

A Probability and Statistics Refresher

Daniel A. Menascé

Department of Computer Science

www.cs.gmu.edu/faculty/menasce.html

© 1999–2001 Menascé. All Rights Reserved.

1

Review of Basic Probability Results

- The HTTP log of a Web server was analyzed and file sizes were collected:

File	File Size (KB)
1	2.3
2	3.7
3	10.4
4	2.2
5	7.3
6	102.0
7	2.9
8	4.0
9	30.0
10	1.2
11	3.4
12	20.0
13	3.5
14	9.0
15	2.8

© 1999–2001 Menascé. All Rights Reserved.

2

Sample vs. Population

- Population: all possible file sizes (very large or infinite)
- Sample: a finite number of file sizes.
- Q: How we present measured data?

Probability 101

- Random Variable (r.v.): a variable that takes one of a specified set of values with a given probability.
 - X: size of file retrieved from a Web server
- Cumulative Distribution Function (CDF):

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

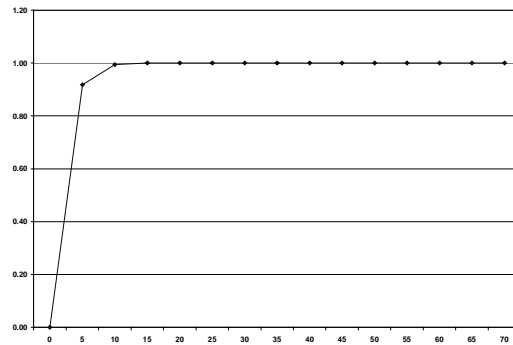
Probability 101

- Properties of a CDF:

$$0 \leq F_X(x) \leq 1$$

$$\lim_{x \rightarrow \infty} F_X(x) = 1$$

$$b > a \Rightarrow F_X(b) \geq F_X(a)$$



© 1999–2001 Menascé. All Rights Reserved.

5

Probability 101

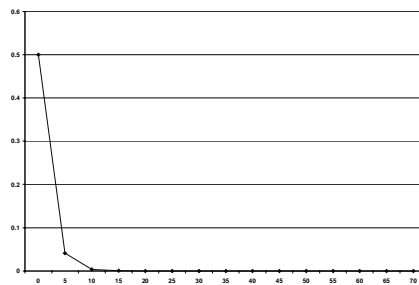
- Probability Density Function (pdf)

$$f_X(x) = \frac{dF_X(x)}{dx}$$

- Properties of the pdf:

$$\int_{x=a}^{x=b} f_X(x) dx = P[a \leq X \leq b]$$

$$\int_{x=0}^{x=\infty} f_X(x) dx = 1$$



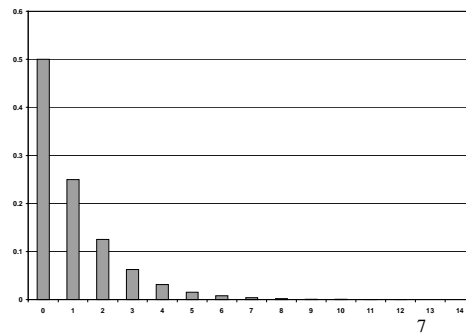
© 1999–2001 Menascé. All Rights Reserved.

6

Probability Mass Function

- For discrete r.v.s:
 - Values: x_1, \dots, x_n
 - Probabilities: p_1, \dots, p_n
- Property:

$$\sum_{i=1}^n p_i = 1$$



© 1999–2001 Menascé. All Rights Reserved.

Probability 101

- Mean or Expected Value

$$E[X] = \mu = \sum_{k=1}^n x_k \times p_k = \int_{x=0}^{\infty} x f_X(x) dx$$

- Variance

$$\sigma^2 = \text{Var}(X) = E[(x - \mu)^2] = \sum_{k=1}^n (x_k - \mu)^2 \times p_k$$

$$= \int_{x=0}^{\infty} (x - \mu)^2 f_X(x) dx$$

- Standard Deviation

σ

© 1999–2001 Menascé. All Rights Reserved.

8

Probability 101

- Coefficient of Variation (COV):

$$COV = \frac{\sigma}{\mu}$$

- α -percentile: value of x at which the PDF takes the value α . E.g., if the 90-percentile of the file size is 2Kbytes, then

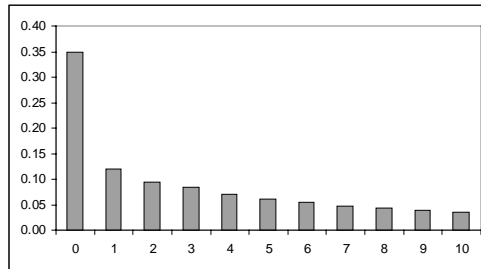
$$P[X \leq 2KB] = 0.9$$

- Median: 50-percentile.

Discrete Probability Distributions

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	

© 1999–2001 Menascé. All Rights Reserved.

11

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

© 1999–2001 Menascé. All Rights Reserved.

12

Central Moments of a Discrete Random Variable

- k-th central moment:

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

- The variance is the second central moment:

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

© 1999–2001 Menascé. All Rights Reserved.

13

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

© 1999–2001 Menascé. All Rights Reserved.

14

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.

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

Discrete Random Variables

- Binomial
- Hypergeometric
- Negative Binomial
- Geometric
- Poisson

The Binomial Distribution

- Distribution: based on carrying out 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.

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

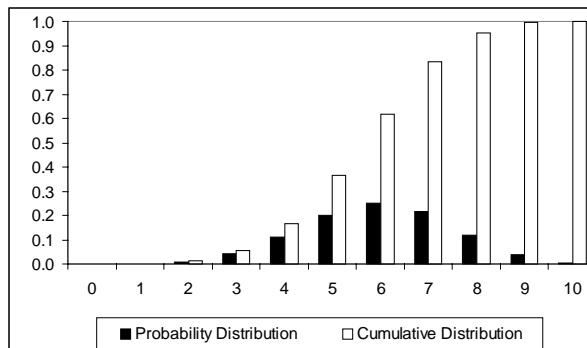
© 1999–2001 Menascé. All Rights Reserved.

17

The Binomial Distribution

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

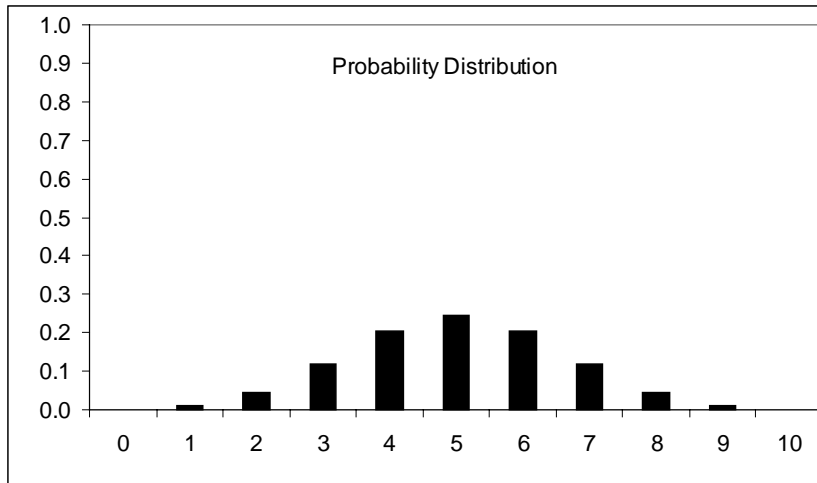
Number of Attempts (k)	Probability k successful attempts in n	Cumulative
0	0.000105	0.000105
1	0.001573	0.001678
2	0.010617	0.012295
3	0.042467	0.054762
4	0.111477	0.166239
5	0.200658	0.366897
6	0.250823	0.617719
7	0.214991	0.832710
8	0.120932	0.953643
9	0.040311	0.993953
10	0.006047	1.000000



© 1999–2001 Menascé. All Rights Reserved.

18

Shape of the Binomial Distribution

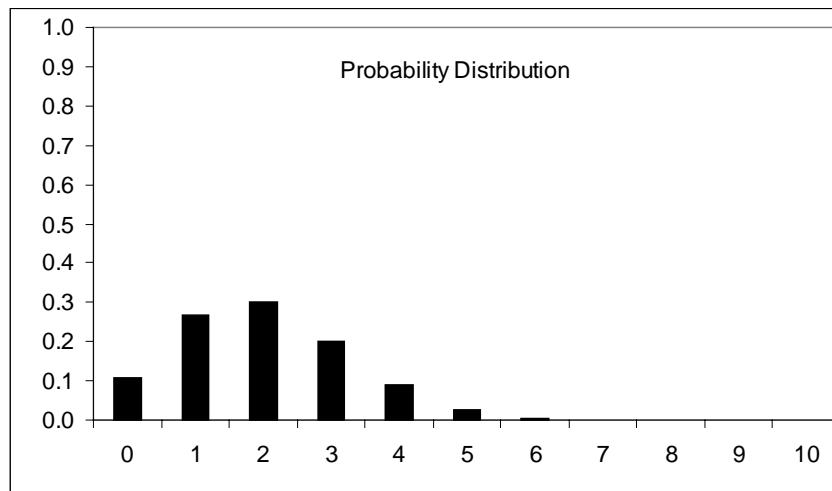


$p = 0.5$ symmetric for any n .

© 1999–2001 Menascé. All Rights Reserved.

19

Shape of the Binomial Distribution

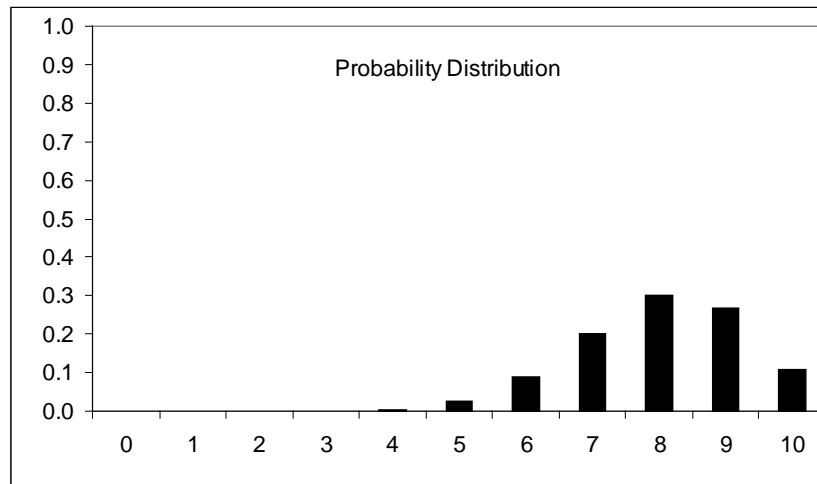


$p = 0.2$ right skewed

© 1999–2001 Menascé. All Rights Reserved.

20

Shape of the Binomial Distribution



$p = 0.8$ left skewed

© 1999–2001 Menascé. All Rights Reserved.

21

Moments of the Binomial Distribution

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

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

© 1999–2001 Menascé. All Rights Reserved.

22

Hypergeometric Distribution

- Binomial was based on experiments with equal success probability.
- Hypergeometric: not all experiments have the same success probability.
- Given a sample size of n out of a population of size N with A known successes in the population, the probability of k successes is

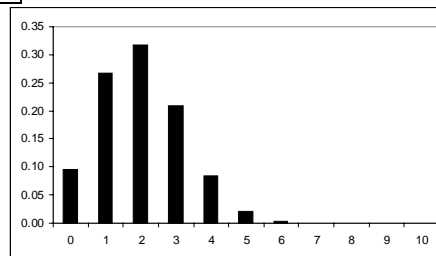
$$P[X = k] = \frac{\overbrace{\binom{A}{k} \binom{N-A}{n-k}}^{\substack{\text{choose } k \text{ successes out of } A \\ \text{successes in the population}} \quad \substack{\text{choose } (n-k) \text{ failures from} \\ N-A \text{ failures in the population}}}^{\substack{\text{choose } k \text{ successes out of } A \\ \text{successes in the population}}} \underbrace{\binom{N}{n}}_{\text{total \# of possible samples}}$$

© 1999–2001 Menascé. All Rights Reserved.

23

Hypergeometric Distribution

No. successes in sample k	sample size n	no. successes in population A	population size N	
0	20	10	100	0.09511627
1	20	10	100	0.26793316
2	20	10	100	0.31817063
3	20	10	100	0.20920809
4	20	10	100	0.08410730
5	20	10	100	0.02153147
6	20	10	100	0.00354136
7	20	10	100	0.00036793
8	20	10	100	0.00002300
9	20	10	100	0.00000078
10	20	10	100	0.00000001



© 1999–2001 Menascé. All Rights Reserved.

24

Moments of the Hypergeometric

- Average: $\frac{nA}{N}$
- Standard Deviation: $\sqrt{\frac{nA(N-A)}{N^2}} \sqrt{\frac{N-n}{N-1}}$
- If the sample size is less than 5% of the population, the binomial is a good approximation for the hypergeometric.

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.

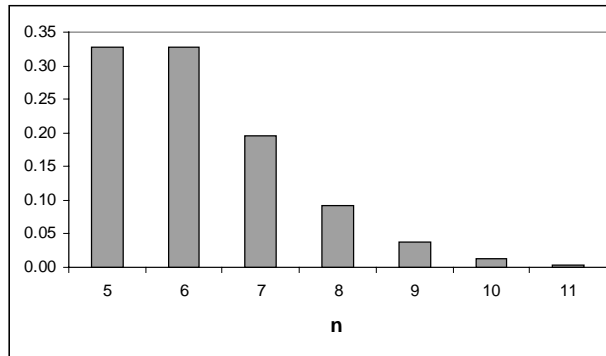
$$P[X = n] = \binom{n-1}{k-1} (1-p)^{n-k} p^k$$

$$\frac{S}{1} \quad \frac{F}{2} \quad \frac{F}{3} \quad \frac{S}{4} \quad \dots \quad \frac{F}{n-1} \quad \frac{S}{n}$$

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



© 1999–2001 Menascé. All Rights Reserved.

27

Moments of the Negative Binomial Distribution

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

© 1999–2001 Menascé. All Rights Reserved.

28

Geometric Distribution

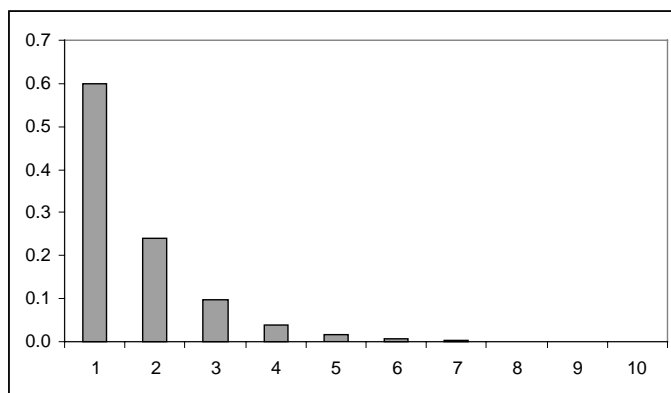
- Special case of the negative binomial with $k=1$.
- Probability that the first success occurs after n trials is

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

Geometric Distribution

Success probability

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: $\frac{1}{p}$
- Standard Deviation: $\sqrt{\frac{1-p}{p^2}}$
- 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: $P[X = k] = \frac{\lambda^k e^{-\lambda}}{k!} \quad k = 0, 1, \dots, \infty$

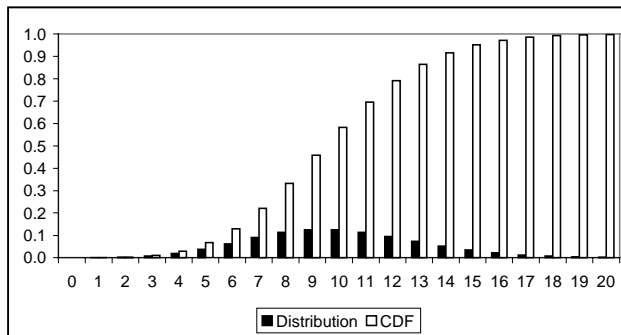
- Counting arrivals in an interval of duration t :

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

- Average=Variance= 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



Continuous Random Variables

© 1999–2001 Menascé. All Rights Reserved.

35

Relevant Functions

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

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

- Cumulative distribution function (CDF):

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

- Tail of the distribution (reliability function):

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

© 1999–2001 Menascé. All Rights Reserved.

36

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

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

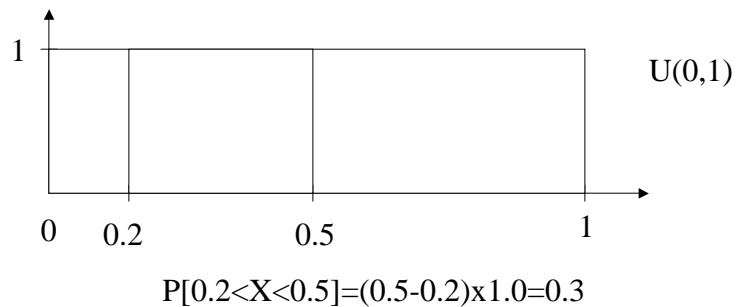
The Uniform Distribution

- pdf: $f_X(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$

- Mean: $\mu = \frac{a+b}{2}$

- Variance: $\sigma^2 = \frac{(b-a)^2}{12}$

The Uniform Distribution



© 1999–2001 Menascé. All Rights Reserved.

39

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

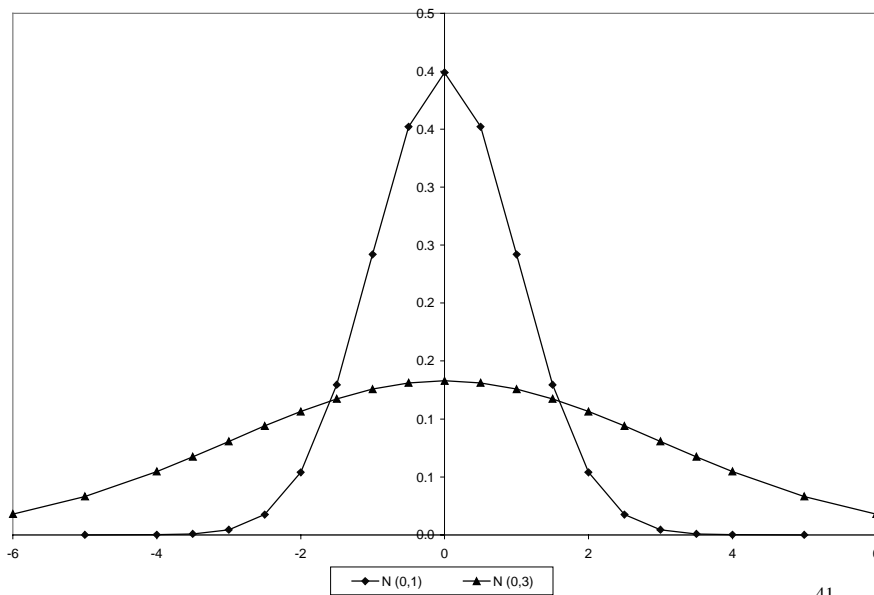
- Many natural phenomena follow a normal distribution.
- The normal distribution can be used to approximate the binomial and the Poisson distributions.
- Two parameters: mean and standard deviation.

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

© 1999–2001 Menascé. All Rights Reserved.

40

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



41

The Standard Normal Distribution

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

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

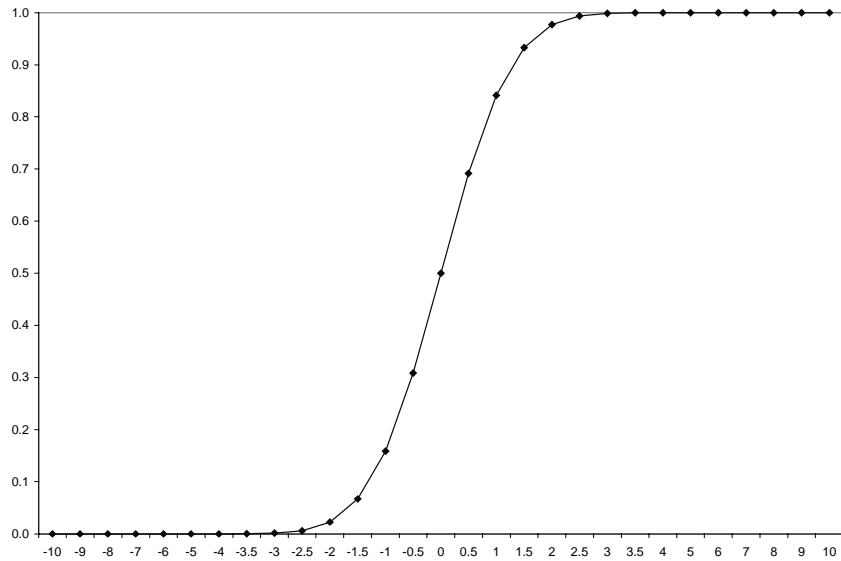
standard normal score \swarrow

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

© 1999–2001 Menascé. All Rights Reserved.

42

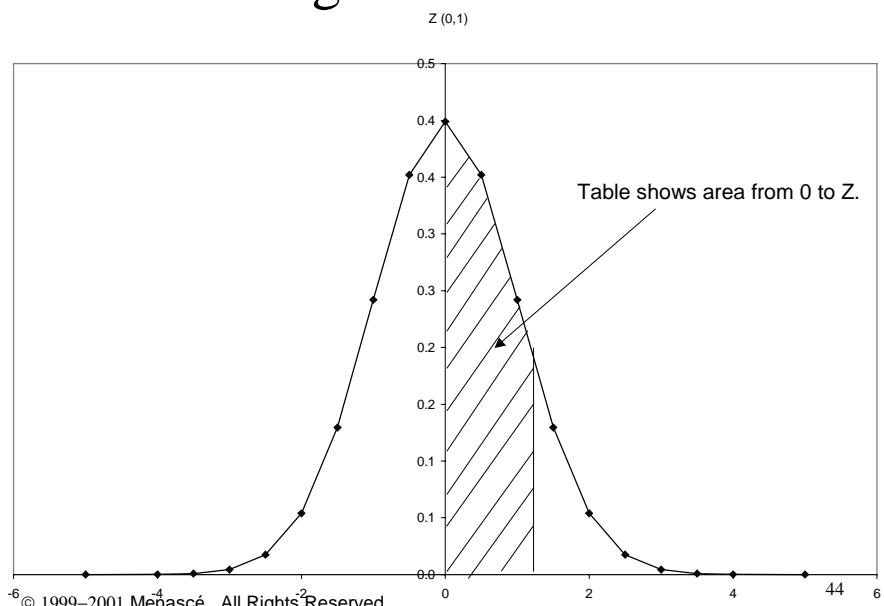
Normal CDF



© 1999–2001 Menascé. All Rights Reserved.

43

Using Normal Tables



© 1999–2001 Menascé. All Rights Reserved.

The Normal as an Approximation to the Binomial Distribution

- The normal can approximate the binomial if the variance of the binomial

$$np(1-p) \geq 10$$

- Binomial: $\mu = np$

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

- Transformation: $Z = \frac{X - np}{\sqrt{np(1-p)}}$

The Normal as an Approximation to the Binomial Distribution

- Consider a binomial r.v. X with average 50 and variance 25. What is $P[50 \leq X \leq 60]$?

- Transformation: $Z = \frac{X - 50}{\sqrt{25}} = \frac{60 - 50}{5} = 2.0$

- Using the table, the area between 50 and 60 for $Z=2.0$ is 0.4772. So,

$$P[50 \leq X \leq 60] = 0.4772$$

The Normal as an Approximation to the Poisson Distribution

- The normal can approximate the Poisson if the $\lambda > 5$.

- Poisson: $\mu = \lambda$
 $\sigma = \sqrt{\lambda}$

- Transformation: $Z = \frac{X - \lambda}{\sqrt{\lambda}}$

© 1999–2001 Menascé. All Rights Reserved.

47

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.

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

© 1999–2001 Menascé. All Rights Reserved.

48

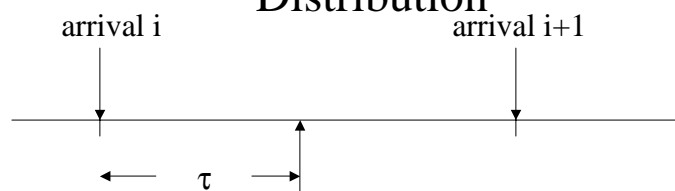
The Exponential Distribution

- Mean and Standard Deviation:

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

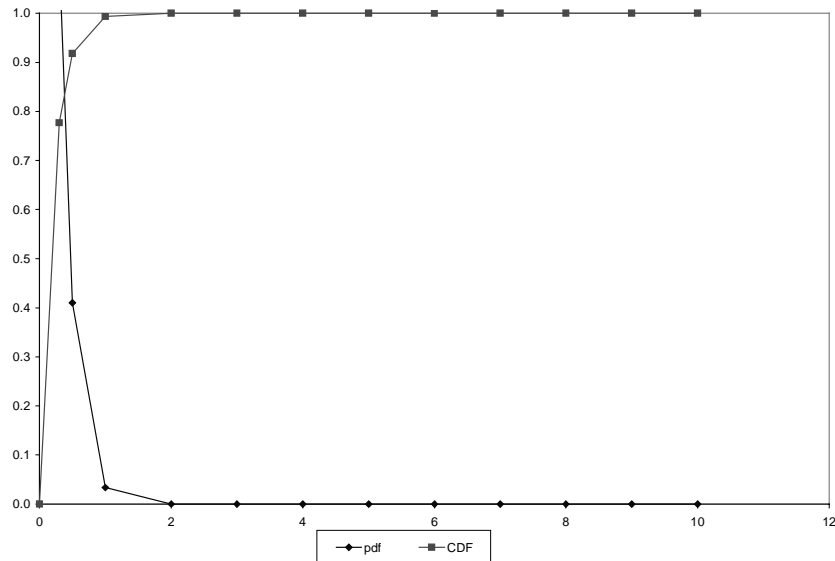
- The COV is 1. The exponential is the only continuous r.v. with COV=1.
- The exponential distribution is “memoryless.” The distribution of the residual time until the next arrival is also exponential with the same mean as the original distribution.

Memoryless Property of the Exponential Distribution



$$\begin{aligned}
 P[\tilde{t} \leq t + \tau \mid \tilde{t} > \tau] &= \frac{P[\tau < \tilde{t} \leq t + \tau]}{P[\tilde{t} > \tau]} \\
 &= \frac{P[\tilde{t} \leq t + \tau] - P[\tilde{t} \leq \tau]}{P[\tilde{t} > \tau]} \\
 &= \frac{1 - e^{-\lambda \cdot (t + \tau)} - (1 - e^{-\lambda \cdot \tau})}{1 - (1 - e^{-\lambda \cdot \tau})} \\
 &= 1 - e^{-\lambda \cdot t}
 \end{aligned}$$

Exponential Distribution



© 1999–2001 Menascé. All Rights Reserved.

51

Pareto Distribution

- A case of a heavy-tailed distribution.
- The probability of large values is not negligible.

$$f_X(x) = \frac{a}{x^{1+a}} \quad a > 0, \quad x \geq 1$$

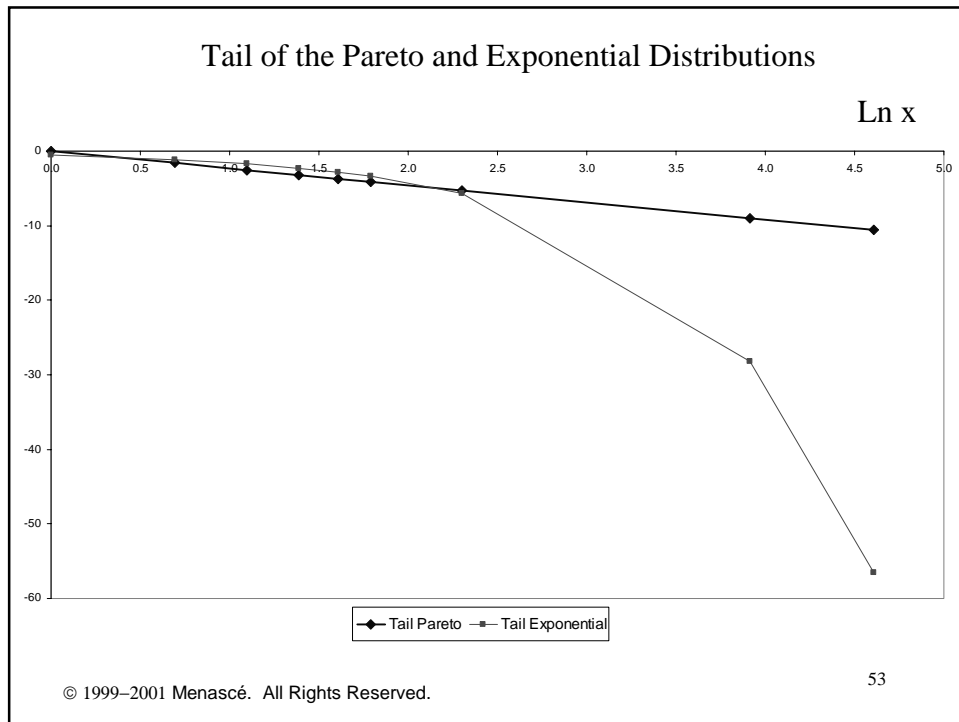
$$F_X(x) = 1 - \frac{1}{x^a} \quad a > 0, \quad x \geq 1$$

- Mean: $\frac{a}{a-1} \quad a > 1$

- Variance: $\frac{a}{(a-1)^2(a-2)} \quad a > 2$

© 1999–2001 Menascé. All Rights Reserved.

52



Sample Statistics

© 1999–2001 Menascé. All Rights Reserved.

54

Sample Statistics

- Sample Mean

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

- Sample Variance

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

File	File Size (KB)
1	2.3
2	3.7
3	10.4
4	2.2
5	7.3
6	102.0
7	2.9
8	4.0
9	30.0
10	1.2
11	3.4
12	20.0
13	3.5
14	9.0
15	2.8

$$\bar{x} = 13.65$$

$$s^2 = 659.59$$

$$s = 25.68$$

© 1999–2001 Menascé. All Rights Reserved.

55

Confidence Intervals

- When analyzing measurements one cannot make a definite statement such as “the average measured response time is 0.65 sec.”
- What we can say is something to the effect of “with 90% confidence, the measured response time is in the interval (0.62,0.68).”
- The interval (0.62, 0.68) is the confidence interval and 90% is the confidence level.

© 1999–2001 Menascé. All Rights Reserved.

56

Confidence Interval Estimation of the Mean

- Known population standard deviation.
- Unknown population standard deviation:
 - Large samples ($n > 30$): sample standard deviation is a good estimate for population standard deviation. OK to use normal distribution.

Central Limit Theorem

- If the observations in a sample are independent and come from the same population that has mean μ and standard deviation σ then the sample mean for **large** samples has a normal distribution with mean μ and standard deviation σ/\sqrt{n} .

$$\bar{x} \sim N(\mu, \sigma / \sqrt{n})$$

- The standard deviation of the sample mean is called the *standard error*.

Computing Confidence Intervals

- For large samples ($n > 30$).

A 100 $(1 - \alpha)\%$ confidence interval for the population mean is

$$(\bar{x} - z_{1-\alpha/2} s / \sqrt{n}, \bar{x} + z_{1-\alpha/2} s / \sqrt{n})$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ -quantile of a $N(0,1)$

$$z_{1-\alpha/2} = \text{NORMSINV}(1 - \alpha / 2)$$

MS Excel function

© 1999–2001 Menascé. All Rights Reserved.

59

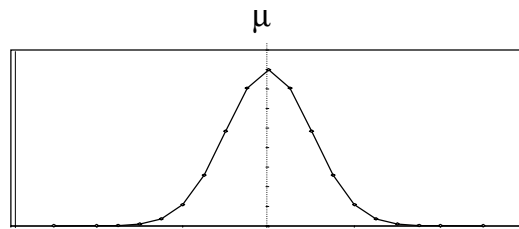
Example of Confidence Interval

File	Resp Time	(xi-xbar)*(xi-xbar)
1	0.650	0.00037636
2	0.540	0.00820836
3	0.620	0.00011236
4	0.570	0.00367236
5	0.620	0.00011236
6	0.680	0.00244036
7	0.590	0.00164836
8	0.550	0.00649636
9	0.625	0.00003136
10	0.675	0.00197136
11	0.645	0.00020736
12	0.673	0.00179776
13	0.667	0.00132496
14	0.700	0.00481636
15	0.654	0.00054756
xbar	0.6306	0.03376360
s2	0.002411686	
s	0.049108917	

	Interval	
90%	0.609742	0.651458
95%	0.605773	0.655427
99%	0.597937	0.663263

© 1999–2001 Menascé. All Rights Reserved.

60



Sample		Interval include μ ?
1		YES
2		YES
3		NO
...		
100		YES

$100(1 - \alpha)$ of the 100 samples include the population mean μ .