More Workload Characterization & Basic Probability and Statistics

Dr. John Mellor-Crummey

Department of Computer Science
Rice University

johnmc@cs.rice.edu

Goals for Today

- **Finish remaining approaches for characterizing workload**
  - —**Markov models**
  - —**clustering**
- **Review basic probability and statistics concepts needed that will be used throughout the rest of the course**

Markov Models

Sometimes, not only the relative frequency but order of service requests is important

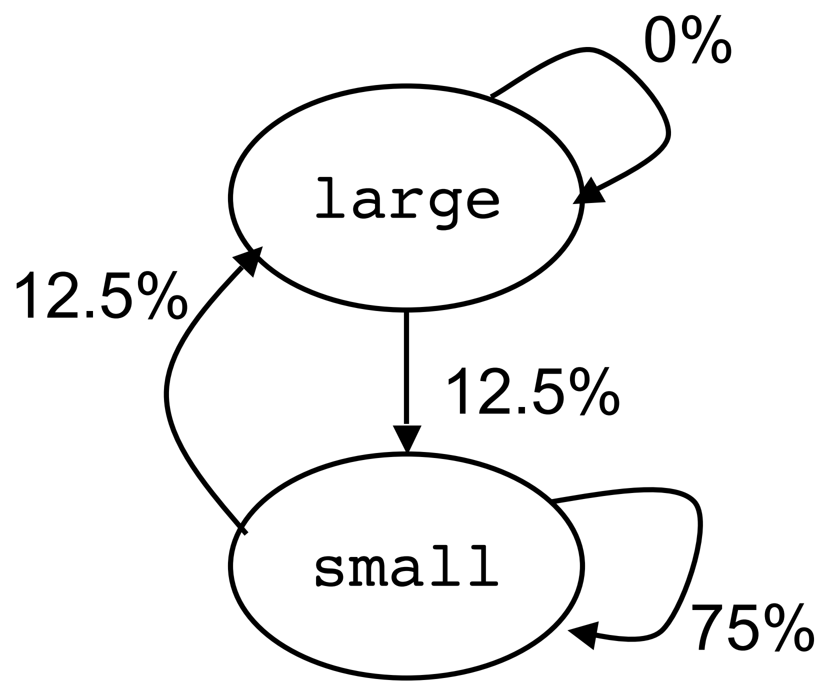
- **First Order Markov Models**
  - —probabilistic models that can generate sequences of states (here, representing service requests)
  - —next state depends only upon current state
  - —completely characterized by a transition probability matrix
  - —transition probability matrix properties
    - a_{ij} = P(system will enter state j | system is in state i)
    - $0 \leq a_{ij} \leq 1$
    - $\sum_{j=1}^{n} a_{ij} = 1$
- **Commonly used in queueing analysis**

First Order Markov Model Example I

- **Modeling packets on network**
  - —small packets = 87.5%; large packets = 12.5% (1 in 8 is large)
  - —large packet always followed by a small packet

Pairwise percentages

	next packet	
current packet	small	large
small	75%	12.5%
large	12.5%	0%



First Order Markov Model Example I

- **Modeling packets on network**
  - —small packets = 87.5%; large packets = 12.5% (1 in 8 is large)
  - —large packet always followed by a small packet

Pairwise percentages

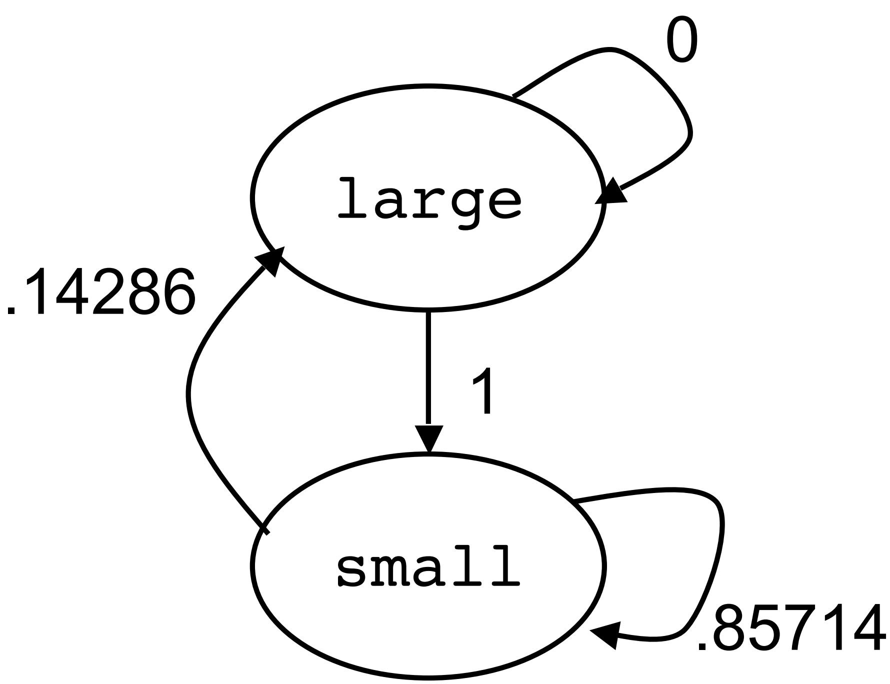
	next packet	
current packet	small	large
small	75%	12.5%
large	12.5%	0%

Transition Matrix

	next packet	
current packet	small	large
small	75/87.5 = .85714	12.5/87.5 = .14286
large	1	0

First Order Markov Model Example I

- **Modeling packets on network**
  - —small packets = 87.5%; large packets = 12.5% (1 in 8 is large)
  - —large packet always followed by a small packet



Transition Matrix

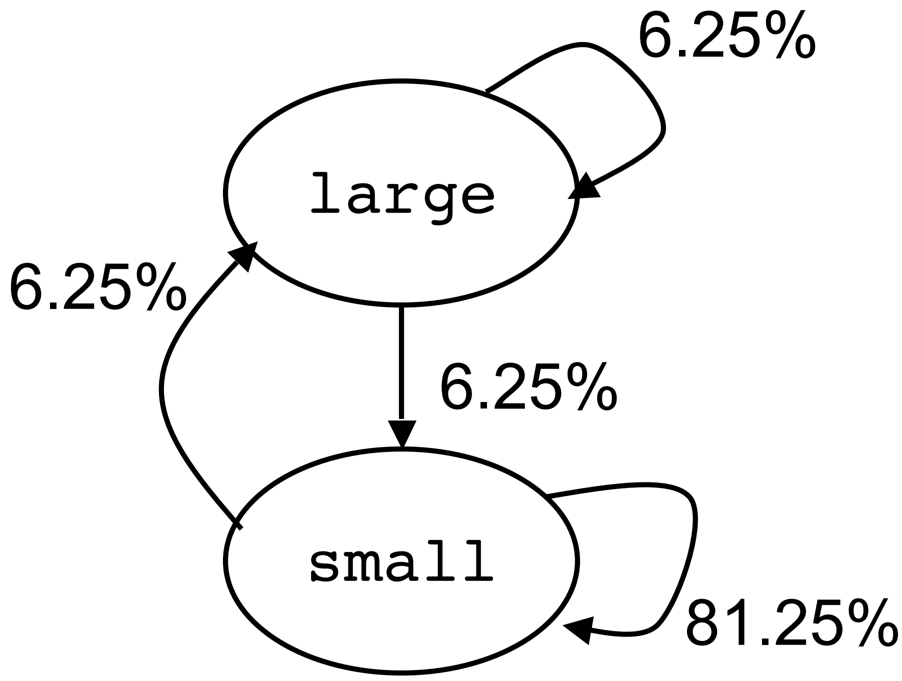
current packet	next packet: small	next packet: large
small	75/87.5 = .85714	12.5/87.5 = .14286
large	1	0

First Order Markov Model Example II

- **Modeling packets on network**
  - —small packets = 87.5%; large packets = 12.5% (1 in 8 is large)
  - —large packet followed by a small packet 50% of the time

Pairwise percentages

	next packet	
current packet	**small**	**large**
small	**81.25%**	**6.25%**
large	**6.25%**	**6.25%**



First Order Markov Model Example 2

- **Modeling packets on network**
  - —small packets = 87.5%; large packets = 12.5% (1 in 8 is large)
  - —large packet followed by a small packet half the time

Pairwise percentages

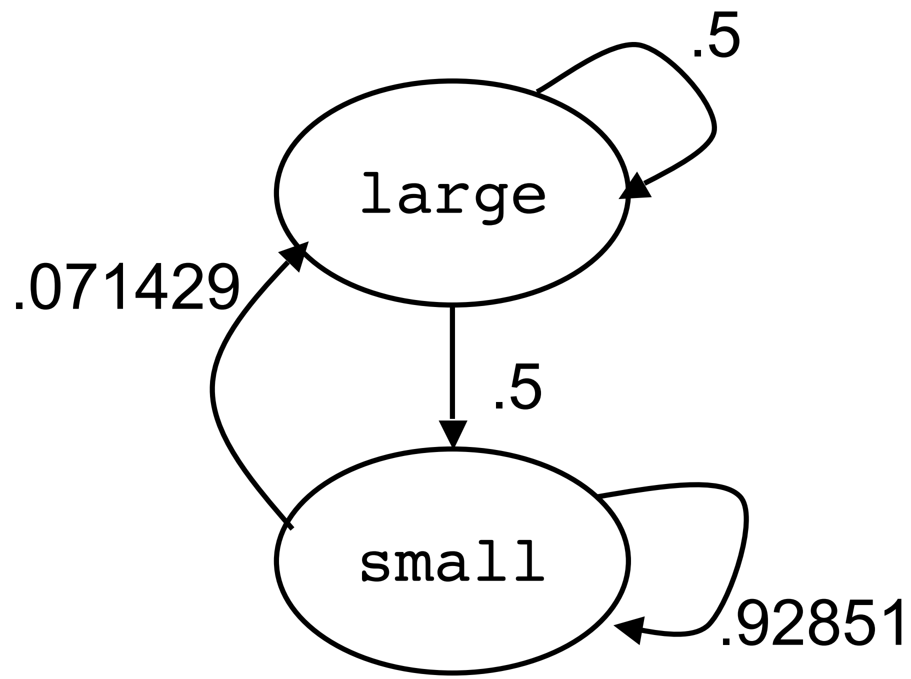
current packet	next packet: small	next packet: large
small	81.25%	6.25%
large	6.25%	6.25%

Transition Matrix

current packet	next packet: small	next packet: large
small	81.25/87.5 = .928571	6.25/87.5 = .071429
large	6.25/12.5 = .5	6.25/12.5 = .5

First Order Markov Model Example 2

- **Modeling packets on network**
  - **small packets = 87.5%; large packets = 12.5% (1 in 8 is large)**
  - **large packet followed by a small packet half the time**



Transition Matrix

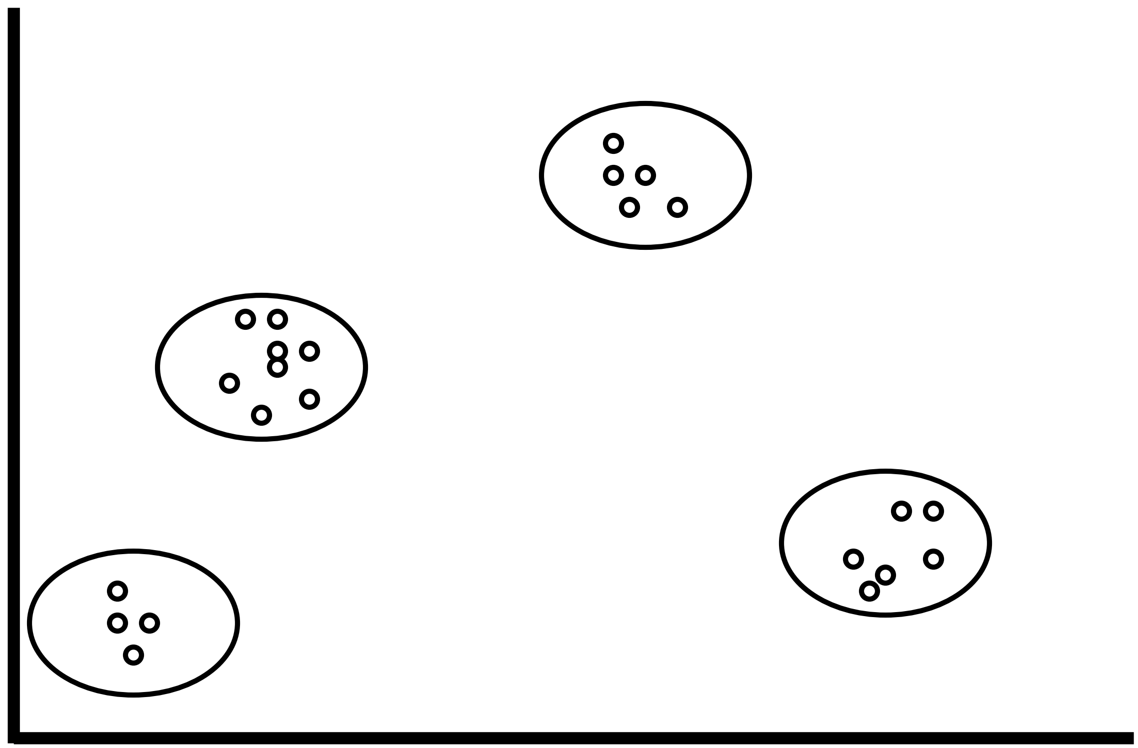
	next packet	
current packet	**small**	**large**
small	**81.25/87.5 = .928571**	**6.25/87.5 = .071429**
large	**6.25/12.5 = .5**	**6.25/12.5 = .5**

First Order Markov Model Properties

- **Finite**: The model consists of a finite number of states
- **Memory-less**: The next state depends only upon the present state, not past states
- **Absorbing state:** (enter, but never leave) any state with a 1 on the main diagonal in the transition probability matrix
- **Time independent**: Transition matrix probabilities do not vary over time

Clustering

- **Workload often consists of large number of components**
- **Want to classify components into small number of clusters whose members are similar**



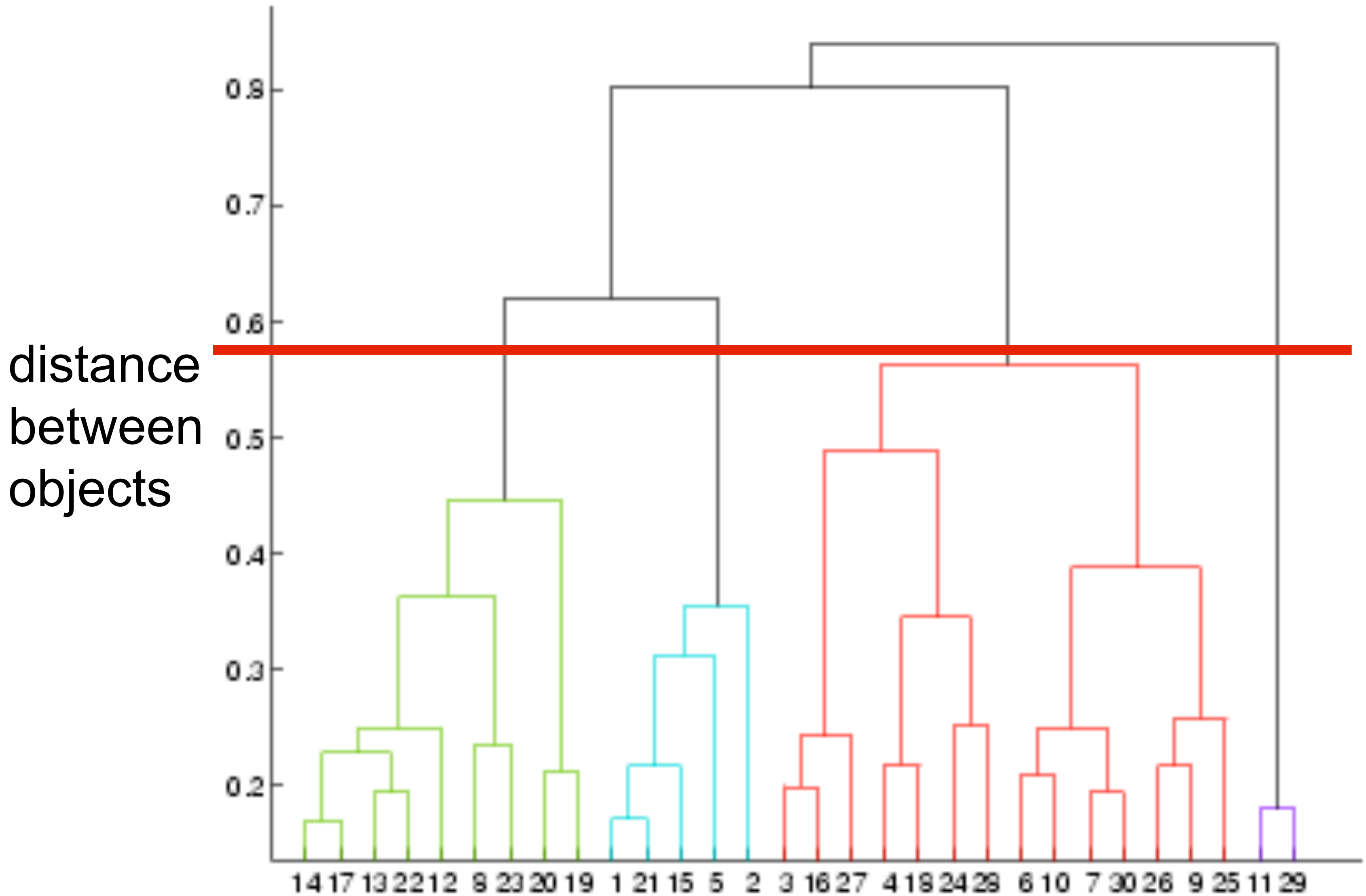
How to Cluster - I

- **Select workload components to cluster**
  - —e.g. random sample of all components of interest
- **Select workload parameters for clustering components**
  - —select those that (1) impact performance, and (2) vary significantly
- **Transform parameters to minimize skew, if necessary**
  - —e.g. log of parameter value if ratio rather than absolute value is important
- **Remove outliers**
  - —outliers affect normalization and thus can affect clustering
  - —remove if they do not consume significant fraction of system resources
- **Scale observations (e.g. normalize to 0 mean, unit variance)**
- **Select distance metric, e.g. Euclidean, Manhattan distance**

How to Cluster - II

- **Select clustering algorithm, e.g.**
  - —**nearest neighbor: min dist between objects in two clusters**
  - —**furthest neighbor: max dist between objects in two clusters**
  - —**average: avg dist betw all pairs (a,b) of objects $a \in$ cluster i, $b \in$ cluster j**
  - —**centroid: Euclidean distance between centroid of two clusters**
- **Perform clustering**
  - —**start with n clusters, using minimum spanning tree to merge**
  - —**represent with dendrogram**

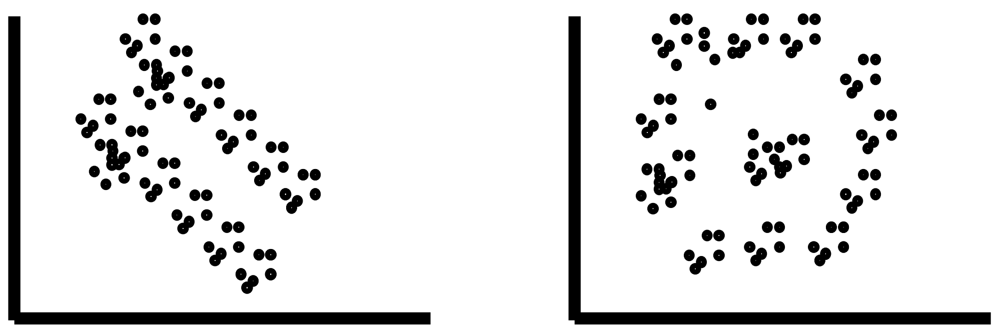
Dendrogram



From Matlab Statistics Toolbox Help: dendrogram

How to Cluster - III

- **Interpret results**
  - —discard small clusters, particularly if insignificant resources used
  - —interpret clusters in functional terms: what do they mean?
  - —select one or more representatives from each cluster for further study
- **Potential problems**
  - —minimizing intracluster variance may not always give the natural partitioning

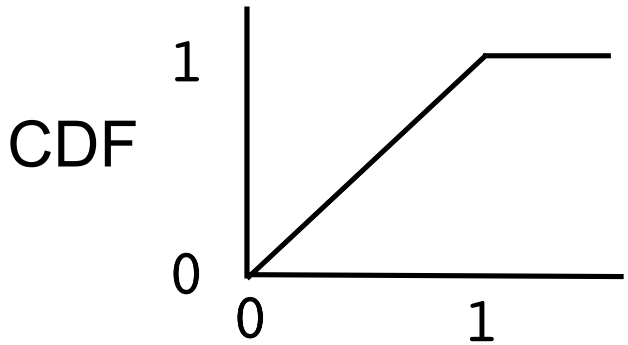


Basic Probability and Statistics Concepts

Basic Probability and Statistics - I

- **Independent events**
  - **—occurrence of one event does not in any way affect the probability of another**
- **Random variable**
  - **—takes one of a specified set of values with a specified probability**
- **Cumulative distribution function (CDF)**
  - **—CDF of a random variable maps a given value a to the probability of the variable taking a value ≤ a**

$$F_x(a) = P(x \le a)$$



CDF of a uniform random variable $0 < x \le 1$

Basic Probability and Statistics - II

- **For a continuous random variable x**
  - —**probability density function (pdf) = the derivative of the CDF**

$$f(x) = \frac{dF(x)}{dx}$$

$$P(x_1 < x \le x_2) = F(x_2) - F(x_1) = \int_{x_1}^{x_2} f(x)dx$$

- **For a discrete random variable x $\in$ {x_1, x_2, …, x_n}**
  - —**discrete probabilities {p_1, p_2, …, p_n} such that P(x = x_i) = p_i**
  - —**probability mass Function (pmf) maps x_i to p_i**

$$f(x_i) = p_i$$

$$P(x_1 < x \le x_2) = F(x_2) - F(x_1) = \sum_{x_1 < x_i \le x_2} p_i$$

Mean or Expected Value

Mean or expected value $$\mu = E(x) = \sum_{i=1}^{n} p_i x_i = \int_{-\infty}^{+\infty} x f(x)\,dx$$

form for a
discrete random variable

form for a
continuous random variable

Variance σ^2 and Standard Deviation σ

- **Form for a discrete variable**

$$\sigma^2 = Var(x) = E((x-\mu)^2) = \sum_{i=1}^{n} p_i (x_i - \mu)^2$$

- **Form for a continuous variable**

$$\sigma^2 = Var(x) = E((x-\mu)^2) = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x) dx$$

- **Standard deviation** $\sigma = \sqrt{Var(x)}$

Assessing Variation

- **Coefficient of variation C.O.V.**

$$\text{C.O.V.} = \frac{\text{standard deviation}}{\text{mean}} = \frac{\sigma}{\mu}$$

- **Covariance of random variables x and y with means μ_x and μ_y**

$$Cov(x,y) = \sigma_{xy}^2 = E((x-\mu_x)(y-\mu_y)) = E(xy) - E(x)E(y)$$

- **For independent random variables x and y, Cov(x,y) = 0**
- **Correlation: normalized covariance**

$$Cor(x,y) = \rho_{xy} = \frac{Cov(x,y)}{\sigma_x \sigma_y} = \frac{\sigma_{xy}^2}{\sigma_x \sigma_y}, -1 \le Cor(x,y) \le 1$$

Mean and Variance of Sums

- **Mean of sums for random variables**
  - **—Given**
    - **x_1, x_2, …, x_k are k random variables**
    - **a_1, a_2, …, a_k are arbitrary weights**

$$E(a_1x_1 + a_2x_2 + \ldots + a_kx_k) = a_1E(x_1) + a_2E(x_2) + \ldots + a_kE(x_k)$$

- **Variance of sums for independent variables**

$$Var(a_1x_1 + a_2x_2 + \ldots + a_kx_k) =$$

$$a_1^2Var(x_1) + a_2^2Var(x_2) + \ldots + a_k^2Var(x_k)$$

Quantile, Percentile, Median & Mode

- α -quantile: the x value at which the CDF takes value α
  - —denoted as x_α

$$P(x \le x_\alpha) = F(x_\alpha) = \alpha$$

- 100α -percentile: the x value at which the CDF reaches percentile 100α
- Median = 50-percentile = .5-quantile
- Mode = most likely value
  - —for a discrete variable, the x_i that has the highest probability
  - —for a continuous variable, the x where pdf is maximum

Normal Distribution

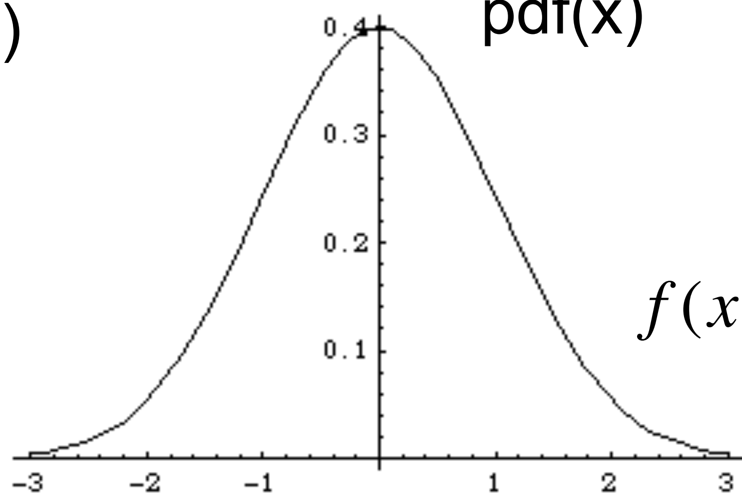
N(μ,σ) most commonly used distribution in data analysis

pdf = $f(x) = \dfrac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, -\infty \le x \le \infty$

μ = mean
σ = std dev

(also known as a Gaussian distribution)

N(μ=0,σ=1)
unit normal
distribution



$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$