PROBABILITY THEORY REVIEW

CSCI-B555

Martha White
DEPARTMENT OF COMPUTER SCIENCE AND INFORMATICS
INDIANA UNIVERSITY, BLOOMINGTON

Fall, 2015

Reminders

- Assignment 1 is due on September 16
- Thought questions 1 are due on September 9
  - Chapters 1 and 2
- CS Colloquium: Friday, 3:00 p.m. in LH 102
  - in general attending talks is good
- Proficiency in LaTex?
- Proficiency in a programming language?
- Anonymous course feedback
- Waiting list

(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)

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

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

Consequences of the Axioms of Probability

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

1. $P(\emptyset) = 0$
2. $P(A^c) = 1 - P(A)$
3. $P(A) = \sum_{i=1}^{k} P\left(A \cap B_i\right)$, where $\{B_i\}_{i=1}^{k}$ is a partition of Ω
4. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

… and everything else.

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

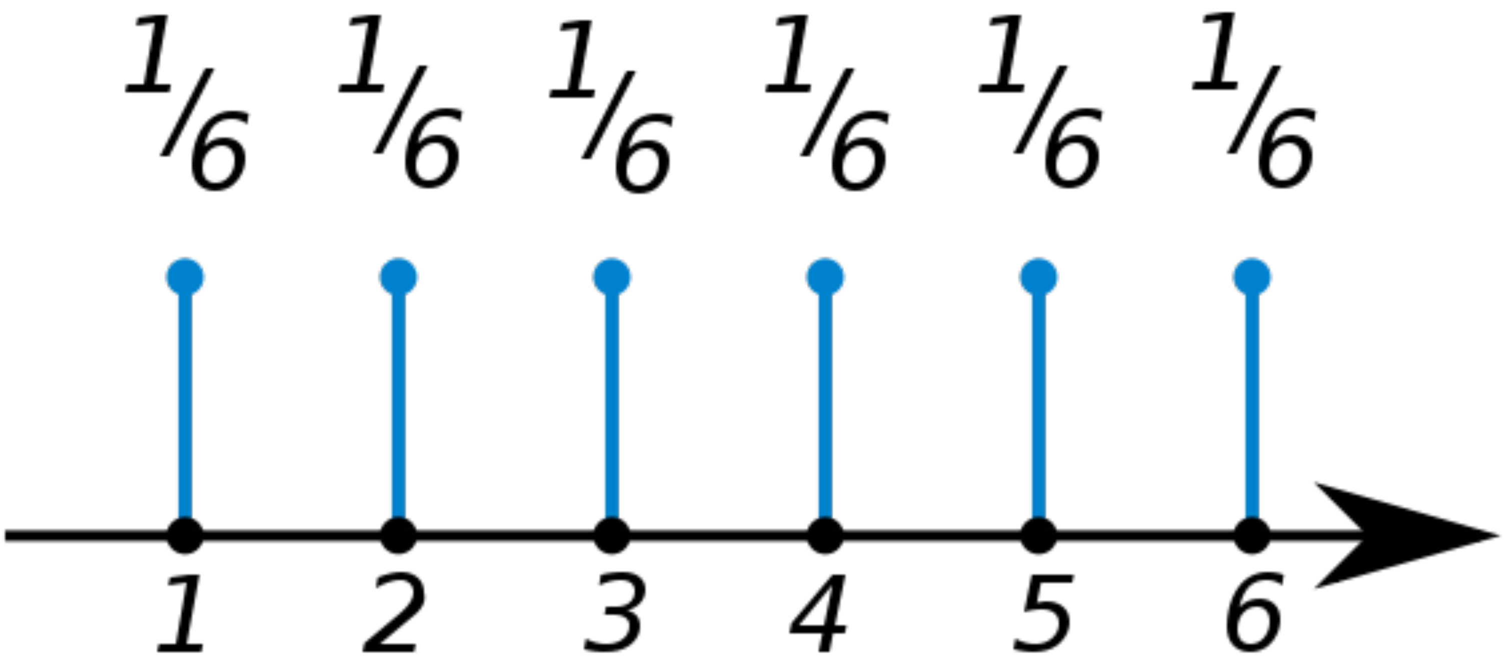
$$P(A) = \sum_{\omega \in A} p(\omega)$$

ARBITRARY PMFS

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

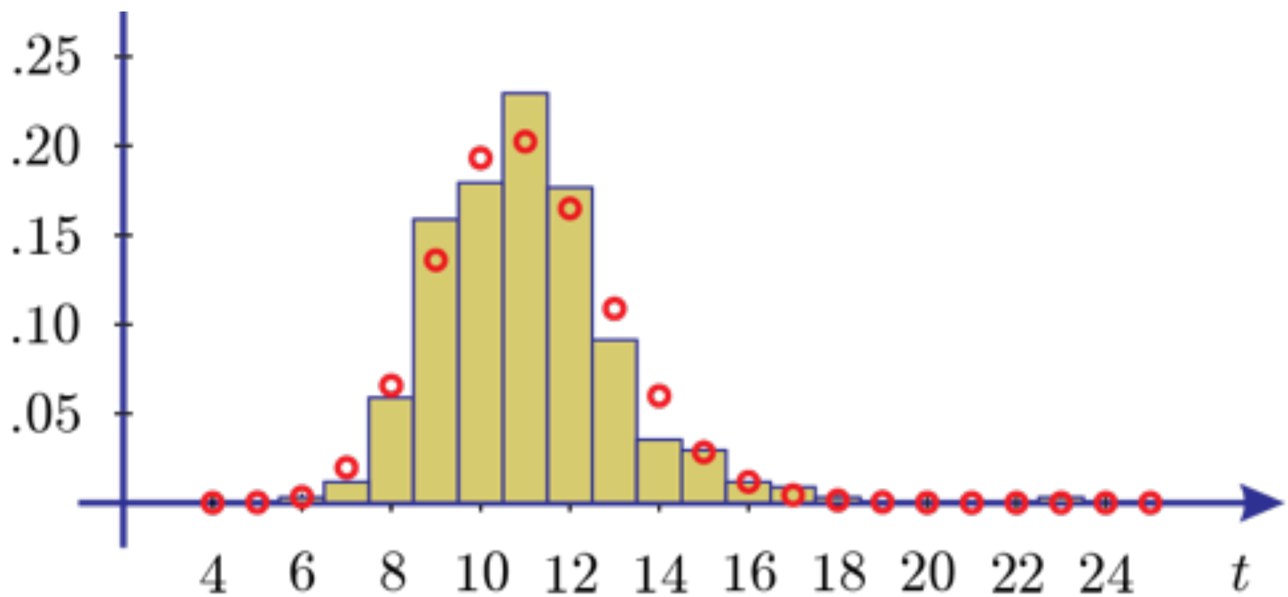
$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

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



EXERCISE: HOW ARE PMFS USEFUL AS A MODEL?

- Recall we modeled commute times using a gamma distribution (continuous time t)
- Instead we could have used a probability table for minutes: count number of times t = 1, 2, 3, … occurs and then normalize probabilities
- Pick t with the largest p(t)



PMFs You've Heard About

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

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

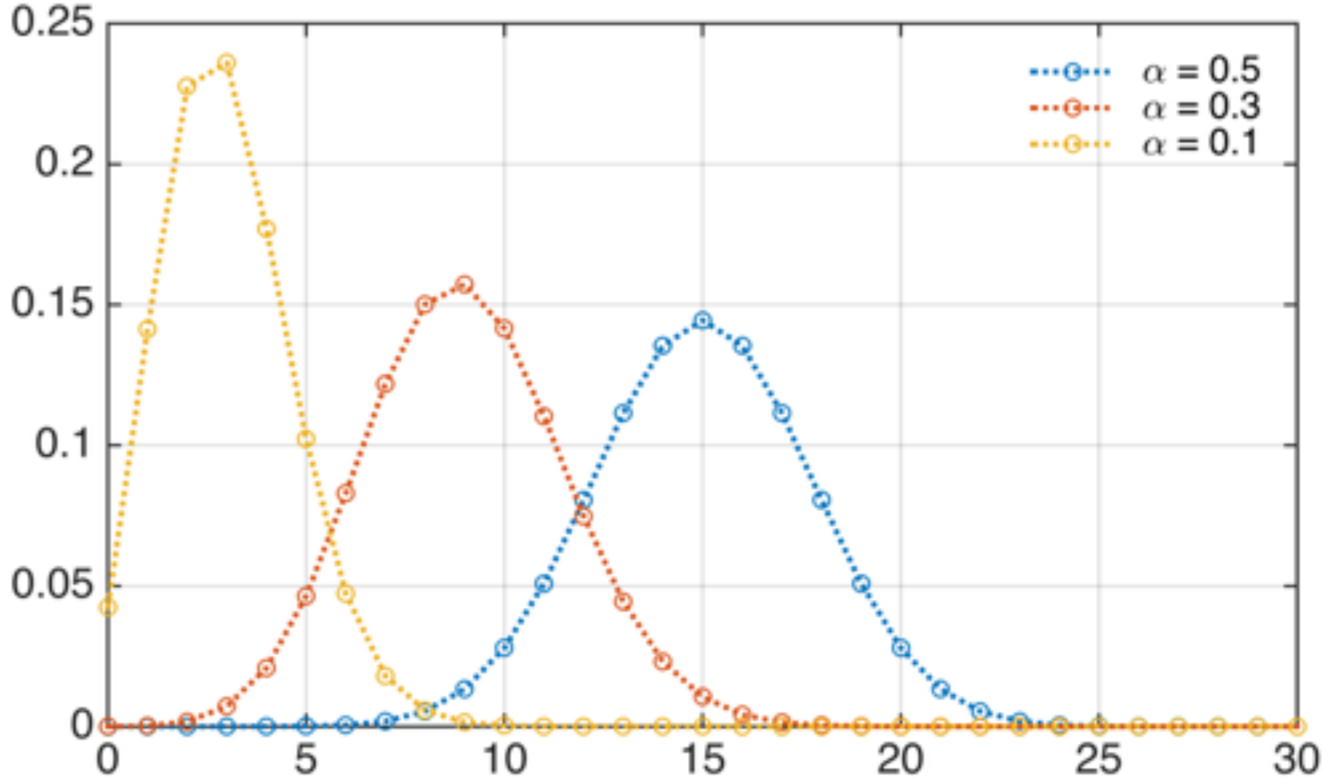
Alternatively, $\Omega = \{0,1\}$

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

PMFs You've Heard About

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

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

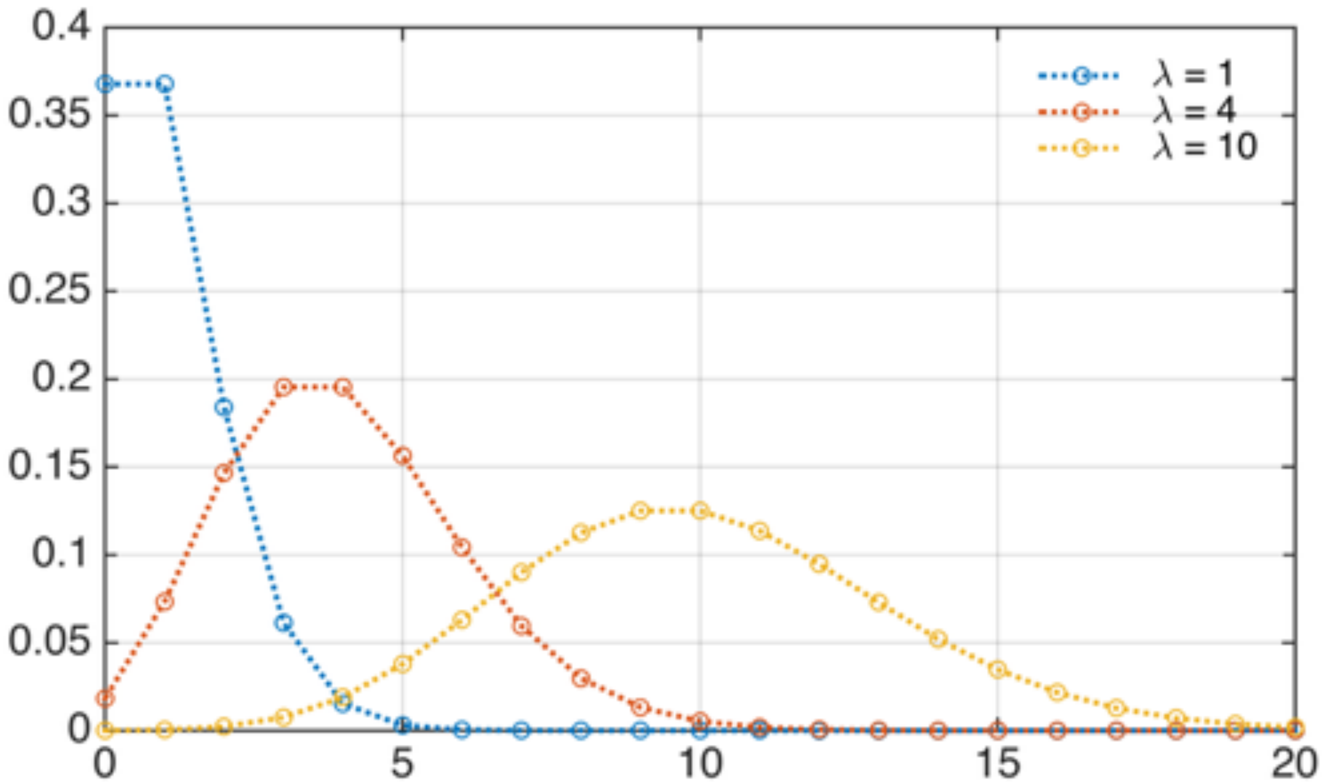


$n = 30$

PMFs You've Heard About

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

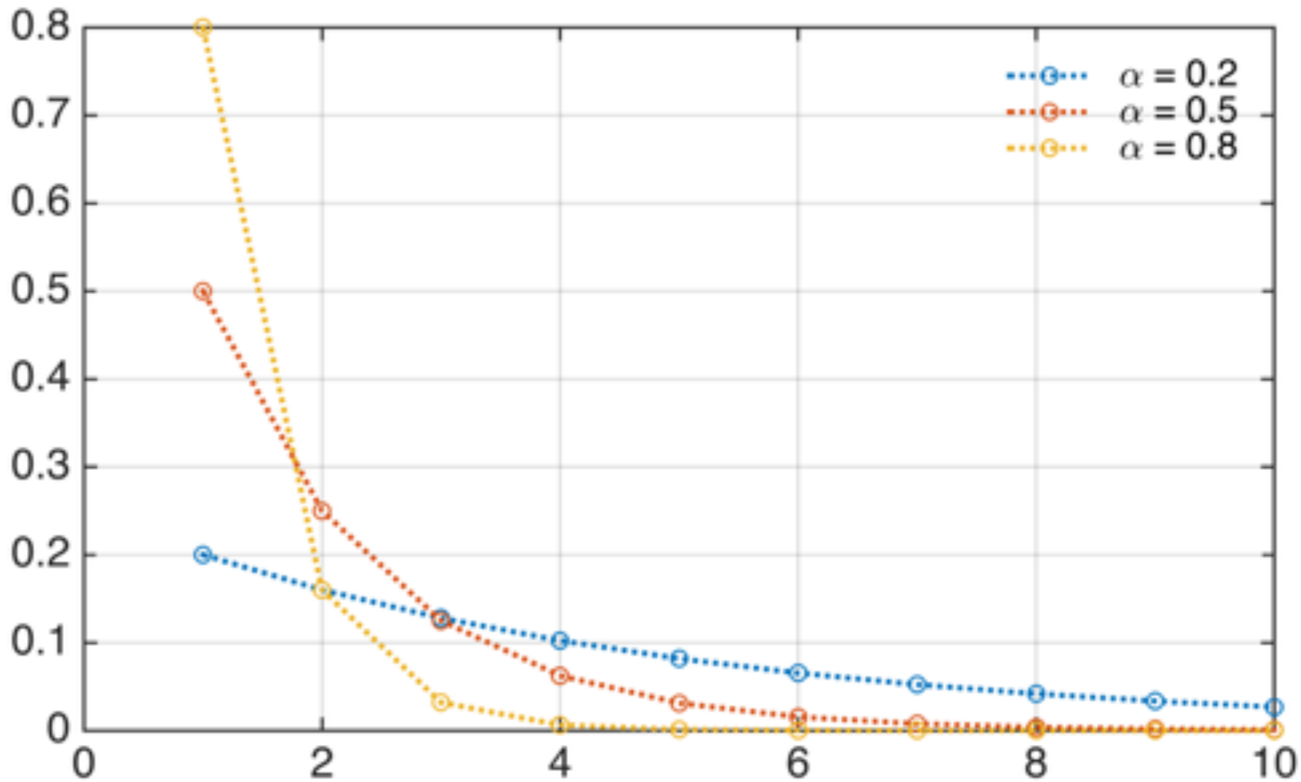
$$p(k) = \frac{\lambda^k e^{-\lambda}}{k!} \qquad \forall k \in \Omega$$



PMFs You've Heard About

Geometric distribution: $\Omega = \{1, 2, \ldots\} \quad \alpha \in (0, 1)$

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



Probability Density Functions

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

Probability density function:

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

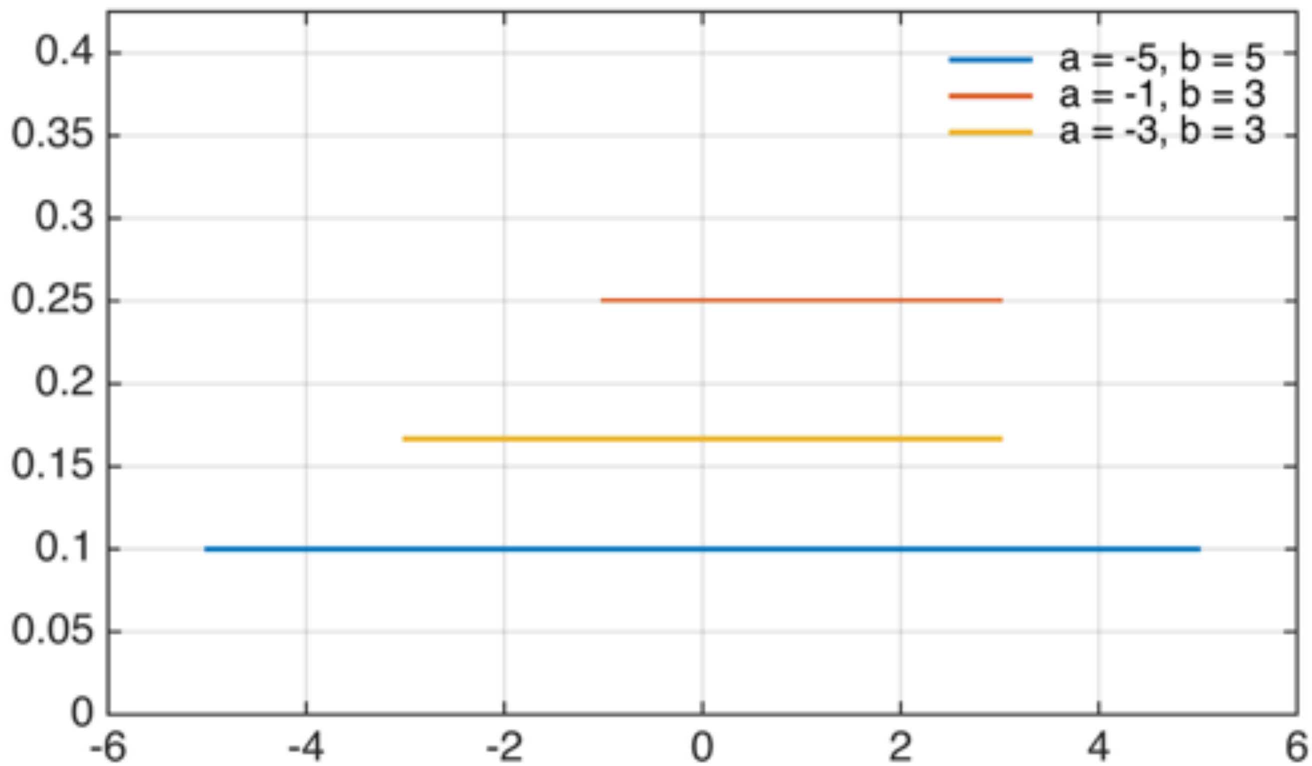
The probability of any event $A \in \mathcal{F}$ is defined as

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

PDFs You've Heard About

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

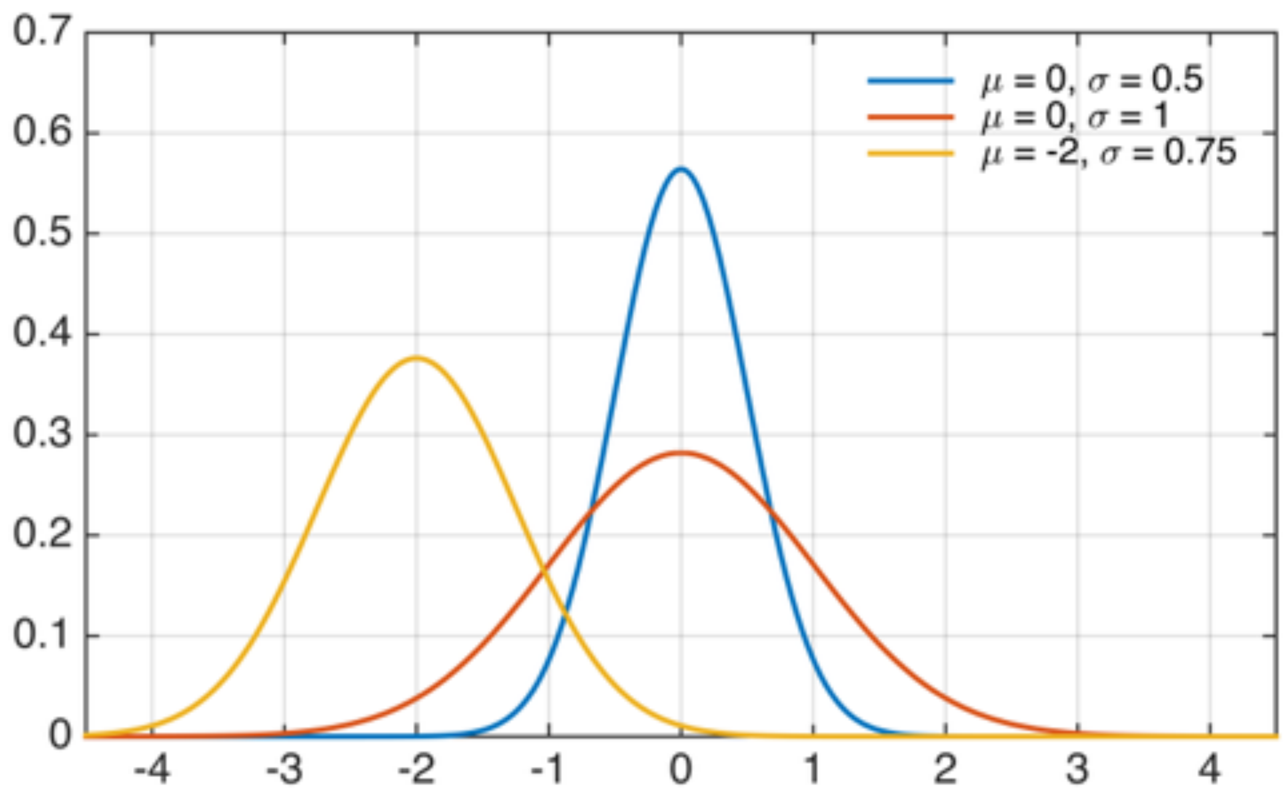
$$p(\omega) = \frac{1}{b-a} \qquad \forall \omega \in [a, b]$$



PDFs You've Heard About

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

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

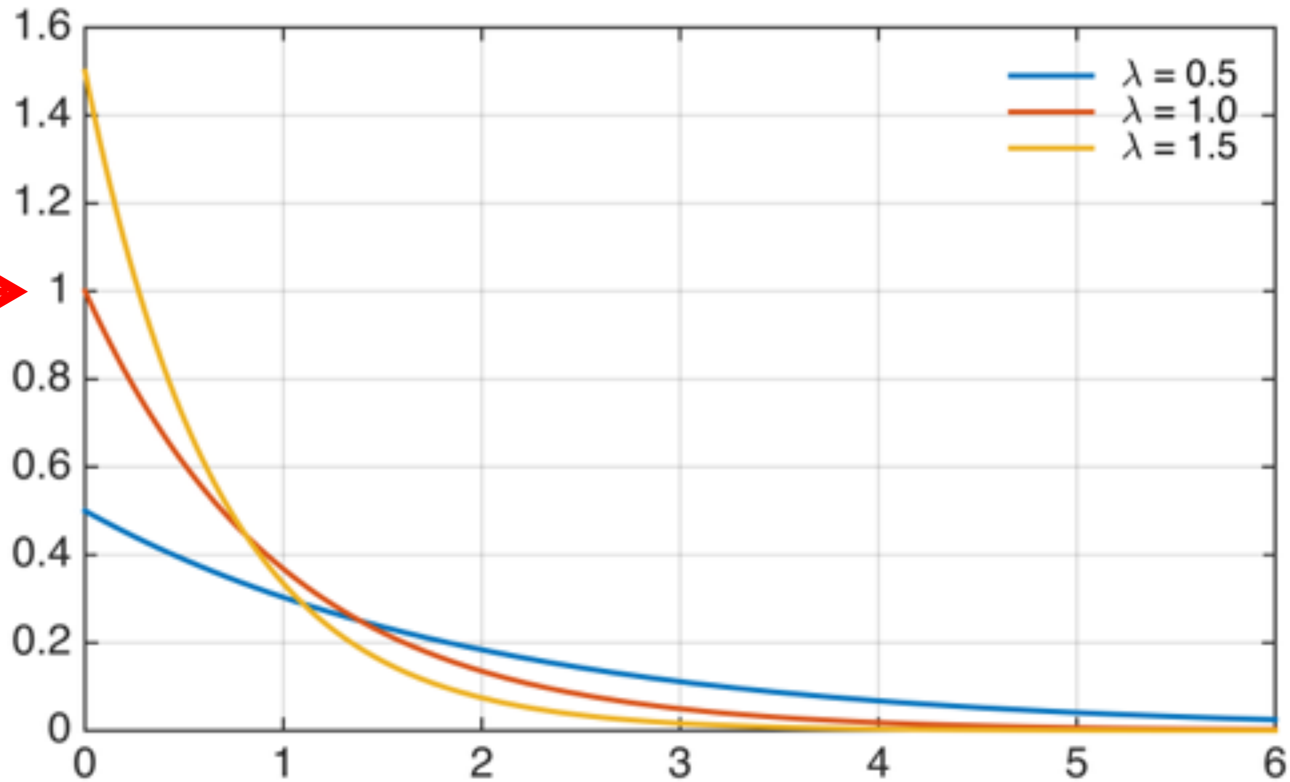


PDFs You've Heard About

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

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

1 is here!



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

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

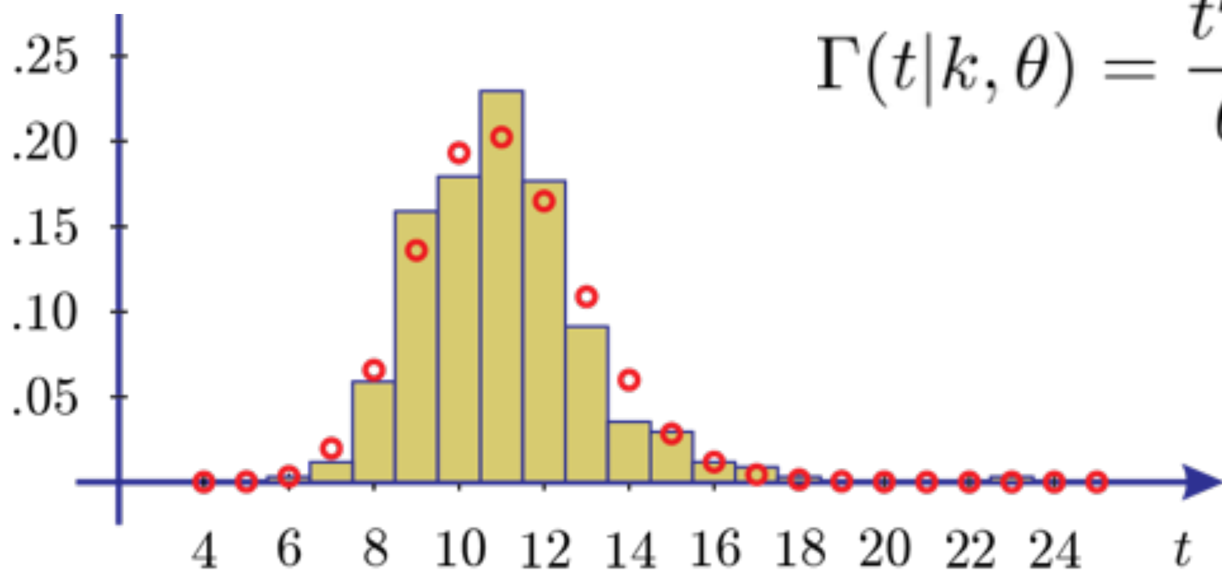
$$\begin{aligned} P(A) &= \int_{x}^{x+\Delta x} p(\omega) d\omega \\ &\approx p(x)\Delta x \end{aligned}$$

EXERCISE: UTILITY OF PDFS AS A MODEL

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

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

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



Multidimensional PMFs

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

Probability mass function:

1. $p : \Omega_1 \times \Omega_2 \times \ldots \times \Omega_k \rightarrow [0, 1]$
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 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

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

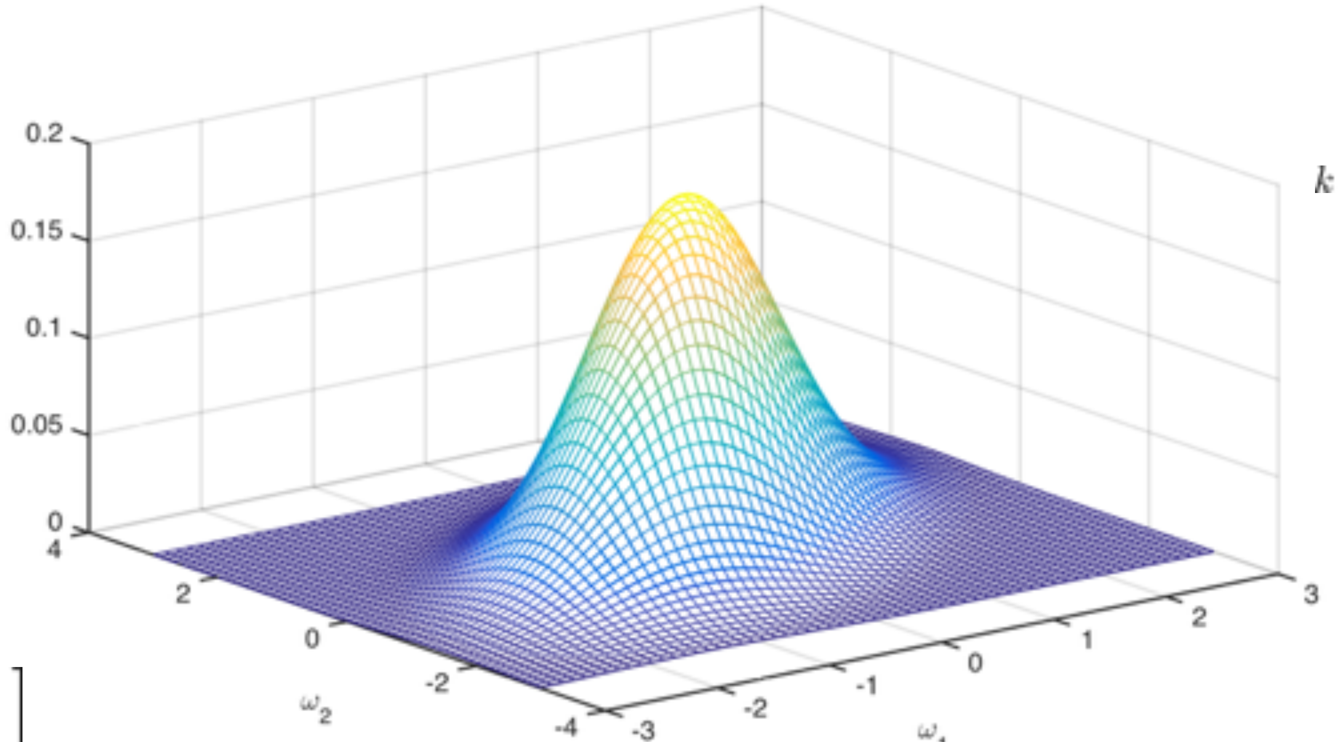
$\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_k)$

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 $\boldsymbol{\Sigma}$

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



$k = 2$

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

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

Elementary Conditional Probabilities

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

B = event that already occurred

The probability that any event $A \in \mathcal{F}$ has also occurred is

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

where $P(B) > 0$.

Bayes' rule:

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

Chain Rule

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

Chain rule

$$P(A_1 \cap A_2 \ldots \cap A_k) = P(A_1)P(A_2|A_1)\ldots P(A_k|A_1 \cap A_2 \ldots \cap A_{k-1})$$

where $\{A_i\}_{i=1}^{k}$ is a collection of k events

Sum Rule, Product Rule

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

Sum rule:

$$\boxed{P(A) = \sum_{i=1}^{k} P\left(A \cap B_i\right)}$$

where $\{B_i\}_{i=1}^{k}$ is a partition of Ω

Product rule:

$$\boxed{P(A \cap B) = P(A|B) \cdot P(B)}$$

where $P(B) > 0$

EXERCISE: MORE POWERFUL PMFS

- 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?