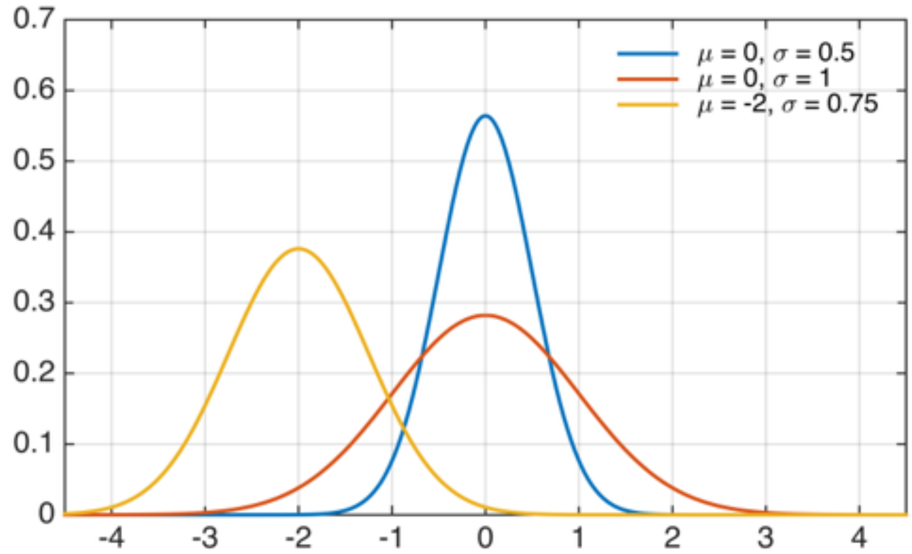
- Let X be a Bernoulli random variable (i.e., 0 or 1 with probability alpha)
- Let Y be a random variable in {10, 11, …, 1000}
- p(y | X = 0) and p(y | X = 1) are different distributions
- Two types of books: fiction (X=0) and non-fiction (X=1)
- Let Y corresponds to number of pages
- Distribution over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



What do we know about p(y)?

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

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

Chain Rule

Conditional probability distribution:

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

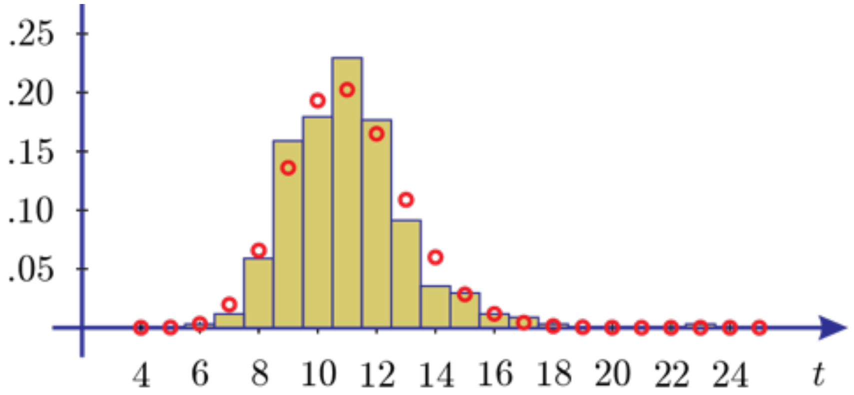
This leads to:

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

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

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?

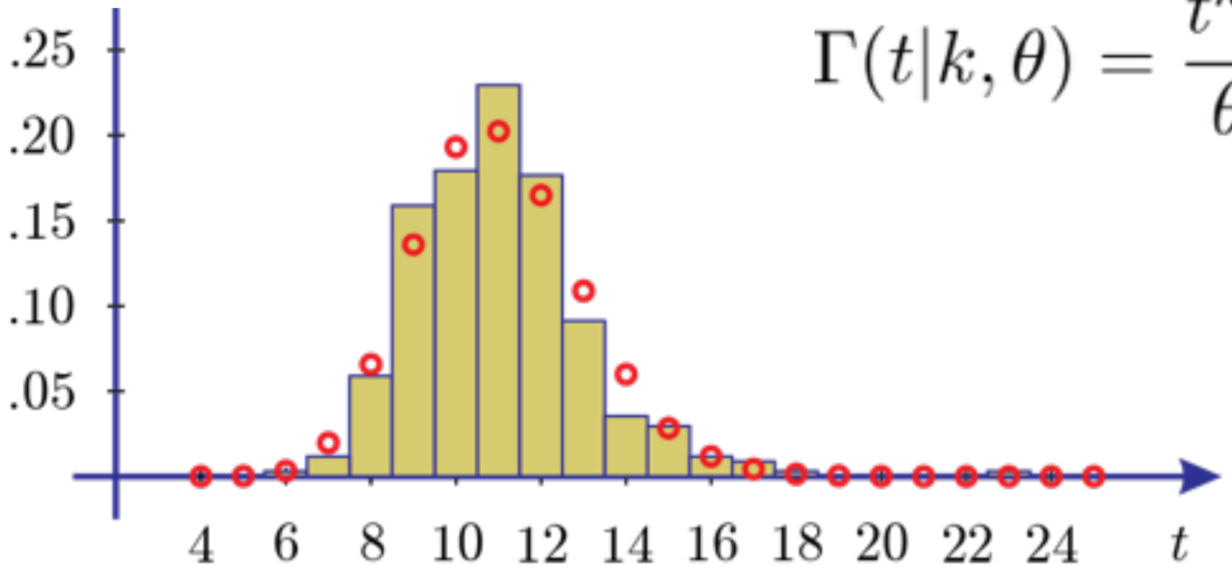


Exercise: Adding in auxiliary information

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

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

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



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:

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

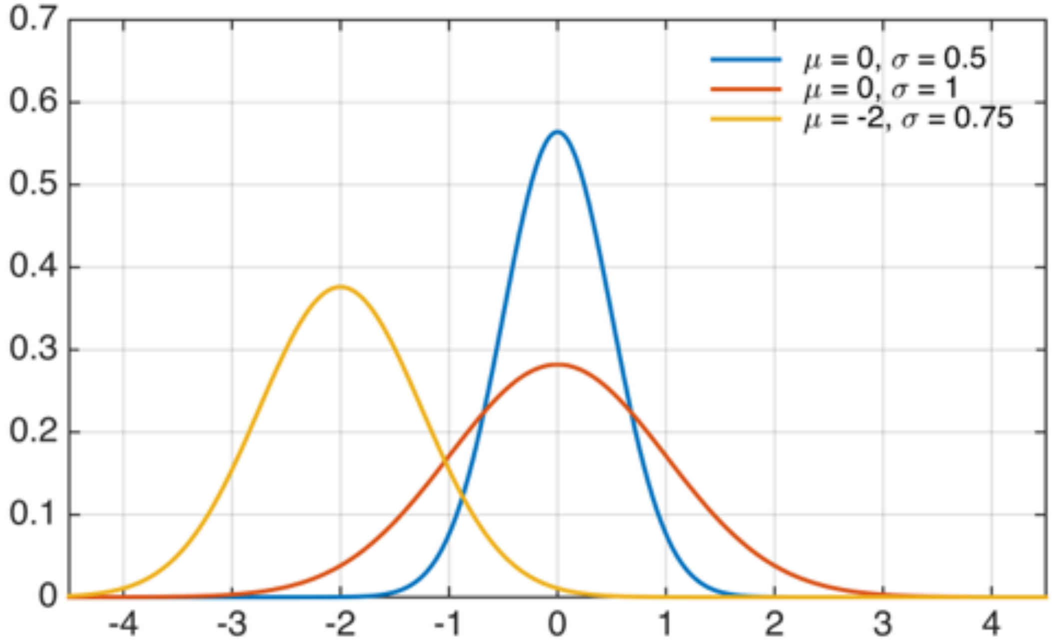
Exercise: what if we had k random variables, $X_1, \ldots, X_k$?

Conditional Independence Examples

- 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)

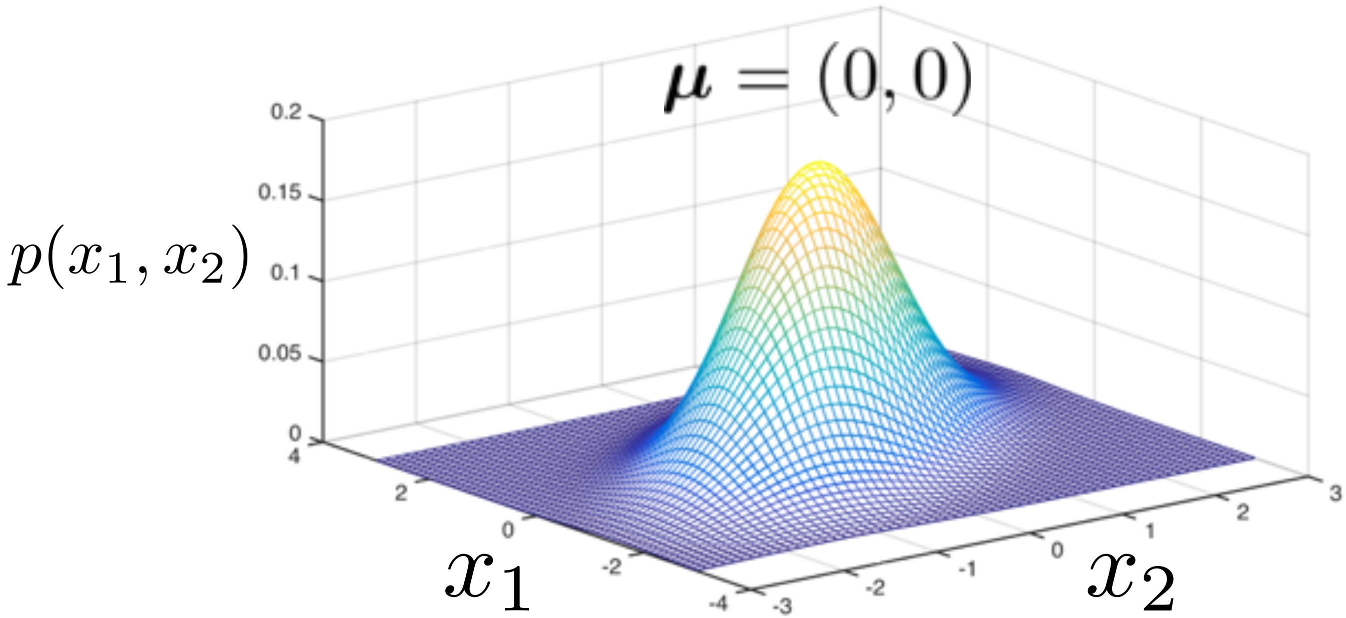
$$\mathbb{E}[X] = \begin{cases} \sum_{x \in \mathcal{X}} x p(x) & X : \text{discrete} \\ \\ \int_{\mathcal{X}} x p(x) dx & X : \text{continuous} \end{cases}$$



Expected value for multivariate

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

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



Conditional Expectations

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

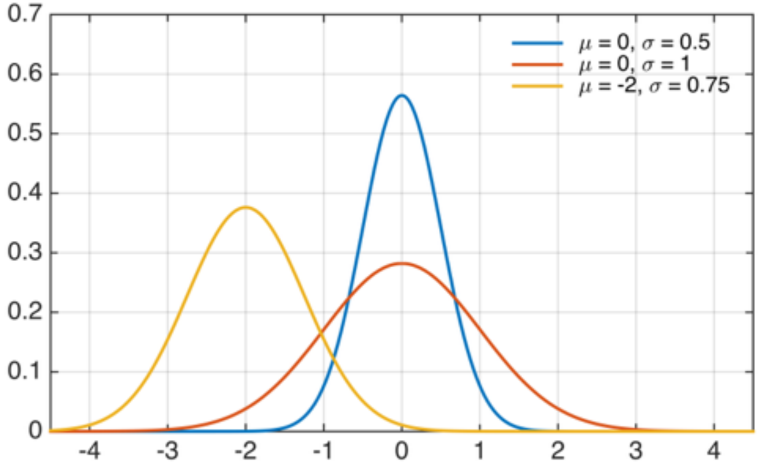
Different expected value, depending on which x is observed

Exercise: RVs, PDFs and Uncertainty

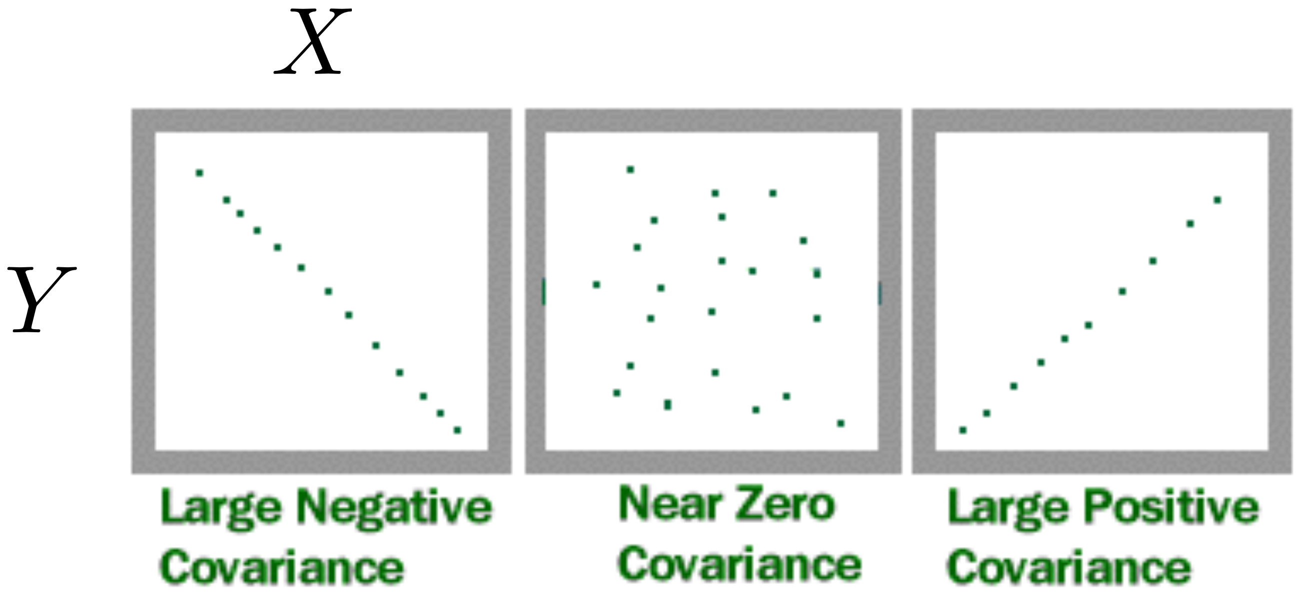
- In ML, common strategy to assume trying to learn a deterministic function, from noisy measurements
- Denoised "truth": y = f(x)
- Noisy observation: f(x) + noise
  - one common assumption is the noise N is a Gaussian RV
  - E[f(x) + noise] = f(x) + E[noise] = f(x) = E[Y | x]
- For a sample x of RV X:

$$N \sim \mathcal{N}(0, \sigma^2)$$

$$Y = f(x) + N \sim \mathcal{N}(f(x), \sigma^2)$$



Covariance



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

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

Covariance for more than two dimensions

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

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

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

Covariance for more than two dimensions

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

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

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

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. $\mathrm{Cov}[\boldsymbol{X}+\boldsymbol{Y}] = \mathrm{V}[\boldsymbol{X}] + \mathrm{V}[\boldsymbol{Y}] + 2\mathrm{Cov}[\boldsymbol{X}, \boldsymbol{Y}]$

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 $\frac{1}{n}\sum_{i=1}^{n} x_i = \mu$

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

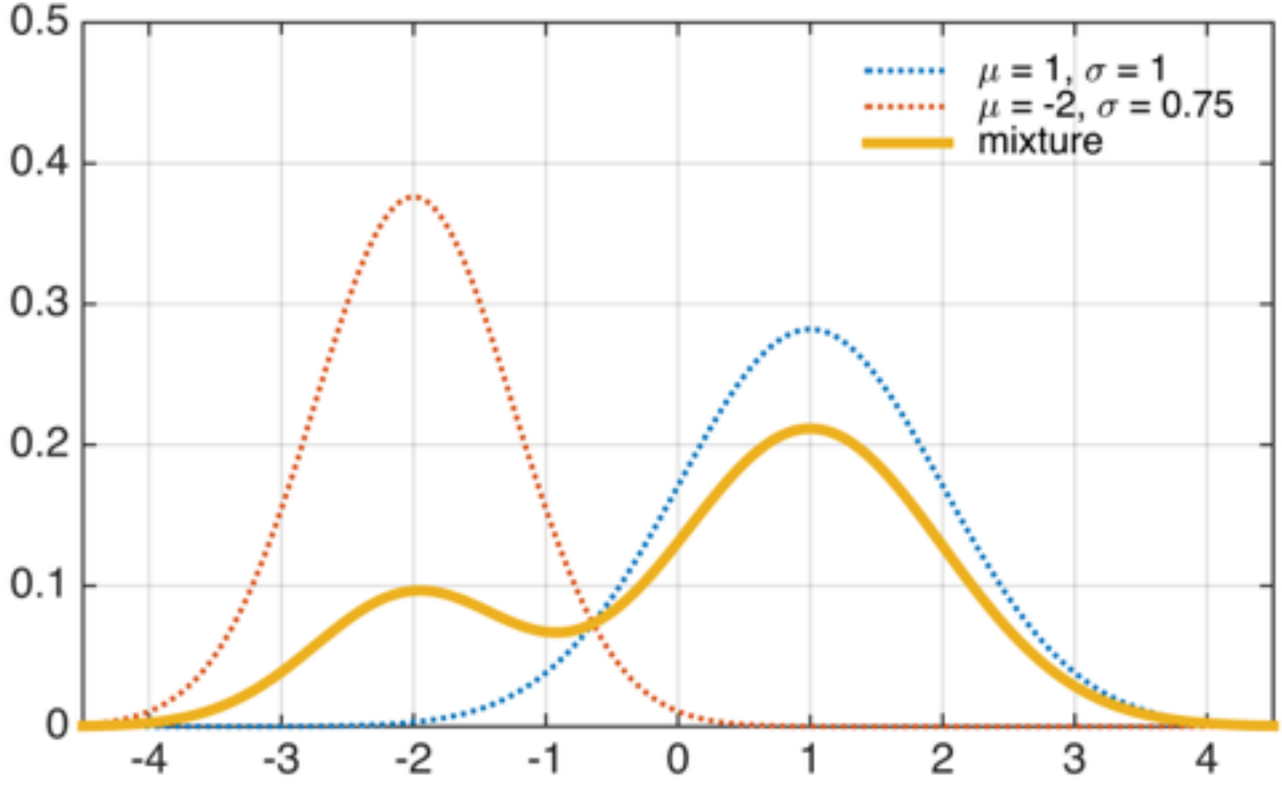
$$\sum_{i=1}^{m} w_i = 1$$

MIXTURES OF GAUSSIANS

Mixture of $m = 2$ Gaussian distributions:

$w_1 = 0.75,\ w_2 = 0.25$

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



Summary: Parametric Models

- We will consider many parametric models in machine learning
- To model the data, we can pick a parametric class and do parameter estimation (next)
- Given a model, we can make statements about our data
  - predict target given inputs (conditional probs)
  - find underlying structure of data
  - find explanatory variables
  - …
- We will incrementally generalize the types of models we consider to model our data