L- 1: Overview of the course & Descriptive Statistics

“*When you can measure what you are speaking about and express in numbers, you know something about it ;but when you cannot measure it, cannot express it in numbers, your knowledge is of meagre and unsatisfactory kind”*

Lord Kelvin

“Statistical thinking will be one day as necessary for efficient citizenship as the ability to read and write”

H G Wells

Lies

damn lies

Statistics

Analytics

- The term “ Analytics”

- Disciplines
  - Statistics
  - Machine Learning
  - Biology
  - Kernel Methods

Importance of Data

- Importance : For any analytical exercise Data are key ingredients
  Replace intuition with data driven decisions

- For example consider the following cases:

- Medical treatment
- Industry
- Power generation
- Crime detection
- Cognitive assessment

Model _requirements

- Business relevance
- Statistical performance
- Interpretable
- Justifiability
- Operational efficiency
- Economic cost
- Regulation and legislation

Statistics?

Procedures for organising, summarizing, and interpreting information

- ✓ Standardized techniques used by scientists
- ✓ Vocabulary & symbols for communicating about data

Two main branches:

- *Descriptive statistics*
- *Inferential statistics*

Basic tools / concepts in analysis

- Mean
- Median
- Mode
- Range
- Variance / Standard deviation
- Coefficient of variation
- Mean Deviation

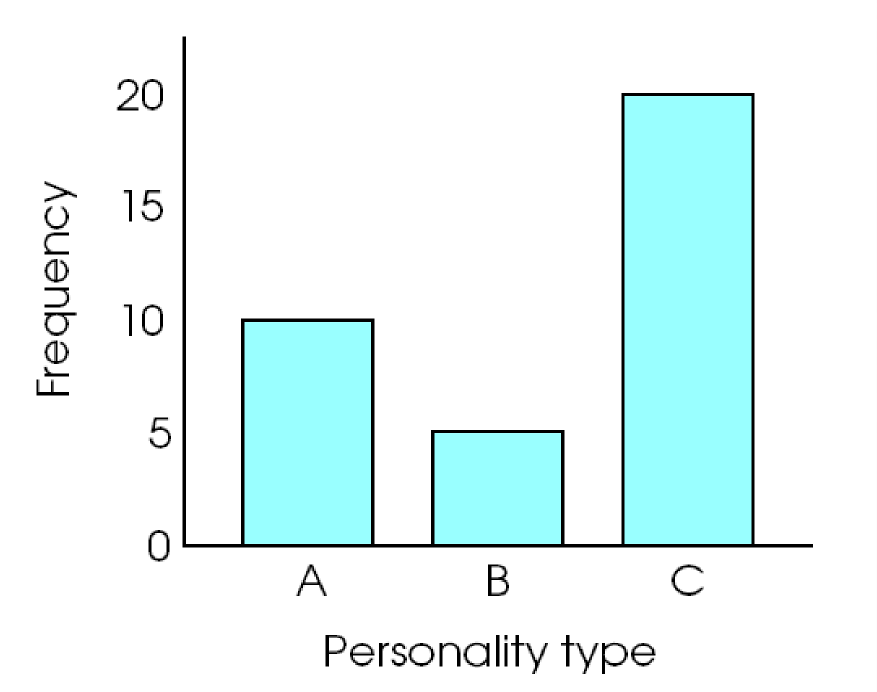
Statistical graphs of data

A picture is worth a thousand words!

- Bar chart / graph
- Histograms
- Pie chart
- Pareto chart / diagram
- Frequency polygons
- Scatter plots
- Time series plot

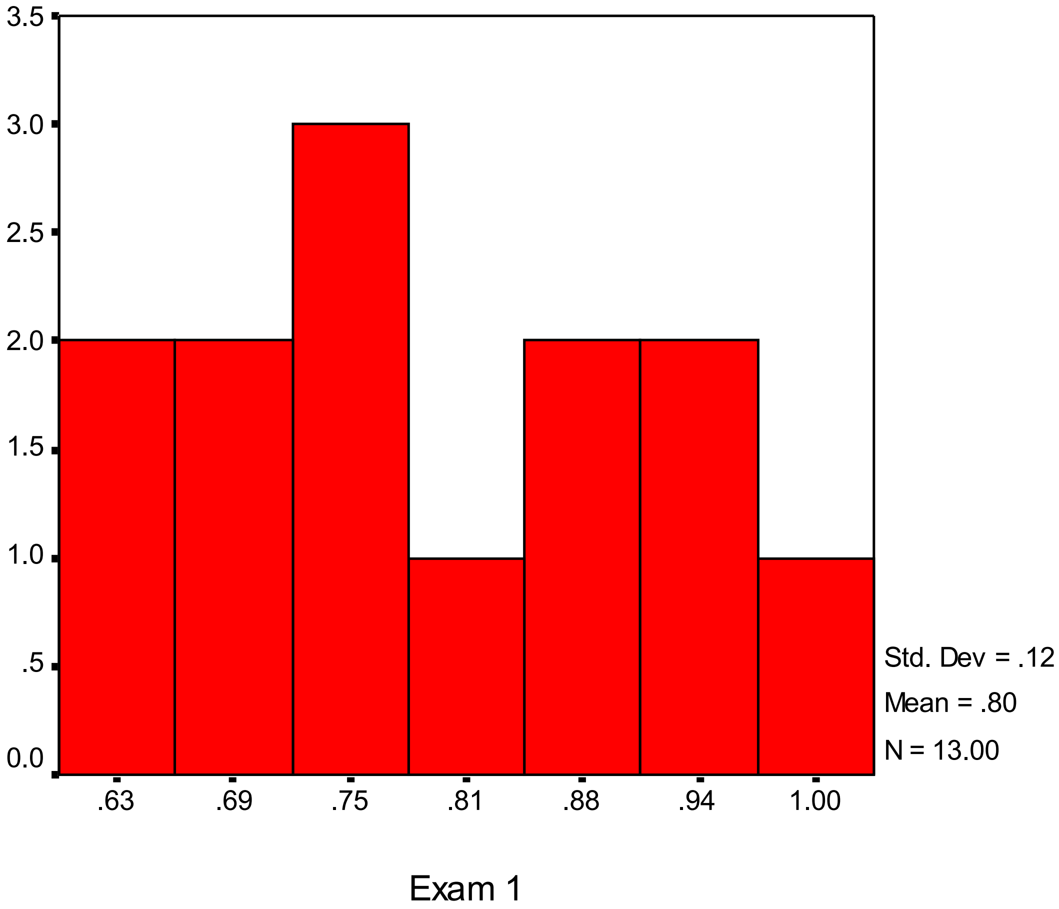
Bar Graphs

Useful for showing two samples side-by-side

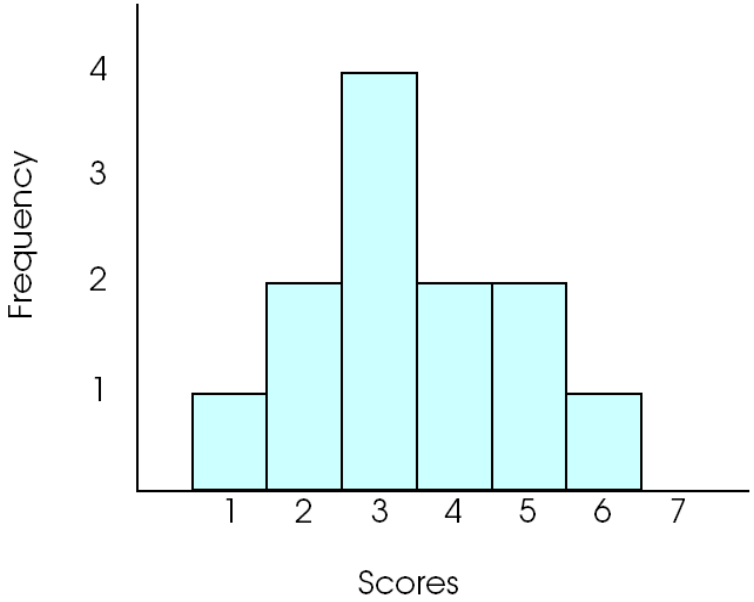


Histograms

- Univariate histograms



Histograms



X	*f*
6	1
5	2
4	2
3	4
2	2
1	1

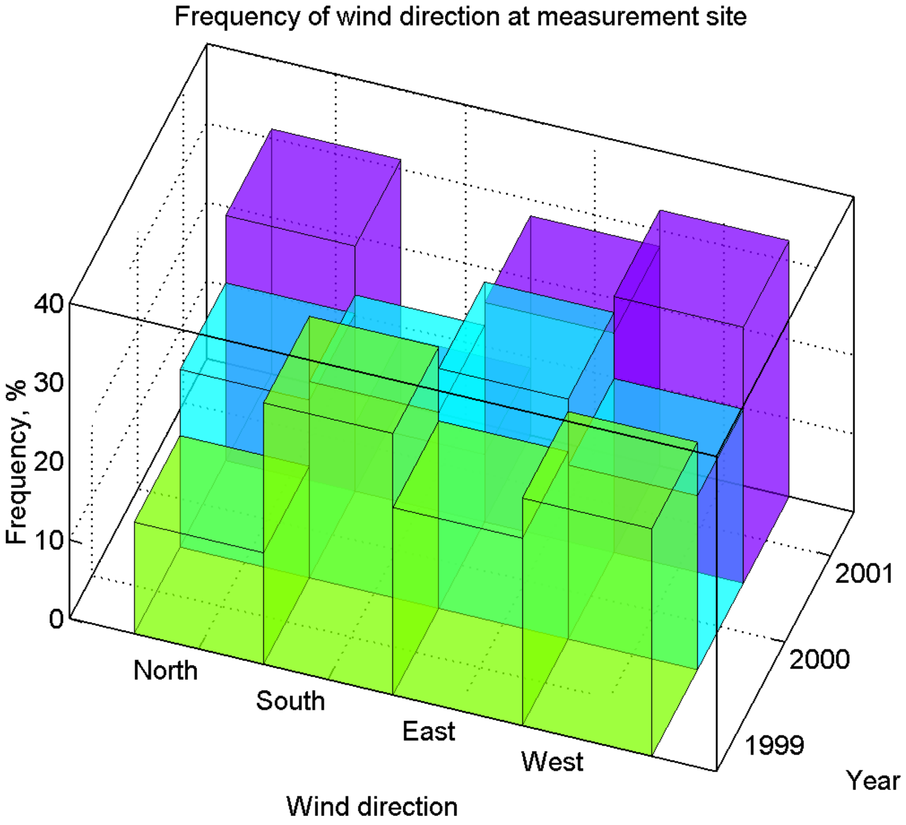
f on *y* axis (could also plot *p* or %)

X values (or midpoints of class intervals) on *x* axis

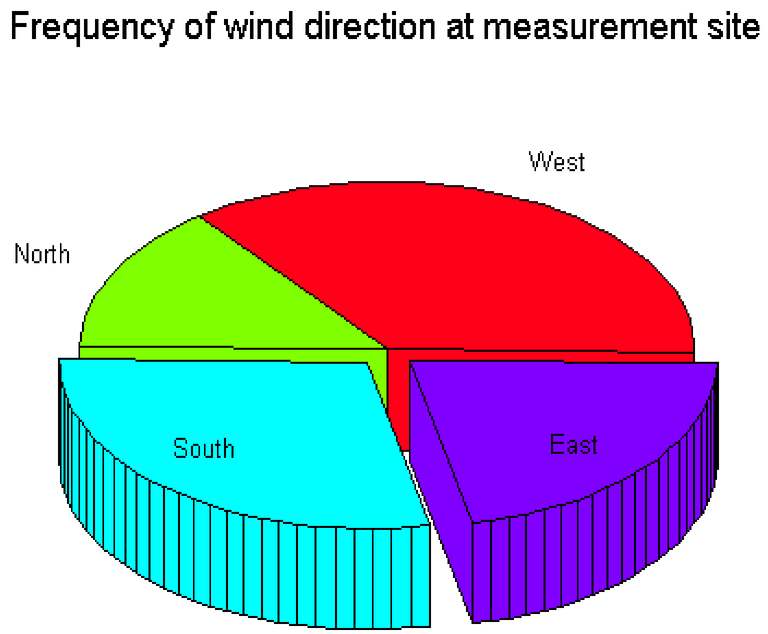
Plot each *f* with a bar, equal size, touching

No gaps between bars

Bivariate histogram

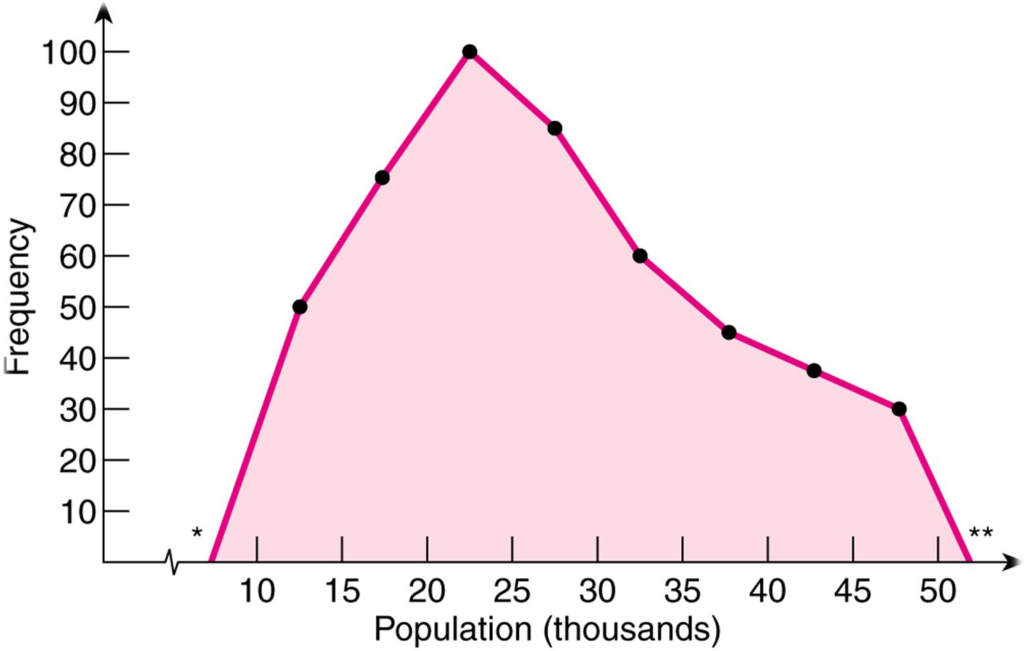


Graphing the data – Pie charts



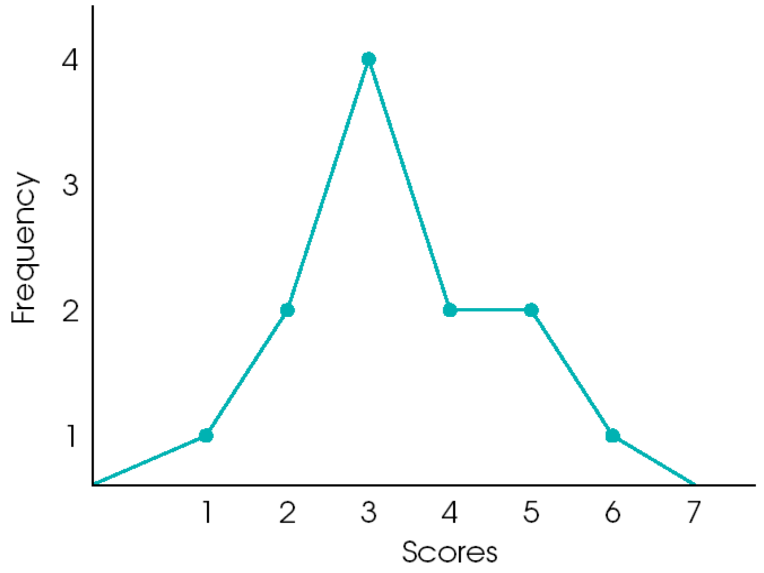
Frequency Polygons

- Frequency Polygons
  - Depicts information from a frequency table or a grouped frequency table as a **line graph**



* 4 cities had populations of less than 10,000.
** 5 cities had populations of 50,000 or greater.

Frequency Polygon



X	*f*
6	1
5	2
4	2
3	4
2	2
1	1

A smoothed out histogram

Make a point representing *f* of each value

Connect dots

Anchor line on x axis

Useful for comparing distributions in two samples (in this case, plot *p* rather than *f*)

!!!!

A famous statistician would never travel by airplane, because she had studied air travel and estimated the probability of there being a bomb on any given flight was 1 in a million, and she was not prepared to accept these odds.

One day a colleague met her at a conference far from home.

"How did you get here, by train?"

"No, I flew"

"What about the possibility of a bomb?"

"Well, I began thinking that if the odds of one bomb are 1:million, then the odds of TWO bombs are (1/1,000,000) x (1/1,000,000) = 10^{-12}. This is a very, very small probability, which I can accept. So, now I bring my own bomb along!"

Random Experiment

Term "**random experiment**" is used to describe any action whose outcome is not known in advance. Here are some examples of experiments dealing with statistical data:

- Tossing a coin
- Counting how many times a certain word or a combination of words appears in the text of the “King Lear” or in a text of Confucius
- counting occurrences of a certain combination of amino acids in a protein database.
- pulling a card from the deck

Sample spaces, sample sets and events

The ***sample space*** of a random experiment is a set S that includes all possible outcomes of the experiment.

For example, if the experiment is to throw a die and record the outcome, the sample space is S = { 1,2,3,4,5,6}

- Discrete sample spaces.
- Continuous sample spaces

The set of possible outcomes S describes an event that always occurs.

Each outcome is represented by a **sample point** in the sample space.

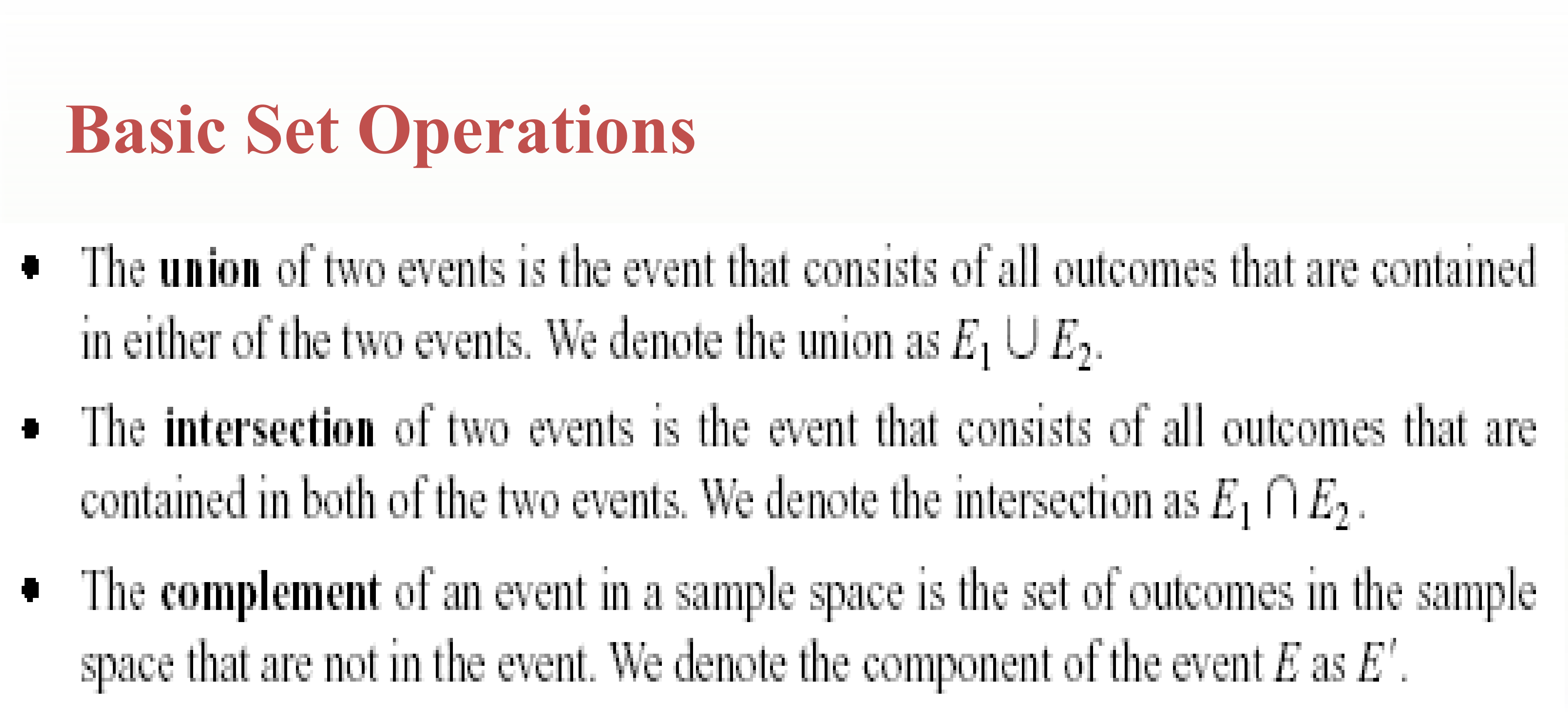
There is more than one way to view and experiment, so an experiment may have more than one associated sample space.

In tossing a die, one sample space is {1,2,3,4,5,6}, while two others are {odd, even} and {less then 3.5, more then 3.5}

Events

An **event** is a subset of the sample space of a random experiment.

An event is a set of outcomes of the experiment. This includes the *null* (empty) set of outcomes and the set of *all* outcomes. Each time the experiment is run, a given event *A* either *occurs*, if the outcome of the experiment is an element of *A*, or *does not occur*, if the outcome of the experiment is not an element of *A*.



Basic Set Operations

- The **union** of two events is the event that consists of all outcomes that are contained in either of the two events. We denote the union as $E_1 \cup E_2$.
- The **intersection** of two events is the event that consists of all outcomes that are contained in both of the two events. We denote the intersection as $E_1 \cap E_2$.
- The **complement** of an event in a sample space is the set of outcomes in the sample space that are not in the event. We denote the component of the event E as E'.

Mutually Exclusive Events

Two events are mutually exclusive if they can not occur at the same time. Which are mutually exclusive?

- Draw an Ace and draw a heart from a standard deck of 52 cards
- It is raining and I show up for class
- Dr. Li is an easy teacher and I fail the class
- Dr. Beaubouef is a hard teacher and I ace the class.

Independent & Dependent

Events are either

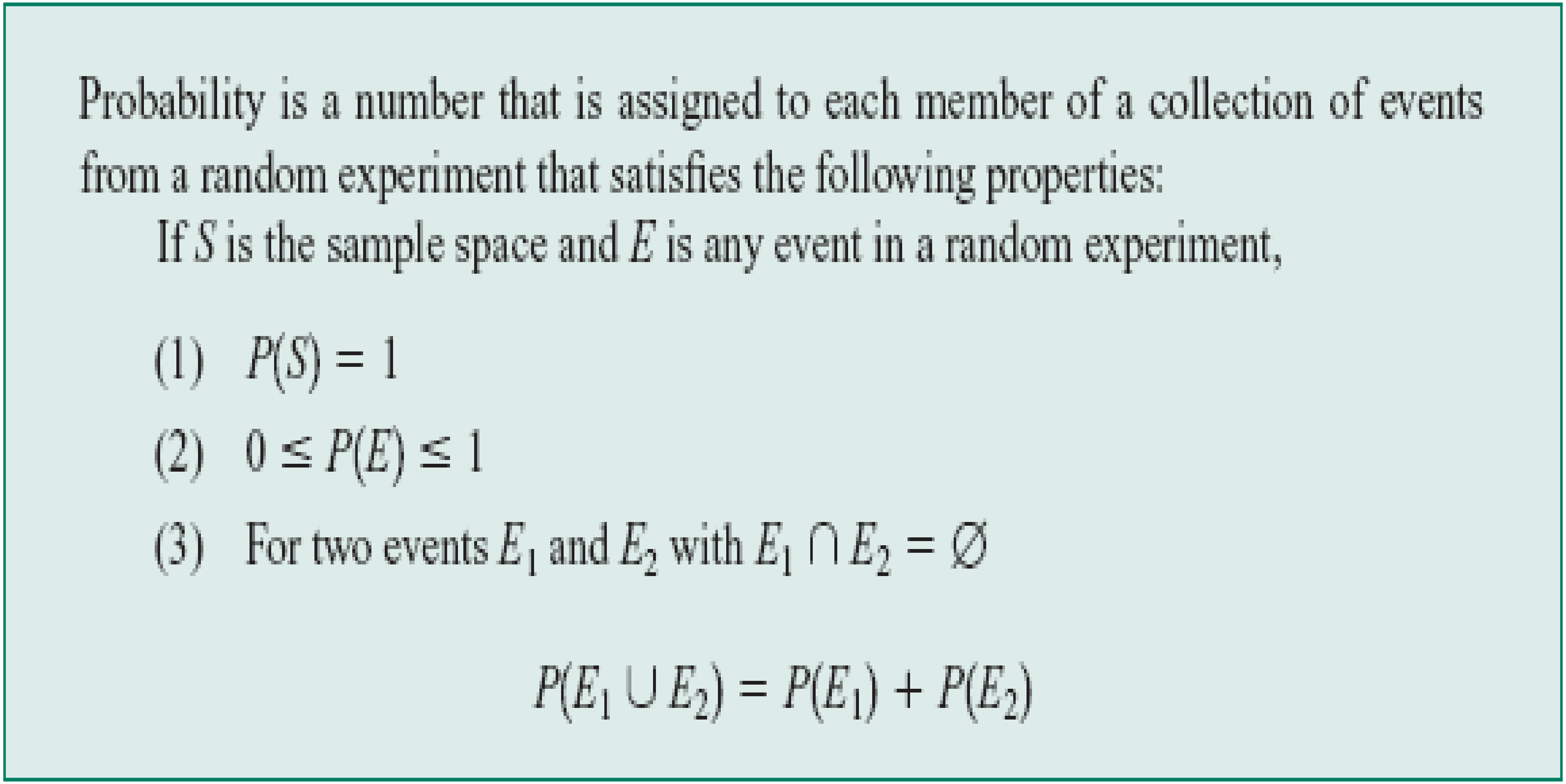
- ✓*independent* (the occurrence of one event has no effect on the probability of occurrence of the other) or
- ✓*dependent* (the occurrence of one event gives information about the occurrence of the other)

Random experiment

Consider the random experiment of dropping a Styrofoam cup onto the floor from a height of four feet. The cup hits the ground and eventually comes to rest. It could land upside down, right side up, or it could land on its side. We represent these possible outcomes of the random experiment by the following.

Probability

Axioms of Probability



Probability is a number that is assigned to each member of a collection of events from a random experiment that satisfies the following properties:

If S is the sample space and E is any event in a random experiment,

(1) $P(S) = 1$

(2) $0 \leq P(E) \leq 1$

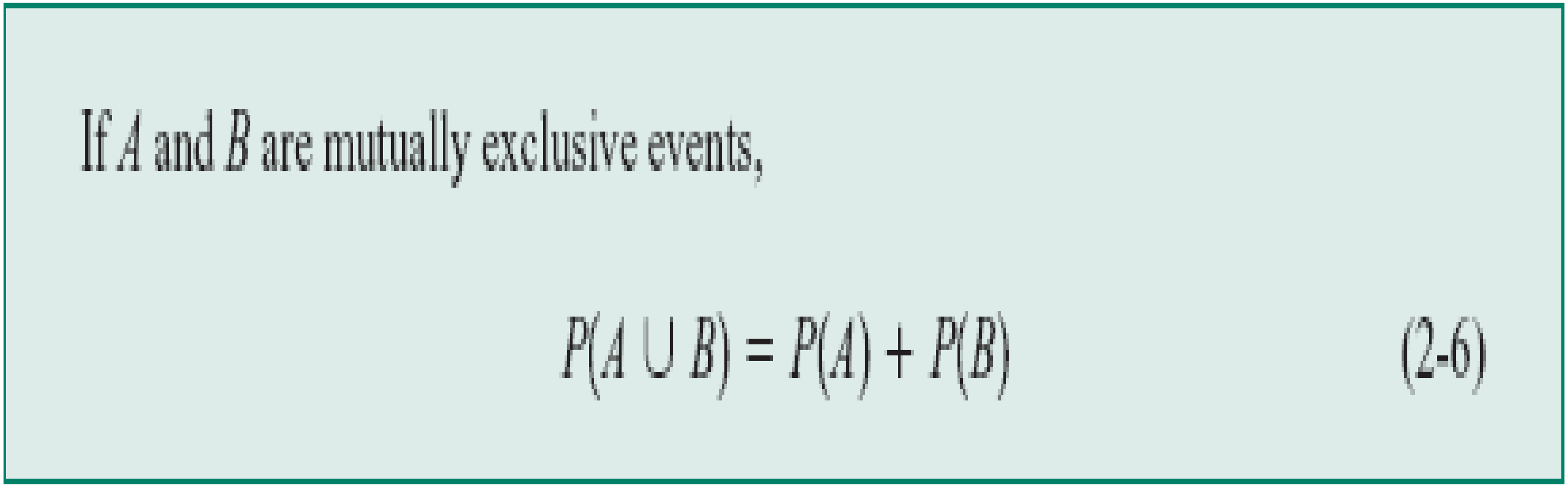
(3) For two events E_1 and E_2 with $E_1 \cap E_2 = \varnothing$

$$P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

Probability of a Union

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \qquad (2\text{-}5)$$

Mutually Exclusive Events



If *A* and *B* are mutually exclusive events,

$$P(A \cup B) = P(A) + P(B) \qquad (2\text{-}6)$$

Three Events

$$\begin{aligned} P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) \\ - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C) \quad (2\text{-}7) \end{aligned}$$

1

The sales manager of an e commerce company says that 80% of those who visit their website for the first time do not buy any mobile. If a new customer visits the website, what is the probability that the customer would buy mobile

2

	Blue	Black	Brown	Total
Software prog	35	25	20	80
Project Mgrs	7	8	5	20
Total	42	33	25	100

If an employee is selected at random , what is the probability that he is a software prog?

………………………………………………..,what is the probability that he is wearing a blue trouser

3

A Survey conducted by a bank revealed that 40% of the accounts are savings accounts and 35% of the accounts are current accounts and the balance are loan accounts.

- ➢ What is the probability that an account taken at random is a loan account ?
- ➢ What is the probability that an account taken at random is NOT savings account ?
- ➢ What is the probability that an account taken at random is NOT a current account
- ➢ What is the probability that an account taken at random is a current account or a loan account?

4

From a Hospital data it is found that 45% of the patients are having high B.P. Also it was found that 35% of these patients having high B P is also having diabetes.

What is the probability that a patient having high BP is also diabetic

Conditional Probability

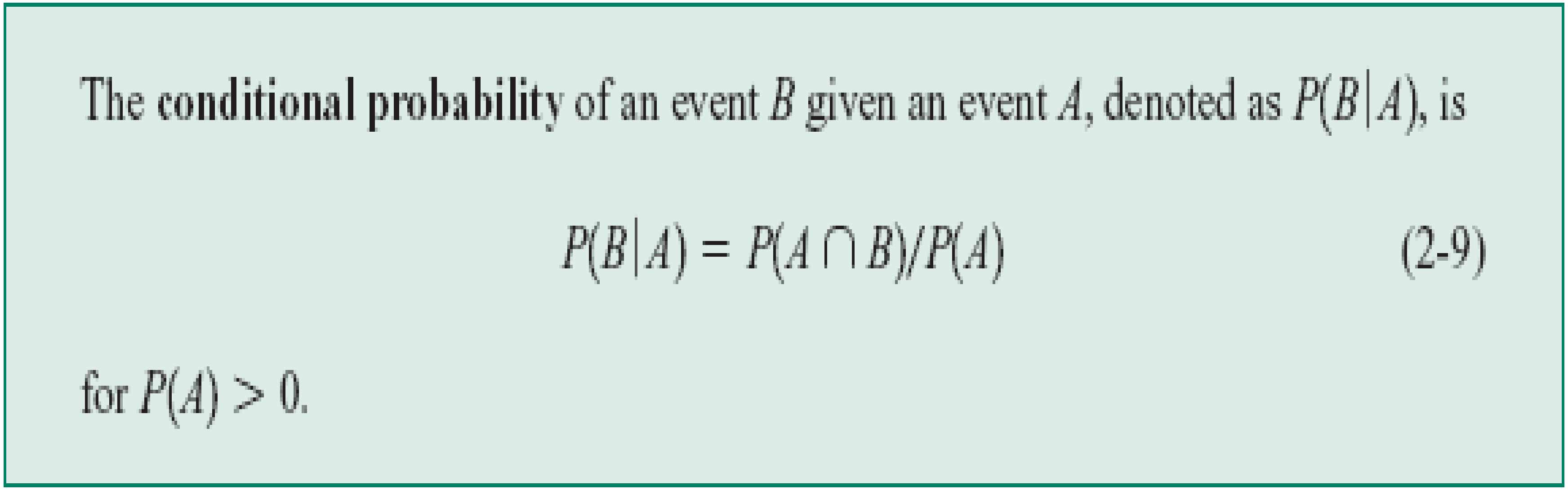
The probability of event B given that event A has occurred P(B|A) or, the probability of event A given that event B has occurred P(A|B)

5

	Actually purchased			
Planned to purchase	YES	NO	TOTAL	
YES	200	50	250	
NO	100	650	750	
TOTAL	300	700	1000	

Conditional Probability

Definition



The conditional probability of an event B given an event A, denoted as $P(B \mid A)$, is

$$P(B \mid A) = P(A \cap B)/P(A) \qquad (2\text{-}9)$$

for $P(A) > 0$.

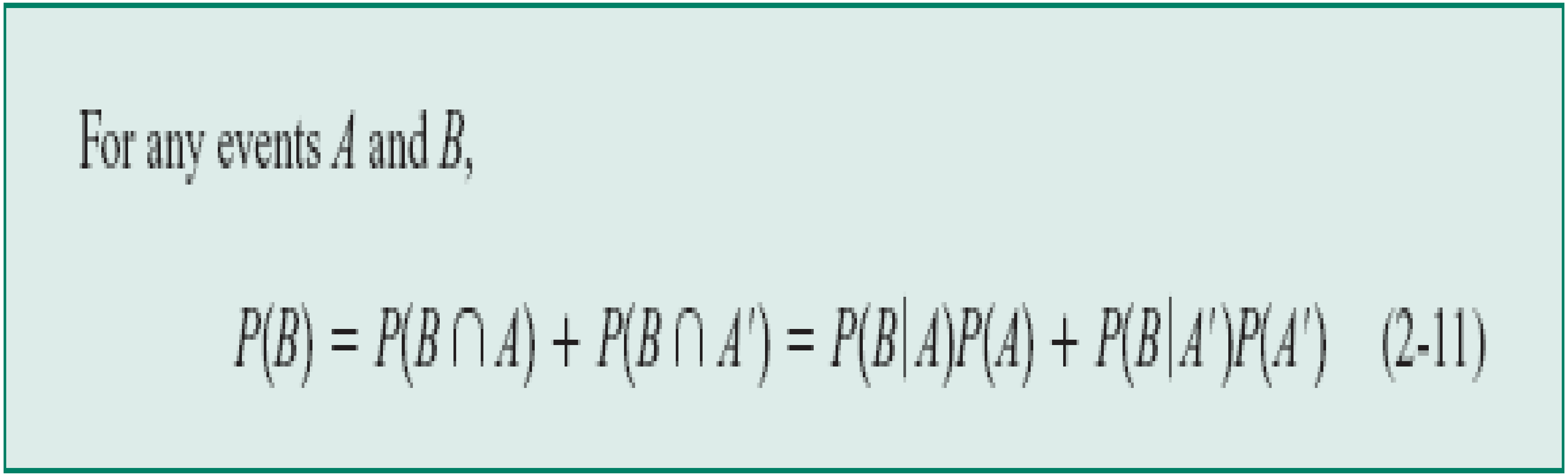
Multiplication and Total Probability Rules

Multiplication Rule

$$P(A \cap B) = P(B|A)P(A) = P(A|B)P(B) \qquad (2\text{-}10)$$

Multiplication and Total Probability Rules

Total Probability Rule (two events)

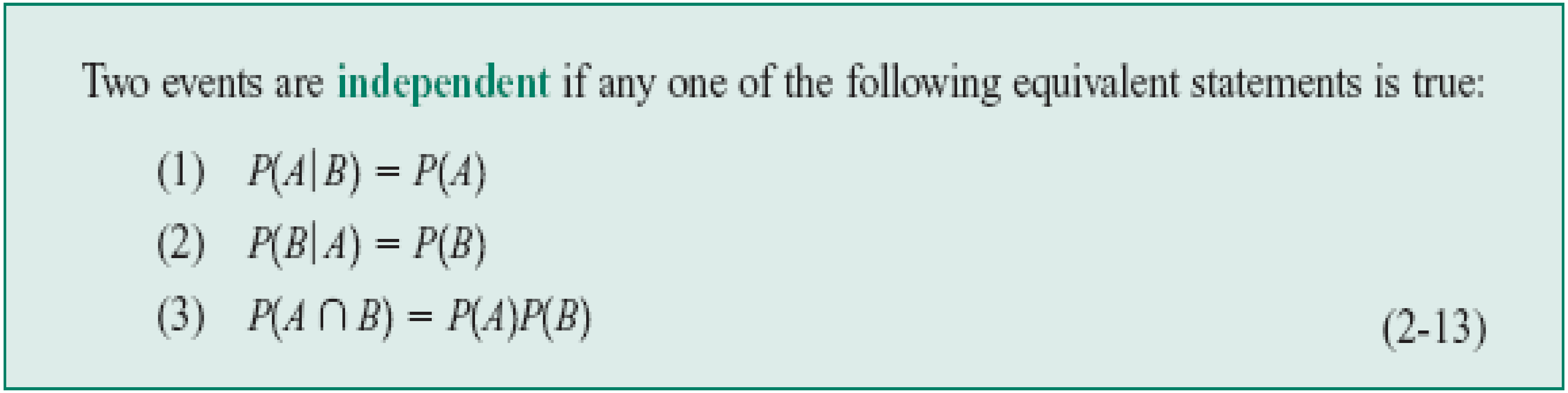


For any events A and B,

$$P(B) = P(B \cap A) + P(B \cap A') = P(B|A)P(A) + P(B|A')P(A') \quad (2\text{-}11)$$

Independence

Definition (two events)



Two events are **independent** if any one of the following equivalent statements is true:

(1) $P(A|B) = P(A)$

(2) $P(B|A) = P(B)$

(3) $P(A \cap B) = P(A)P(B)$ (2-13)

6

Toss a six-sided die twice. The sample space consists of all ordered pairs (i; j) of the numbers 1; 2; : : : ; 6, that is, S = {(1; 1); (1; 2); : : : ; (6; 6)}..

Let A = {outcomes match}

and B = {sum of outcomes at least 8}.

Then find P(A),P(B),P(A/B) and P(B/A)

7

Three persons A,B and C are competing for the post of CEO of a company. The chances of they becoming CEO are 0.2,0.3 and 0.4 respectively.

The chances of they taking employees beneficial decisions are 0.50,0.45 and 0.6 respectively

What are the chances of having employees beneficial decisions after having new CEO

Bayes' Theorem

Definition

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad \text{for} \quad P(B) > 0 \qquad (2\text{-}15)$$

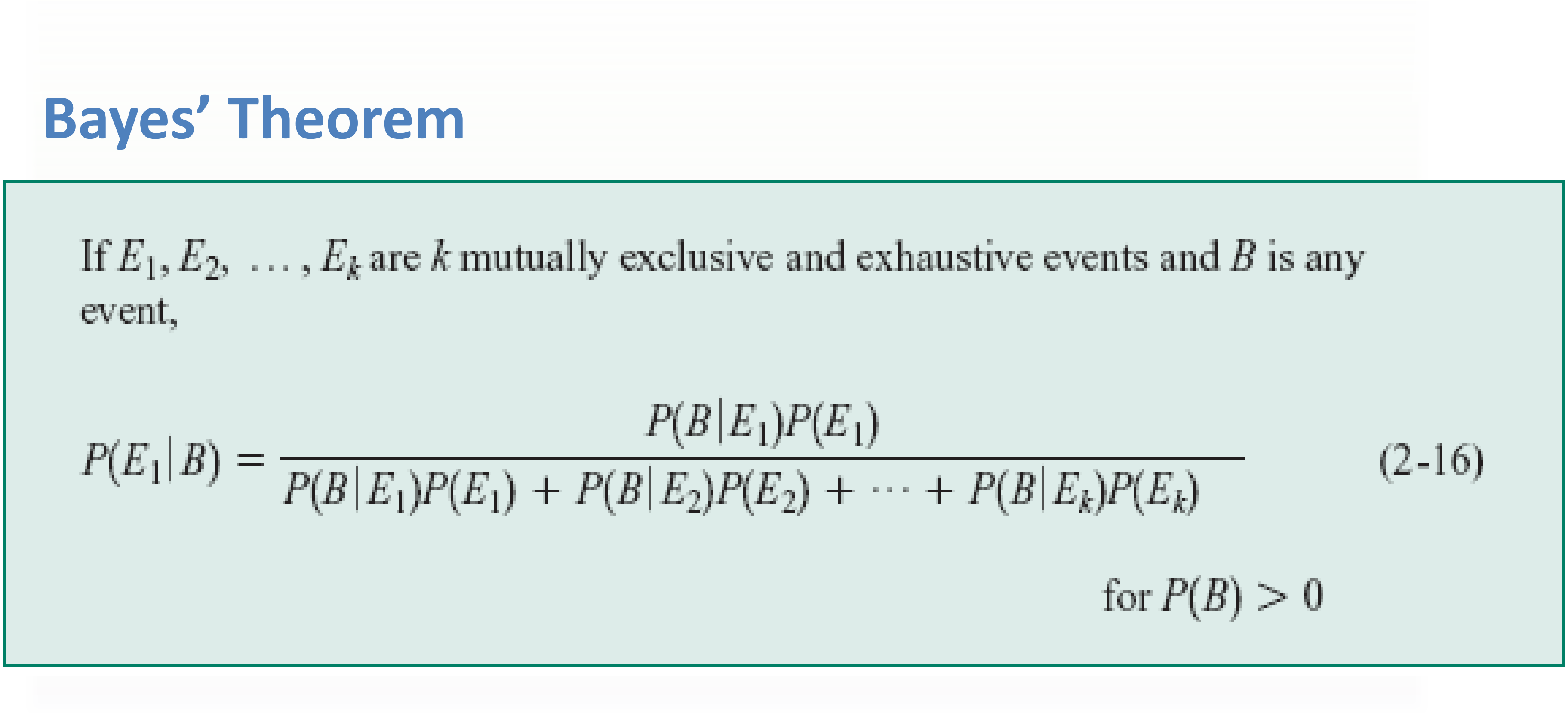
$$P(B) = P(E_1 \cap B) + P(E_2 \cap B) + \cdots + P(E_n \cap B)$$

For each

$$P(E_i \cap B) = P(B \mid E_i)P(E_i)$$

$$\begin{aligned} P(B) &= P(E_1 \cap B) + P(E_2 \cap B) + \cdots + P(E_n \cap B) \\ &= P(B \mid E_1)P(E_1) + P(B \mid E_2)P(E_2) + \cdots + P(B \mid E_n)P(E_n) \\ &= \sum_{i=1}^{n} P(B \mid E_i)P(E_i) \end{aligned}$$

Bayes’ Theorem



Bayes’ Theorem

If $E_1, E_2, \ldots, E_k$ are k mutually exclusive and exhaustive events and B is any event,

$$P(E_1|B) = \frac{P(B|E_1)P(E_1)}{P(B|E_1)P(E_1) + P(B|E_2)P(E_2) + \cdots + P(B|E_k)P(E_k)} \qquad (2\text{-}16)$$

$$\text{for } P(B) > 0$$

Applications

- Diagnostic tests in medicine
- Telecommunication
- Customer service
- Trouble shooting in engineering processes & systems

Example 1

A Component is tested for its stipulated quality , but the test is not infalliable. If the component is good,70% of the time , test gives positive indication i.e. 70% of the time the test classifies good item as good. If the component is defective,80% of the time , test gives negative indication implying that the component is bad. If in the manufacturing process, the percentage of defective components is 20,then find

- ➢ probability that the component is good and test gives positive indication
- ➢ …….the component is not good and test gives negative indication
- ➢ …….the component is good given that the test is positive

Example 2

Technicians regularly make repairs when breakdowns occur on an automated production line. Janak, who services 20% of the breakdowns, makes an incomplete repair 1 time in 20.Tarun ,who services 60% of the breakdowns ,makes an incomplete repair 1 time in 10 Gautham, who services 15% of the breakdowns, makes an incomplete repair 1 time in 10 and Prasad ,who services 5% of the breakdowns, makes an incomplete repair 1 time in 20.For the next problem with the production line diagnosed as being due to an initial repair that was incomplete, what is the probability that this initial repair was made by Janak?

Solution

Let A be the event that the initial repair was incomplete

B_1 that the repair was made by Janak

B_2 that it was made by Tarun ,

B_3 that it was made by Gautham,

B_4 that it was made by Prasad,

$$P(B_1/A) = \frac{P(B_1)P(A/B_1)}{P(B_1)P(A/B_1)+P(B_2)P(A/B_2)+P(B_3)P(A/B_3)+P(B_4)P(A/B_4)}$$

$$= \frac{(0.20)(0.05)}{(0.20)(0.05)+(0.60)(0.10)+(0.15)(0.10)+(0.05)(0.05}$$

$$= 0.114$$

Random Variables

We now introduce a new term

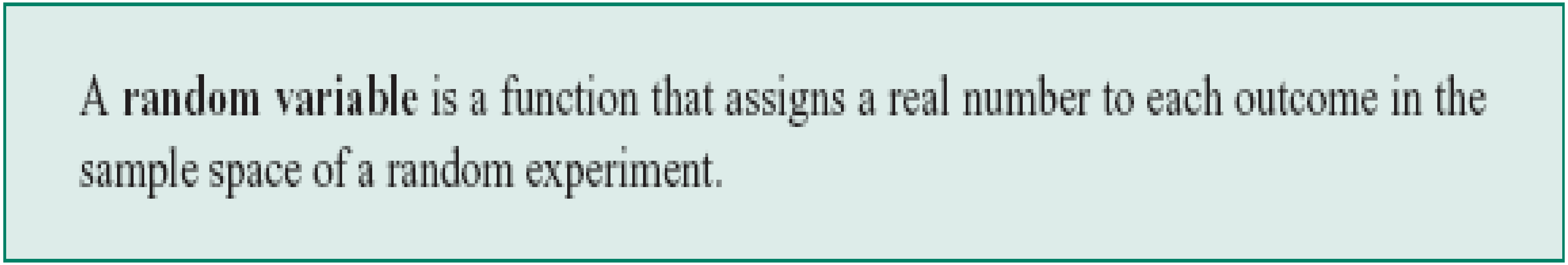
Instead of saying that the possible outcomes are 1,2,3,4,5 or 6, we say that ***random variable*** X can take values {1,2,3,4,5,6}. **A random variable is an expression whose value is the outcome of a particular experiment.**

The random variables can be either *discrete* or *continuous*.

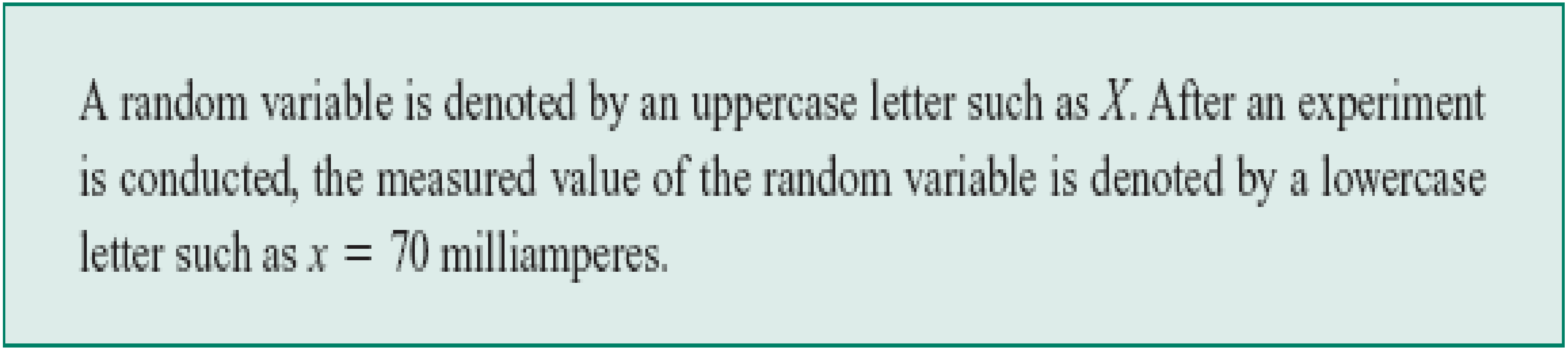
It's a convention to use the upper case letters (*X,Y*) for the names of the random variables and the lower case letters (*x,y*) for their possible particular values.

Random Variables

Definition



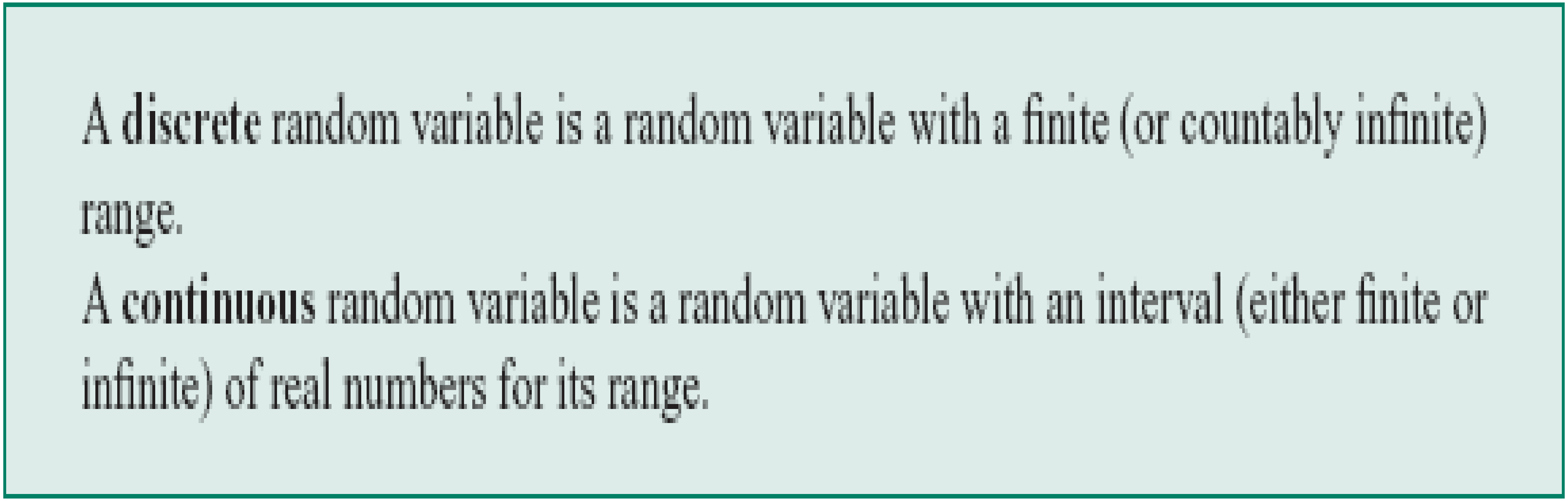
A **random variable** is a function that assigns a real number to each outcome in the sample space of a random experiment.



A random variable is denoted by an uppercase letter such as X. After an experiment is conducted, the measured value of the random variable is denoted by a lowercase letter such as $x = 70$ milliamperes.

Random Variables

Definition

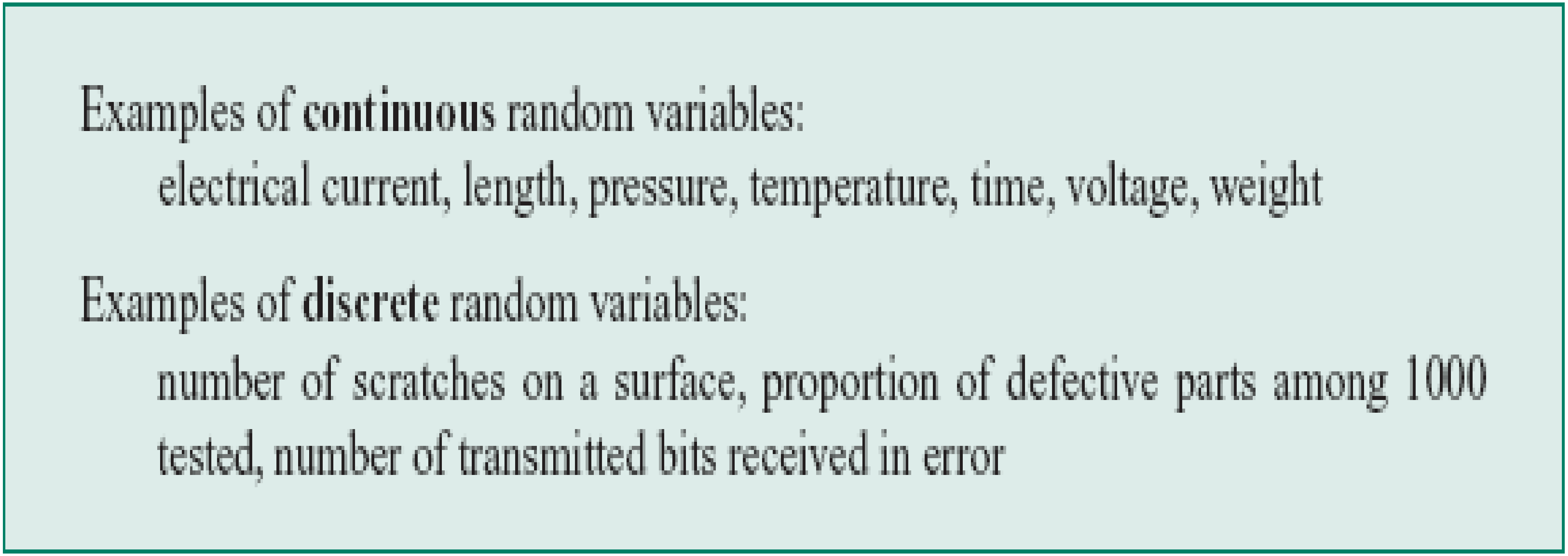


A discrete random variable is a random variable with a finite (or countably infinite) range.

A continuous random variable is a random variable with an interval (either finite or infinite) of real numbers for its range.

Random Variables

Examples of Random Variables



Examples of **continuous** random variables:

electrical current, length, pressure, temperature, time, voltage, weight

Examples of **discrete** random variables:

number of scratches on a surface, proportion of defective parts among 1000 tested, number of transmitted bits received in error

The Probability Function for discrete random variables

We assigned a probability 1/6 to each face of the dice. In the same manner, we should assign a probability 1/2 to the sides of a coin.

What we did could be described as *distributing the values of probability* between different elementary events:

$$P(X=x_k)=p(x_k),k=1,2,\ldots$$

It is convenient to introduce the *probability function p(x)* :

$$P(X=x)=p(x)$$

Continuous distribution and the probability density function

A random variable X is said to have a ***continuous distribution*** with ***density function*** *f(x)* if for all *a≤b we have*

$$P(a \leq X \leq b) = \int_a^b f(x)dx \qquad (1.15)$$

$$\int_\Omega f(x) = 1 \qquad (1.16)$$

$$P(E) = \int_E f(x)dx \qquad (1.17)$$

Examples:

1. The uniform distribution on (a,b):

We are picking a value at random from (a,b).

$$f(x)=\begin{cases} {}^{1}\!/_{b-a}, & a<x<b \\ 0 & otherwise \end{cases} \qquad (1.18)$$

2. The exponential distribution

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0 & otherwise \end{cases} \qquad (1.19)$$

Expected Value

$$E(X) = \sum_{i=1}^{n} X_i P(X_i)$$

Variance

$$\sigma^2 = \sum_{i=1}^{n} [X_i - E(X)]^2 P(X_i)$$

1

Toss a coin 3 times. The sample space is

S = {HHH; HTH; THH; TTH; HHT; HTT; THT; TTT}

Mean

Variance

Binomial Distribution

n = number of trials ,x = number of successes , p = probability of success

q = probability of failure

Probability of x successes in n trials

$$= \frac{\text{n!}}{\text{r!(n - x)!}} p^{x} q^{n-x}$$

$$\mu = np$$

$$\sigma^{2} = np(1 - p)$$

2

Roll 12 dice simultaneously, and let X denote the number of 6's that appear.

Then find $P(7<= X <= 9)$.

3

A recent national study showed that approximately 44.7% of college students have used Wikipedia as a source in at least one of their term papers.

Let X equal the number of students in a random sample of size n = 31 who have used Wikipedia as a source.

How is X distributed?

Find the probability that X is equal to 17.

Find the probability that X is at most 13.

Find the probability that X is between 16 and 19, inclusive.

Find mean and variance

The Poisson Distribution

$$P(X) = \frac{\lambda^{x} e^{-\lambda}}{X!}$$

Expected value = λ
Variance = λ

Problem

On the average, five cars arrive at a particular car wash every hour. Let X count the number of cars that arrive from 10AM to 11AM. (mean = 5)

What is the probability that no car arrives during this period?

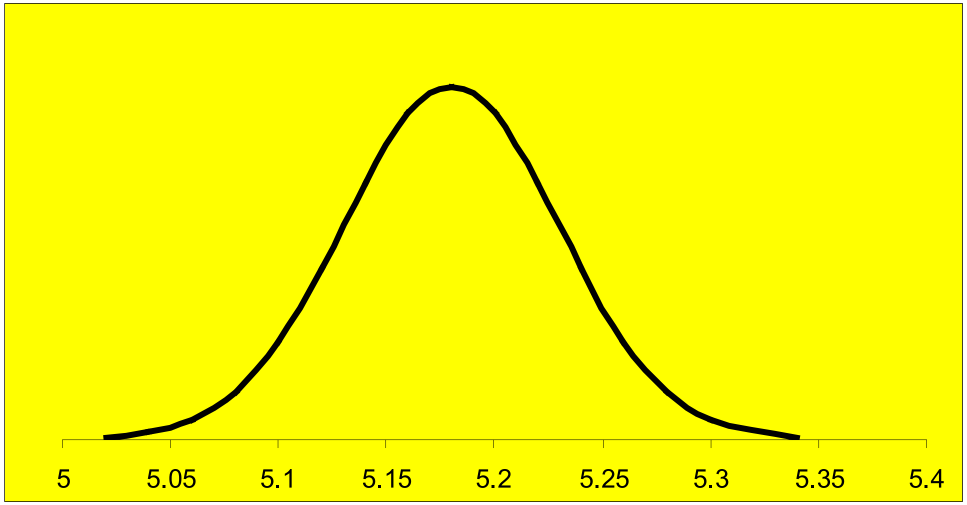
Problem

Suppose the car wash is in operation from 8AM to 6PM, and we let Y be the number of customers that appear in this period. Since this period covers a total of 10 hours, from (lambda = 50).

What is the probability that there are between 48 and 50 customers, inclusive?

Normal Distribution

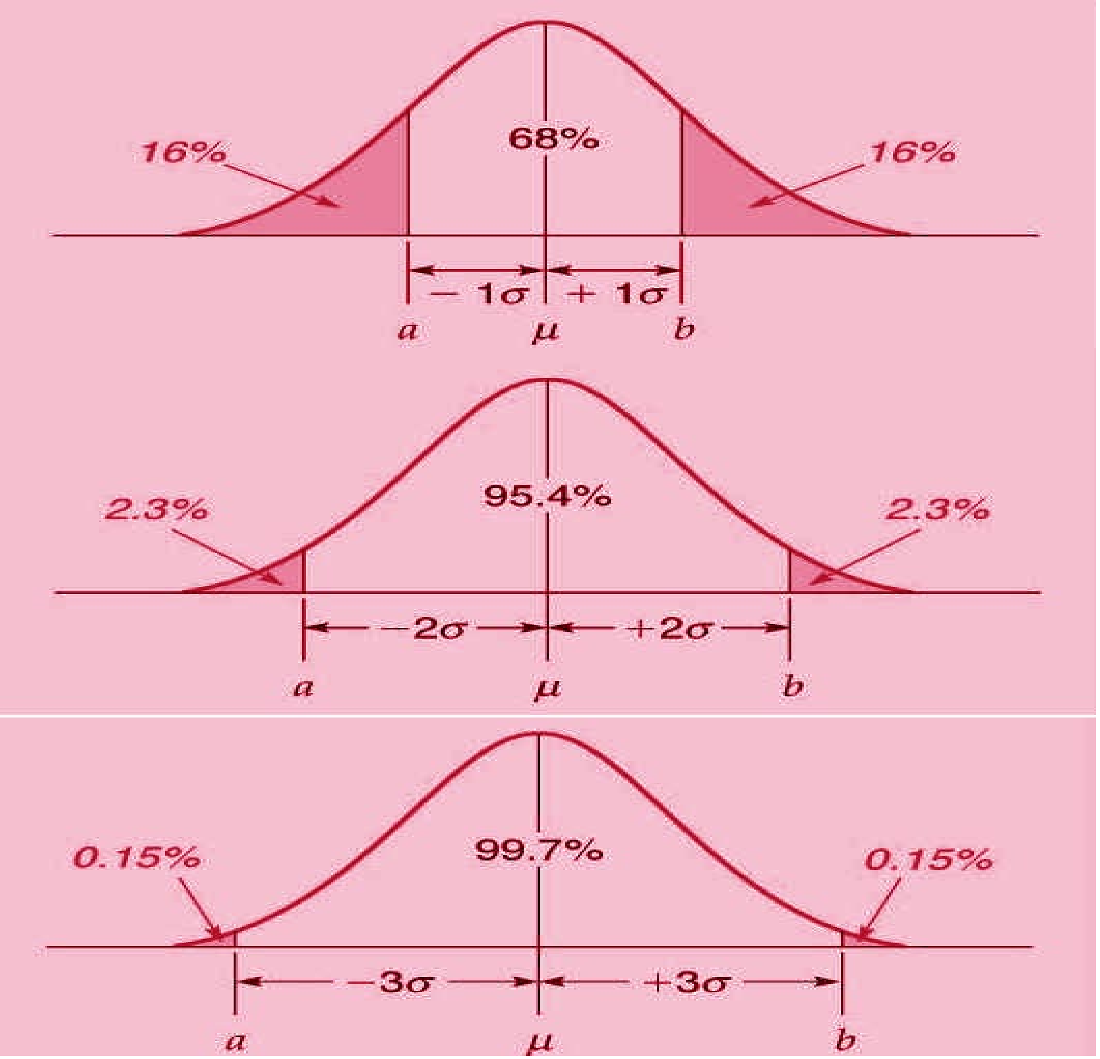
Probability density function - *f*(X)



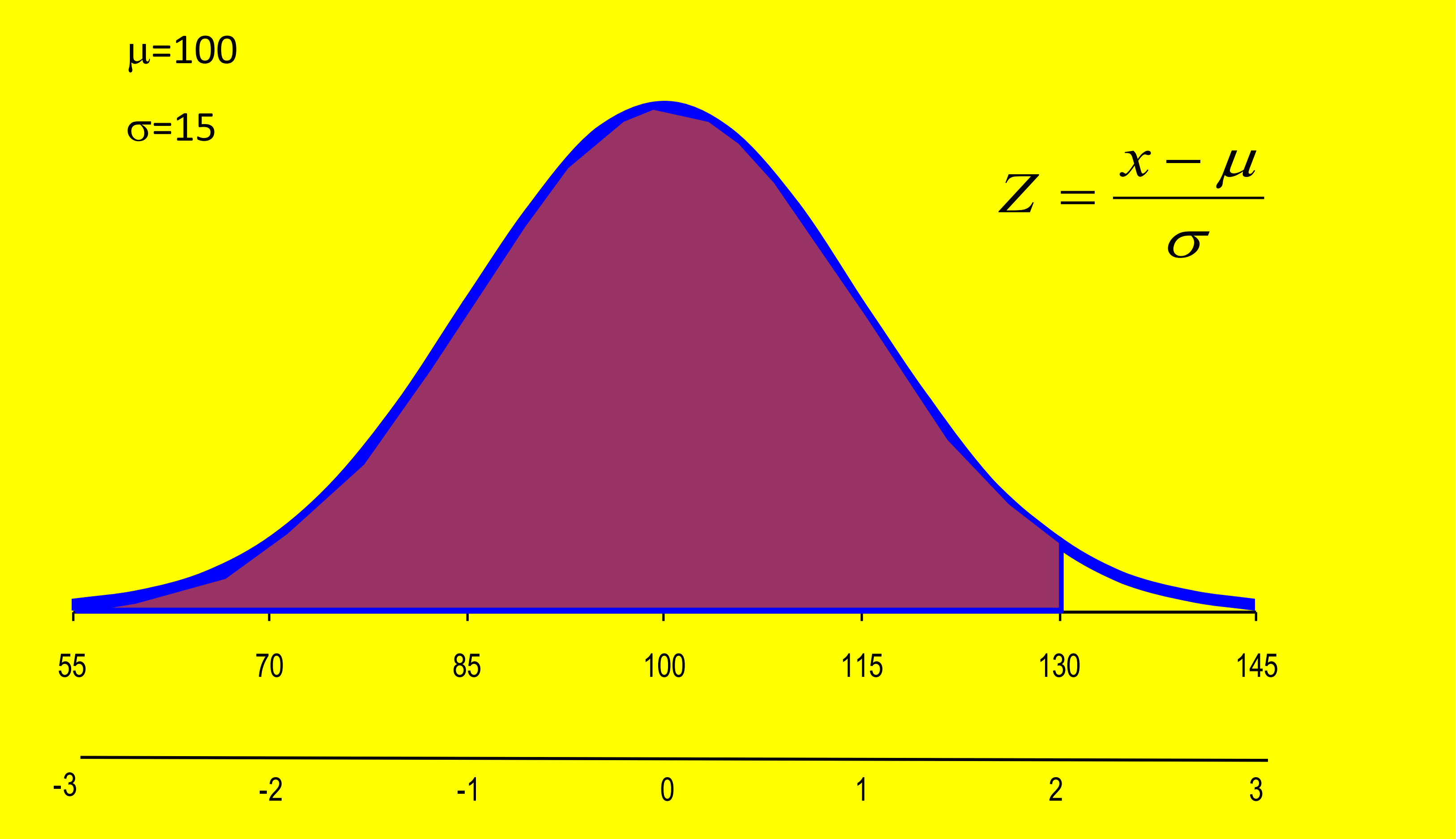
$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1/2(X-\mu)^2}{\sigma^2}}$$

Three Common Areas Under the Curve

Three Normal distributions with different areas



Standard Normal Distribution



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

Thanks