The Abdus Salam
International Centre for Theoretical Physics

United Nations
Educational, Scientific and
Cultural Organization

IAEA
International Atomic Energy Agency

2130-6

Preparatory School to the Winter College on Optics and Energy

1 - 5 February 2010

INTRODUCTION TO STATISTICAL METHODS

I. Ashraf Zahid

Quaid-I-Azam University
Pakistan

INTRODUCTION TO STATISTICAL METHODS

IMRANA ASHRAF
QUAID-I-AZAM UNIVERSITY
ISLAMABAD
PAKISTAN

Why Statistical Methods

> “The true logic of this world is in the calculus of probabilities”
>
> (James Clark Maxwell)

Why Statistical Methods *(Contd..)*

Chance is a word which is in common use in everyday living

Example:

- The radio reports speaking of tomorrow's weather may say: “There is a sixty percent chance of rain.”
- You might say: “ There is a small chance that I shall live to be one hundred years old.”
- A physicist might ask the question: “What is the chance that a particular Geiger Counter will register twenty counts in next ten seconds?”

Why Statistical Methods *(Contd..)*

By chance we mean some thing like a guess.

Why do we make a guess?

- We make guesses when we wish to make a judgment but have incomplete information or uncertain knowledge.

- We want to make a guess as to what things are, or what things are likely to happen.

- Often we wish to make a guess because we have to make a decision.

Statistics is

The study of how to:

- collect
- organize
- analyze
- interpret

numerical information from data

Individuals

The people or objects included in a study

Variable

The characteristic of an individual to be measured or observed

Types of Variables

- **Quantitative** variables are numerical measurements
  - example: number of siblings

- **Qualitative** variables place individuals into a category or group
  - example: siblings ,brand of computer

Population Data

- **The variable is from *every* individual of interest**

- Example: incomes of all residents of a country

Sample Data

- **The variable is from *only some* of the individuals of interest**

- Example: incomes of selected residents

Descriptive Statistics

Involves methods of organizing, picturing, and summarizing information from samples or populations.

Inferential Statistics

Involves methods of using information from a sample to draw conclusions regarding the population.

Data

A graphical display should

- Show the data.
- Induce the viewer to think about the substance of the graphic.
- Avoid distorting the message.

Data *(Contd..)*

Frequency Table

- Partitions data into classes or intervals
- Shows how many data values are in each class
- Each data value falls into exactly one class
- Shows the limits of each class
- Shows the frequency of each data value
- Shows the midpoint of each class

Averages and Variation

Measures of Central Tendency

- Mode

- Median

- Mean

The Mode

- The value that occurs most frequently in a data set

Find the mode:

6, 7, 2, 3, 4, 6, 2, 6

The mode is 6.

And for

6, 7, 2, 3, 4, 5, 9, 8

There is no mode for this data.

The Median

- The central value of an ordered distribution

To find the median of raw data:

- Order the data from smallest to largest.
- For an odd number of values pick the middle value.

or

- For an even number of values compute the average of the middle two values

The Median

Find the median:

Data: 5, 2, 7, 1, 4, 3, 2

Rearrange: 1, 2, 2, 3, 4, 5, 7

The median is 3.

And for,

Data: 31, 57, 12, 22, 43, 50

Rearrange: 12, 22, 31, 43, 50, 57

The median is the average of the middle two values = $\frac{31+43}{2} = 37$

The Median

Finding the median for a large data set

For an ordered data set of *n* values:

Position of the middle value =

$$\frac{n+1}{2}$$

The Mean

- An average that uses the exact value of each entry

- Sometimes called the arithmetic mean

The mean of a collection of data is found by:

- summing all the entries
- dividing by the number of entries

$$\textbf{mean} = \frac{\textbf{sum of all entries}}{\textbf{number of entries}}$$

The Mean

Find the Mean:

6, 7, 2, 3, 4, 5, 2, 8

$$\text{mean}=\frac{6+7+2+3+4+5+2+8}{8}=\frac{37}{8}=4.625\approx 4.6$$

The Mean

Notations for mean

Sample mean

$$\bar{\mathbf{x}}$$

Population mean

$$\boldsymbol{\mu}$$

Greek letter (mu)

The Mean

Sample mean

$$\bar{x} = \frac{\sum x}{n}$$

Population mean

$$\mu = \frac{\sum x}{N}$$

Weighted Average

- An average where more importance or weight is assigned to some of the numbers

If x is a data value and w is the weight assigned to that value

$$\text{Weighted average} = \frac{\Sigma xw}{\Sigma w}$$

Weighted Average

Calculating a Weighted Average

In a pageant, the interview is worth 30% and appearance is worth 70%. Find the weighted average for a contestant with an interview score of 90 and an appearance score of 80.

$$\text{Weighted average} = \frac{0.30(90) + 0.70(80)}{0.30 + 0.70}$$

$$= \frac{27 + 56}{1.00} = 83$$

Measures of Variation

- Range
- Standard Deviation
- Variance

The Range

- The difference between the largest and smallest values of a distribution

Find the range:

10, 13, 17, 17, 18

The range = largest minus smallest

= 18 minus 10 = 8

The Standard Deviation

- A measure of the average variation of the data entries from the mean

Standard deviation of a sample

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

mean of the sample

n = sample size

The Standard Deviation

To calculate standard deviation of a sample

- Calculate the mean of the sample.
- Find the difference between each entry (*x*) and the mean. These differences will add up to zero.
- Square the deviations from the mean.
- Sum the squares of the deviations from the mean.
- Divide the sum by $(n - 1)$ to get the variance.
- Take the square root of the variance to get the standard deviation.

The Variance

- The square of the standard deviation

Variance of a Sample

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

Find the standard deviation and variance

x	$x-\bar{x}$		$(\mathbf{x}-\bar{\mathbf{x}})^2$
30	4		16
26	0	Sum = 0	0
$\underline{22}$	-4		$\underline{16}$
78	Mean = 26		32

Find the standard deviation and variance *(Contd..)*

The variance

$$s^2 = \frac{\Sigma \ (x - \overline{x})^2}{n - 1} = 32 \div 2 \ = 16$$

The standard deviation

$$s = \ \sqrt{16} = 4$$

Population Mean

$$\text{population mean} = \mu = \frac{\sum x}{N}$$

where N = number of data values in the population

Population Standard Deviation

$$\sigma = \sqrt{\frac{\sum (x - \overline{x})^2}{N}}$$

where N = number of data values in the population

Coefficient Of Variation:

- A measurement of the relative variability (or consistency) of data.

$$\mathbf{CV} = \frac{s}{\bar{x}} \cdot 100 \quad \text{or} \quad \frac{\sigma}{\mu} \cdot 100$$

Coefficient Of Variation:

- CV is used to compare variability or consistency

A sample of newborn infants had a mean weight of 6.2 pounds with a standard deviation of 1 pound.

A sample of three-month-old children had a mean weight of 10.5 pounds with a standard deviation of 1.5 pound.

Which (newborns or 3-month-olds) are more variable in weight?

Coefficient Of Variation:

To compare variability, compare Coefficient of Variation

- For newborns:

CV = 16% ⇦ Higher CV: more variable

- For 3-month-olds:

CV = 14% ⇦ Lower CV: more consistent

Coefficient Of Variation:

Use Coefficient of Variation

- To compare two groups of data, to answer:
- Which is more consistent?
- Which is more variable?

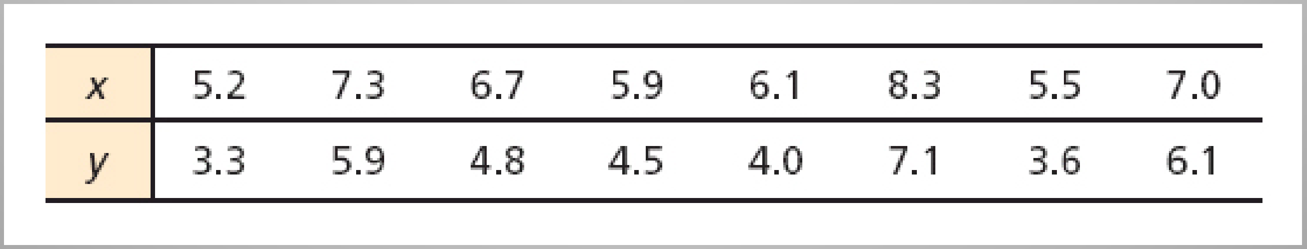
Correlation and Regression

Scatter Diagram

- A graph in which data pairs (*x*, *y*) are plotted as individual points on a grid with horizontal axis *x* and vertical axis *y*
- We call *x* the explanatory variable.
- We call *y* the response variable.

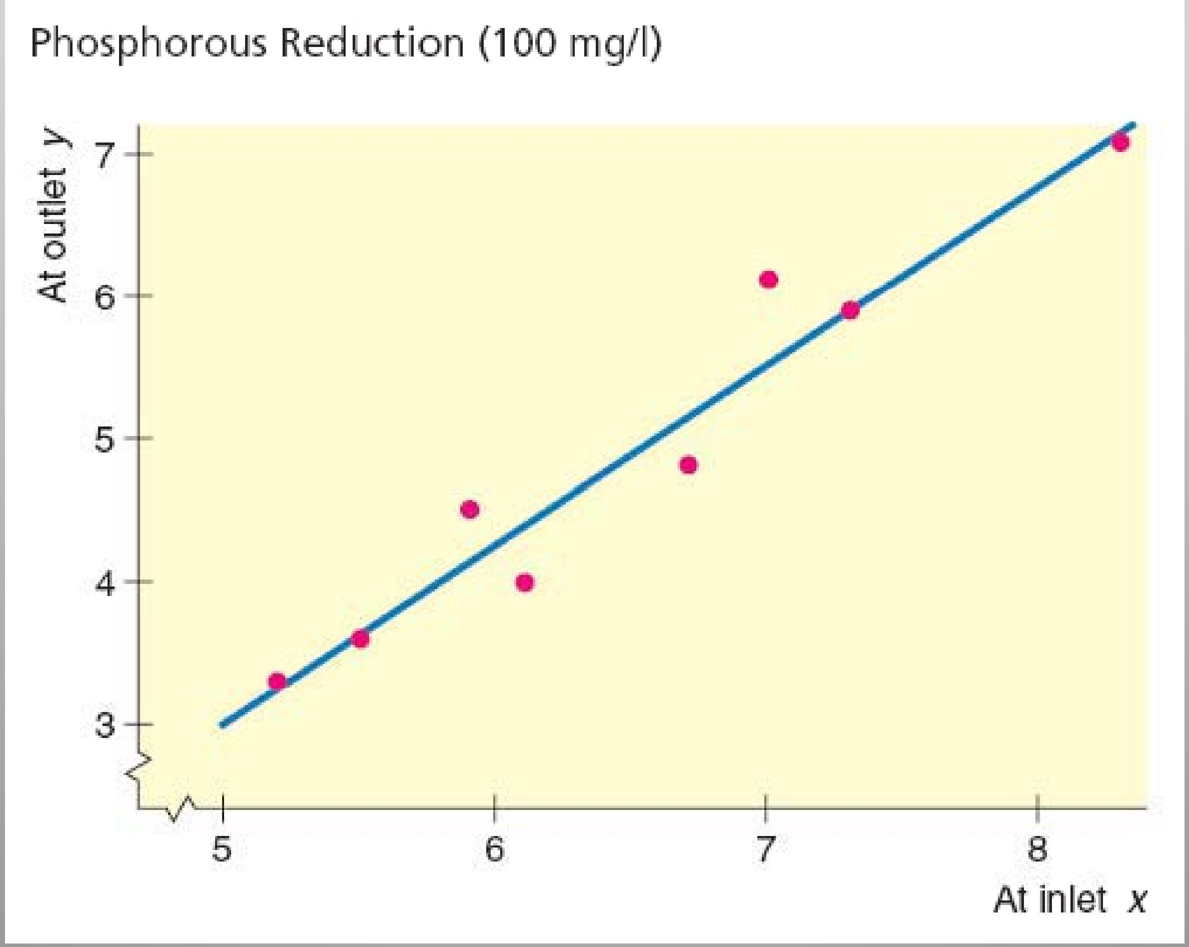
Paired data

- x = phosphorus concentration at inlet
- y = phosphorus concentration at outlet



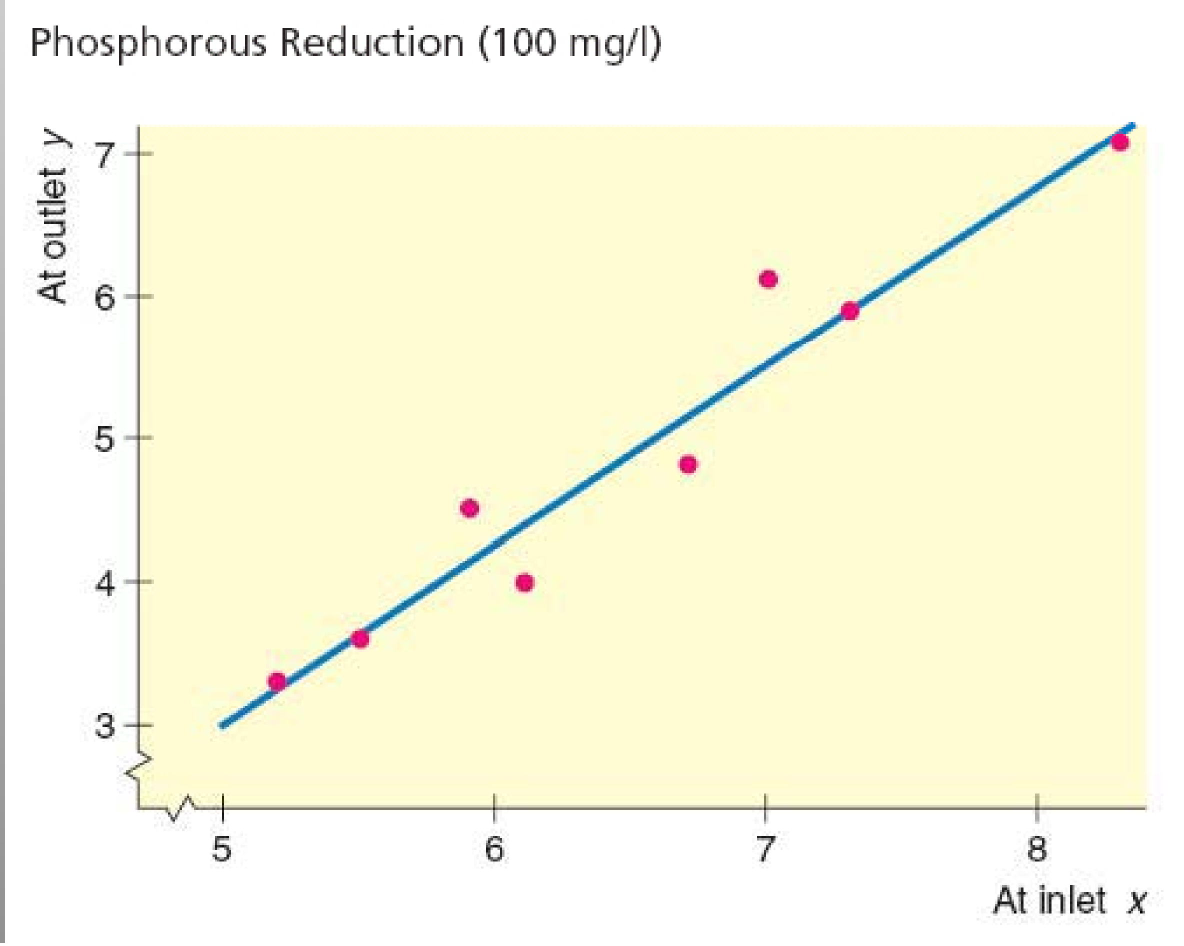
x	5.2	7.3	6.7	5.9	6.1	8.3	5.5	7.0
y	3.3	5.9	4.8	4.5	4.0	7.1	3.6	6.1

Scatter Diagram

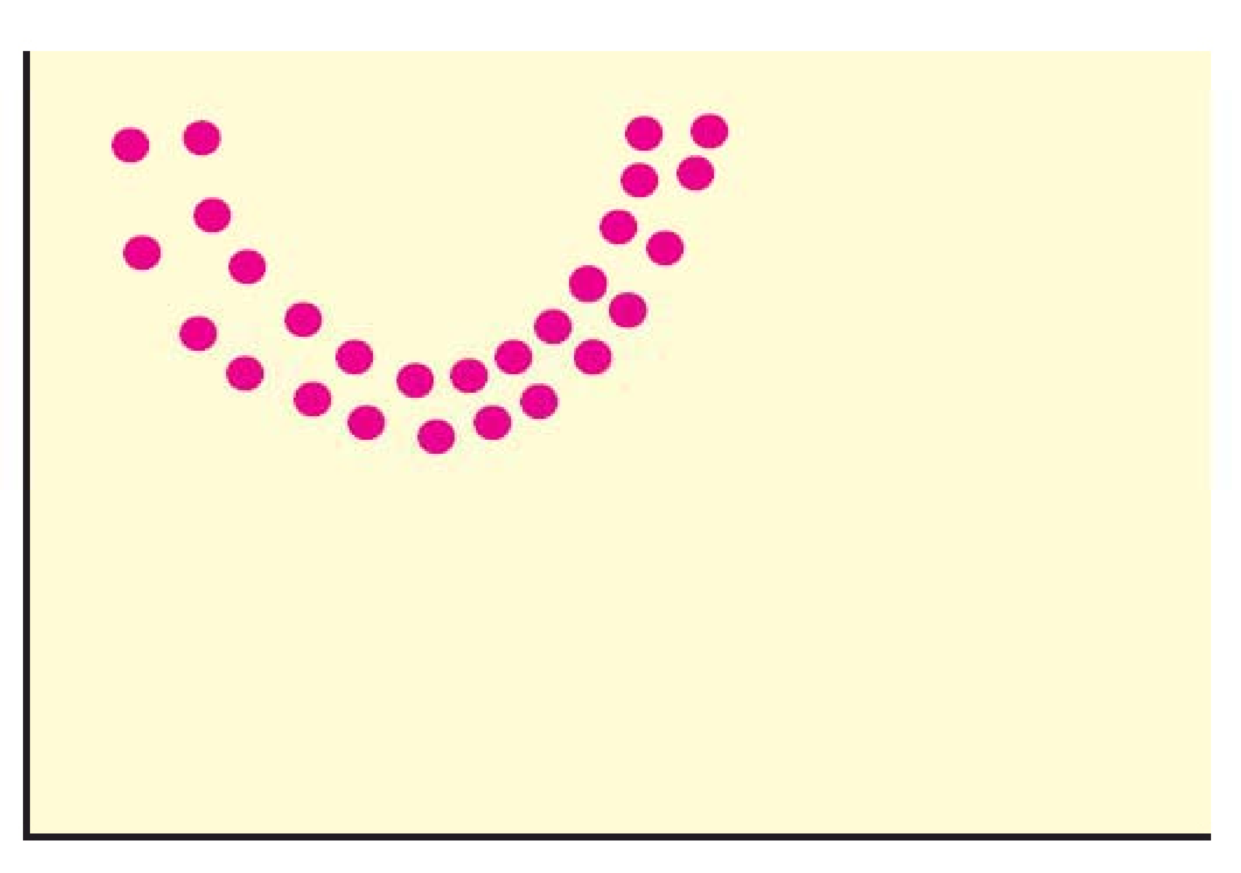


Linear Correlation

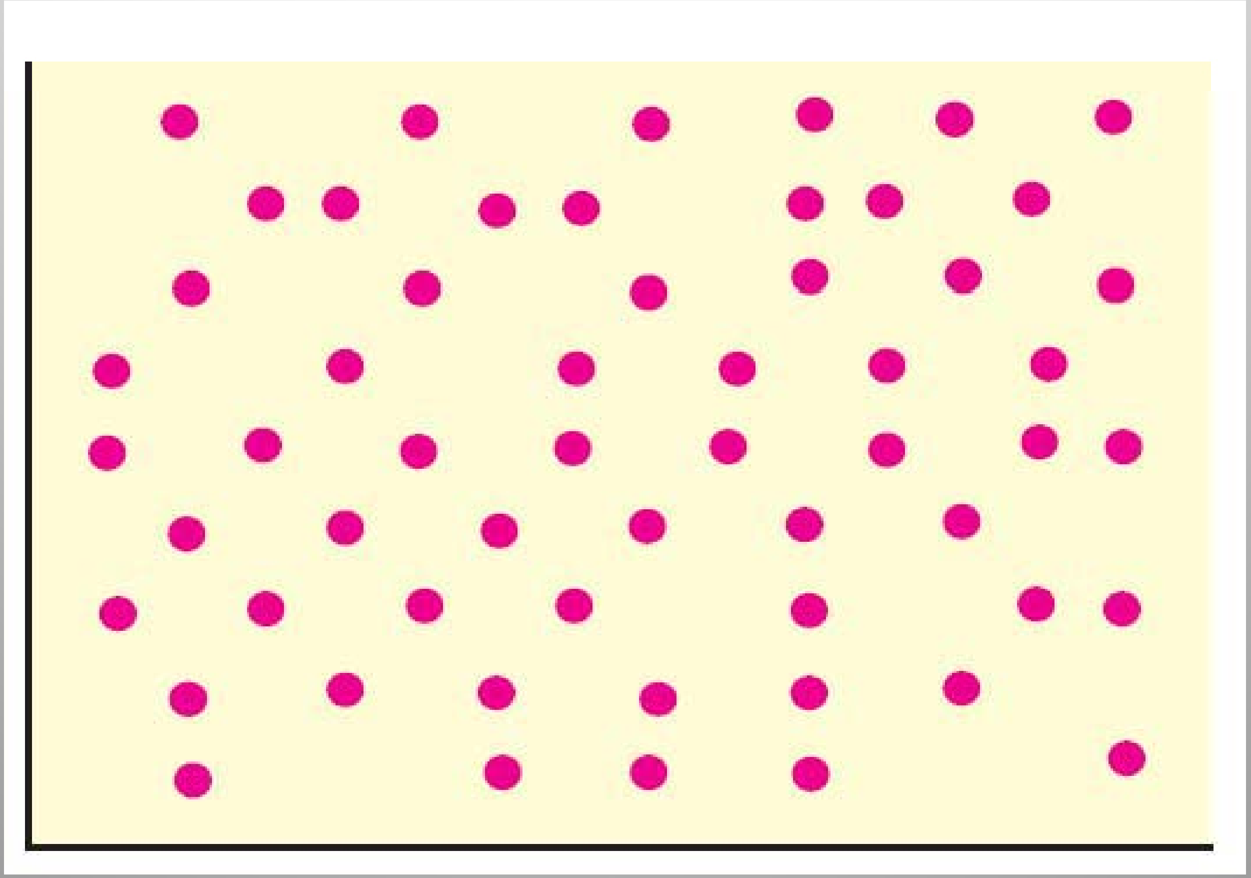
- The general trend of the points seems to follow a straight line segment.



Non-Linear Correlation

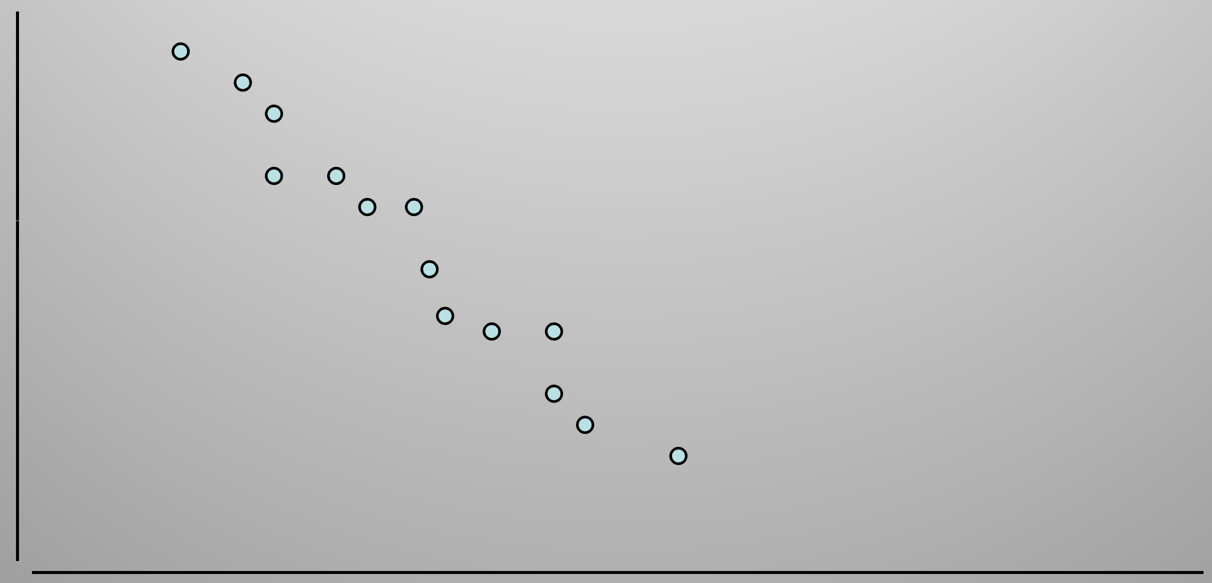


No Linear Correlation

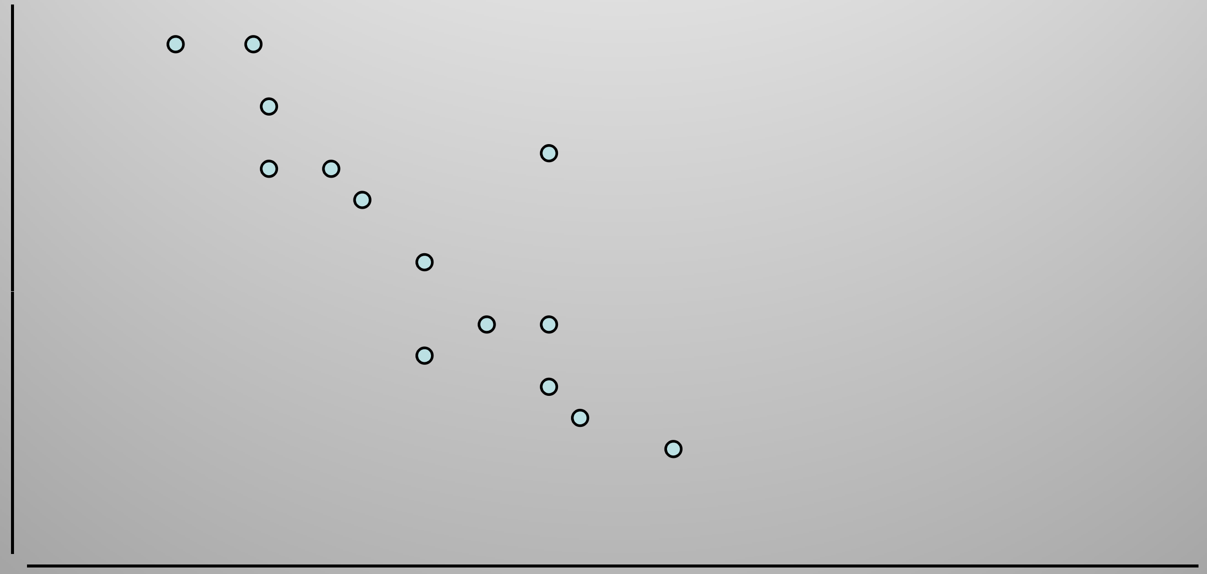


High Linear Correlation

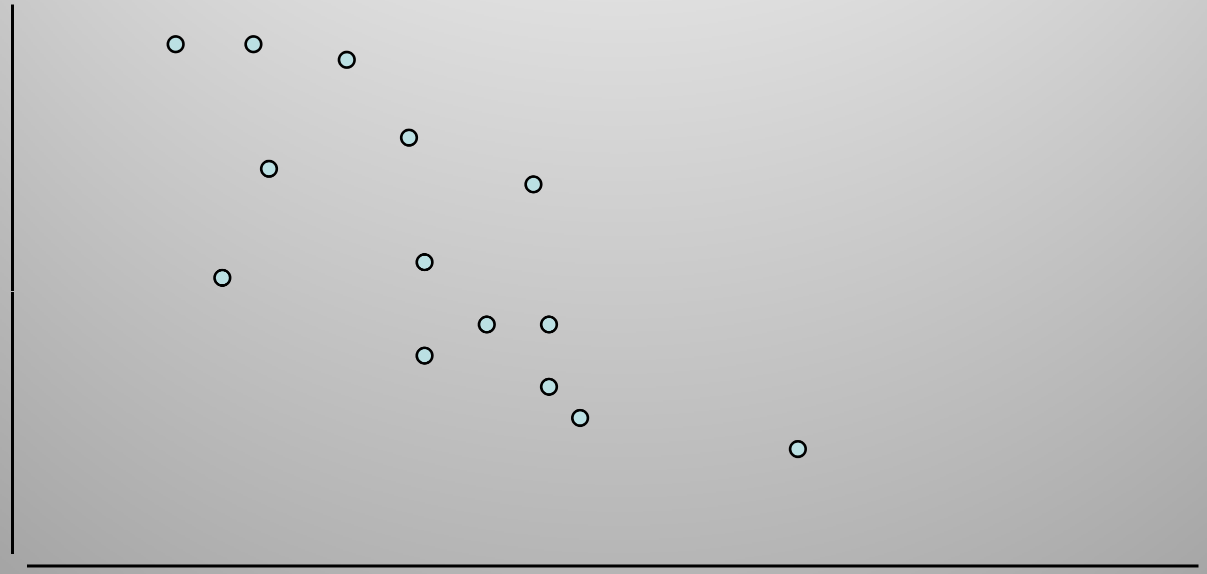
- Points lie close to a straight line.



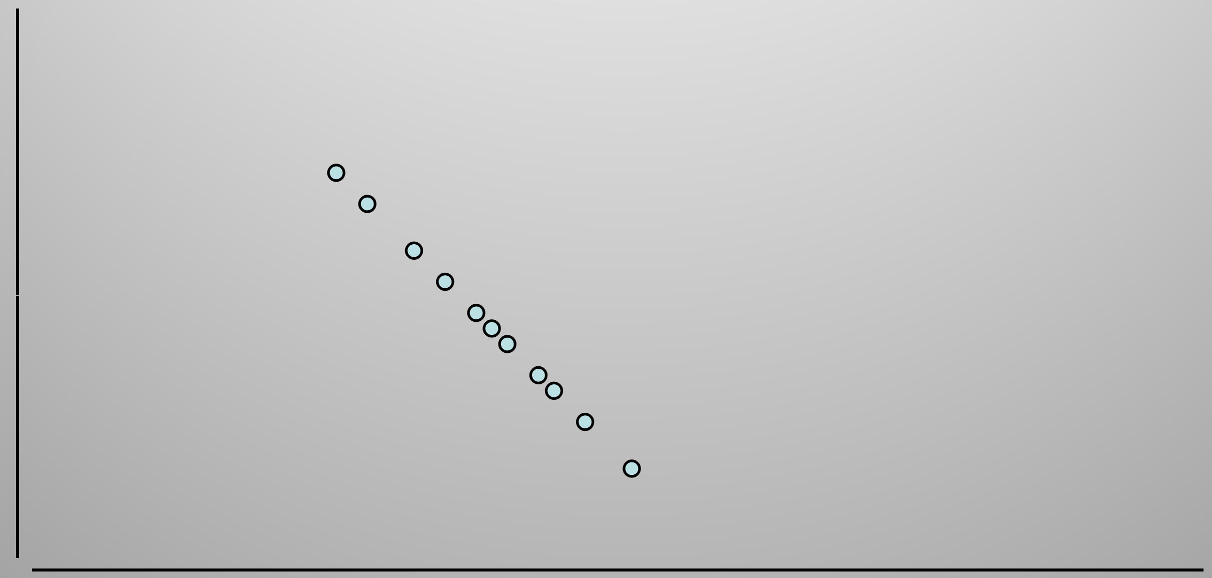
Moderate Linear Correlation



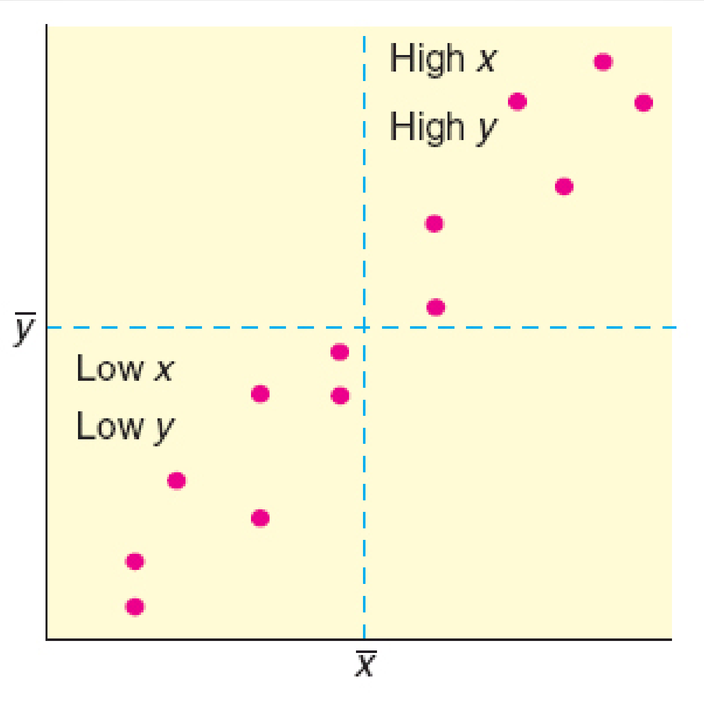
Low Linear Correlation



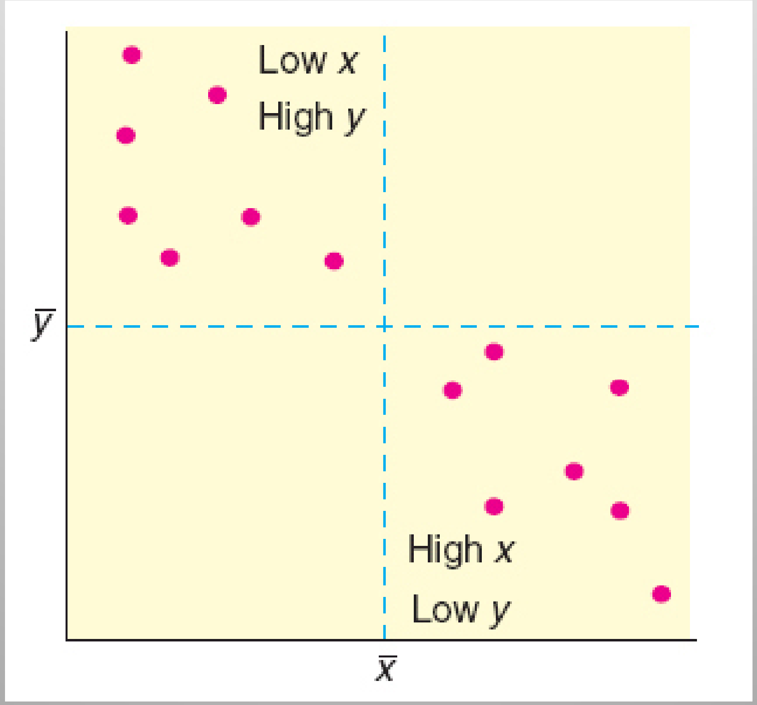
Perfect Linear Correlation



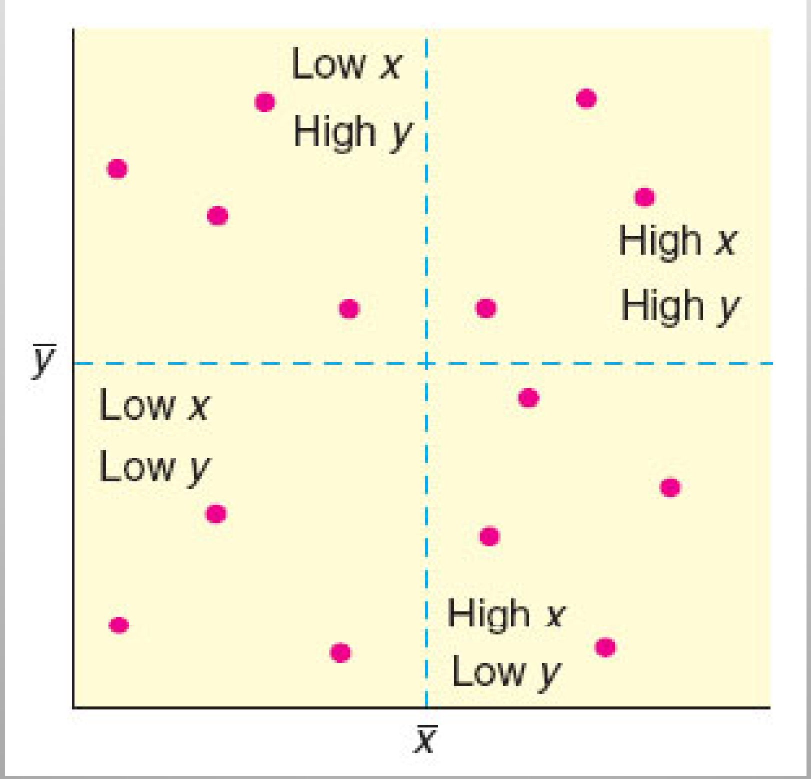
Positive Linear Correlation



Negative Linear Correlation



Little or No Linear Correlation



Questions Arising

- Can we find a relationship between x and y?
- How strong is the relationship?

The Correlation Coefficient (*r*)

- A numerical measurement that assesses the strength of a linear relationship between two variables *x* and *y*

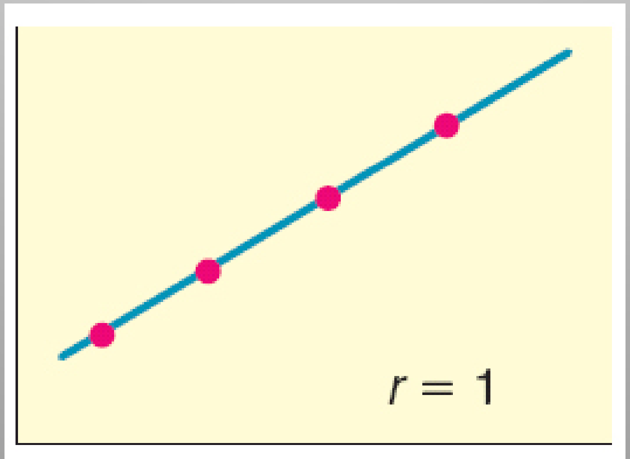
Properties of the Correlation Coefficient *r*

- Also called the Pearson product-moment correlation coefficient, *r* is a unitless measurement between –1 and 1.

- That is $-1 \underline{\leq} r \underline{\leq} 1$.

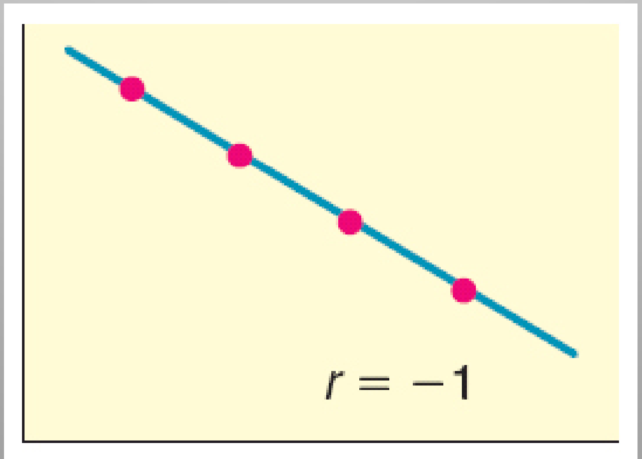
Properties of the Correlation Coefficient *r*

- If *r* = 1, there is a perfect positive correlation.
- Positive values of *r* imply that as *x* increases, *y* tends to increase



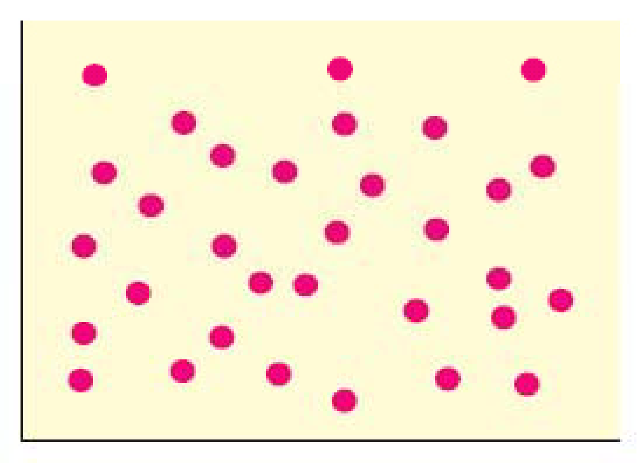
Properties of the Correlation Coefficient *r*

- If $r = -1$, there is a perfect negative correlation.
- Negative values of r imply that as x increases, y tends to decrease



Properties of the Correlation Coefficient *r*

- If $r = 0$, there is no linear correlation.



Properties of the Correlation Coefficient r

- The closer r is to – 1 or +1, the better a line describes the relationship between the two variables x and y.

- The value of r does not change when either variable is converted to different units.
- The value of r is the same regardless of which variable is the explanatory variable and which variable is the response variable. In other words, the value of r is the same for the pairs (x, y) as for the pairs (y, x).

Computing the Correlation Coefficient *r*

- Obtain a random sample of n data pairs (*x*, *y*).
- Using the data pairs, compute Σx, Σy, Σx^2, Σy^2, and Σxy.
- Use the following formula:

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}}$$

Computing r

x (Miles)	y (Min.)	x^2	y^2	xy
2	6	4	36	12
5	9	25	81	45
12	23	144	529	276
7	18	49	324	126
7	15	49	225	105
15	28	225	784	420
10	19	100	361	190
$\Sigma x = 58$	$\Sigma y = 118$	$\Sigma x^2 = 596$	$\Sigma y^2 = 2340$	$\Sigma xy = 1174$

Computing r

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2) - \left(\sum y\right)^2}}$$

$$= \frac{7(1174) - (58)(118)}{\sqrt{7(596) - 58^2}\sqrt{7(2340) - 118^2}}$$

$$\approx 0.975$$

An *r* value of 0.975 indicates a strong positive correlation between the variables *x* and *y*

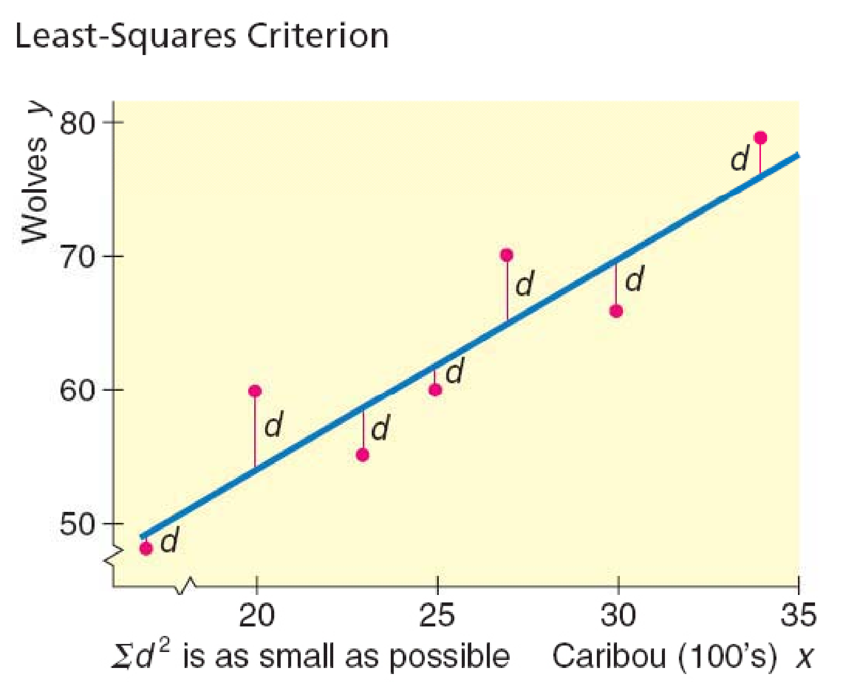
When there appears to be a linear relationship between *x* and *y*:

- Attempt to “fit” a line to the scatter diagram.

The Least Squares Line

- The sum of the squares of the vertical distances from the points to the line is made as small as possible.

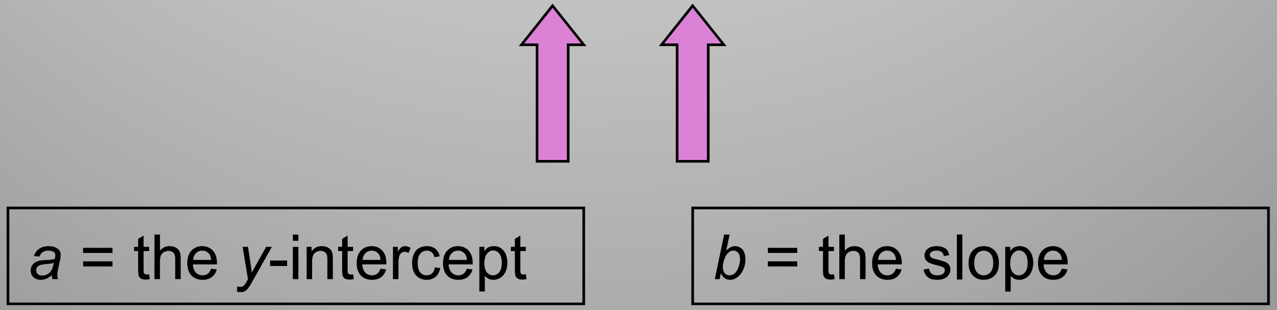
Least Squares Criterion



- **The sum of the squares of the vertical distances from the points to the line is made as small as possible.**

Equation of the Least Squares Line

$$\hat{y} = a + bx$$



a = the *y*-intercept

b = the slope

Finding the Equation of the Least Squares Line

- Obtain a random sample of *n* data pairs (*x*, *y*).
- Using the data pairs, compute Σx, Σy, Σx^2, Σy^2, and Σxy.
- Compute the sample means $\bar{\mathbf{x}}$ **and** $\bar{\mathbf{y}}$**.**

Finding the Slope

- Use the following formula:

$$\mathbf{slope = b = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\sum x^2 - \left(\sum x\right)^2}}$$

Finding the y-intercept

$$y - \text{intercept} = a = \bar{y} - b\bar{x}$$

where $\bar{y} = \text{mean of } y \text{ values}$

and $\bar{x} = \text{mean of } x \text{ values}$

and $b = \text{the slope}$

Find the Least Squares Line

X (Miles Traveled)	y (Minutes)	x^2	xy
2	6	4	12
5	9	25	45
12	23	144	276
7	18	49	126
7	15	49	105
15	28	225	420
10	19	100	190
$\Sigma x = 58$	$\Sigma y = 118$	$\Sigma x^2 = 596$	$\Sigma xy = 1174$

Finding the Slope

$$\text{slope} = b = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\sum x^2 - \left(\sum x\right)^2}$$

$$= \frac{7(1174) - (58)(118)}{7(596) - 59^2}$$

$$\approx 1.70$$

Finding the y-intercept

$$\bar{y} = \text{mean of y values} = \frac{118}{7} = 16.857143$$

$$\bar{x} = \text{mean of x values} = \frac{58}{7} = 8.2857143$$

$$\text{y} - \text{int ercept} = a = \bar{y} - b\bar{x} =$$

$$16.857143 - 1.700495\,(\,8.2857143\,)$$

$$= 2.7673273 \approx 2.77$$

The equation of the least squares line is:

$$\hat{y} = a + bx$$

$$\hat{y} = 2.77 + 1.70x$$

Probability

- Probability is a numerical measurement of likelihood of an event.
- The probability of any event is a number between zero and one.
- Events with probability close to one are more likely to occur.
- Events with probability close to zero are less likely to occur.

Probability Notation

If A represents an event,

$P(A)$

represents the probability of A.

$$\text{If } P(A) = 1$$

Event A is certain to occur

$$\text{If } P(A) = 0$$

Event A is certain not to occur

Three methods to find probabilities:

- Intuition
- Relative frequency
- Equally likely outcomes

Intuition Method of Determining Probability

- Incorporates past experience, judgment, or opinion.
- Is based upon level of confidence in the result
- Example: “I am 95% sure that I will attend the party.”

Probability as Relative Frequency

Probability of an event =

the fraction of the time that the event occurred in the past =

$$\frac{f}{n}$$

where f = frequency of an event

n = sample size

Example of Probability as Relative Frequency

If you note that 57 of the last 100 applicants for a job have been female, the probability that the next applicant is female would be:

$$\frac{57}{100}$$

Equally Likely Outcomes

- No one result is expected to occur more frequently than any other.

$$\textbf{Probability of an event} = \frac{\textbf{Number of outcomes favorable to event}}{\textbf{Total number of outcomes}}$$

When rolling a die, the probability of getting a number less than three = $\frac{2}{6} = \frac{1}{3}$

Law of Large Numbers

- In the long run, as the sample size increases and increases, the relative frequencies of outcomes get closer and closer to the theoretical (or actual) probability value.

- Any random activity that results in a definite outcome is called Statistical Observation

Event

- A collection of one or more outcomes of a statistical experiment or observation

Simple Event

- An outcome of a statistical experiment that consists of one and only one of the outcomes of the experiment

Sample Space

- The set of all possible distinct outcomes of an experiment
- The sum of all probabilities of all simple events in a sample space must equal one.

Sample Space for the rolling of an ordinary die:

1, 2, 3, 4, 5, 6

For the experiment of rolling an ordinary die:

- *P*(even number) = $\frac{3}{6} = \frac{1}{2}$
- *P*(result less than five) = $\frac{4}{6} = \frac{2}{3}$
- *P*(not getting a two) = $\frac{5}{6}$

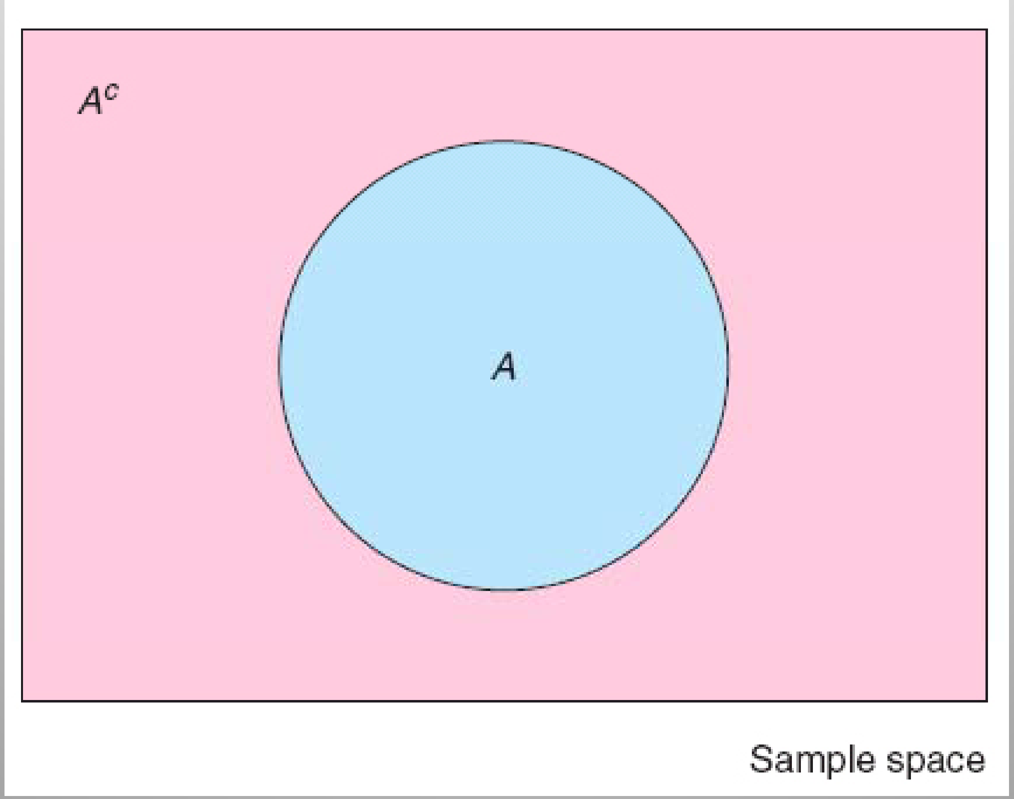
Complement of Event *A*

- The event that *A* does not occur

- Notation for the complement of event *A*:

$$\mathbf{A}^{\mathbf{c}}$$

Event *A* and its complement A^c



Probability of a Complement

- P(event A does not occur) =

$$P(A^c) = 1 – P(A)$$

- So, $P(A) + P(A^c) = 1$

If the probability that it will snow today is 30%,

P(It will not snow) = 1 – P(snow) =

1 – 0.30 = 0.70

Probability Related to Statistics

- Probability makes statements about what will occur when samples are drawn from a known population.
- Statistics describes how samples are to be obtained and how inferences are to be made about unknown populations.

Independent Events

- The occurrence (or non-occurrence) of one event does not change the probability that the other event will occur.

If events *A* and *B* are independent,

- $P(A \text{ and } B) = P(A) \cdot P(B)$

Conditional Probability

- If events are dependent, the fact that one occurs affects the probability of the other.

- *P*(*A*, *given B*) equals the probability that event *A* occurs, assuming that *B* has already occurred.

The Multiplication Rules:

- For independent events:

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

- For any events:

$$P(A \text{ and } B) = P(A) \cdot P(B, \text{given } A)$$

$$P(A \text{ and } B) = P(B) \cdot P(A, \text{given } B)$$

For independent events: $P(A \text{ and } B) = P(A) \cdot P(B)$

When choosing two cards from two separate decks of cards, find the probability of getting two fives.

P(two fives) =

P(5 from first deck and 5 from second) =

$$\frac{1}{13} \cdot \frac{1}{13} = \frac{1}{169}$$

For dependent events: *P*(*A and B*) = *P*(*A*) · *P*(*B, given A*)

When choosing two cards from a deck without replacement, find the probability of getting two fives.

P(two fives) =

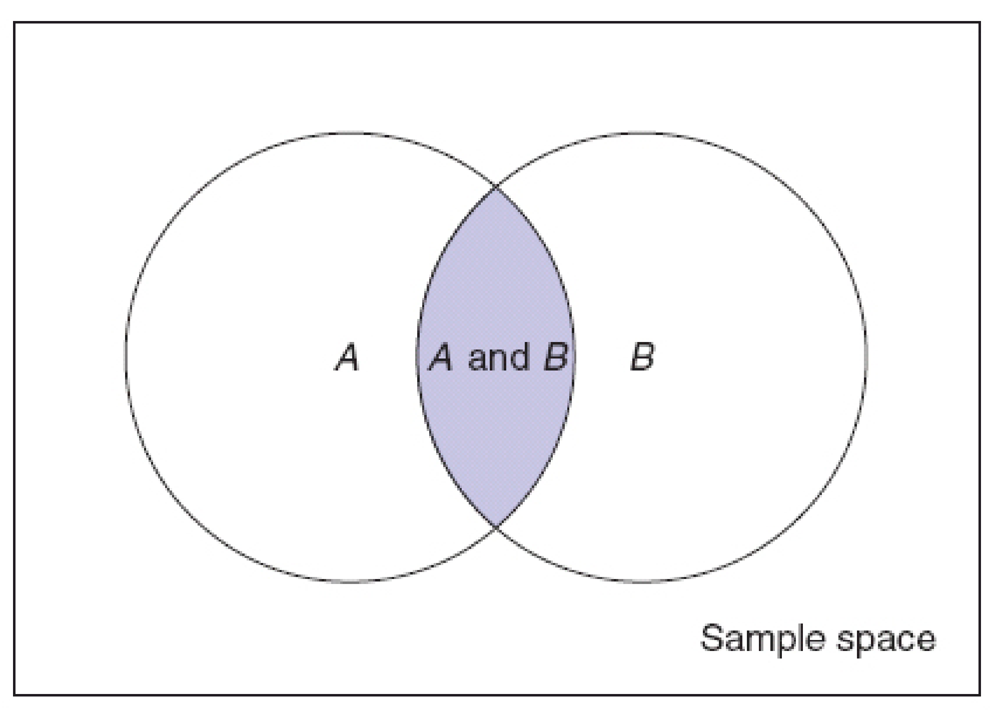
P(5 on first draw and 5 on second) =

$$\frac{4}{52} \cdot \frac{3}{51} = \frac{12}{2652} = \frac{1}{221}$$

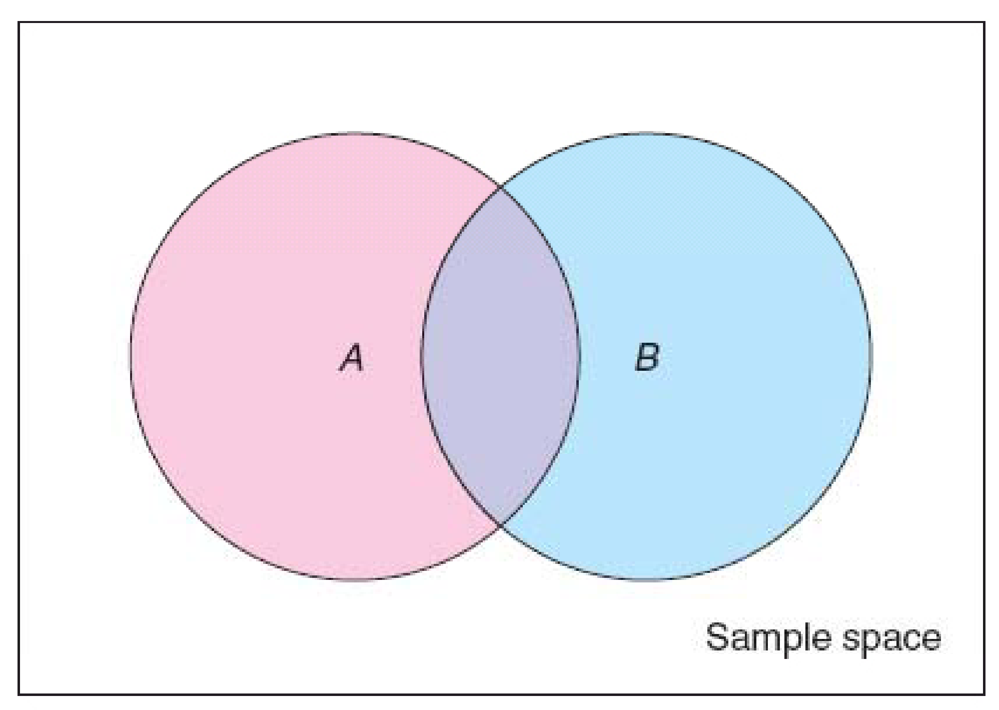
“*And*” versus “*or*”

- *And* means both events occur together.
- *Or* means that at least one of the events occur.

The Event *A and B*



The Event *A or B*



General Addition Rule

For any events A and B,

$$P(A \text{ or } B) =$$

$$P(A) + P(B) - P(A \text{ and } B)$$

General Addition Rule

When choosing a card from an ordinary deck, the probability of getting a five or a red card:

$$P(5\ ) + P(\text{red}) - P(5\ \textit{and}\ \text{red}) =$$

$$\frac{4}{52}+\frac{26}{52}-\frac{2}{52}=\frac{28}{52}=\frac{7}{13}$$

When choosing a card from an ordinary deck, the probability of getting a five or a six:

$$P(5\ ) + P(6) - P(5\ \textit{and}\ 6) =$$

$$\frac{4}{52}+\frac{4}{52}-\frac{0}{52}=\frac{8}{52}=\frac{2}{13}$$

Mutually Exclusive Events

- Events that are disjoint, cannot happen together.

For any *mutually exclusive* events *A* and *B*,

$$P(A \text{ or } B) = P(A) + P(B)$$

When rolling an ordinary die:

$$P(4 \text{ or } 6) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$$

Survey results:

Education:	Males	Females	Row totals
College Graduates	54	62	116
Not College Graduates	31	40	71
Column totals	85	102	187(Grand total)

P(male *and* college grad) = ?

Survey results:

Education:	Males	Females	Row totals
College Graduates	54	62	116
Not College Graduates	31	40	71
Column totals	85	102	187(Grand total)

$$P(\text{male } \mathit{and} \text{ college grad}) = \frac{\mathbf{54}}{\mathbf{187}}$$

Survey results:

Education:	Males	Females	Row totals
College Graduates	54	62	116
Not College Graduates	31	40	71
Column totals	85	102	187(Grand total)

P(male *or* college grad) = ?

Survey results:

Education:	Males	Females	Row totals
College Graduates	54	62	116
Not College Graduates	31	40	71
Column totals	85	102	187(Grand total)

$$P(\text{male } \textit{or} \text{ college grad}) = \frac{\mathbf{147}}{\mathbf{187}}$$

Survey results:

Education:	Males	Females	Row totals
College Graduates	54	62	116
Not College Graduates	31	40	71
Column totals	85	102	187(Grand total)

P(male, *given* college grad) = ?

Survey results:

Education:	Males	Females	Row totals
College Graduates	54	62	116
Not College Graduates	31	40	71
Column totals	85	102	187(Grand total)

$$P(\text{male, } \textit{given} \text{ college grad}) = \frac{\mathbf{54}}{\mathbf{116}}$$

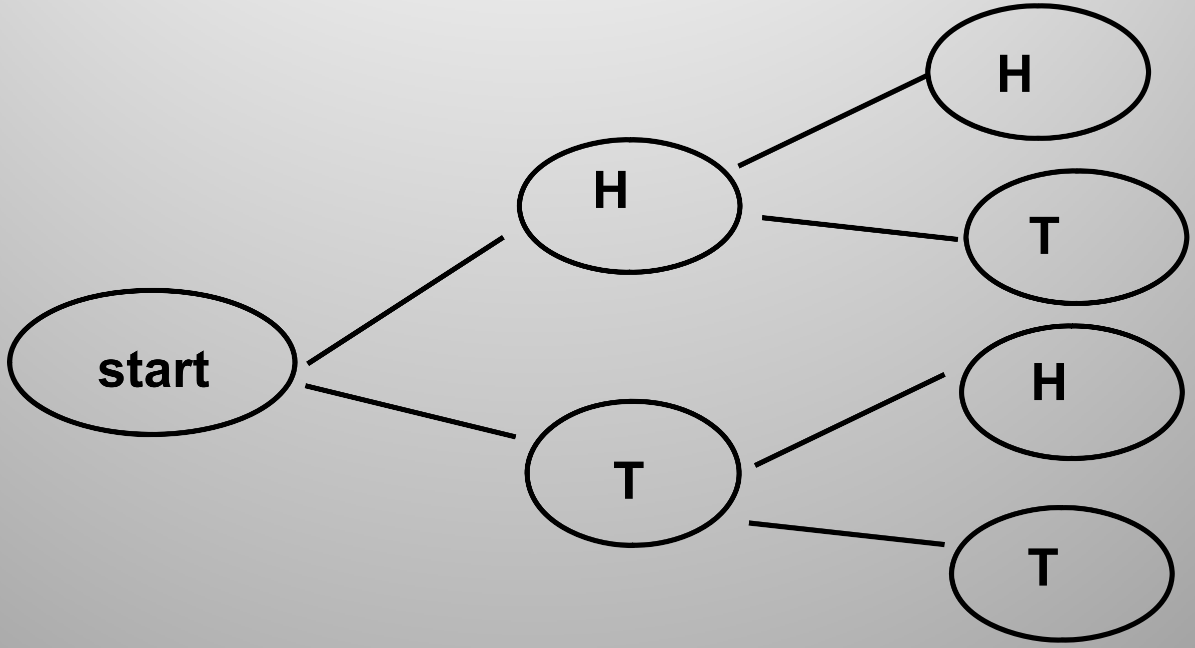
Counting Techniques

- Tree Diagram
- Multiplication Rule of Counting
- Permutations
- Combinations

Tree Diagram

- a visual display of the total number of outcomes of an experiment consisting of a series of events

Tree diagram for the experiment of tossing two coins



Find the number of paths without constructing the tree diagram:

Experiment of rolling two dice, one after the other and observing any of the six possible outcomes each time .

Number of paths = 6 x 6 = 36

Multiplication Rule of Counting

For any series of events, if there are

n_1 possible outcomes for event E_1

and n_2 possible outcomes for event E_2, etc.

then there are

$n_1 \times n_2 \times \cdots \times n_m$ possible outcomes

for the series of events E_1 through E_m.

Area Code Example

In the past, an area code consisted of 3 digits, the first of which could be any digit from 2 through 9.

The second digit could be only a 0 or 1.

The last could be any digit.

How many different such area codes were possible?

$$8 \cdot 2 \cdot 10 = 160$$

Ordered Arrangements

In how many different ways could four items be arranged in order from first to last?

$$4 \bullet 3 \bullet 2 \bullet 1 = 24$$

Factorial Notation

- *n*! is read "*n* factorial"
- *n*! is applied only when *n* is a whole number.
- *n*! is a product of *n* with each positive counting number less than *n*

Permutations

- A Permutation is an arrangement in a particular order of a group of items.

- There are to be no repetitions of items within a permutation.

Listing Permutations

How many different permutations of the letters a, b, c are possible?

Solution: There are six different permutations:

abc, acb, bac, bca, cab, cba.

Listing Permutations

How many different two-letter permutations of the letters a, b, c, d are possible?

Solution: There are twelve different permutations:

ab, ac, ad, ba, ca, da, bc, bd, cb, db, cd, dc.

Permutation Formula

The number of ways to *arrange in order n* distinct objects, taking them *r* at a time, is:

$$\mathbf{P_{n,r} = \frac{n!}{(n-r)!}}$$

where *n* and *r* are whole numbers and *n* > *r*.

Another Notation for Permutations

$$_n P_r$$

Find $P_{7,\,3}$

$$P_{7,3} = \frac{7!}{(7-3)!} = \frac{7!}{4!} = \frac{5040}{24} = 210$$

Application of Permutations

A teacher has chosen eight possible questions for an upcoming quiz. In how many different ways can five of these questions be chosen and arranged in order from #1 to #5?

Solution: $P_{8,5} = \frac{8!}{3!} = 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 = 6720$

Combinations

A combination is a grouping
in no particular order
of items.

Combination Formula

The number of combinations of *n* objects taken *r* at a time is:

$$C_{n,r} = \frac{n!}{(n-r)!r!} = {}_{n}C_{r} \quad \text{or} \quad \binom{n}{r}$$

where *n* and *r* are whole numbers and *n* > *r*.

Find $C_{9,\,3}$

$$C_{9,3} = \frac{9!}{3!(9-3)!} = \frac{9!}{3!6!} = \frac{362880}{6(720)} = 84$$

Application of Combinations

A teacher has chosen eight possible questions for an upcoming quiz. In how many different ways can five of these questions be chosen if order makes no difference?

Solution: $C_{8,5} = \dfrac{8!}{5!3!} = 56$

Determining the Number of Outcomes of an Experiment

- If the experiment consists of a series of stage with various outcomes, use the multiplication rule or a tree diagram.
- If the outcomes consist of ordered subgroups of *r* outcomes taken from a group of *n* outcomes use the permutation rule.
- If the outcomes consist of non-ordered subgroups of *r* items taken from a group of *n* items use the combination rule.

Pair of Dice

- For one dice, the probability of any face coming up is the same, 1/6. Therefore, it is equally probable that any number from one to six will come up.

- For two dices, what is the probability that the total will come up 2, 3, 4, etc up to 12?

- To calculate the probability of a particular outcome, count the number of all possible results. Then count the number that give the desired outcome. The probability of the desired outcome is equal to the number that gives the desired outcome divided by the total number of outcomes. Hence, 1/6 for one dice.

Pair of Dice *(Contd..)*

List all possible outcomes (36) for a pair of dice.

Total	Combinations	How Many
2	1+1	1
3	1+2, 2+1	2
4	1+3, 3+1, 2+2	3
5	1+4, 4+1, 2+3, 3+2	4
6	1+5, 5+1, 2+4, 4+2, 3+3	5
7	1+6, 6+1, 2+5, 5+2, 3+4, 4+3	6
8	2+6, 6+2, 3+5, 5+3, 4+4	5
9	3+6, 6+3, 4+5, 5+4	4
10	4+6, 6+4, 5+5	3
11	5+6, 6+5	2
12	6+6	1

Sum = 36

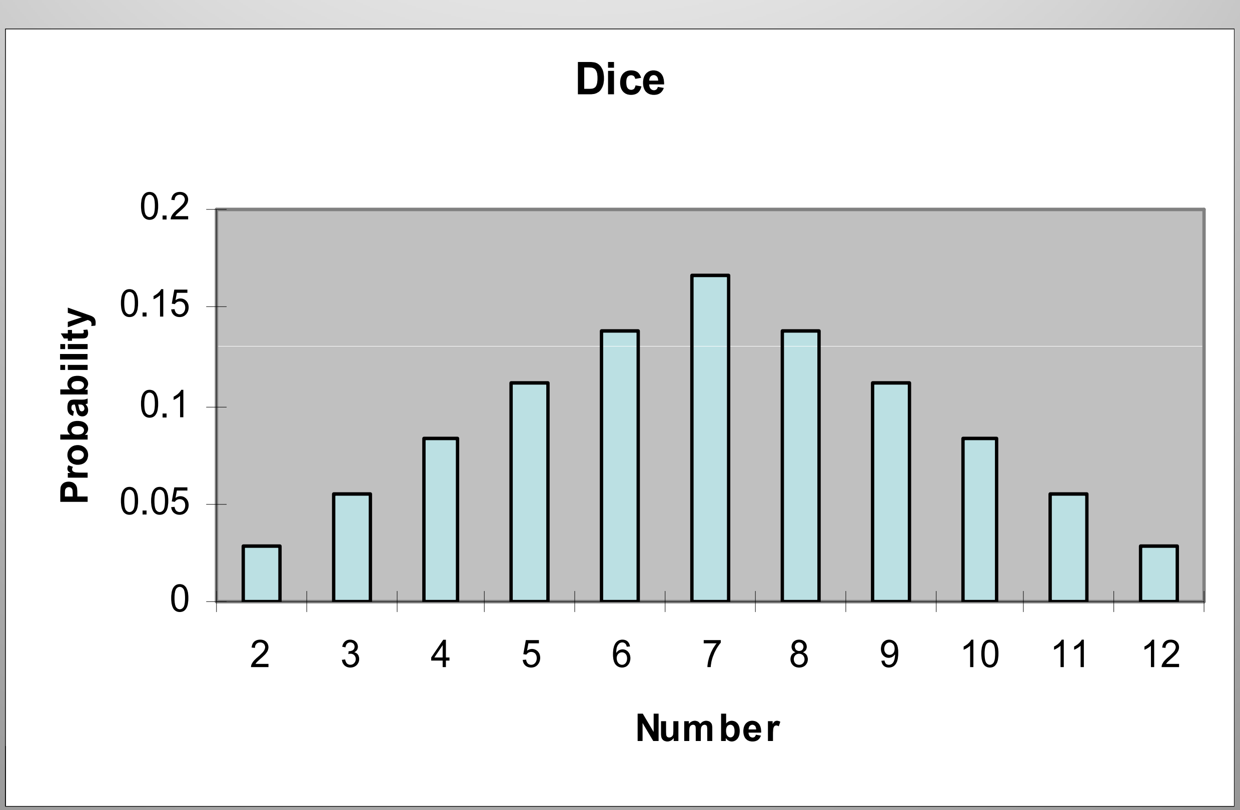
Pair of Dice *(Contd..)*

Probabilities for Two Dice

Total	2	3	4	5	6	7	8	9	10	11	12
Prob.	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$
%	2.8	5.6	8.3	11	14	17	14	11	8.3	5.6	2.8

Pair of Dice *(Contd..)*

Probabilities for Two Dice



Microstates and Macrostates

- Each possible outcome is called a “microstate”.
- The combination of all microstates that give the same number of spots is called a “macrostate”.
- The macrostate that contains the most microstates is the most probable to occur.

Combining Probabilities

•If a given outcome can be reached in two (or more) mutually exclusive ways whose probabilities are p_A and p_B, then the probability of that outcome is: $p_A + p_B$.

This is the probability of having either *A* or *B*.

•If a given outcome represents the combination of two independent events, whose individual probabilities are p_A and p_B, then the probability of that outcome is: $p_A \times p_B$.

This is the probability of having both *A* and *B*.

Combining Probabilities

Example

- Paint two faces of a die red. When the die is thrown, what is the probability of a red face coming up?

$$p = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

- Throw two normal dice. What is the probability of two sixes coming up?

$$p(2) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$

Complications

- *p* is the probability of success. (1/6 for one die)
- *q* is the probability of failure. (5/6 for one die)

$$p + q = 1, \quad \text{or} \quad q = 1 - p$$

- When two dice are thrown, what is the probability of getting only one six?

Probability of the six on the first die and not the second is:

$$pq = \frac{1}{6} \times \frac{5}{6} = \frac{5}{36}$$

Probability of the six on the second die and not the first is the same, so:

$$p(1) = 2pq = \frac{10}{36} = \frac{5}{18}$$

Complications *(Contd..)*

- Probability of no sixes coming up is:

$$p(0) = qq = \frac{5}{6} \times \frac{5}{6} = \frac{25}{36}$$

- Probability of two sixes coming up is

$$p(2) = pp = \frac{5}{6} \times \frac{5}{6} = \frac{25}{36}$$

•The sum of all three probabilities is:

$$p(2) + p(1) + p(0) = 1$$

$$p(2) + p(1) + p(0) = 1$$

$$p^2 + 2pq + q^2 =1$$

$$(p + q)^2 = 1$$

The exponent is the number of dice (or tries).Is this general?

Three Dice

$$(p + q)^3 = 1$$

$$p^3 + 3p^2q + 3pq^2 + q^3 = 1$$

$$p(3) + p(2) + p(1) + p(0) = 1$$

It works! It must be general!

$$(p + q)^N = 1$$

Binomial Distribution

Probability of *n* successes in *N* attempts

$$(p+q)^N = 1$$

$$P(n) = \frac{N!}{n!(N-n)!} p^n q^{N-n}$$

where, $q = 1 - p$.

Mean of Binomial Distribution

average of a set of values, or distribution

$$\overline{n} = \sum_n P(n)n$$

where

$$P(n) = \frac{N!}{n!(N-n)!} p^n q^{N-n}$$

$$\text{Notice}: p \frac{\partial}{\partial p} P(n) = P(n)n$$

Mean of Binomial Distribution *(Contd..)*

$$\overline{n} = \sum_n P(n)n = \sum_n p\frac{\partial}{\partial p}P(n)$$

$$\overline{n} = p\frac{\partial}{\partial p}\sum_n P(n) = p\frac{\partial}{\partial p}(p+q)^N$$

$$\overline{n} = pN(p+q)^{N-1} = pN(1)^{N-1}$$

$$\overline{n} = pN$$

Standard Deviation (*s*)

Standard deviation is a measure of the variability or dispersion

$$\sigma = \sqrt{\overline{\left(n - \bar{n}\right)^2}}$$

$$\sigma^2 = \overline{\left(n - \bar{n}\right)^2} = \sum_n P(n)\left(n - \bar{n}\right)^2$$

$$\overline{\left(n - \bar{n}\right)^2} = \overline{n^2 - 2n\bar{n} + \bar{n}^2} = \overline{n^2} - 2\bar{n}\bar{n} + \bar{n}^2$$

$$\sigma^2 = \overline{n^2} - \bar{n}^2$$

Standard Deviation (*s*) *(Contd..)*

$$\overline{n^2} = \sum_n P(n) n^2 = \left(p\frac{\partial}{\partial p}\right)^2 \sum_n P(n)$$

$$\overline{n^2} = \left(p\frac{\partial}{\partial p}\right)\left(p\frac{\partial}{\partial p}\right)(p+q)^N = \left(p\frac{\partial}{\partial p}\right) pN(p+q)^{N-1}$$

$$\overline{n^2} = p\left[N(p+q)^{N-1} + pN(N-1)(p+q)^{N-2}\right]$$

$$\overline{n^2} = pN[1+pN-p] = pN[q+pN]$$

$$\sigma^2 = \overline{n^2} - \bar{n}^2$$

$$\sigma^2 = pN[q+pN] - (pN)^2$$

$$\sigma^2 = Npq + (pN)^2 - (pN)^2 = Npq$$

$$\sigma = \sqrt{Npq}$$

For a Binomial Distribution

$$\overline{n} = pN$$

$$\sigma = \sqrt{Npq}$$

$$\frac{\sigma}{\overline{n}} = \sqrt{\frac{q}{Np}}$$

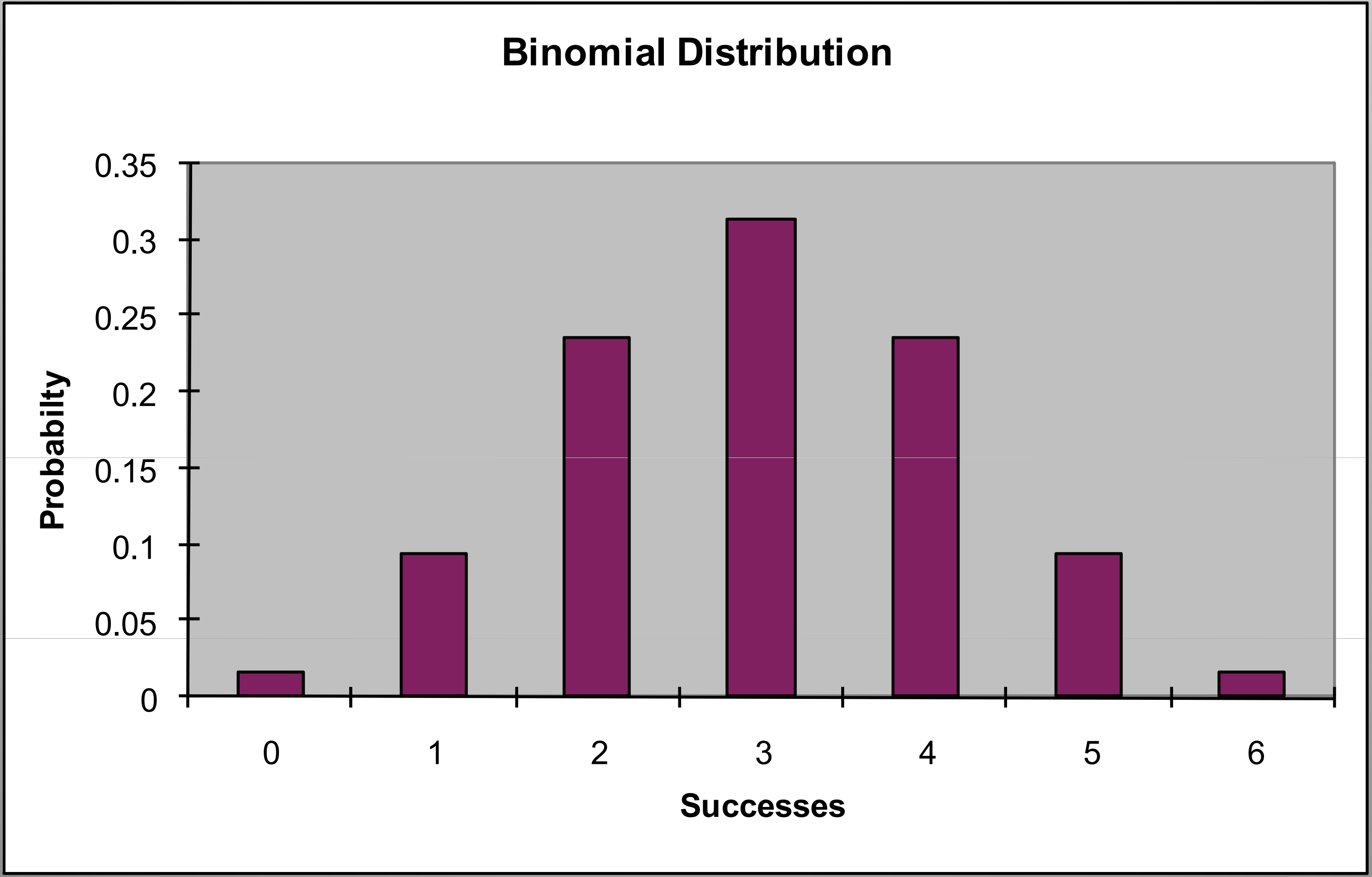
Coins

Toss 6 coins. Probability of *n* heads:

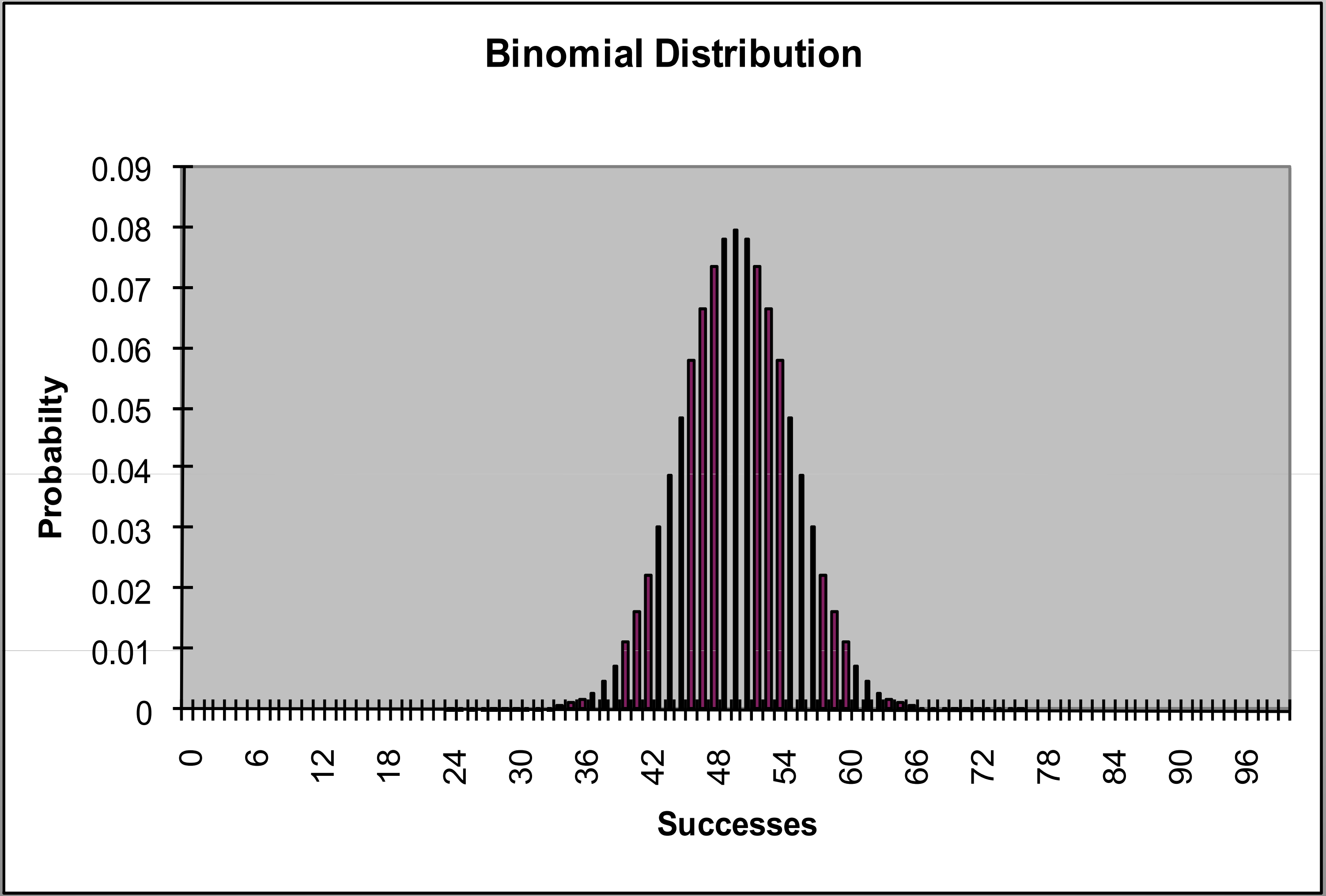
$$P(n) = \frac{N!}{n!(N-n)!} p^n q^{N-n} = \frac{6!}{n!(6-n)!} \left(\frac{1}{2}\right)^n \left(\frac{1}{2}\right)^{6-n}$$

$$P(n) = \frac{6!}{n!(6-n)!} \left(\frac{1}{2}\right)^6$$

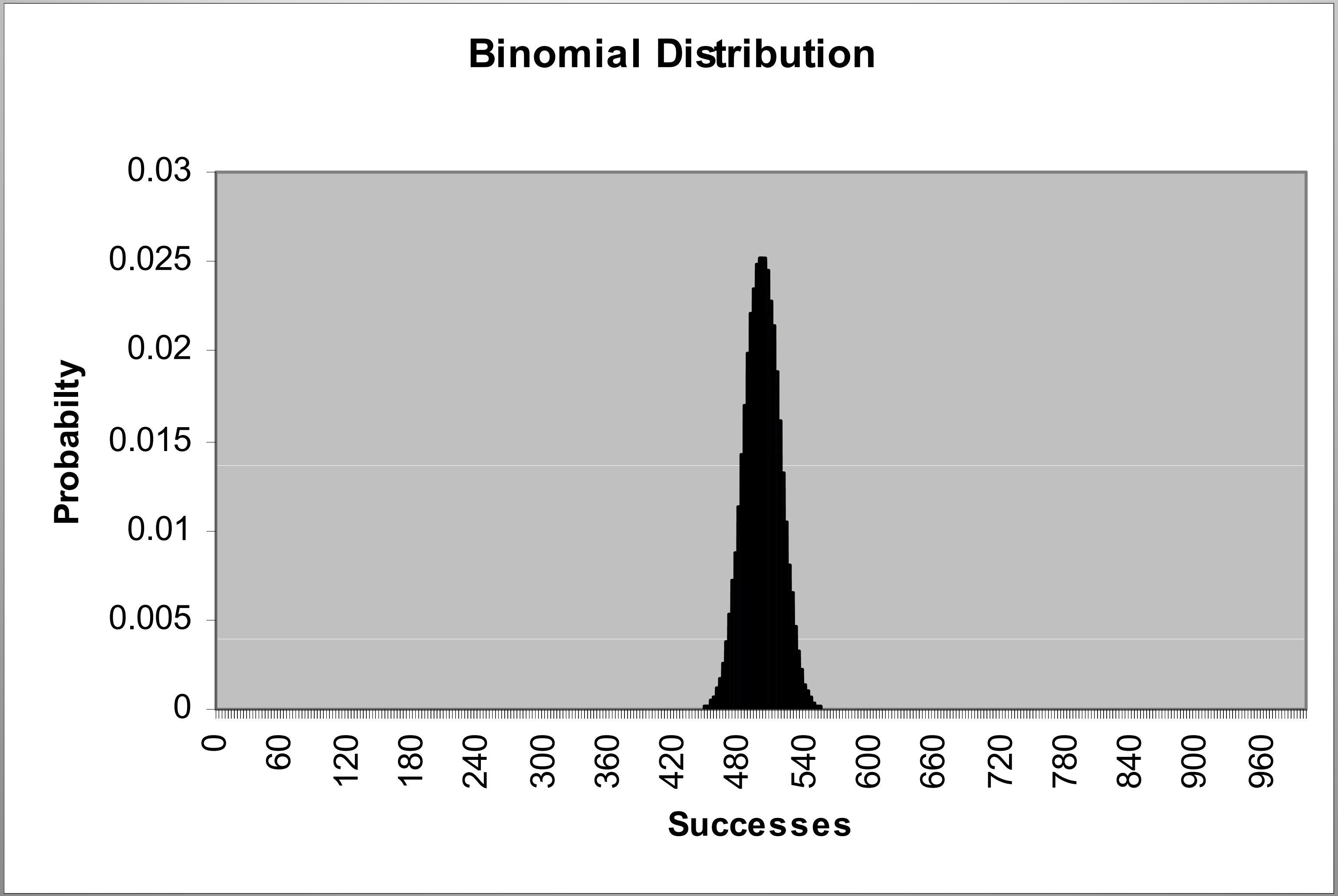
For Six Coins



For 100 Coins



For 1000 Coins



Continuous variables

- If x is discrete, $P(x_i)$ then gives the frequency at each point x_i. If x is continuous, this interpretation is not possible and only probability of finding x in finite intervals (x and $x+dx$) have meaning. The distribution $P(x)$ is then continuous density such that the probability is $P(x)dx$.
- Very often it is desired to know the probability of finding x between certain limits $P(x_1 \leq x \leq x_2)$. This is given by the cumulative or integral distribution.

$$P(x_1 \leq x \leq x_2) = \int_{x_1}^{x_2} P(x)dx$$

P(x)is continuous

$$P(x_1 \leq x \leq x_2) = \sum_{i=1}^{2} P(x_i)$$

P(x) is discrete

By convention the probability distribution is normalized to 1

$$\int P(x)dx = 1$$

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

Poisson Distribution

The ***Poisson distribution*** occurs as the limiting form of the binomial distribution when

$$N \longrightarrow \infty$$

$$p \longrightarrow 0$$

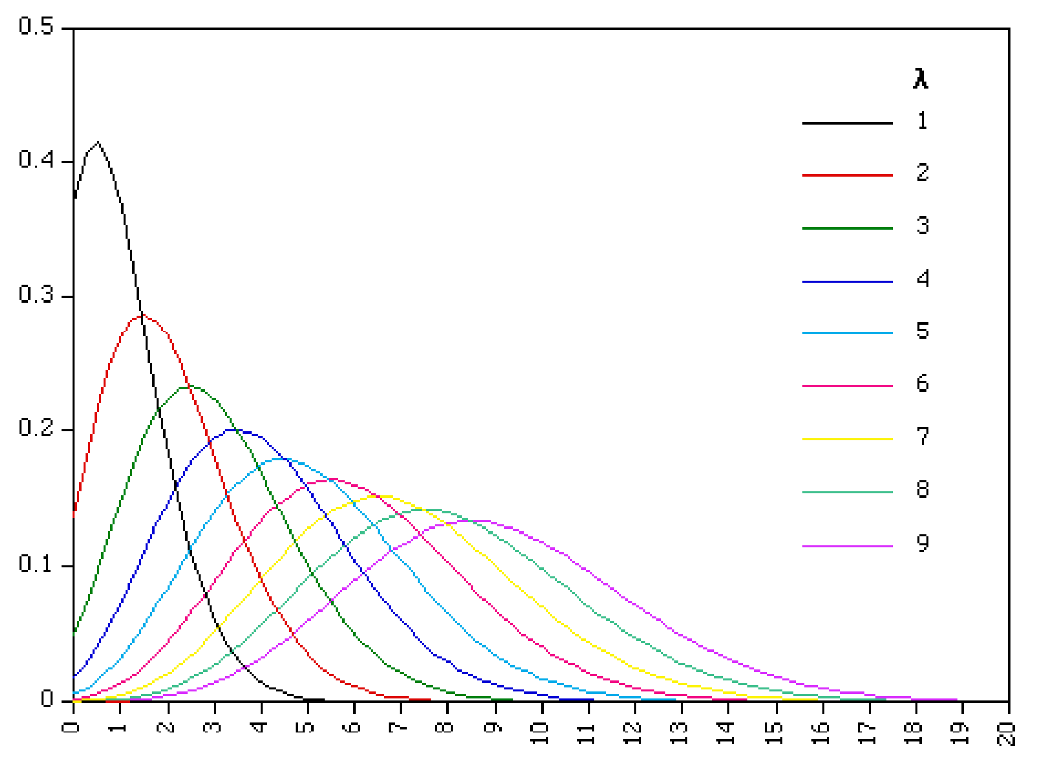
Such as $Np = constant = \mu t = \lambda$ remains finite. Then

$$P_r(t) = \frac{(\mu t)^r}{r!} e^{-\mu t} = \frac{\lambda^r}{r!} e^{-\lambda}$$

This is the probability of observing *r* independent events in a time interval *t*, when the counting rate is μ and the expected number of events in the time interval is λ.

Poisson Distribution *(Contd..)*

- The Poisson distribution is discrete. It essentially describe processes for which the single trial probability of success is very small but in which the number of trials is so large that there is nevertheless a reasonable rate of event.



Two important examples of such processes are radioactive decay and particle reaction.

Poisson Distribution *(Contd..)*

Example

Consider a typical radioactive source such as ^{137}Cs which has a half-life of 27 years. The probability per unit time for a single nucleus to decay is then

$\lambda = ln2/27 = 0.026/year = 8.2 \times 10^{-10} s^{-1}$

However, ever a 1 μg sample of ^{137}Cs will contain about 10^{15} nuclei. Since each nucleus constitutes a *trial*, the mean number of decays from the sample will be

$\mu = Np = 8.2 \times 10^{5}$ decays/s

This satisfies the limiting conditions describing above, so that the probability of observing *r* decays is given by formula for Poisson distribution.

Poisson Distribution *(Contd..)*

It is important to remember that if the rate of the basic process changes (as a function of time or of position), then the observed distribution of events may not follow the Poisson distribution.

Example

The number of people who die while operating computers each year is not Poisson distribution since although the probability of dying may remain constant, the number of people who operate computers increases from year to year.

•An important feature of the Poisson distribution is that it depends on only *one parameter*: μ

We also can find that $\sigma^2 = \mu$

that is the variance of the Poisson distribution is equal to the mean. The standard deviation is then

$$\sigma = \sqrt{\mu}$$

Gaussian or Normal Distribution

The ***Gaussian or normal distribution*** plays a central role in all of statistics and is the most ubiquitous distribution in all the sciences. Measurement errors and instrumental errors are generally described by this probability distribution.

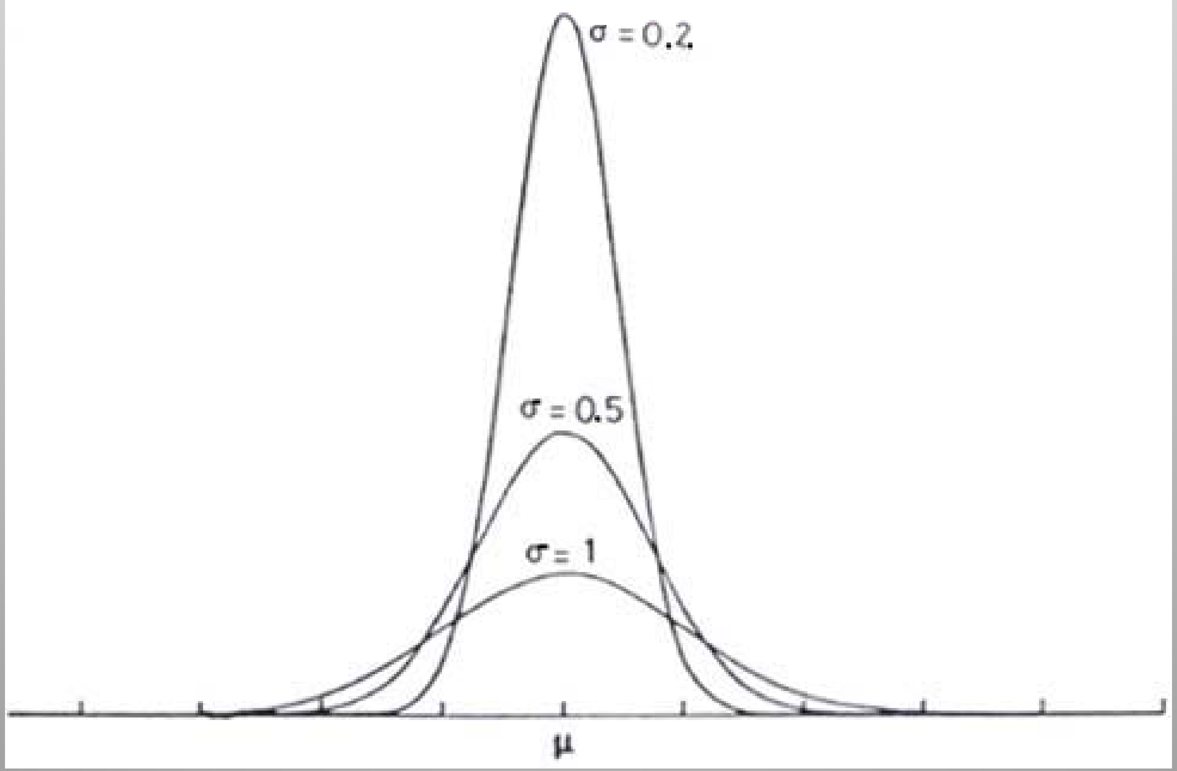
The Gaussian is a continuous, symmetric distribution whose density is given by

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right)$$

The two parameters μ and σ^2 can be shown to correspond to the mean and variance of the distribution.

Gaussian or Normal Distribution *(Contd..)*

- The Gaussian distribution for various σ. The significance of σ. A measure of the distribution width is clearly seen.



The standard deviation corresponds to the half width of the peak at about 60% of the full height. In some applications, however, the full width at half maximum (*FWHM*) is often used instead.

Gaussian or Normal Distribution *(Contd..)*

- Probability (*P*) of *y* being in the range [*a*, *b*] is given by an integral

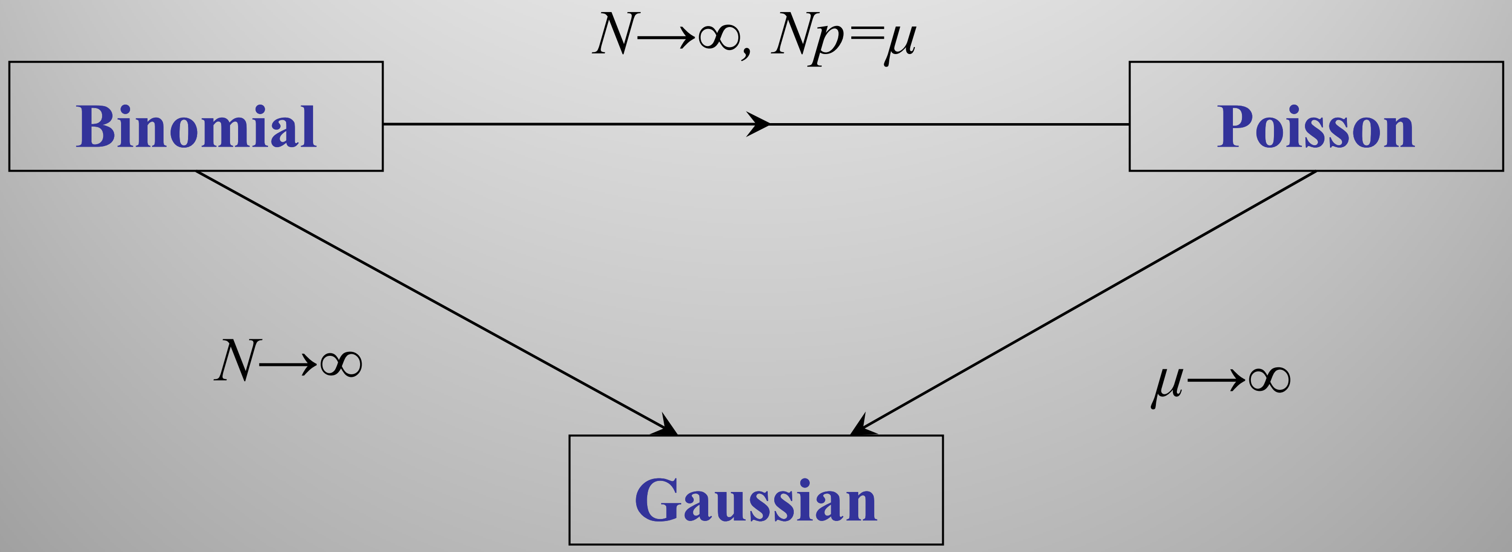
$$P(a < y < b) = \int_a^b p(y)dy = \frac{1}{\sigma\sqrt{2\pi}} \int_a^b e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy$$

• The integral for arbitrary *a* and *b* cannot be evaluated analytically. The value of the integral has to be looked up in a table

•The total area under the curve is normalized to one by the σ√(2π) factor.

$$P(-\infty < y < \infty) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy = 1$$

The relationship among the basic distributions



$N\rightarrow\infty,\ Np=\mu$

Binomial

Poisson

$N\rightarrow\infty$

$\mu\rightarrow\infty$

Gaussian

THANK YOU