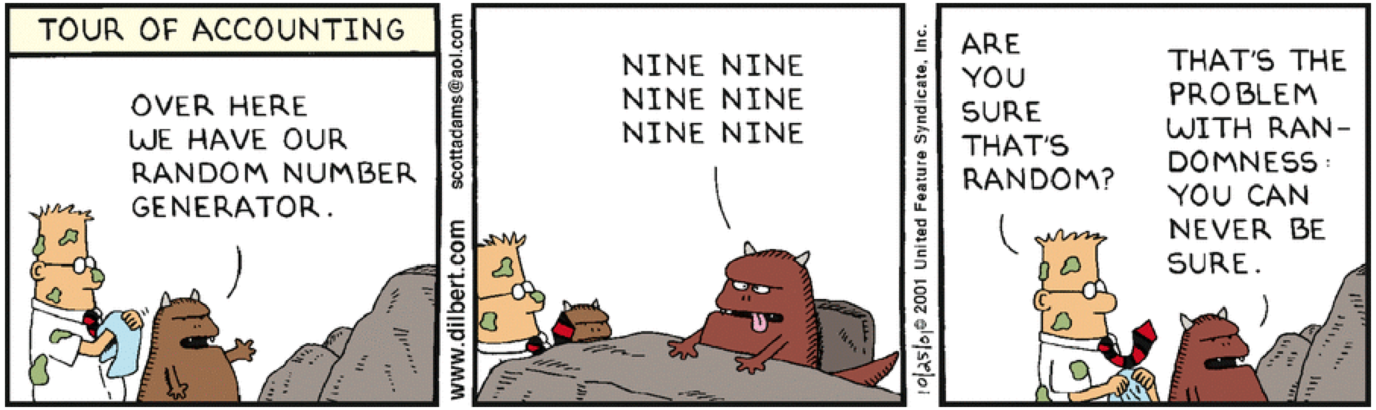
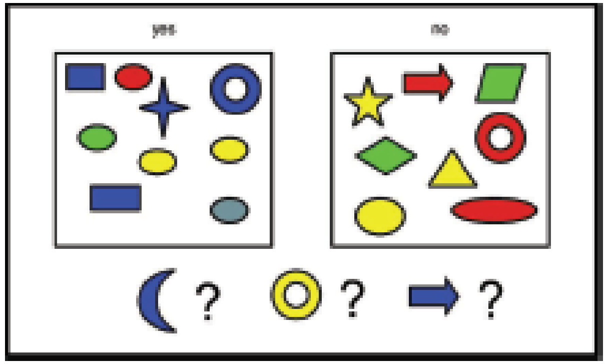
Random Variables



Need for Probabilistic machine Learning and Artificial Intelligence

- Probability theory provides a mathematical framework to handle uncertainty



(a)

D features (attributes)

N cases

Color	Shape	Size (cm)
Blue	Square	10
Red	Ellipse	2.4
Red	Ellipse	20.7

Label
1
1
0

(b)

Probability in machine learning

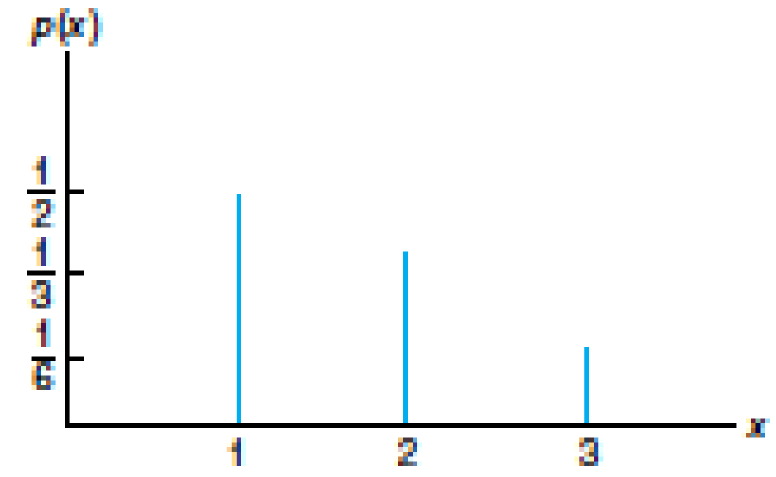
- Probability theory can be applied to any problem involving uncertainty.
- In machine learning, uncertainty comes in many forms:
  - what is the best prediction about the future given some past data?
  - what is the best model to explain some data
  - what measurement should I perform next?

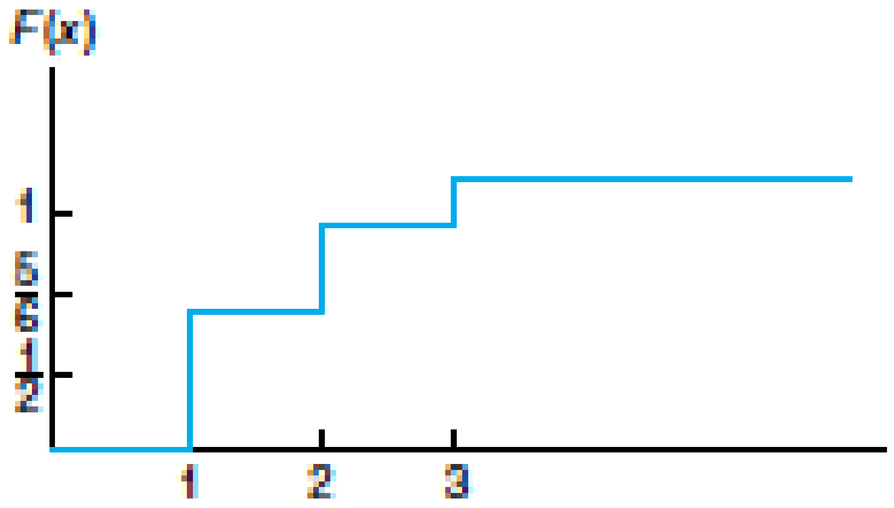
Discrete random Variables

- Discrete RV
  - Possible values form a countable set which is either a finite set or a countably infinite set.
    - e.g. {0,1}, number of heads {0,…,N},
    - number of goals in a football match {0,1,…}
  - probability mass function $P\{X = a\} = p(a)$
    - $p(x_i) >= 0, i = 1, 2, . . .$
    - $p(x) = 0$, all other values of x
    - sum $p(x_i) = 1$

Cumulative Distributionfunction

- Cumulative distribution function F(x) = P(X <= x)
- Define characteristics of a random variable
- Useful for sampling
- Discrete RV





Cumulative Distribution function

$F(x) = P(X <= x)$

In the coin-tossing experiment, the probability of heads equals p and the probability of tails equals q. We define the random variable x such that X(h) = 1 X(t) = 0. Find the cumulative distribution function F(x)

In the die experiment, we assign to the six outcomes the numbers X(i) = 10i.

Whats P(X < 35)

Plot F(x)

Continuous Random variables

- X takes values from a uncountable set
  - Time until next arrival [0, infty)
- Characterized Probability density function f(x)

$$P\{a \le X \le b\} = \int_a^b f(x)\,dx$$

- Probability that X = [a,b]

$$1 = P\{X \in (-\infty, \infty)\} = \int_{-\infty}^{\infty} f(x)\,dx$$

- Cumulative distribution function F(x) is continuous everywhere

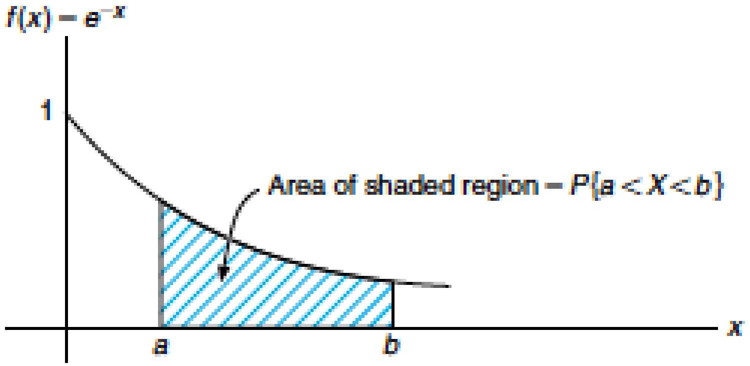
$$F(a) = P\{X \in (-\infty, a]\} = \int_{-\infty}^{a} f(x)\,dx$$

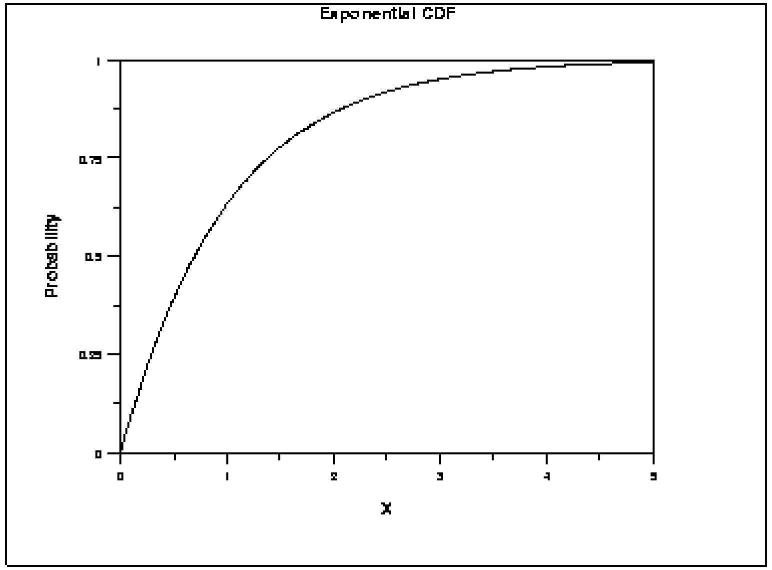
- Probability that X = a is 0!

$$\int_a^a f(x)\,dx = 0$$

- CDF vs PDF

$$\frac{d}{da}F(a) = f(a)$$





Example

- Suppose that X is a continuous random variable whose probability density function is given by

$$f(x) = \begin{cases} C(4x - 2x^2) & 0 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

  - (a) What is the value of C?
  - (b) Find P{X > 1}.

Discrete Distributions : Bernoulli

- Let $X \in \{0, 1\}$ be a binary random variable, with probability of "success" θ, X has a Bernoulli distribution, $X \sim \mathrm{Ber}(\theta)$
  - Coin toss, Rain or not

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

$$\mathrm{Ber}(x|\theta) = \theta^{\mathbb{I}(x=1)}(1-\theta)^{\mathbb{I}(x=0)}$$

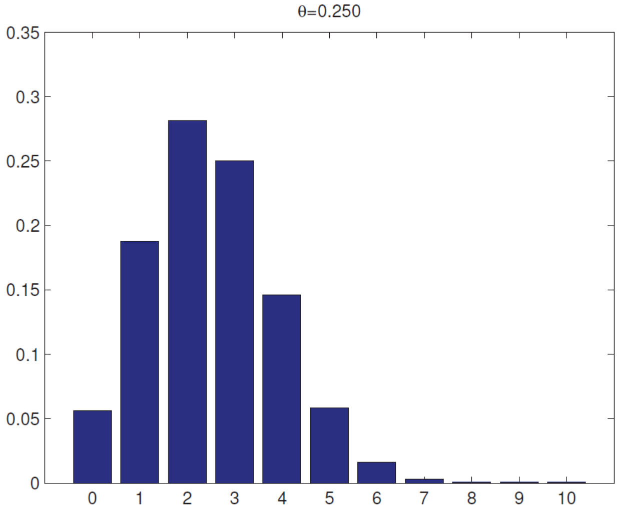
Discrete Distributions : Binomial

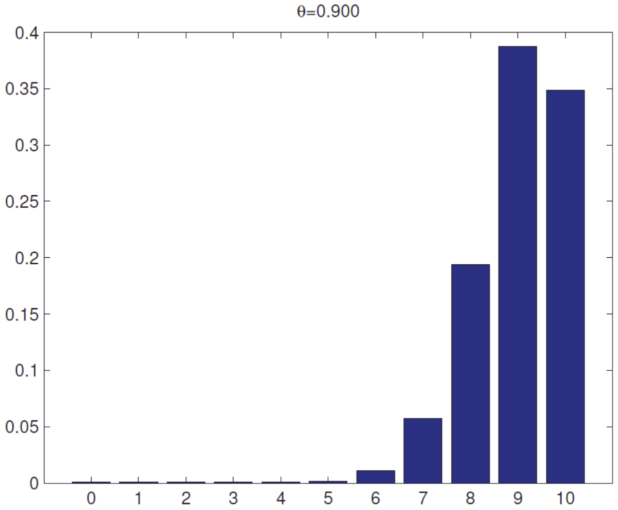
- Suppose we toss a coin n times. Let X ∈ {0, . . . , n} be the number of heads. If the probability of heads is θ , then we say X has a binomial distribution, written as X ∼ Bin(n, θ) .

$$\mathrm{Bin}(k|n,\theta) \triangleq \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

$$\binom{n}{k} \triangleq \frac{n!}{(n-k)!k!}$$

$$\text{mean} = n\,\theta, \quad \text{var} = n\theta(1-\theta)$$





Discrete Distributions

- Categorical/Multinoulli distribution
- Model the outcomes of tossing a K -sided die
  $x \sim \mathrm{Cat}(\theta)$, $p(x = j|\theta) = \theta_j$.

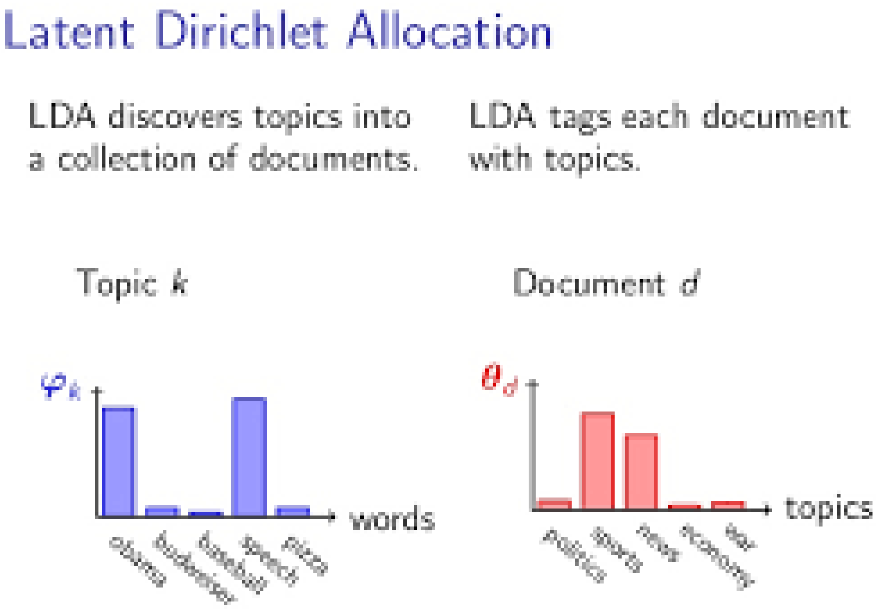
$$\mathrm{Mu}(\mathbf{x}|1, \boldsymbol{\theta}) = \prod_{j=1}^{K} \theta_j^{\mathbb{I}(x_j=1)}$$

Discrete Distributions

- Multinomial distribution : Models the outcome of n dice rolls
- let x = (x1, . . . , xk) be a random vector, where xj number of times side j of the die occurs.

$$\mathrm{Mu}(\mathbf{x}|n,\boldsymbol{\theta}) \triangleq \binom{n}{x_1 \ldots x_K} \prod_{j=1}^{K} \theta_j^{x_j}$$

- Probabilistic topic model
- Text classification



Poisson distribution

"My husband always loves your Poisson distribution – it's something to do with him being a mathematician."

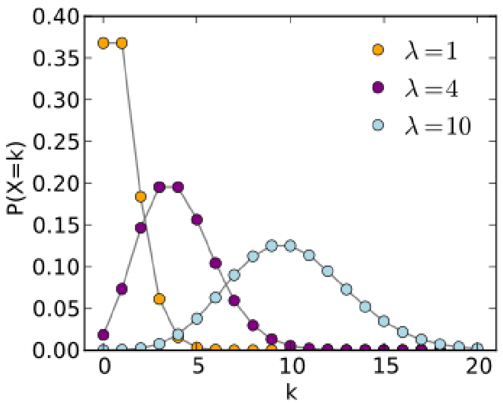
- Model number of events occurring in a fixed interval of time/space

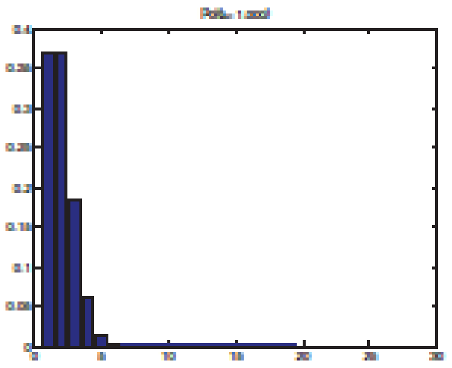
$$P(k \text{ events in interval}) = e^{-\lambda}\frac{\lambda^k}{k!}$$

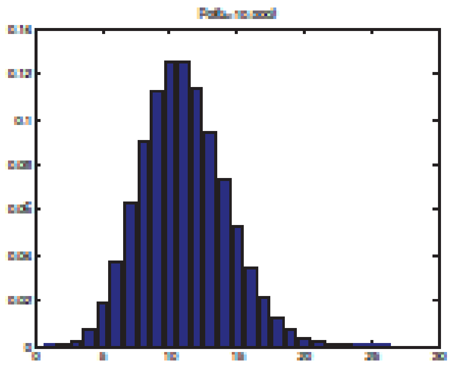
- λ is the average (mean) number of events per interval, k = 0, 1, 2, …, events occur independently, rate is a constant.
- Models rare events
  - Number of misprints on a page of a book.
  - average number of goals in a World Cup soccer match is approximately 2.5 ; λ = 2.5.

$$P(k \text{ goals in a match}) = \frac{2.5^k e^{-2.5}}{k!}$$

  - Number of wrong telephone numbers that are dialed in a day.



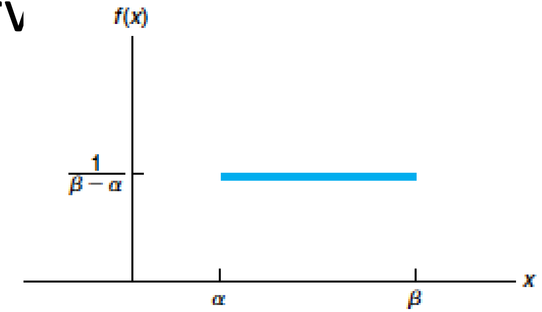


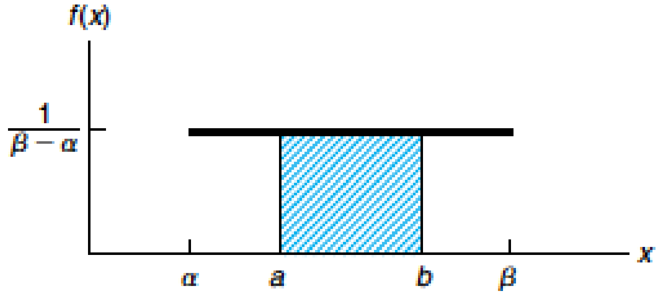


Uniform Random Variables

- Uniform random variable : X is said to be uniformly distributed over the interval $[\alpha, \beta]$

$$f(x) = \begin{cases} \dfrac{1}{\beta - \alpha} & \text{if } \alpha \le x \le \beta \\ 0 & \text{otherwise} \end{cases}$$





- Probability that X lies in [a,b]

$$P\{a < X < b\} = \frac{1}{\beta - \alpha} \int_a^b dx = \frac{b - a}{\beta - \alpha}$$

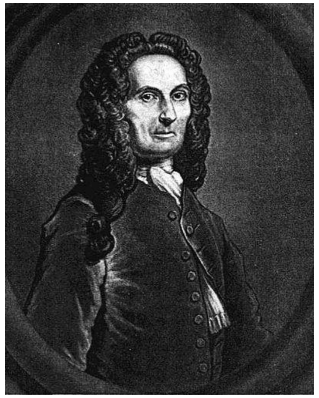
Uniform Random Variables

- Uniform random variable : X is said to be uniformly distributed over the interval [α, β]

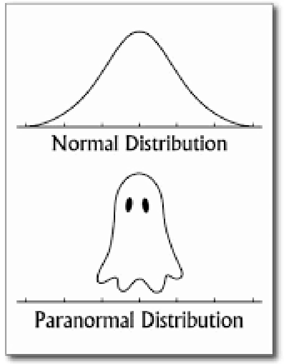
$$f(x) = \begin{cases} \dfrac{1}{\beta - \alpha} & \text{if } \alpha \le x \le \beta \\ 0 & \text{otherwise} \end{cases}$$

Buses arrive at a specified stop at 15-minute intervals starting at 7 A.M. That is, they arrive at 7, 7:15, 7:30, 7:45, and so on. If a passenger arrives at the stop at a time that is uniformly distributed between 7 and 7:30, find the probability that he waits

- (a) less than 5 minutes for a bus;
- (b) at least 12 minutes for a bus.

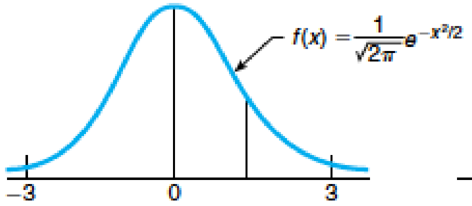


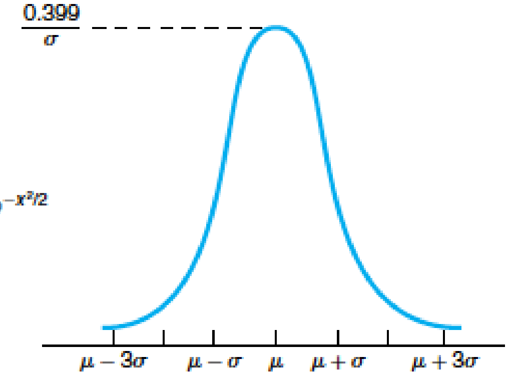
Normal Random Variables



- 1809 Gauss published his monograph "Theoria motus corporum coelestium in sectionibus conicis solem ambientium"
- All distributions of frequency other than normal are ‘abnormal’- Pearson
- A random variable is said to be normally distributed with parameters μ and σ2, X ∼ N(μ, σ2)

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}, \qquad -\infty < x < \infty^*$$

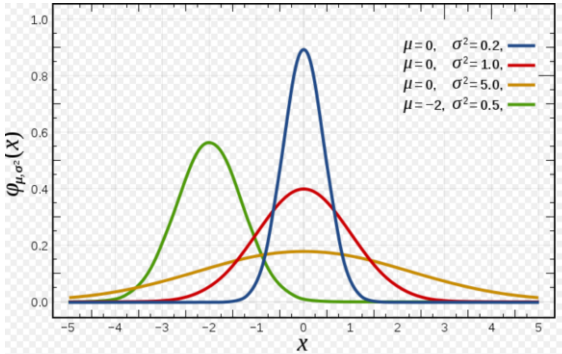


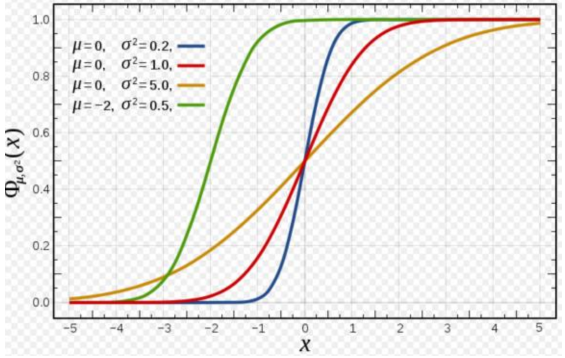


- μ = E [X] is the mean (and mode), and σ2 = var [X] is the variance.

$$\Phi(x; \mu, \sigma^2) \triangleq \int_{-\infty}^{x} \mathcal{N}(z|\mu, \sigma^2) dz$$

-
- CDF of the Gaussian



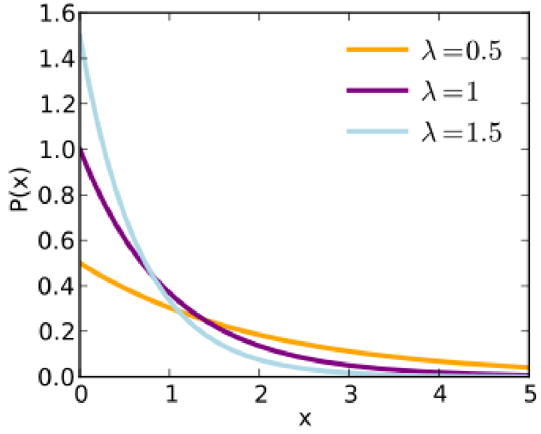


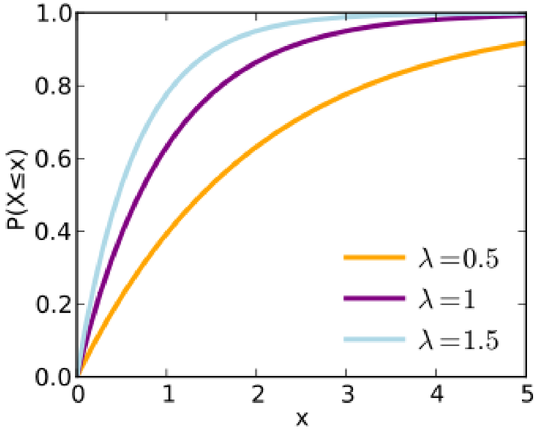
Exponential Random variables

- Distribution of the amount of time until some specific event occurs.
  - the amount of time until an earthquake occurs, a new war breaks out
- X is exponentially distributed with rate parameter λ > 0

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \qquad F(x) = P\{X \leq x\} \qquad 1 - e^{-\lambda x},$$

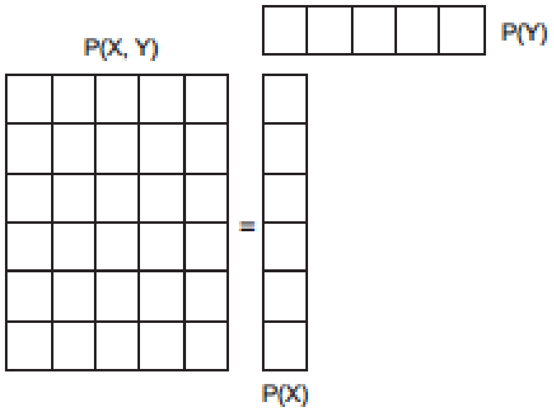
- Suppose that a number of miles that a car can run before its battery wears out is exponentially distributed with an average value of 10,000 miles. If a person desires to take a 5,000-mile trip, what is the probability that she will be able to complete her trip without having to replace her car battery?





Joint Probability Distributions

- p(x1 , . . . , xD) : models the (stochastic) relationships between the variables.
- Discrete variables : multi-dimensional array, number of parameters is O(K^D)
- Covariance between measures the degree to which X and Y are (linearly) related.



$$X \perp Y \iff p(X, Y) = p(X)p(Y)$$

Joint Probability Distributions

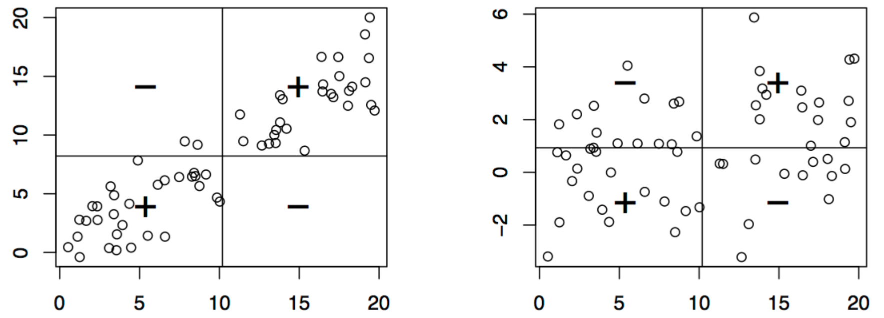
- Covariance between measures the degree to which X and Y are (linearly) related.

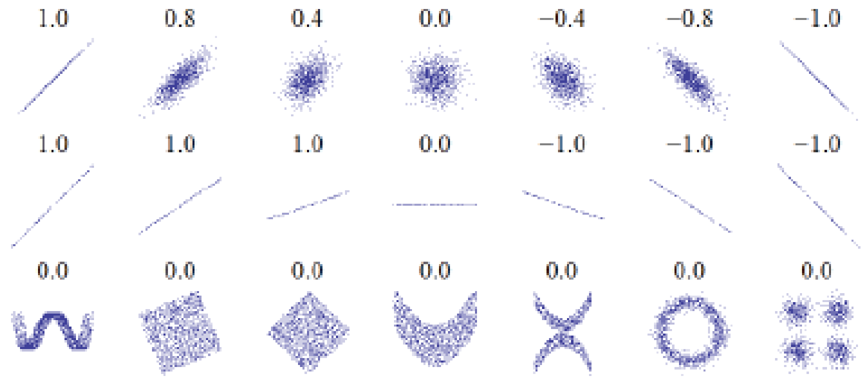
$$\mathrm{corr}[X,Y] \triangleq \frac{\mathrm{cov}[X,Y]}{\sqrt{\mathrm{var}[X]\,\mathrm{var}[Y]}}$$

$$\mathrm{cov}[X,Y] \triangleq \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

$$\mathrm{cov}[\mathbf{x}] \triangleq \mathbb{E}\left[(\mathbf{x}-\mathbb{E}[\mathbf{x}])(\mathbf{x}-\mathbb{E}[\mathbf{x}])^T\right]$$

$$= \begin{pmatrix} \mathrm{var}[X_1] & \mathrm{cov}[X_1,X_2] & \cdots & \mathrm{cov}[X_1,X_d] \\ \mathrm{cov}[X_2,X_1] & \mathrm{var}[X_2] & \cdots & \mathrm{cov}[X_2,X_d] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}[X_d,X_1] & \mathrm{cov}[X_d,X_2] & \cdots & \mathrm{var}[X_d] \end{pmatrix}$$



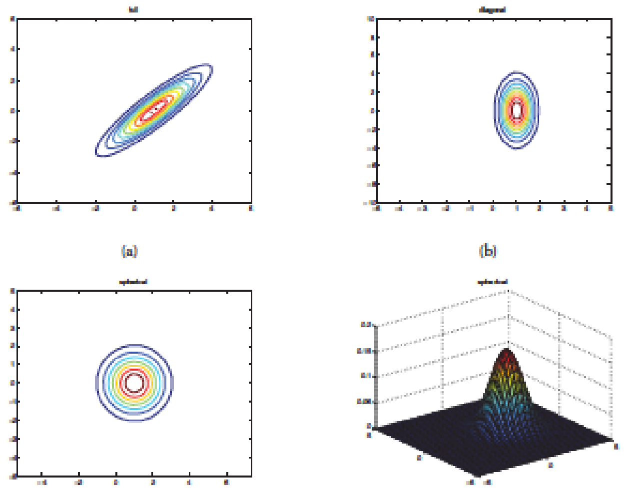


- independence imply uncorrelation but uncorrelation does not imply independence
  - X = Unif[−1, 1] and Y = X^2

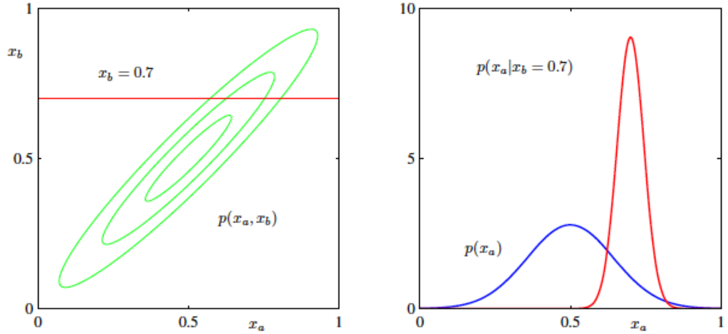
Multivariate Gaussian

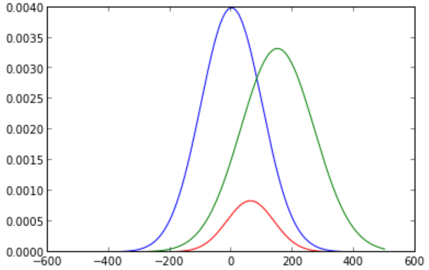
mean covariance

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

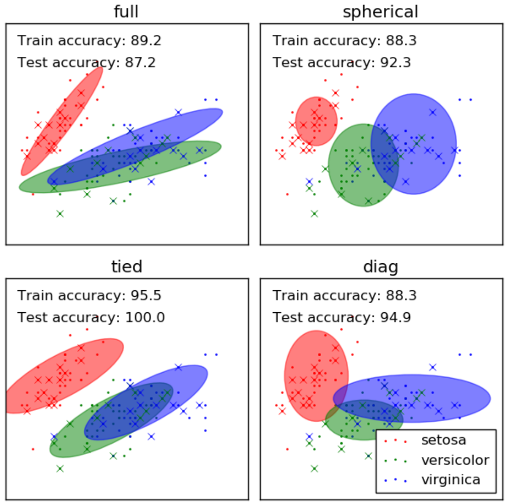


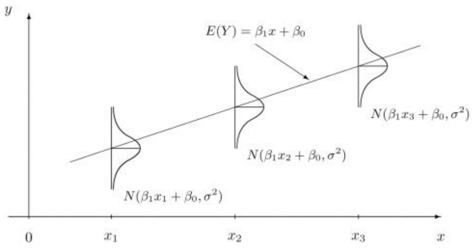
- Marginal and conditional distributions are Gaussian,
- product of Gaussians are Gaussian
- Gaussian mixture model

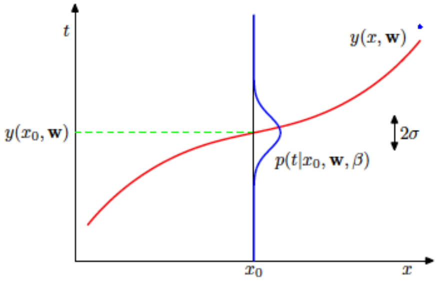


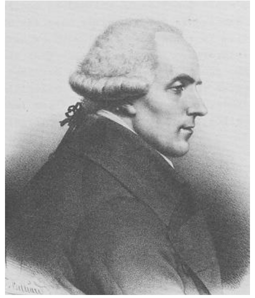


$$p(x) = \sum_{i=0}^{k} \pi_i N(x|\mu_k, \Sigma_k)$$





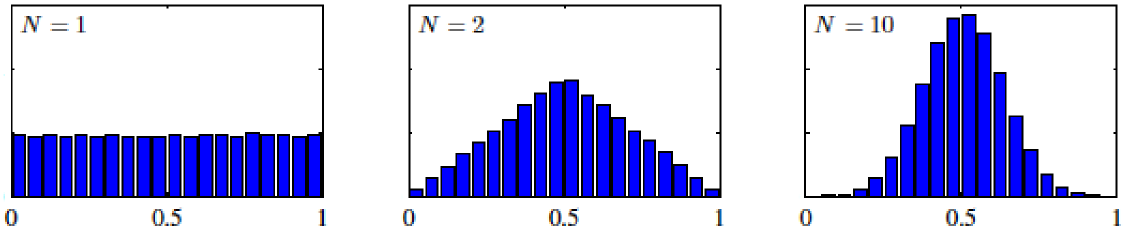




Central Limit Theorem

- Distribution of sum independent and identically distributed random variables
- Zn is standard normal

$$S_N = \sum_{i=1}^{N} X_i \qquad p(S_N = s) = \frac{1}{\sqrt{2\pi N\sigma^2}} \exp\left(-\frac{(s - N\mu)^2}{2N\sigma^2}\right)$$



$$Z_N \triangleq \frac{S_N - N\mu}{\sigma\sqrt{N}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{N}}$$

Limit Theorems

- Markov Inequality : X is a random variable that takes only nonnegative values, then for any value a > 0

$$P\{X \geq a\} \leq \frac{E[X]}{a}$$

- X is a random variable with mean μ and variance σ2, then, for any k > 0

$$P\{|X - \mu| \geq k\} \leq \frac{\sigma^2}{k^2}$$

- Useful when only mean, or both the mean and the variance, and not distribution of X

Limit Theorems Example

- Suppose we know that the number of items produced in a factory during a week is a random variable with mean 500.
- (a) What can be said about the probability that this week's production will be at least 1000?
- (b) If the variance of a week's production is known to equal 100, then what can be said about the probability that this week's production will be between 400 and 600?

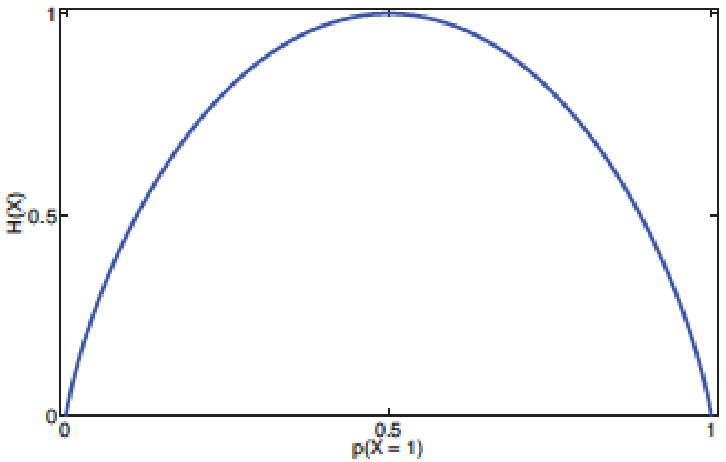
Entropy

$$\mathbb{H}(X) \triangleq -\sum_{k=1}^{K} p(X=k)\log_2 p(X=k)$$

- measure of its uncertainty
- [0. 25,0. 25,0. 2,0. 15,0. 15] vs [0. 2,0. 2,0. 2,0. 2,0. 2]
- maximum entropy is the uniform distribution
- compactly representing data(short codewords to highly probable bit strings)
- natural language, common words (“a”, “the”, “and”) are short
- Bernoulli r.v. for what value of θ, entropy is maximum ?
- Many models in ML such as MEMM, CRFs are based on maximum entropy principle - choose the simplest model

Entropy

- Entropy of a Bernoulli Random variable



Probability Distribution Summary

- X : Discrete
  - Binary valued scalar (0/1) : Bernoulli
  - Binary valued vector (one of K): Multinoulli/categorical
  - Multivalued scalar (M of N): Binomial
  - Multivalued vector (M1, M2, … MK) : Multinomial
  - Integer valued scalar (1 to infinity) : Poisson
- X : continous, real valued
  - Interval [a,b] : Uniform, Interval [0,1] : Beta
  - non-negative (0,infinity) : Exponential, Gamma
  - real line (-infinity, infinity) : Normal, students, Laplace
  - Vector : Real valued : Gaussian ; Simplex : Dirichlet