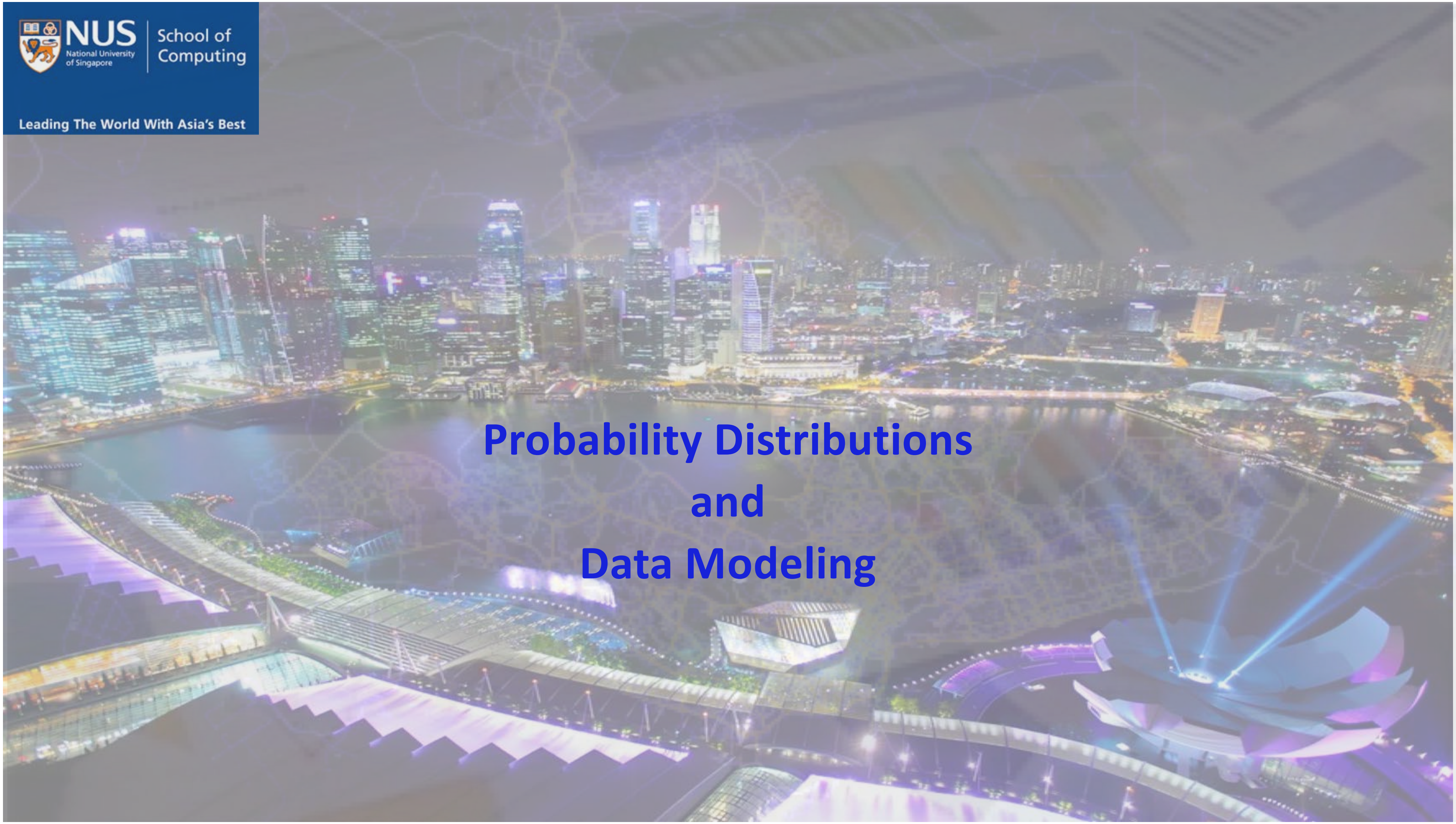
NUS National University of Singapore
School of Computing
Leading The World With Asia's Best



Probability Distributions and Data Modeling

Learning objectives

- Explain the concept of probability and provide examples of 3 definitional perspectives of probability
- Use probability rules and formulas to perform probability calculations
- Explain differences between discrete and continuous random variable
- Describe what is a normal distribution and its properties
- Compute normal probabilities from normal and standard normal distributions

Basic Concepts of Probability

- Probability: likelihood that an outcome occurs
- Experiment: process that results in an outcome
- Outcome of an experiment: result that we observe
- Sample space: collection of all possible outcomes of an experiment
- Events: a collection of one or more outcomes from a sample space

Basic Concepts of Probability

- Probabilities may be defined from one of three perspectives:
  - Classical definition: probabilities can be deduced from theoretical arguments
  - Relative frequency definition: probabilities are based on empirical data
  - Subjective definition: probabilities are based on judgment and experience

Basic Concepts of Probability

Examples of Classical Definition of Probability

Eg 1: Suppose we roll 2 dice

- Sample space: {(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),
(2,1),(2,2),(2,3),(2,4),(2,5),(2,6),
(3,1),(3,2),(3,3),(3,4),(3,5),(3,6),
(4,1),(4,2),(4,3),(4,4),(4,5),(4,6),
(5,1),(5,2),(5,3),(5,4),(5,5),(5,6),
(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)}
- Probability die rolls sum to three = 2/36

Eg 2: Suppose two consumers try a new product

- Assume 4 equally likely outcomes:
  1. like, like
  2. like, dislike
  3. dislike, like
  4. dislike, dislike
- Probability at least one dislikes product = 3/4

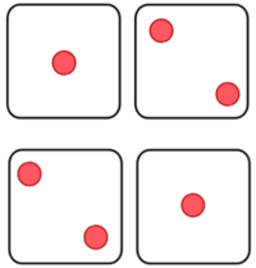
Events

Die Sum	Frequency
2	1
3	2
4	3
5	4
6	5
7	6
8	5
9	4
10	3
11	2
12	1
Sum	36

Sample space

First Die

Second Die



Basic Concepts of Probability

Two Basic Facts Govern Probability

- Probability associated with any outcome must be between 0 and 1.

  $0 \leq P(O_i) \leq 1$ for each outcome O_i

- Sum of the probabilities over all possible outcomes must be equal to 1.

  $$P(O_1) + P(O_2) + \ldots + P(O_n) = 1$$

(n is the number of outcomes in the sample space.)

Rule 1. Probability of any event is the sum of probability of outcomes that comprise that event

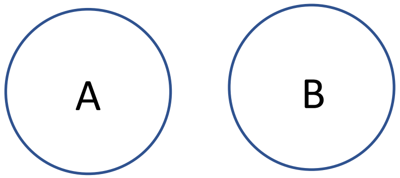
Probability Rules

Rule 2: Probability of complement of an event A is $P(A^c)=1-P(A)$

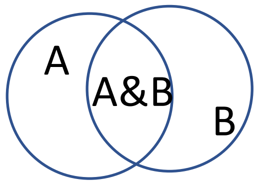
For an event A, $P(A^c)+ P(A) = 1$

Therefore, $P(A^c)=1-P(A)$

Rule 3: If events A and B are mutually exclusive, then P(A or B) = P(A)+P(B)



Rule 4: If two events A and B are not mutually exclusive, then P(A or B) = P(A) + P(B) – P(A and B).



Conditional Probability

$$P(A|B) = P(A \text{ and } B)/P(B)$$

Example: Conditional Probability in Marketing (Apple Purchase History file)

- Data shows the first and second purchases for a sample of 200 customers

- Probability of purchasing an iPad given already purchased an iMac = 2/13

	A	B
1	Apple Purchase History	
2		
3	First Purchase	Second Purchase
4	iPod	iMac
5	iPhone	MacBook
6	iMac	iPhone
7	iPhone	iPod
8	iPod	iPhone
9	MacBook	iPod
10	iPhone	MacBook

Count of Second Purchase	Colu					
Row Labels	iMac	iPad	iPhone	iPod	MacBook	Grand Total
iMac		2	3	2	6	13
iPad	1		1	2	10	14
iPhone	3	4		14	21	42
iPod	3	12	12		30	57
MacBook	8	16	26	24		74
Grand Total	15	34	42	42	67	200

Random Variables

- random variable: numerical description of an experimental outcome
  - discrete random variable: number of possible outcomes can be counted
    - e.g. outcomes of dice rolls, whether a customer likes or dislikes a product, number of hits on a Web site link today
  - continuous random variable: outcomes over one or more continuous intervals of real numbers
    - e.g. weekly change in DJIA (Dow Jones Industrial Average), daily temperature, time between machine failures

Probability distribution

- Probability distribution:
  - a characterization of possible values that a random variable may assume along with the probability of assuming these values
- Discrete vs continuous probability distribution
  - depending on nature of random variable it models
- 3 perspectives for developing
  - theoretical arguments
  - empirical data – empirical probability distribution
  - using subjective values and expert judgement

Discrete Probability Distributions

Probability Mass Function (Discrete Random Variables)

~ mathematical function *f(x)* specifying the probability of the random variable *X*.

x_i represents the i th value of *X*.

Properties:

$$0 \leq f(x_i) \leq 1 \qquad \text{for all } i$$

$$\sum_i f(x_i) = 1$$

Cumulative distribution function (CDF): $F(x) = P(X \leq x)$

Discrete Probability Distributions

<u>Example: Probability Mass Function for Rolling Two Dice</u>

$f(x_2) = 1/36$

$f(x_3) = 2/36$

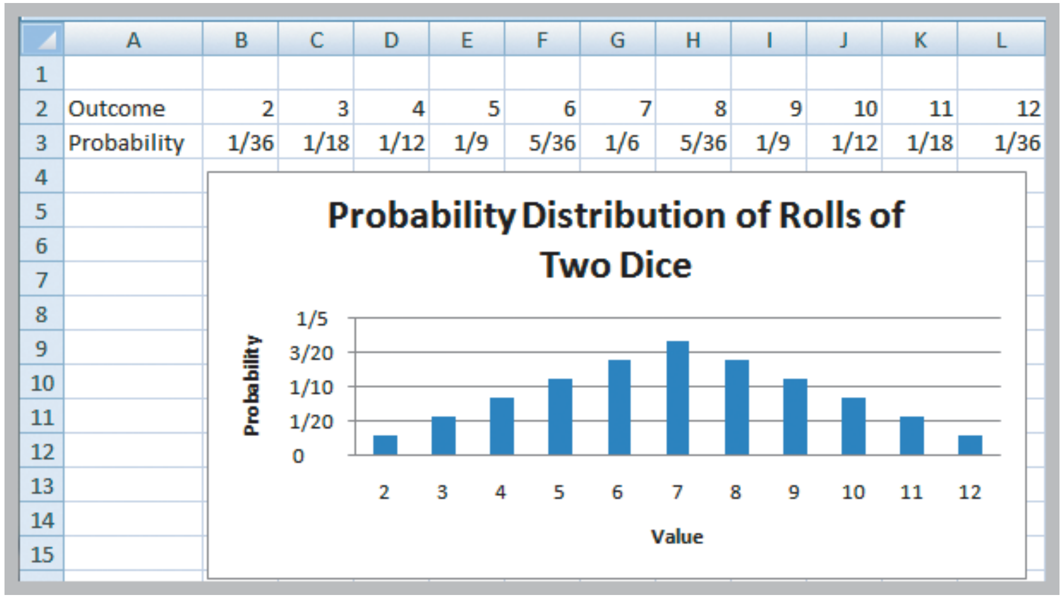
$f(x_4) = 3/36$

$f(x_5) = 4/36$

$f(x_6) = 5/36$

:

$f(x_{12}) = 1/36$



	A	B	C	D	E	F	G	H	I	J	K	L
1												
2	Outcome	2	3	4	5	6	7	8	9	10	11	12
3	Probability	1/36	1/18	1/12	1/9	5/36	1/6	5/36	1/9	1/12	1/18	1/36

Discrete Probability Distributions

Example: Using the Cumulative Distribution Function

- Probability of rolling between 4 and 8:

$= P(4 \leq X \leq 8)$

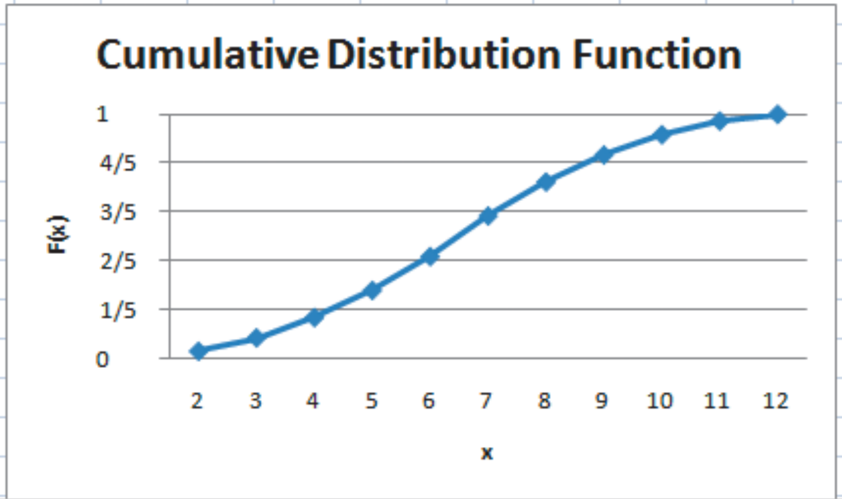
$= P(3 < X \leq 8)$

$= F(x_8) - F(x_3)$

$=13/18 - 1/12$

$= 23/36$

	A	B	C	D	E	F	G	H	I	J	K	L
18	Outcome	2	3	4	5	6	7	8	9	10	11	12
19	Probability	1/36	1/18	1/12	1/9	5/36	1/6	5/36	1/9	1/12	1/18	1/36
20	Cumulative Probability	1/36	1/12	1/6	5/18	5/12	7/12	13/18	5/6	11/12	35/36	1



Continuous Probability Distributions

Change in DJIA	Subjective Probability
-20%	0.01
-15%	0.05
-10%	0.08
-5%	0.15
0%	0.2
5%	0.25
10%	0.18
15%	0.06
20%	0.02

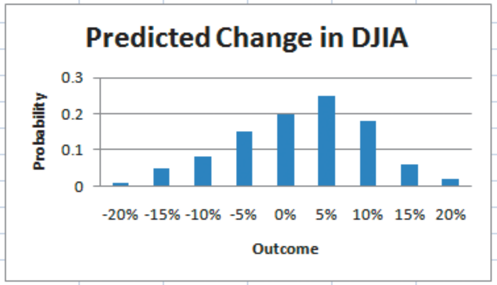
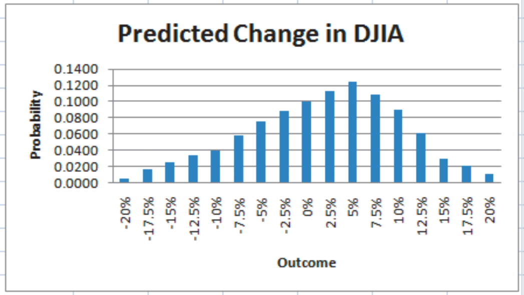


Figure 5.6

Change in DJIA using 5% increments

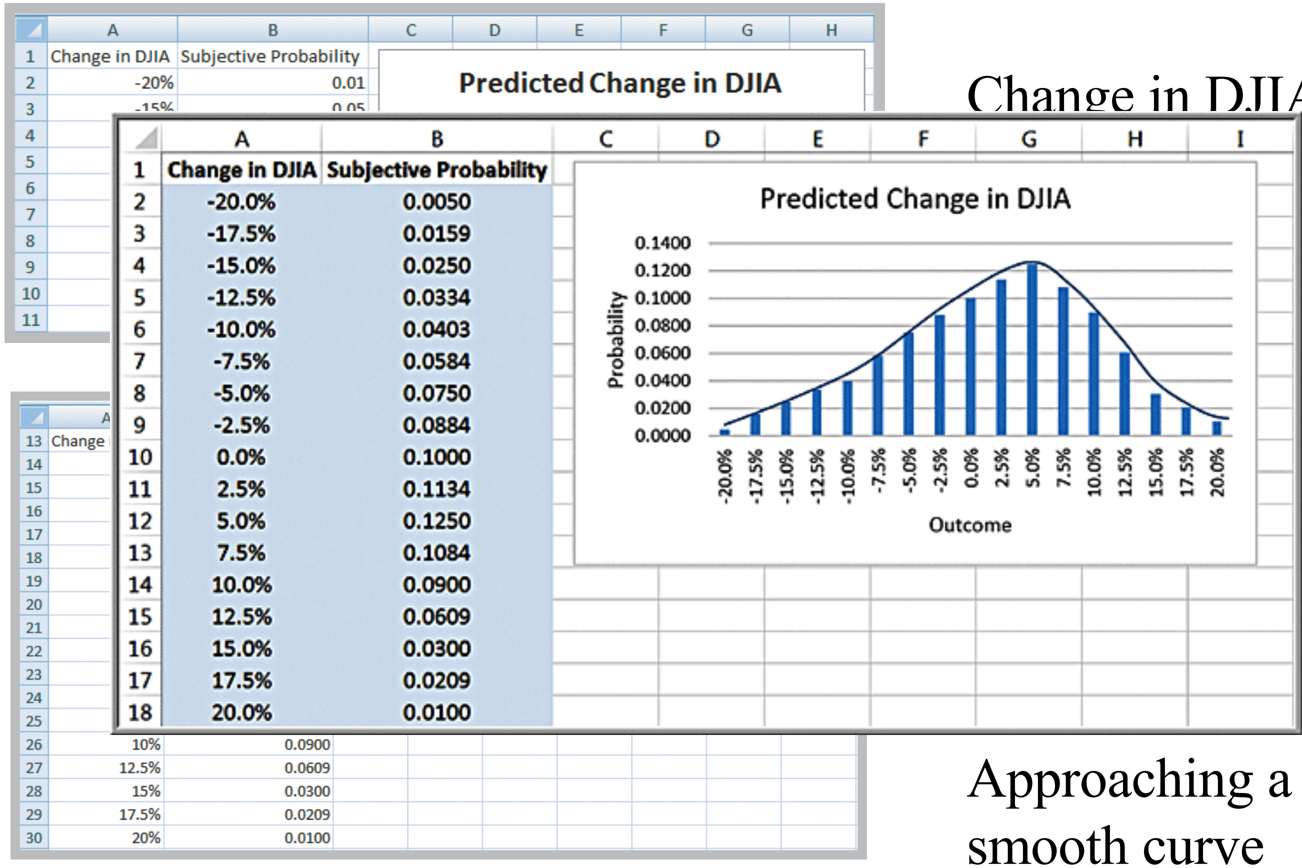
Change in DJIA	Subjective Probability
-20%	0.0050
-17.5%	0.0159
-15%	0.0250
-12.5%	0.0334
-10%	0.0400
-7.5%	0.0584
-5%	0.0750
-2.5%	0.0884
0%	0.1000
2.5%	0.1134
5%	0.1250
7.5%	0.1084
10%	0.0900
12.5%	0.0609
15%	0.0300
17.5%	0.0209
20%	0.0100



Change in DJIA using 2.5% increments

Approaching a smooth curve

Continuous Probability Distributions



	Change in DJIA	Subjective Probability
2	-20.0%	0.0050
3	-17.5%	0.0159
4	-15.0%	0.0250
5	-12.5%	0.0334
6	-10.0%	0.0403
7	-7.5%	0.0584
8	-5.0%	0.0750
9	-2.5%	0.0884
10	0.0%	0.1000
11	2.5%	0.1134
12	5.0%	0.1250
13	7.5%	0.1084
14	10.0%	0.0900
15	12.5%	0.0609
16	15.0%	0.0300
17	17.5%	0.0209
18	20.0%	0.0100

Change in DJIA

A

ts

Approaching a smooth curve

Continuous Probability Distributions

Probability density function

- A curve described by a mathematical function f(x) that characterizes a continuous random variable

Properties of a probability density function

- $f(x) \geq 0$ for all values of x
- Total area under the density function equals 1
- $P(X = x) = 0$
- Probabilities are only defined over an interval (eg. $P(a \leq X \leq b)$, $P(X<c)$, $P(X>a)$)
- $P(a \leq X \leq b)$ is the area under the density function between *a* and *b*

$$P(a \leq X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a)$$

Commonly Used Distributions

- Discrete Random Variables
  - Bernoulli Distribution
  - Binomial Distribution
  - Poisson Distribution

    …
- Continuous Random Variables
  - Uniform Distribution
  - Normal Distribution
  - (Student's) *t*-distribution
  - Exponential Distribution

    …

Why do we need to know about these distributions?

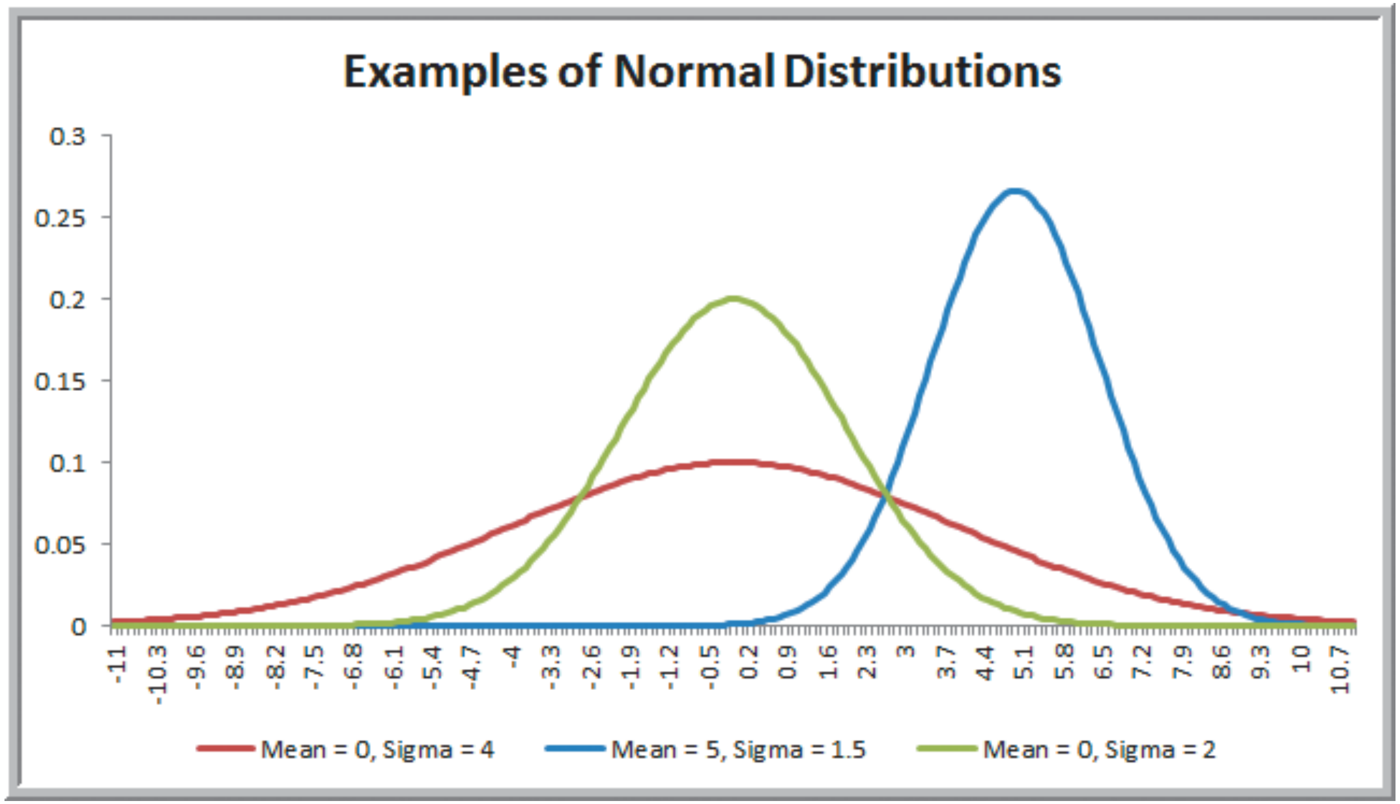
Working knowledge of common families of probability distributions:

1) helps you to understand underlying process that generates sample data
2) useful in building decision models with theoretical distribution of data
3) helps to compute probabilities of occurrence of outcomes to assess risk and make decisions

Normal Distribution

- f(x) is a bell-shaped curve
- Characterized by 2 parameters
  - μ (mean)
  - σ (standard deviation)
- Properties
  1. Symmetric
  2. Mean = Median = Mode
  3. Range of X is unbounded
  4. Empirical rules apply (i.e., the area under the density function within $\pm$ 2 standard deviation is 95.4%, and that within $\pm$3 standard deviation is 99.7%)

Normal Distributions



Normal Distributions

Example

The distribution for customer demand (units per month) is normal with:

mean = 750

stdev = 100

Find the probability that demand will be:

a) at most 900 units/month

b) exceed 700 units/month

c) be between 700 and 900 units/month

Normal Distributions

Probability of Demand be at most 900 units/month

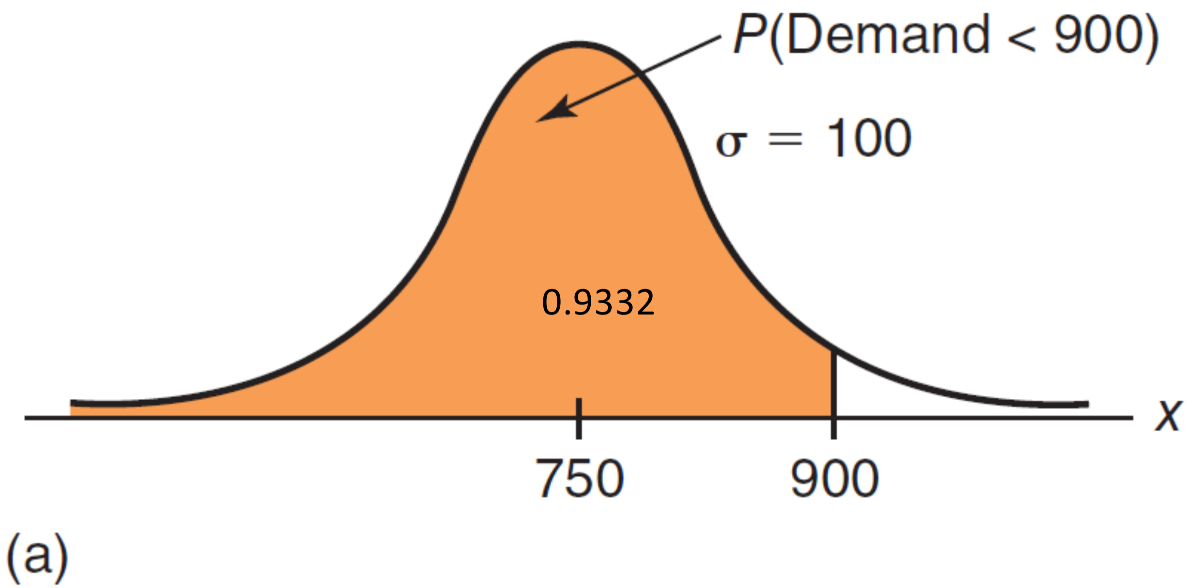
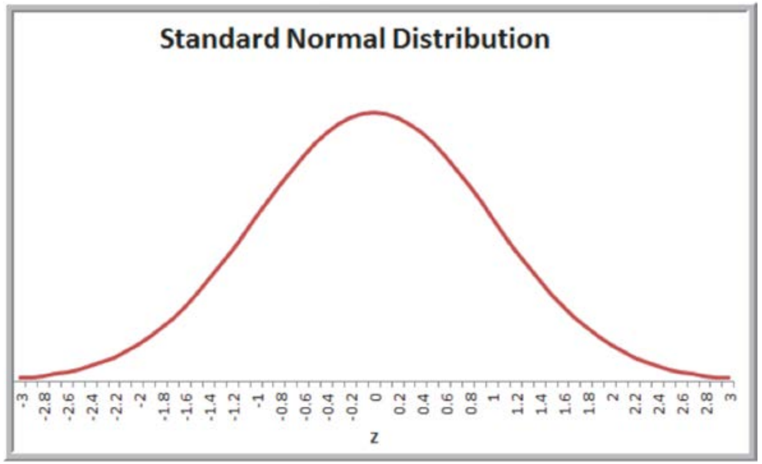


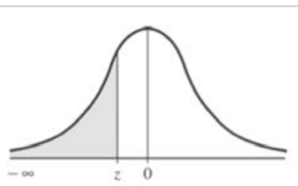
Figure 5.20a

Standard Normal Distribution

- A standard normal distribution is a normal distribution with a mean of 0 and standard deviation of 1
  - Standard normal random variable is denoted by *Z*
  - Scale along the z-axis represents the number of standard deviations from the mean of zero



Using Standard Normal Distribution Tables



- Table 1 of Appendix A

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.9	.00005	.00005	.00004	.00004	.00004	.00004	.00004	.00004	.00003	.00003
−3.8	.00007	.00007	.00007	.00006	.00006	.00006	.00006	.00005	.00005	.00005
−3.7	.00011	.00010	.00010	.00010	.00009	.00009	.00008	.00008	.00008	.00008
−3.6	.00016	.00015	.00015	.00014	.00014	.00013	.00013	.00012	.00012	.00011
−3.5	.00023	.00022	.00022	.00021	.00020	.00019	.00019	.00018	.00017	.00017
−3.4	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024
−3.3	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035

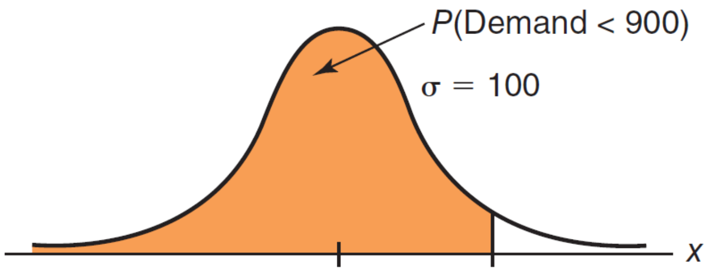
- We may compute probabilities for any normal random variable *X* having a mean μ and standard deviation σ by converting it to a standard normal random variable *Z*:

$$z = \frac{(x - \mu)}{\sigma}$$

Computing Probabilities with Standard Normal Tables

What is the probability that demand will be at most 900 units/month?

- $z = (900 - 750)/100 = 1.50$
- Using Table 1 in Appendix A, we find:
  - $P(X < 900) = P(Z < 1.50) = 0.9332$



(a)

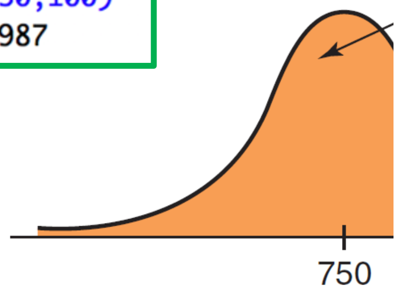
z	.00
0.0	.5000
0.1	.5398
0.2	.5793
0.3	.6179
0.4	.6554
0.5	.6915
0.6	.7257
0.7	.7580
0.8	.7881
0.9	.8159
1.0	.8413
1.1	.8643
1.2	.8849
1.3	.9032
1.4	.9192
1.5	.9332

Computing Probabilities with Standard Normal Tables

- what is the probability that demand will be at most 900 units/month?
- $z = (900 - 750)/100 = 1.50$
- Using Table 1 in Appendix A, w
  - $P(X < 900) = P(Z < 1.50) = 0.9332$

z	.00
0.0	.5000

```
> pnorm(900,750,100)
[1] 0.9331927987
```



(a)

The Normal Distribution

Description

Density, distribution function, quantile function and random generation for the normal distribution with mean equal to `mean` and standard deviation equal to `sd`.

Usage

```
dnorm(x, mean = 0, sd = 1, log = FALSE)
pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
rnorm(n, mean = 0, sd = 1)
```

Arguments

`x, q`	vector of quantiles.
`p`	vector of probabilities.
`n`	number of observations. If `length(n) > 1`, the length is taken to be the number required.
`mean`	vector of means.
`sd`	vector of standard deviations.
`log, log.p`	logical; if TRUE, probabilities p are given as log(p).
`lower.tail`	logical; if TRUE (default), probabilities are $P[X \le x]$ otherwise, $P[X > x]$.

Calculating probability for question b

- Probability that demand will exceed 700 units, or P(X > 700).

= 1 – P(X<=700) = 1 – P(z<=-0.5) = 1-0.3085 = 0.6915

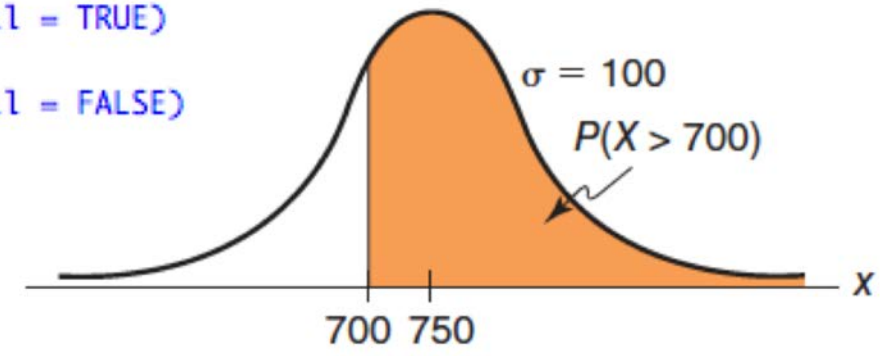
OR

= 1-pnorm(700,750,100) = 1 - 0.3085 = 0.6915

z	.00
−1.4	.0808
−1.3	.0968
−1.2	.1151
−1.1	.1357
−1.0	.1587
−0.9	.1841
−0.8	.2119
−0.7	.2420
−0.6	.2743
−0.5	.3085
−0.4	.3446
−0.3	.3821
−0.2	.4207
−0.1	.4602
−0.0	.5000

```
> 1-pnorm(700,750,100)
[1] 0.6914624613
```

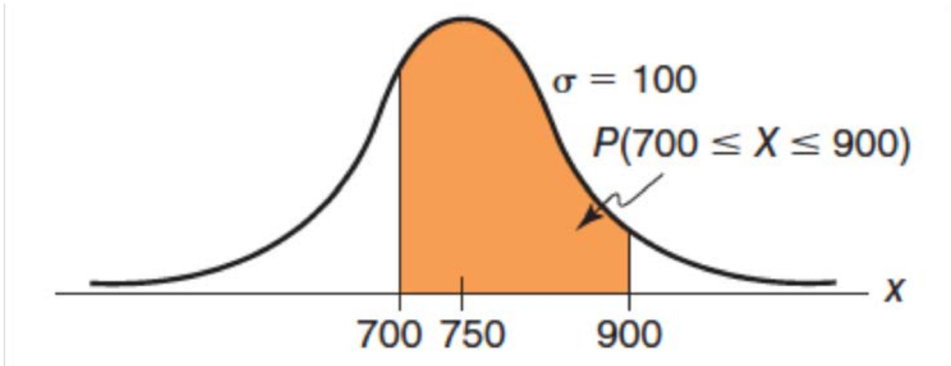
```
> pnorm(700,750,100,lower.tail = TRUE)
[1] 0.3085375
> pnorm(700,750,100,lower.tail = FALSE)
[1] 0.6914625
```



Calculating probability for question c

- Probability that demand will be between 700 and 900, or P(700 <= *X* <= 900):
  - =pnorm (900,750,100) - pnorm(700,750,100) =0.9332 - 0.3085 = 0.6247

```
> pnorm (900,750,100) - pnorm(700,750,100)
[1] 0.62465526
```



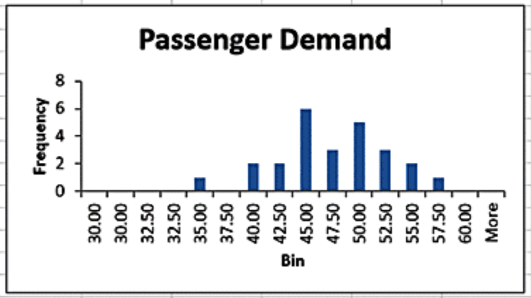
Data Modeling and Distribution Fitting

- sample data limits our ability to predict uncertain events
  - potential values outside the range of the sample data are not included.
- better to identify the underlying probability distribution from which sample data come by "fitting" a theoretical distribution to the data and verifying the goodness of fit statistically
  - Examine a histogram for clues about the distribution's shape
  - Look at summary statistics such as the mean, median, standard deviation, coefficient of variation, and skewness

Analyzing Airline Passenger Data – example

- Sample data on passenger demand for 25 flights

	A	C	D	F	G
1	**Airline Passengers**				
2					
3	**Sample Data**	*Bin*	*Frequency*		
4	36	30.00	0	*Passengers*	
5	55	30.00	0		
6	47	32.50	0	Mean	45.68
7	45	32.50	0	Standard Error	1.043584
8	48	35.00	1	Median	45
9	43	37.50	0	Mode	45
10	42	40.00	2	Standard Deviation	5.217918
11	56	42.50	2	Sample Variance	27.22667
12	40	45.00	6	Kurtosis	0.707219
13	47	47.50	3	Skewness	0.823163
14	44	50.00	5	Range	22
15	46	52.50	3	Minimum	36
16	53	55.00	2	Maximum	58
17	45	57.50	1	Sum	1142
18	44	60.00	0	Count	25
19	45	More	0		
20	45				
21	41				
22	47				
23	46				
24	40				
25	46				
26	42				
27	41				
28	58				



- histogram shows relatively symmetric distribution. The mean, median, and mode are all similar, although there is moderate skewness. Normal distribution seems reasonable.

Goodness of Fit

- fitting data to a probability distribution statistically
- Statistical measures of goodness of fit:
  - Kolmogorov-Smirnov (works well for small samples and only for non-parametric data)
  - Chi-square (need at least 50 data points)
  - Anderson-Darling (puts more weight on the differences between the tails of the distributions)
  - Shapiro-Wilk Test (test data against normal distribution)

Test of normality

- Shapiro-Wilk test
  - R function shapiro.test(x)

```
> shapiro.test(Airline_Passengers$`Sample Data`)

	Shapiro-Wilk normality test

data:  Airline_Passengers$`Sample Data`
W = 0.92081032, p-value = 0.05345947
```

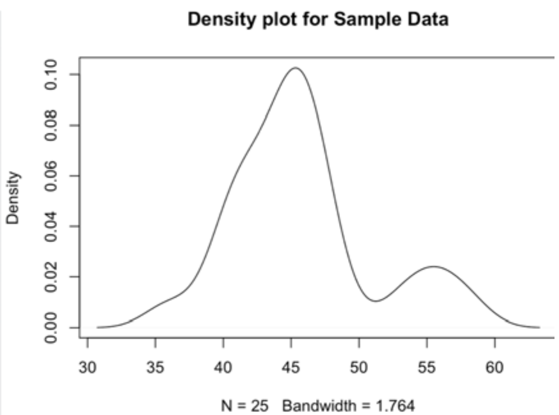
W close to 1; P-value > 0.05 implies that the distribution of the data is not significantly different from normal distribution. In other words, the data does not deviate from normality.

	Sample Data
1	36
2	55
3	47
4	45
5	48
6	43
7	42
8	56
9	40
10	47
11	44
12	46

Showing 1 to 13 of 25 entries

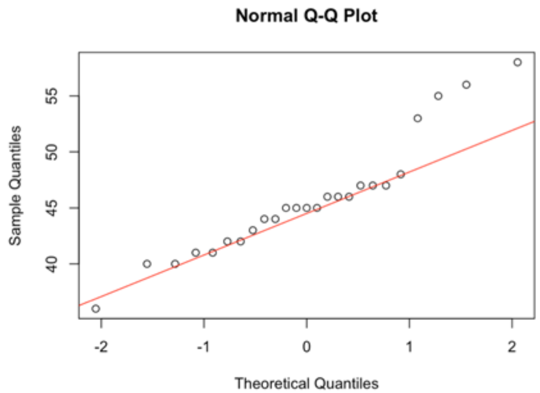
- **Graphical methods**

Density plot




```
plot(density(Airline_Passengers$`Sample Data`))
```

Q-Q plot




```
qqnorm(Airline_Passengers$`Sample Data`)
qqline(Airline_Passengers$`Sample Data`, col = 2)
```