CS 700: Quantitative Methods & Experimental Design in Computer Science

Sanjeev Setia

Dept of Computer Science

George Mason University

About this Class

- Required class for CS Ph.D. students
- Prerequisites:
  - Undergraduate probability & statistics
  - Doctoral status

What you will learn

- Applications of probability and statistical techniques for computer science
  - comparing systems using sample data
  - fitting distributions to sample data
  - confidence interval calculations
  - regression models
  - design of experiments
  - simulation and analysis of simulation results
  - introduction to analytic performance modeling and queuing analysis
  - workload characterization, pitfalls in performance analysis and reporting
  - Back-of-the envelope calculations
- Goal: motivate these techniques with examples from the research literature

Logistics

- Grade: 35% project, 15% Homework assignments 25% midterm, 25% take home final
- Slides, assignments, reading material on class web page http://www.cs.gmu.edu/~setia/cs700/
- Several small assignments related to material discussed in class
  - Not all will be graded, but we will go over solutions in class
- Term project
  - should involve experimentation (measurement, simulation)
  - select a topic in your research area if possible
  - apply techniques discussed in this class

Acknowledgement

These slides are based on presentations created and copyrighted by Prof. Daniel Menasce (GMU)

Review of Probability

Review of Probability Concepts

- ❑ Classical (theoretical) approach:

$$\frac{\text{No. Ways Event } A \text{ Can Occur}}{\text{Total Number of Events}}$$

process has to be known!

- ❑ Empirical approach (relative frequency):

$$\frac{\text{No. Times Result } A \text{ Occurred in the Experiment}}{\text{Total Number of Observations}}$$

- ❑ The relative frequency converges to the probability for a large number of experiments.

Review of Probability Rules

1. A probability is a number between 0 and 1 assigned to an event that is the outcome of an experiment:

$$P[A] \in [0,1]$$

2. Complement of event A.

$$P[A] = 1 - P[\overline{A}]$$

3. If events A and B are mutually exclusive then

$$P[A \text{ or } B] = P[A] + P[B]$$

$$P[A \text{ and } B] = 0$$

Review of Probability Rules (cont'd)

4. If events $A_1, \ldots, A_N$ are mutually exclusive and collectively exhaustive then:

$$\sum_{i=1}^{N} P[A_i] = 1$$

5. If events A and B are not mutually exclusive then: $P[A \text{ or } B] = P[A] + P[B] - P[A \text{ and } B]$

6. Conditional Probability:

$$P[A \mid B] = \frac{P[A \text{ and } B]}{P[B]} = \frac{P[B \mid A]P[A]}{P[B]}$$

Review of Probability Rules (cont'd)

7. If events A and B are independent (i.e., P[A] = P[A|B] and P[B]=P[B|A]) then:

$$P[A \text{ and } B] = P[A] \times P[B]$$

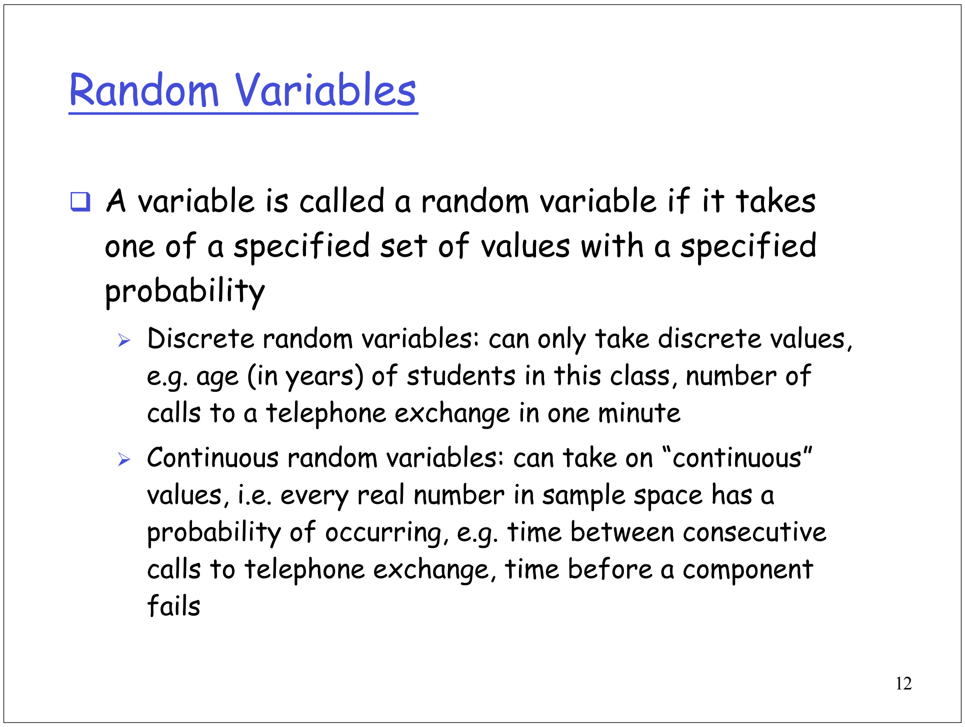
8. Regardless of whether events A and B are independent or not

$$P[A \text{ and } B] = P[A \mid B]P[B] = P[B \mid A]P[A]$$

9. Theorem of Total Probability: if events $A_1, \ldots, A_N$ are mutually exclusive and collectively exhaustive then

$$P[B] = \sum_{i=1}^{N} P[B \mid A_i]P[A_i]$$

Discrete Random Variables

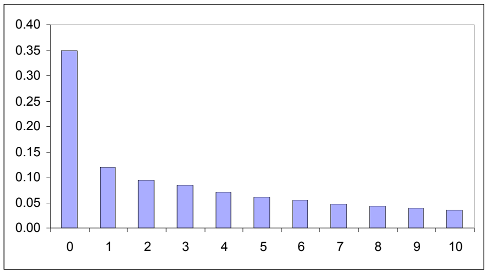


Random Variables

- A variable is called a random variable if it takes one of a specified set of values with a specified probability
  - Discrete random variables: can only take discrete values, e.g. age (in years) of students in this class, number of calls to a telephone exchange in one minute
  - Continuous random variables: can take on “continuous” values, i.e. every real number in sample space has a probability of occurring, e.g. time between consecutive calls to telephone exchange, time before a component fails

Discrete Probability Distribution

- ❑ Distribution: set of all possible values and their probabilities.



Number of I/Os per Transaction	Probability
0	0.350
1	0.120
2	0.095
3	0.085
4	0.070
5	0.060
6	0.054
7	0.048
8	0.043
9	0.040
10	0.035
	1.000

Moments of a Discrete Random Variable

- ❑ Expected Value:

$$\mu = E[X] = \sum_{\forall i} X_i \times P[X_i]$$

- ❑ k-th moment:

$$\mu = E[X^k] = \sum_{\forall i} X_i^k \times P[X_i]$$

Number of I/Os per Transaction	Probability	For First Moment (average)	For Second Moment
0	0.350	0.000	0.000
1	0.120	0.120	0.120
2	0.095	0.190	0.380
3	0.085	0.255	0.765
4	0.070	0.280	1.120
5	0.060	0.300	1.500
6	0.054	0.324	1.944
7	0.048	0.336	2.352
8	0.043	0.344	2.752
9	0.040	0.360	3.240
10	0.035	0.350	3.500
	1.000	**2.859**	**17.673**

mean

second moment

Central Moments of a Discrete Random Variable

- k-th central moment:

$$E[(X-\overline{X})^k] = \sum_{\forall i} (X_i - \overline{X})^k \times P[X_i]$$

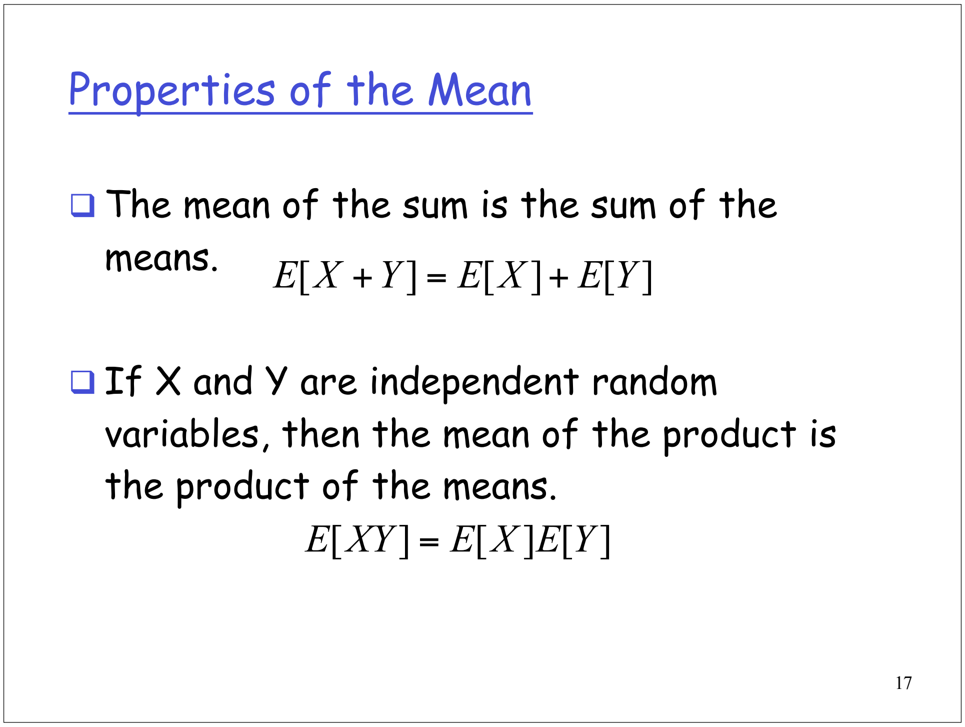
- The variance is the second central moment:

$$\begin{aligned}\sigma^2 = E[(X-\overline{X})^2] &= E[X^2 + (\overline{X})^2 - 2X\overline{X}]\\ &= E[X^2] + (\overline{X})^2 - 2(\overline{X})^2 = \\ &= E[X^2] - (\overline{X})^2\end{aligned}$$

Central Moments of a Discrete Random Variable

Number of I/Os per Transaction	Probability	For First Moment (average)	For Second Moment	For Second Central Moment
0	0.350	0.000	0.000	2.8609
1	0.120	0.120	0.120	0.4147
2	0.095	0.190	0.380	0.0701
3	0.085	0.255	0.765	0.0017
4	0.070	0.280	1.120	0.0911
5	0.060	0.300	1.500	0.2750
6	0.054	0.324	1.944	0.5328
7	0.048	0.336	2.352	0.8231
8	0.043	0.344	2.752	1.1365
9	0.040	0.360	3.240	1.5085
10	0.035	0.350	3.500	1.7848
	1.000	**2.859**	**17.673**	**9.4991**

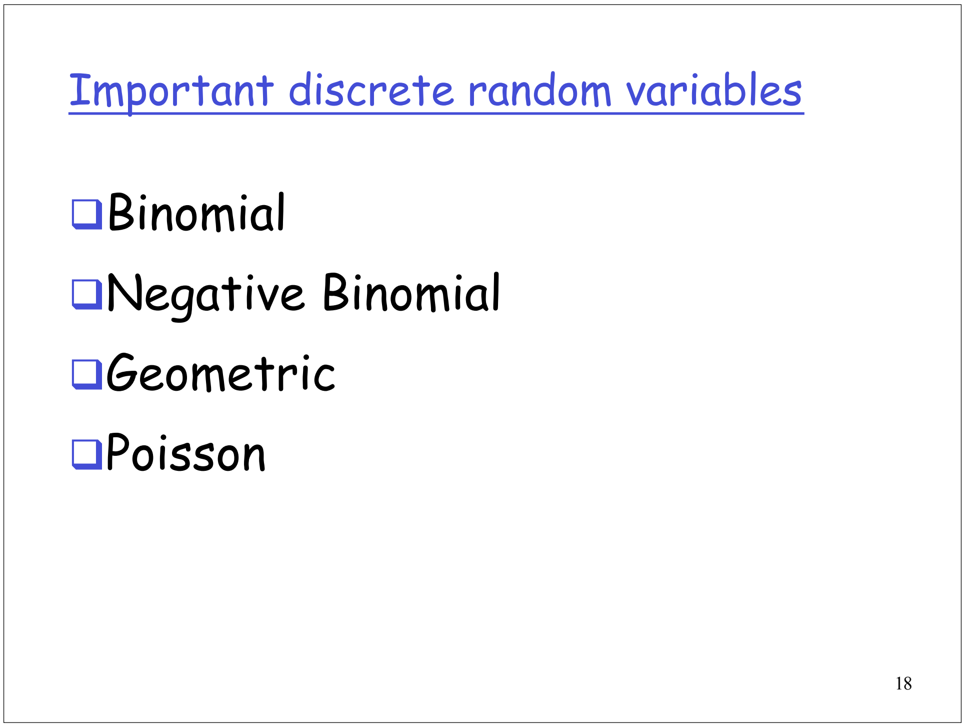
average *variance*



Properties of the Mean

- ❑ The mean of the sum is the sum of the means. $E[X+Y]=E[X]+E[Y]$
- ❑ If X and Y are independent random variables, then the mean of the product is the product of the means.

$$E[XY]=E[X]E[Y]$$



Important discrete random variables

- ❑Binomial
- ❑Negative Binomial
- ❑Geometric
- ❑Poisson

The Binomial Distribution

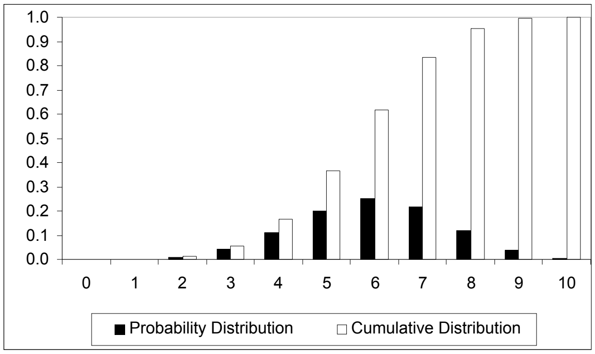
- Distribution: based on carrying out Bernoulli trials (independent experiments with two possible outcomes):
  - Success with probability p and
  - Failure with probability $(1-p)$.
- A binomial r.v. counts the number of successes in n trials.

$$P[X=k]=\binom{n}{k}p^k(1-p)^{n-k} = \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$$

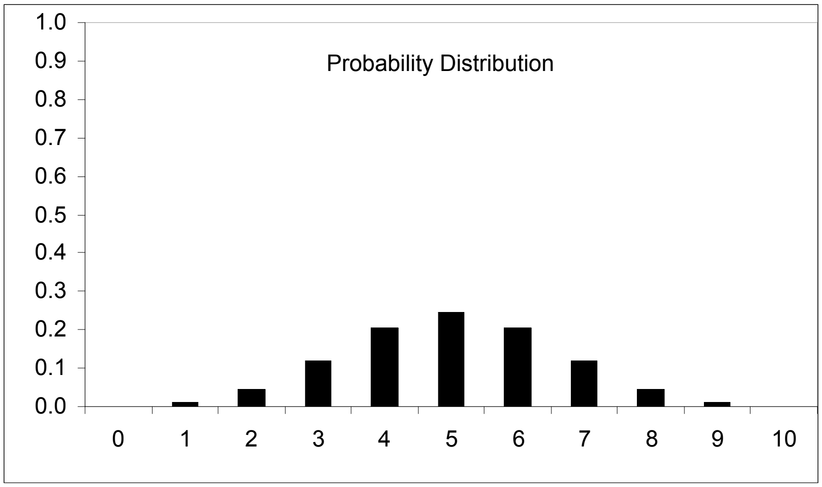
The Binomial Distribution

Success Probability	**0.6**	**(p)**
Number of Attempts	**10**	**(n)**

Number of Attempts (k)	**Probability k successful attempts in n**	**Cumulative**
0	0.000105	0.000105
1	0.001573	0.001678
2	0.010617	0.012295
3	0.042467	0.054762
4	0.111477	0.166239
5	0.200658	0.366897
6	0.250823	0.617719
7	0.214991	0.832710
8	0.120932	0.953643
9	0.040311	0.993953
10	0.006047	1.000000

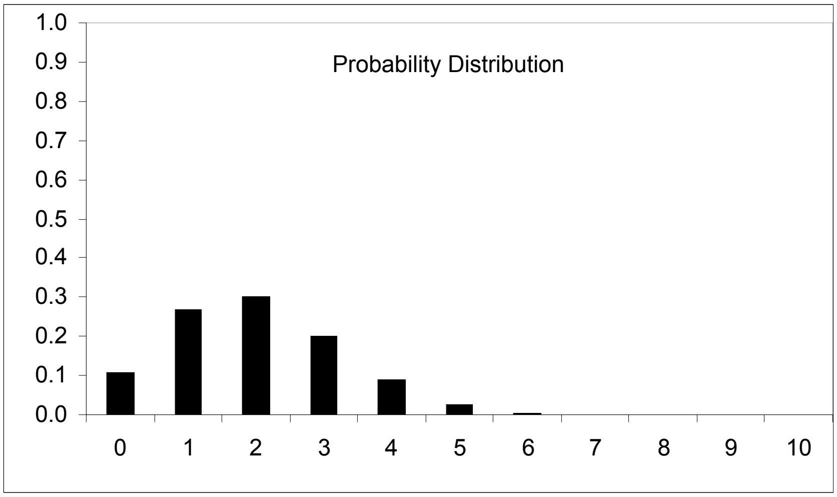


Shape of the Binomial Distribution



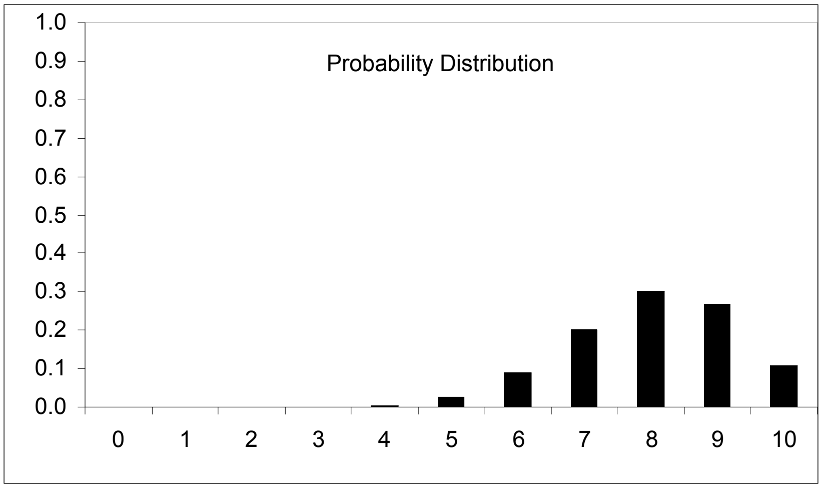
$p = 0.5$ symmetric for any n.

Shape of the Binomial Distribution



$p = 0.2$ right skewed

Shape of the Binomial Distribution



$p = 0.8$ left skewed

Moments of the Binomial Distribution

- ❑ Average: $n\,p$
- ❑ Variance: $np(1-p)$
- ❑ Standard Deviation: $\sqrt{np(1-p)}$
- ❑ Coefficient of Variation:

$$\frac{\sqrt{np(1-p)}}{np} = \sqrt{\frac{1-p}{np}}$$

Negative Binomial Distribution

- Probability of success is equal to *p* and is the same on all trials.
- Random variable X counts the number of trials until the *k*-th success is observed.

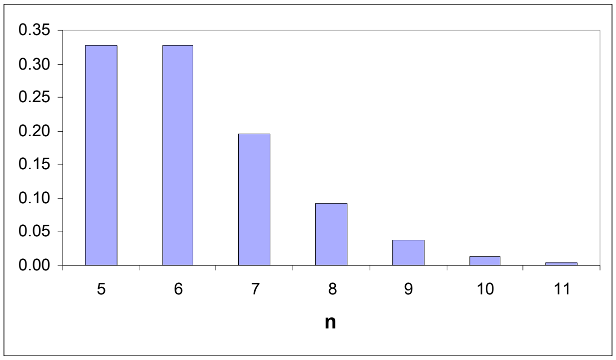
$$P[X=n]=\binom{n-1}{k-1}(1-p)^{n-k}p^{k}$$

S	F	F	S	. . .	F	S
1	2	3	4		n-1	n

Negative Binomial Distribution

Success probability	0.8

k	n	Prob[X=n]
1	1	0.800000
1	2	0.160000
1	3	0.032000
1	4	0.006400
5	5	0.327680
5	6	0.327680
5	7	0.196608
5	8	0.091750
5	9	0.036700
5	10	0.013212
5	11	0.004404



In Excel:
Pr [X=n] = NEGBINOMDIST (n-k,k,p)

Moments of the Negative Binomial Distribution

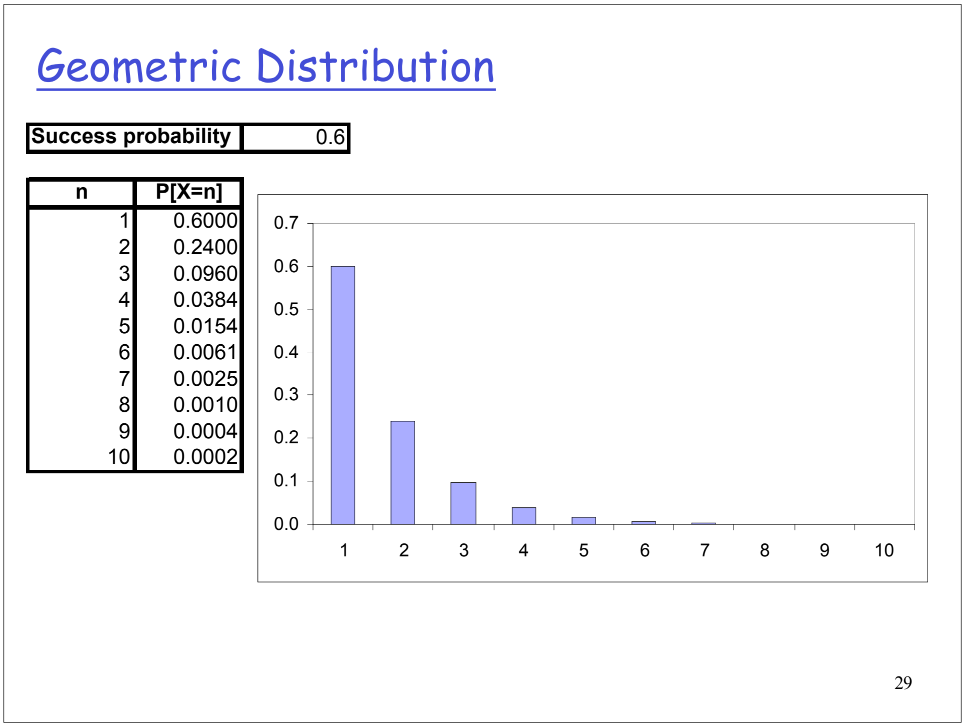
- Average: $\frac{k}{p}$
- Standard Deviation: $\sqrt{\frac{k(1-p)}{p^2}}$
- Coefficient of Variation: $\sqrt{\frac{1-p}{k}}$

Geometric Distribution

- Special case of the negative binomial with *k*=1.
- Probability that the first success occurs after *n* trials is

$$p[X = n] = p(1-p)^{n-1} \qquad n = 1,2,...$$

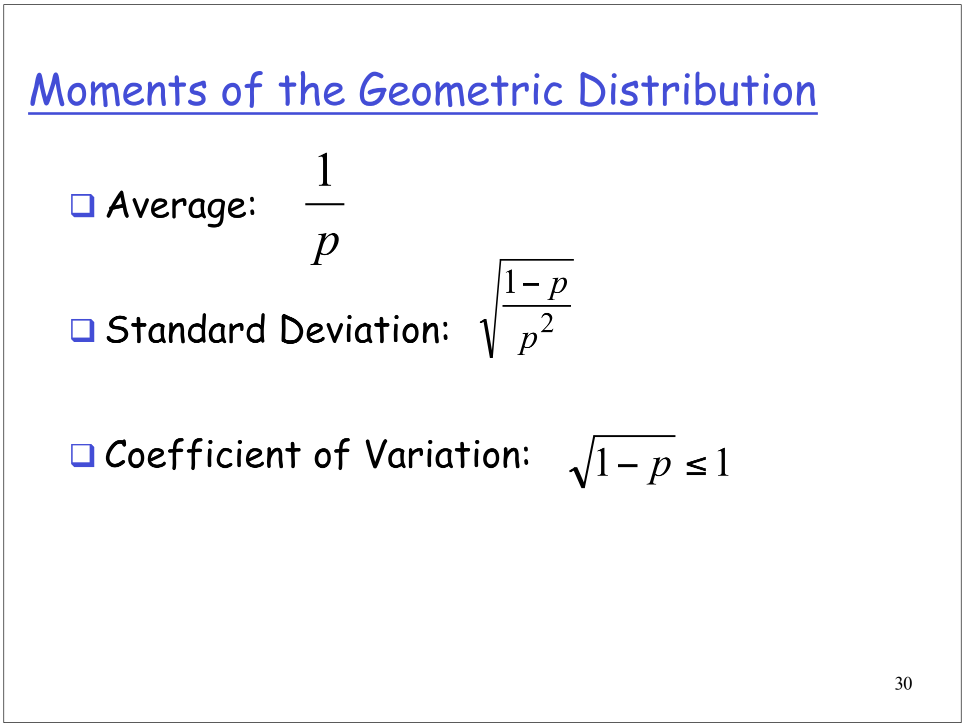
- Discrete random variable with the “memoryless” property



Geometric Distribution

Success probability	0.6

n	P[X=n]
1	0.6000
2	0.2400
3	0.0960
4	0.0384
5	0.0154
6	0.0061
7	0.0025
8	0.0010
9	0.0004
10	0.0002



Moments of the Geometric Distribution

- ❑ Average: $\dfrac{1}{p}$
- ❑ Standard Deviation: $\sqrt{\dfrac{1-p}{p^2}}$
- ❑ Coefficient of Variation: $\sqrt{1-p} \leq 1$

Poisson Distribution

- Used to model the number of arrivals over a given interval, e.g.,
  - Number of requests to a server
  - Number of failures of a component
  - Number of queries to the database.
- A Poisson distribution usually arises when arrivals come from a large number of independent sources.

Poisson Distribution

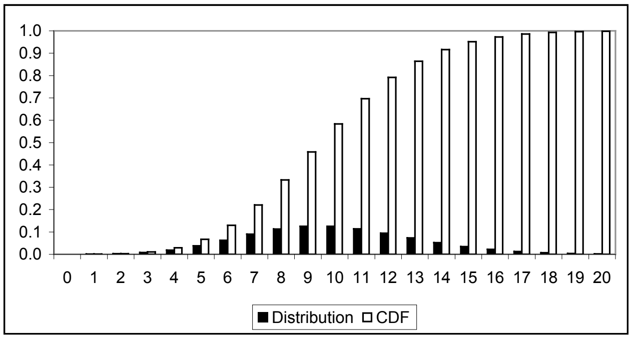
- Distribution: $P[X = k] = \dfrac{\lambda^k e^{-\lambda}}{k!} \qquad k = 0,1,\ldots,\infty$
- Counting arrivals in an interval of duration *t*:

$$P[k \text{ arrivals in } [0,\mathrm{t})] = \frac{(\lambda t)^k e^{-\lambda t}}{k!} \qquad k = 0,1,\ldots,\infty$$

- Average = Standard Deviation = λ

Poisson Distribution

Lambda	10	
K	Poisson Distribution	CDF
0	0.00005	0.0000
1	0.00045	0.0005
2	0.00227	0.0028
3	0.00757	0.0103
4	0.01892	0.0293
5	0.03783	0.0671
6	0.06306	0.1301
7	0.09008	0.2202
8	0.11260	0.3328
9	0.12511	0.4579
10	0.12511	0.5830
11	0.11374	0.6968
12	0.09478	0.7916
13	0.07291	0.8645
14	0.05208	0.9165
15	0.03472	0.9513
16	0.02170	0.9730
17	0.01276	0.9857
18	0.00709	0.9928
19	0.00373	0.9965
20	0.00187	0.9984



In Excel:
P[X=k] = POISSON (k,λ,FALSE)
P[X≤k] = POISSON (k,λ,TRUE)

Continuous Random Variables

Relevant Functions

- Probability density function (pdf) of r.v. X: $f_X(x)$

$$P[a \le X \le b] = \int_a^b f_X(x)dx$$

- Cumulative distribution function (CDF):

$$F_X(x) = P[X \le x]$$

  - pdf is the derivative of the CDF *f(x) = dF(x)/dx*

- Tail of the distribution (reliability function):

$$R_X(x) = P[X > x] = 1 - F_X(x)$$

Moments

- k-th moment: $E[X^k] = \int_{-\infty}^{+\infty} x^k f_X(x)dx$

- Expected value (mean): first moment

$$\mu = E[X] = \int_{-\infty}^{+\infty} x f_X(x)dx$$

- k-th central moment:

$$E[(X-\mu)^k] = \int_{-\infty}^{+\infty} (x-\mu)^k f_X(x)dx$$

- Variance: second central moment

$$\sigma^2 = E[(X-\mu)^2] = \int_{-\infty}^{+\infty} (x-\mu)^2 f_X(x)dx$$

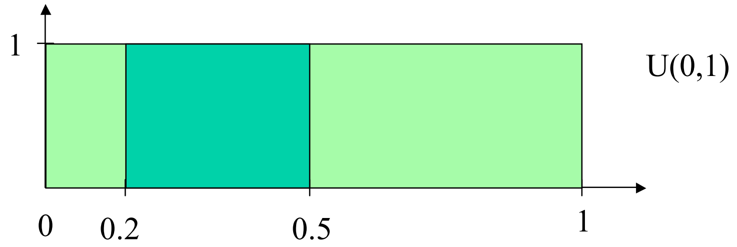
Important continuous distributions

- Uniform
- Exponential
- Normal
- Erlang
- Hypo-exponential
- Hyper-exponential
- Weibull
- Lognormal
- Pareto

The Uniform Distribution

- pdf: $$f_X(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$
- Mean: $\mu = \frac{a+b}{2}$
- Variance: $\sigma^2 = \frac{(b-a)^2}{12}$

The Uniform Distribution



P[0.2<X<0.5]=(0.5-0.2)x1.0=0.3

The Normal Distribution $N(\mu,\sigma)$

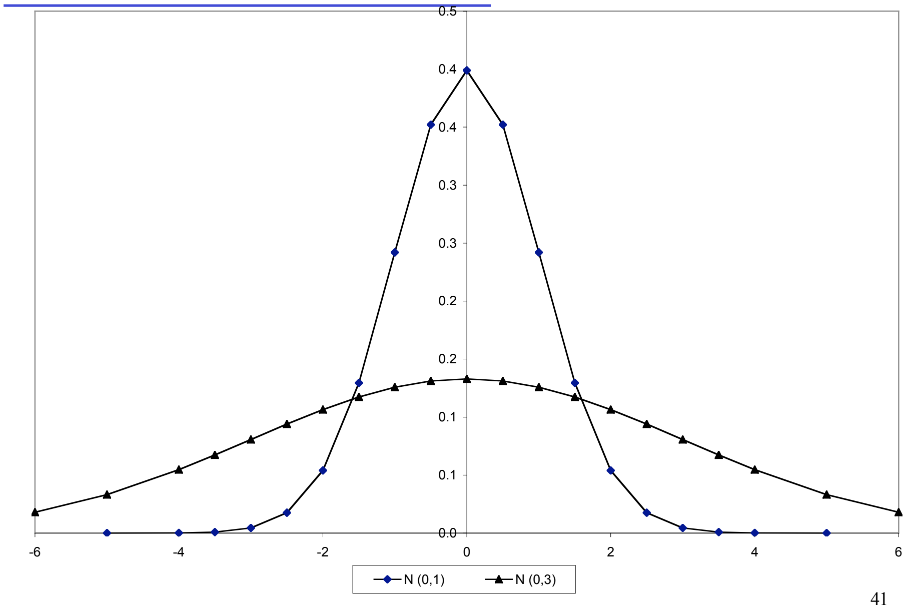
- Important because
  - Many natural phenomena follow a normal distribution (bell curve)
  - Sum of independent normal variables is normally distributed
  - Sum of a large number of independent observations from any distribution tends to have a normal distribution

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(1/2)\left[(x-\mu)/\sigma\right]^2}$$

- Two parameters: mean and standard deviation.

The Normal Distribution

$N(\mu,\sigma)$



The Standard Normal Distribution

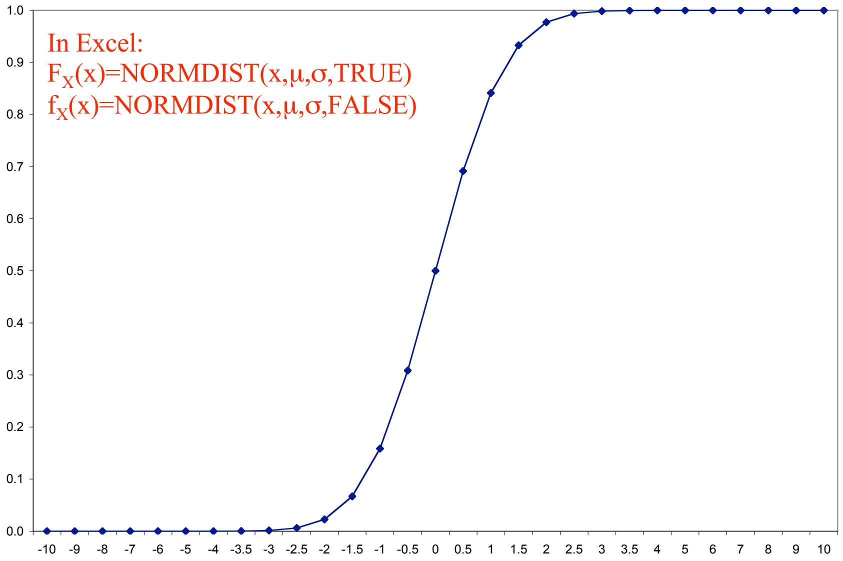
- ❑ To use tables for computing values related to the normal distribution, we need to standardize a normal r.v. as

standard normal score

$$Z = \frac{X - \mu}{\sigma}$$

- ❑ Given X, compute a Z value z.
- ❑ Find the area value in a Table (Prob [0<Z<z]).

Normal CDF



In Excel:
$F_X(x)$=NORMDIST(x,μ,σ,TRUE)
$f_X(x)$=NORMDIST(x,μ,σ,FALSE)

Using Normal Tables

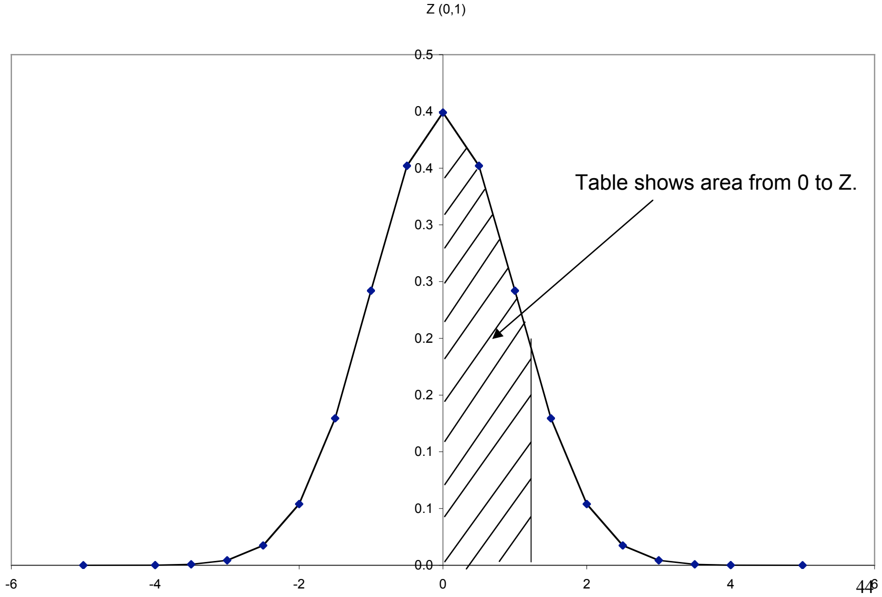
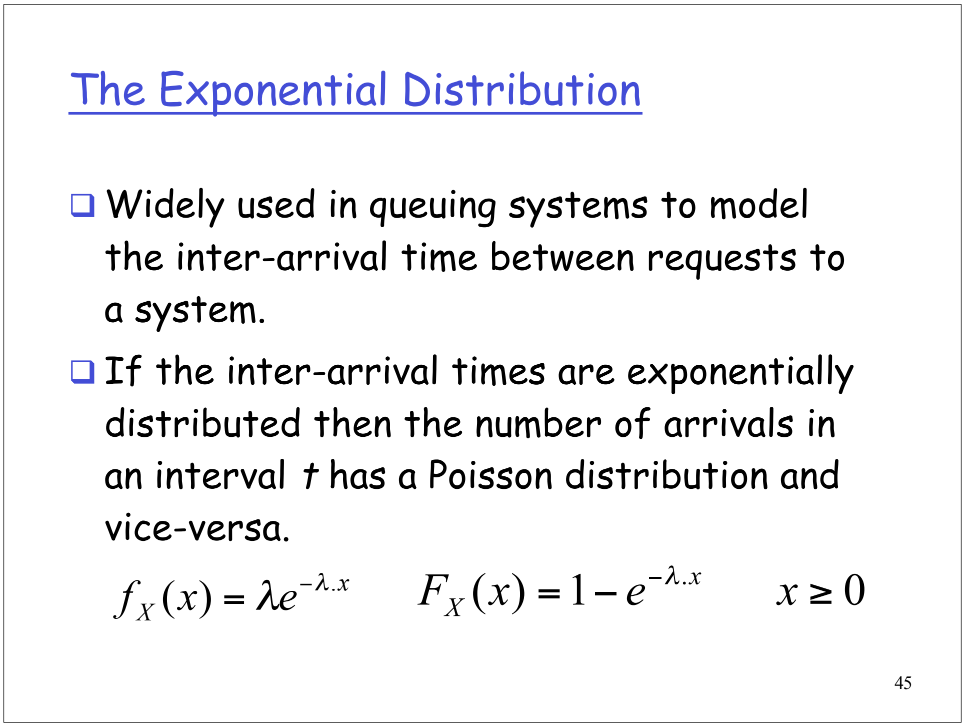


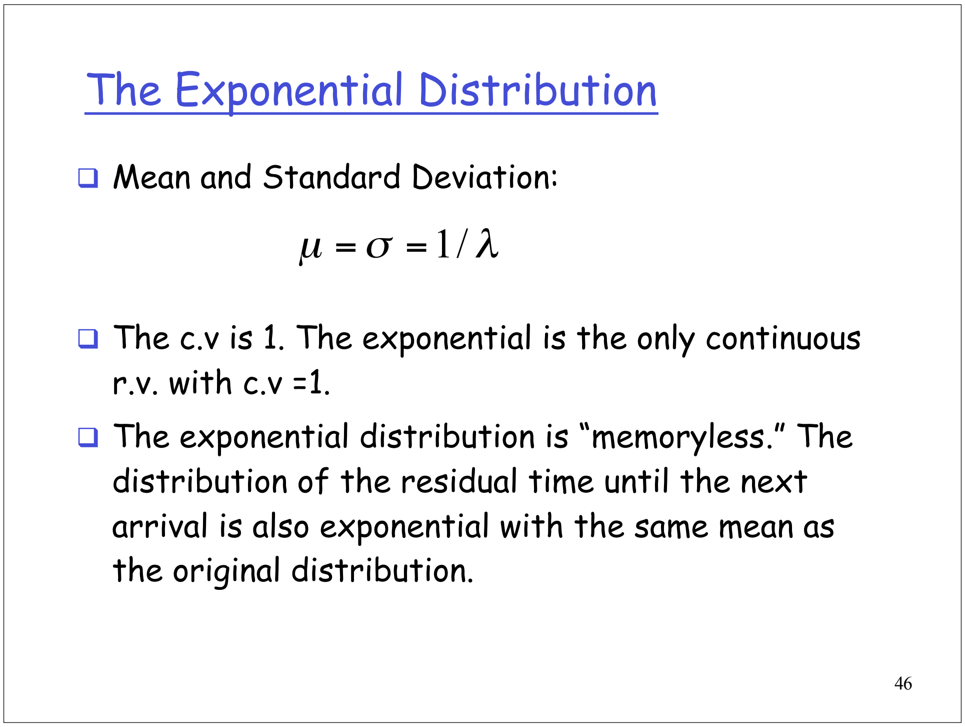
Table shows area from 0 to Z.



The Exponential Distribution

- ❑ Widely used in queuing systems to model the inter-arrival time between requests to a system.
- ❑ If the inter-arrival times are exponentially distributed then the number of arrivals in an interval *t* has a Poisson distribution and vice-versa.

$$f_X(x) = \lambda e^{-\lambda . x} \qquad F_X(x) = 1 - e^{-\lambda . x} \qquad x \geq 0$$



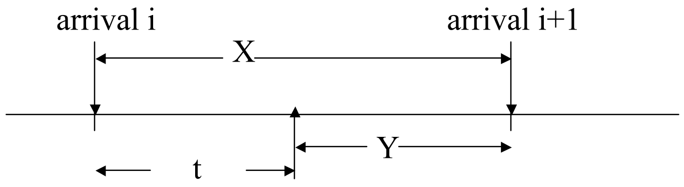
The Exponential Distribution

- ❑ Mean and Standard Deviation:

$$\mu = \sigma = 1/\lambda$$

- ❑ The c.v is 1. The exponential is the only continuous r.v. with c.v =1.
- ❑ The exponential distribution is “memoryless.” The distribution of the residual time until the next arrival is also exponential with the same mean as the original distribution.

Memoryless Property of the Exponential Distribution



$$P[Y \le y \,|\, X > t] = P[X - t \le y \,|\, X > t]$$

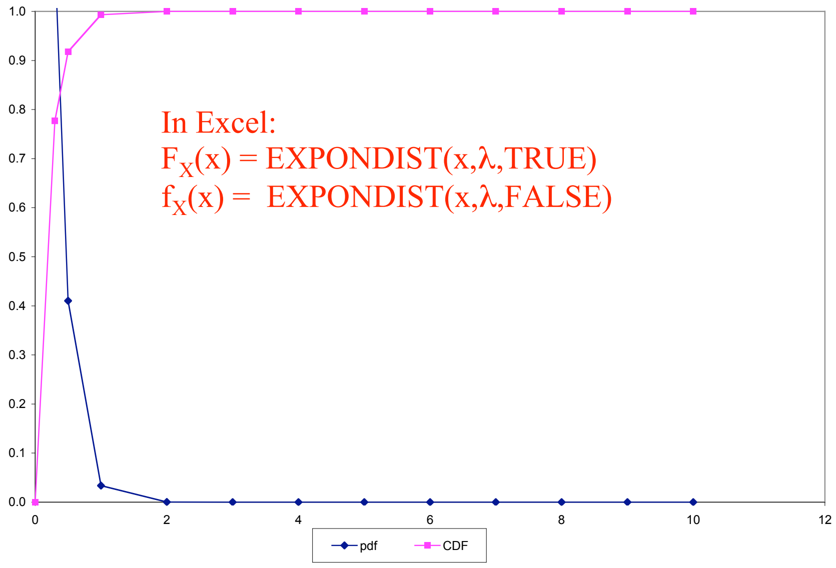
$$P[X \le y + t \mid X > t] = \frac{P[t < X \le y + t]}{P[X > t]}$$

$$= \frac{P[X \le y + t] - P[X \le t]}{P[X > t]}$$

$$= \frac{1 - e^{-\lambda.(y+t)} - (1 - e^{-\lambda.t})}{1 - (1 - e^{-\lambda.t})}$$

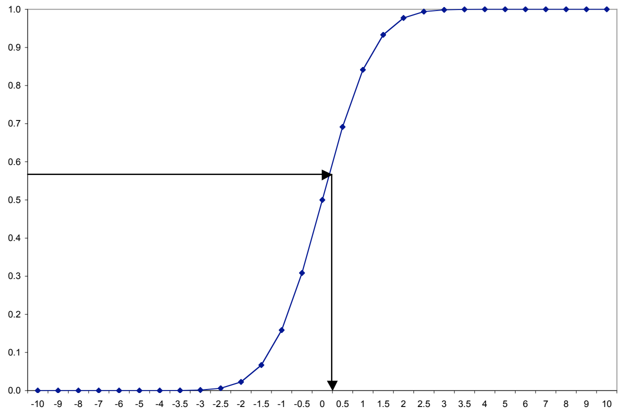
$$= 1 - e^{-\lambda.y}$$

Exponential Distribution



In Excel:
$F_X(x)$ = EXPONDIST(x,λ,TRUE)
$f_X(x)$ = EXPONDIST(x,λ,FALSE)

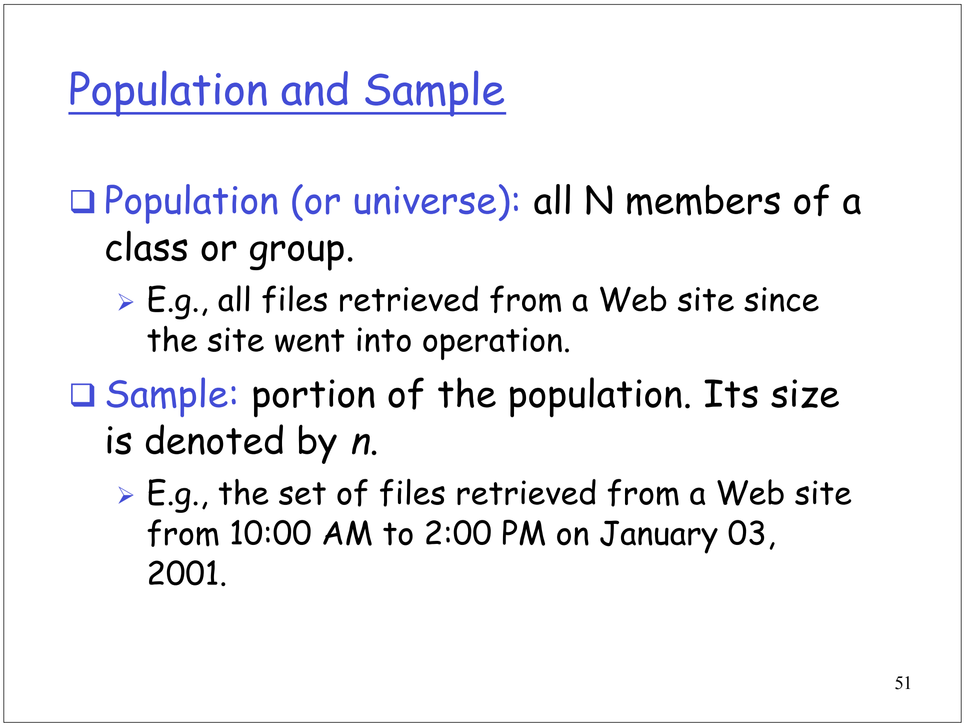
Generation of Random Variables



- randomly generate a number u = U(01,)
- x = F^{-1} (u) where F is the CDF

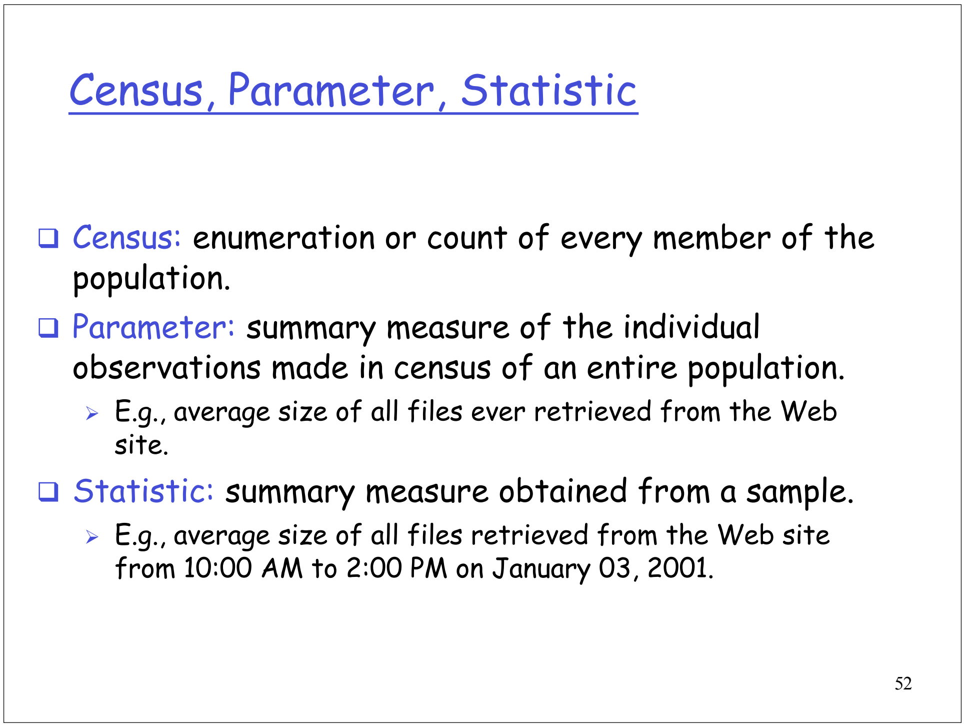
Goals in Studying Statistics

- ❑ Analyze, present, and describe numerical information properly.
- ❑ Draw conclusions about the properties of large populations from sample information (inference).
- ❑ Design experiments to learn about real-world situations.
- ❑ To forecast or predict not-measured values from a set of measurements.



Population and Sample

- □ Population (or universe): all N members of a class or group.
  - E.g., all files retrieved from a Web site since the site went into operation.
- □ Sample: portion of the population. Its size is denoted by *n*.
  - E.g., the set of files retrieved from a Web site from 10:00 AM to 2:00 PM on January 03, 2001.



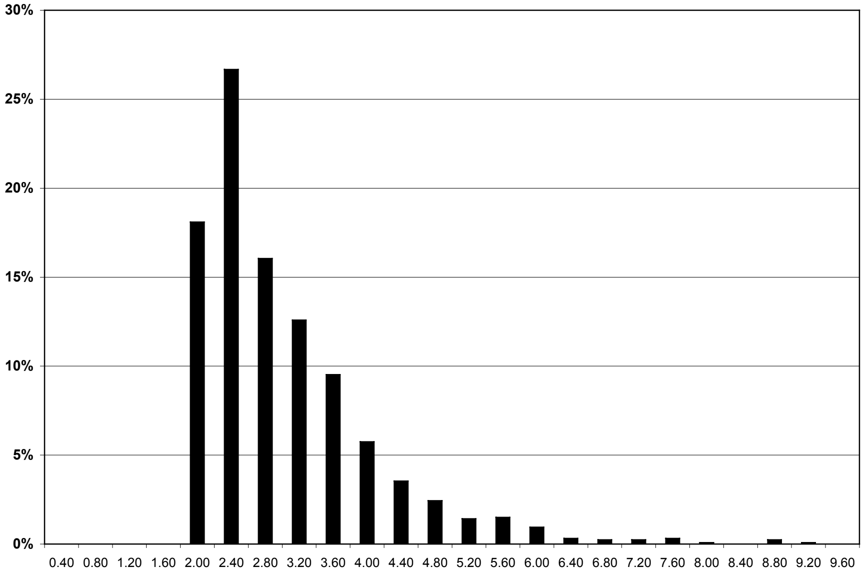
Census, Parameter, Statistic

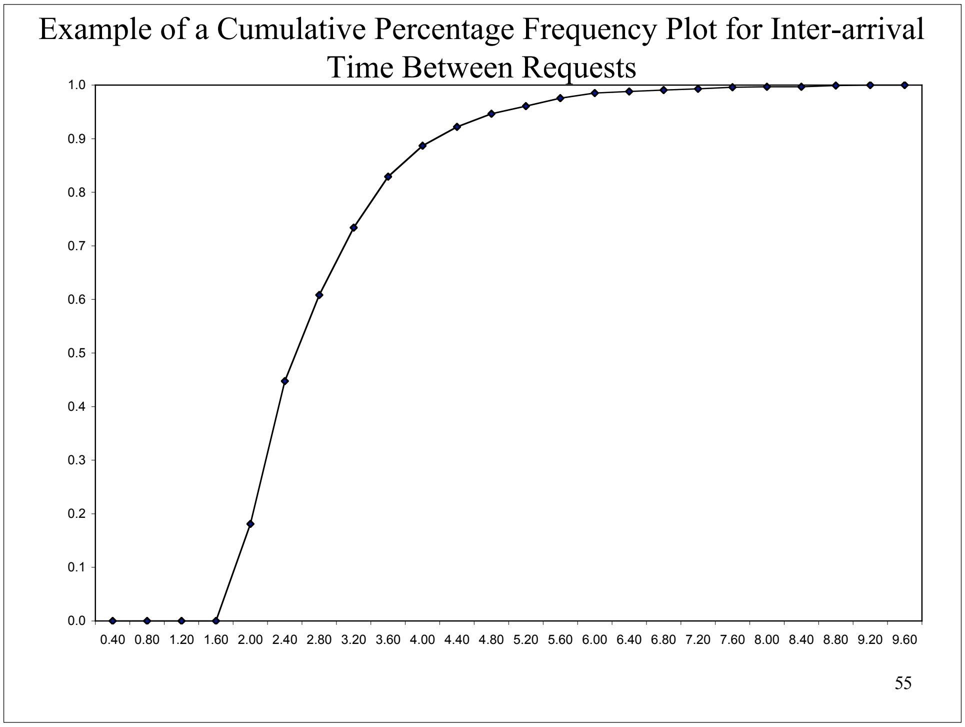
- □ Census: enumeration or count of every member of the population.
- □ Parameter: summary measure of the individual observations made in census of an entire population.
  - E.g., average size of all files ever retrieved from the Web site.
- □ Statistic: summary measure obtained from a sample.
  - E.g., average size of all files retrieved from the Web site from 10:00 AM to 2:00 PM on January 03, 2001.

Visualizing Numerical Data

- ❑ Type of Plots:
  - Percent frequency histograms: show the percentage of occurrences of values in a bin (range of values).
  - Cumulative frequency histograms.
  - Stacked histograms, Gantt charts, Kiviat charts, Schumacher charts
    - o see Chapter 10 of Jain

Example of a Percentage Frequency Histogram for Inter-arrival Time Between Requests





Example of a Cumulative Percentage Frequency Plot for Inter-arrival Time Between Requests