4.1 Basic Concepts of Statistics

Population: entire list of a subset group.

Sample: a subset of a population

Discrete: finite or numeral set

Continuous: interval

As far as sampling is concerned, it is very crucial to select a sample which is not biased. There are several sampling techniques which face this bias.

Suppose that we have a population of 100,000 people and wish to select a sample of 1000 people. If we select the first 1000 in a list, or the youngest 1000 there is certainly a bias in our selection.

Simple Random: Cada miembro tiene igual probabilidad.

Systematic: empezar a un punto al azar y luego tomar cada persona.

Stratified: divide population into subgroups & picking a sample from each.

Quota: taking sample that is convenient for you; ex. students in your class.

There are advantages and disadvantages in each method. Simple random sampling is fair but it may be very time consuming compared to the systematic sampling. In systematic sample though, if there is a periodic pattern in the population there may be a bias. Suppose that the 100000 are in groups of 100 people. If the first person of the group is the leader, then the sampling method of selecting every 100^{th} person may provide a sample of only leaders or no leaders at all.

4.2 Measures of Central Tendency and Spread

Example 1

Find the mean, median, and mode of the following sets of data:

a. 10, 20, 20, 20, 30, 30, 40, 50, 70, 70, 80

1 2 3 4 Median 4 3 2 1

$S = 440$

20 appears the most times, so it's the mode.

Mean: $\frac{440}{11} = 40$

b. 10, 100, 20, 30, 90, 80, 20, 70, 50, 60

10 20 20 30 50 60 70 80 90 100

$S = 530$

• Sort the numbers before taking the median.

Mean: $\frac{530}{10} = 53$ Mode: 20

Median: $\frac{50+60}{2} = 55$

Example 2

a. Find the integers $a \le b \le c$, given that the mean = 4, mode = 5, median = 5.

Order from least to greatest

$b = 5$ $c = 5$

$a = 5$ would say that mean > 5.

$a \le 5 \le c$

Mean needs to be smaller, so c must be 5 for mode.

$\frac{a+5+5}{3} = 4$

$a + 10 = 12$ $a = 2$

b. Find the integers $a \le b \le c \le d$, given that the mean = 5, mode = 7, median = 6.

1, 5, 7, 7

$a = 1$ $c = 7$ $d = 7$

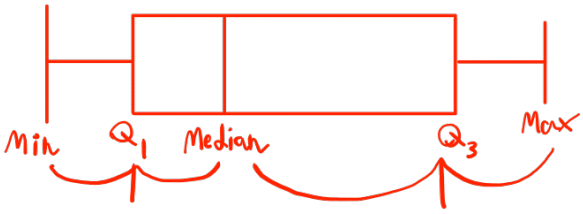
$\frac{b+7}{2} = 6$; $b = 5$

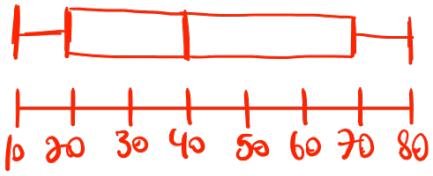
Example 3

a. Create a box and whisker plot for the following data:

Min Q_1 Median Q_3 Max

10, 20, 20, 20, 30, 30, 40, 50, 70, 70, 80





$IQR = Q_3 - Q_1$

b. Determine if there are any outliers in the data set.

$IQR = 70 - 20 = 50$

Lower outliers:	Upper outliers:
$Q_1 - 1.5(IQR)$	$Q_3 + 1.5(IQR)$
$20 - 1.5(50)$	$70 + 1.5(50)$
$20 - 75$	$70 + 75$
$= -55$	$= 145$

• No numbers are below the lower bound and none are above the upper bound, therefore there are no outliers in this dataset.

If our data are $x_1, x_2, \ldots, x_n$

the **variance** is given by (average) $\sigma^2 = \frac{\sum(x_i - \mu)^2}{n}$ $(x_1-\mu)^2+(x_2-\mu)^2+(x_3-\mu)^2+\ldots$

(is the square of)

the **standard deviation** is given by $\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{n}}$

c. Determine the standard deviation and variance for the data.

4.3 Frequency Tables – Grouped Data

Example 1

Find the mean, median, and mode of the following sets of data:

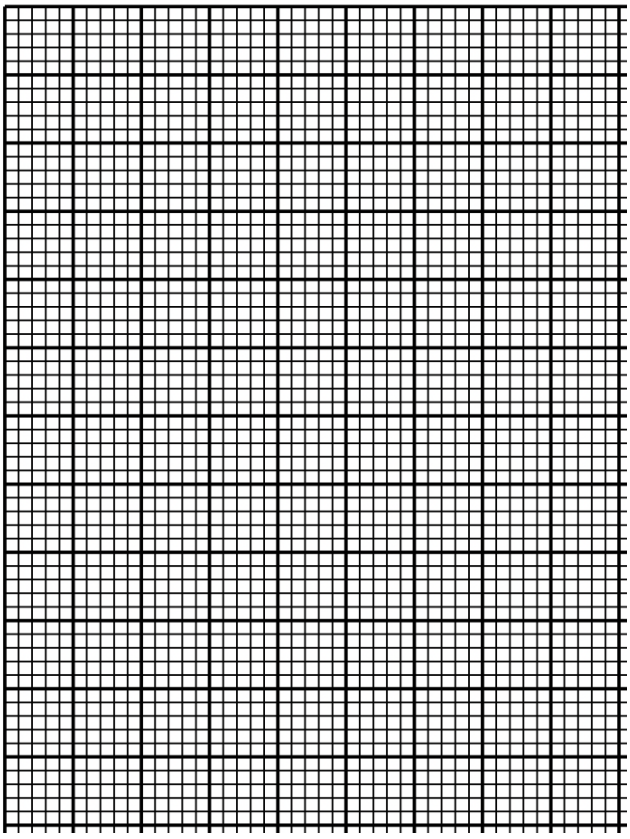
Data x	Frequency f
10	1
20	3
30	2
40	1
50	1
70	2
80	1
	n=11

Example 2

Suppose that 100 students took an exam and obtained scores from 1 to 60, according to the following table:

Score (x)	No of students (frequency f)	Cumulative frequency (cf)
$0<x\leq10$	8	8
$10<x\leq20$	12	20
$20<x\leq30$	10	30
$30<x\leq40$	25	55
$40<x\leq50$	35	90
$50<x\leq60$	10	100

Find the mean, median, mode, standard deviation, variance, and create a box and whisker plot for this data.



4.4 Regression

♦ CHARACTERISTICS OF THE REGRESSION LINE $y=ax+b$

The regression line

- passes through the point $M(\bar{x},\bar{y})$, where

  $\bar{x}$ = the mean of the values of x

  $\bar{y}$ = the mean of the values of y

- separates the points in (almost) two halves: half of the points are above and half below the line.

♦ CHARACTERISTICS OF THE CORRELATION COEFFICIENT r

The correlation between x and y is characterised according to the value of r as follows:

-1	-0.75	-0.5	-0.25	0	0.25	0.5	0.75	1
strong negative correlation		moderate negative correlation		weak negative correlation	very weak or no correlation	weak positive correlation	moderate positive correlation	strong positive correlation

Example 1

Consider the following data:

x	1	2	3	4
y	2	3	7	8

a. Find the correlation coefficient, r, and describe the relationship between x and y.

b. Find the equation $y = ax + b$ of the regression for y on x.

c. Find the equation $x = cy + d$ of the regression line for x on y.

d. Is the equation from part c the inverse of the equation in part b.

Example 2

Gary has found a job in the city and now must find an apartment in which to live. He surveys the monthly rent of several places and their distance from the city center.

Distance(km)	3	6	10	12	15	20
Monthly rent (thousands of rupees)	60	45	32	28	18	15

a. Find the equation $y = ax + b$ of the regression for y on x.

b. Estimate the cost of an apartment which is 8 km from the city center.

c. Explain whether you can use your line of best fit to accurately calculate the cost of an apartment 30 km from the city center.

4.5 Elementary Set Theory

♦ *BASIC NOTIONS*

In elementary set theory, a ***set*** *is just a collection of objects (or* ***elements****). It is usually denoted by a capital letter. For example,*

R = the set of real numbers
Q = the set of rational numbers

When listed, the ***elements*** *of a set are separated by commas "," and included between the symbols { and }. For example,*

$N = \{0,1,2,3,4,\ldots\}$ (i.e. the set of natural numbers)
$Z = \{\ldots,-3,-2,-1,0,1,2,3,\ldots\}$ (i.e. the set of all integers)

or less popular sets, such as

$A = \{1,2,3\}$ (it contains only 3 elements)
$B = \{a,b,c,d\}$ (it contains 4 letters)
$C = \{Chris, Mary, Tom\}$ (it contains 3 names)
etc

To declare that the element a is contained in set B we write

$$a \in B$$

To declare that the element f is not contained in set B we write

$$f \notin B$$

The most trivial set is the ***empty set****. It contains no elements, it is denoted by { } or by the symbol ∅.*

Example 1:

Consider the set $A = \{1, 2, 3\}$.

What are all of the subsets of set A?

Example 2:

Consider the sets $A = \{1, 2, 3\}$ and $B = \{1, 2\}$.

What can you say about the relationship between these sets?

Do not forget that always

$\emptyset \subseteq A$ (The empty set is a subset of any set)

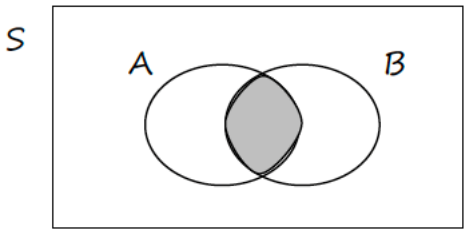
$A \subseteq A$ (Any set is a subset of itself)

Example 3:

Create a Venn diagram for the universal set $S = \{a, b, c, d, e, f, g, h, i, j\}$, set $A = \{a, b, c, d, e\}$, and set $B = \{d, e, f, g\}$.

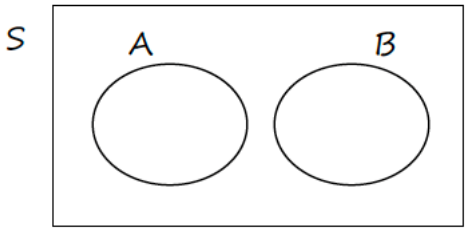
♦ THE INTERSECTION OF A AND B: $A \cap B$ **(A and B)**

It contains the common elements of A and B.



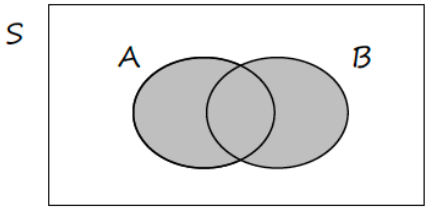
♦ MUTUALLY EXCLUSIVE SETS

If $A \cap B = \varnothing$, then $n(A \cap B) = 0$



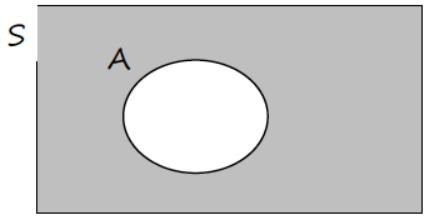
♦ THE UNION OF A AND B: $A \cup B$ **(A or B)**

It contains all the elements that are either in A or in B.



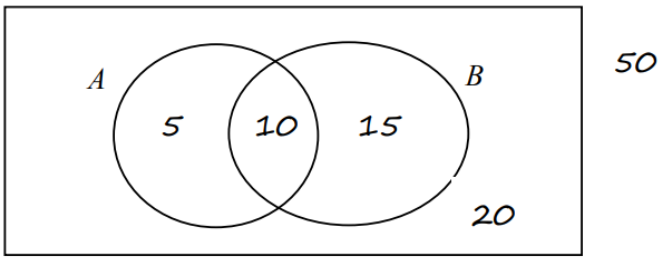
♦ THE COMPLEMENT OF A: A′ **(not A)**

It contains the elements that are not in A.



Example 4:

The following Venn diagram shows the sample space U and the event A and B together with the numbers of elements in the corresponding regions.



Complete the table.

$n(A)$		$n(B)$		$n(A \cap B)$	
$n(A')$		$n(B')$		$n(A \cup B)$	
$n(A' \cap B)$		$n(A \cap B')$		$n(A' \cap B')$	
$n(A' \cup B)$		$n(A \cup B')$		$n(A' \cup B')$	

4.6 Probability

Example 1:

Given that $P(A) = 0.5, P(B) = 0.3, P(A \cup B) = 0.6$, construct a Venn diagram.

Now, write down the following probabilities.

$P(A \cap B') =$	$P(A' \cap B) =$	$P(A' \cap B') =$
$P(A \cup B') =$	$P(A' \cup B) =$	$P(A' \cup B') =$

Example 2:

Consider the following group of 200 people.

	male	female
smoker	40	20
non-smoker	80	60

Determine the probability that if we select a person at random the probability that this person is:

a. male

b. female

c. smoker

d. non-smoker

e. male and smoker

f. male or smoker

Example 3:

Make a table of all the possible outcomes for tossing two dice.

Now find the following probabilities:

a. P (two sixes)

b. P (at least one six)

c. P (exactly one six)

d. P (same score)

e. P (sum of scores = 9)

f. P (sum of scores > 9)

g. P (sum of scores < 9)

4.7 Conditional Probability – Independent Events

Notice the following difference in notation

$P(A)$ means "probability of A"

$P(A|B)$ means "probability of A, given B"

Intuitively, we expect that

"the probability that it will rain in some day"

is different than

"the probability that it will rain in some day, given that this is a day of September"

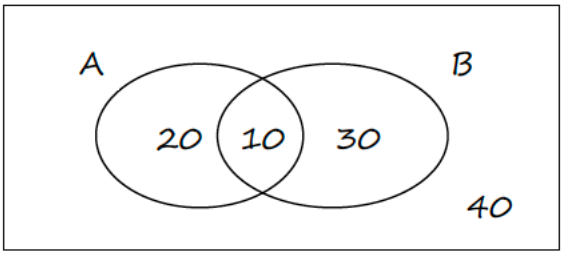
- *FORMAL DEFINITION OF* $P(A|B)$

The conditional probability is given by the formula

$$P(A|B) = \frac{n(A \cap B)}{n(B)} \quad \text{or} \quad P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Example 1:

Consider the Venn diagram:



Determine the following probabilities.

a. P (A | B)

b. P (B | A)

c. P (A' | B)

d. P (A | B')

e. P (A' | B')

Example 2:

Consider the table:

	male	female
smoker	40	20
non-smoker	80	60

Determine the following probabilities.

a. P (smoker | male)

b. P (female | smoker)

c. P (non-smoker | female)

- Many students confuse the terms

  Mutually exclusive events and **Independent events**

  Remember

  Mutually exclusive events means $A \cap B = \varnothing$

  Independent events means $P(A \cap B) = P(A) \cdot P(B)$

- Mind that

  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ holds in general

  $P(A \cap B) = P(A) \cdot P(B)$ holds for independent events

  In particular for independent events, it is sometimes useful to combine these two formulas in the following one

  $$P(A \cup B) = P(A) + P(B) - P(A) \cdot P(B)$$

- Sometimes we know beforehand that two events are independent. Thus, for their combination we can apply the formula $P(A \cap B) = P(A) \cdot P(B)$

Example 3:

Let P(A) = 0.4 and P(B) = 0.3. Find P(A ∪ B) in the following cases:

a) A and B are mutually exclusive

b) A and B are independent

c) P (A ∪ B) = 0.2

d) P (A | B) = 0.2

Example 4:

Let A and B be independent events with P (A) = 0.4 and P (A ∪ B) = 0.7. Find P (B).

4.8 Tree Diagrams

Example 1:

We throw a die.

If we get 1, we stop.

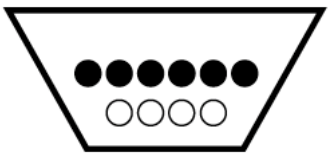
If we get 2, 3, 4 or 5, we toss a coin.

If we get 6, we toss two coins.

Find the probability that only one head is obtained.

Example 2:

A box contains 10 balls: 6 BLACK and 4 WHITE:



We select two balls, one after the other. Determine the following probabilities.

a. P (both black) b. P (only one black) c. P (same color)

Now, we select 3 balls. Determine the following probabilities.

a. P (all Black) b. P (only one black)

Example 3:

In a private school party, 30% of the students wear RED suits, 20% wear GREEN suits and 50% wear BLUE suits. 25% of the RED students, 35% of the GREEN students and 45% of the BLUE students are MALE. Find the probability that a MALE student wears GREEN suit, that is P(GREEN | MALE).

Reverse Given – Bayes Theorem (HL only)

$$P(B \mid A) = \frac{P(B)P(A \mid B)}{P(B)P(A \mid B) + P(B')P(A \mid B')}$$

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

4.9 Distributions – Discrete Random Variables

Example 1:

Consider the table:

x	10	20	30
P(X=x)	a	b	0.5

Given that E(X) = 23, find the values of a and b.

Now, we select one of the numbers 10, 20, 30 at random.

If we select 10, we earn 6 points.

If we select 20, we earn 1 point.

If we select 30, we lose 2 points.

What is the expected number of points in one game?

Example 2:

We throw two dice.

If we obtain TWO SIXES, we earn 15€.

If we obtain ONLY ONE SIX, we earn 1€.

If we obtain NO SIX, we lose 1€.

Find the expected profit in one game.

Example 3:

Consider again:

x	10	20	30
P(X=x)	0.2	0.3	0.5

Determine each of the following.

a. Mode

b. Median

c. Variance

4.10 Binomial Distribution

It is the distribution of a discrete random variable X which takes on the values

$$0, 1, 2, 3, 4, \ldots, n$$

with probability function

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad x = 0,1,2,3,\ldots,n$$

where n, p are two parameters. We will see that the binomial distribution describes a certain type of problems.

Notice: the formula **is not** in the syllabus; Results will be obtained directly by GDC. It is worth to mention though how it works!

♦ DESCRIPTION OF THE PROBLEM

We deal with a game (or any experiment) with two outcomes

SUCCESS with probability p
FAILURE (with probability 1-p)

We play the game n times. Our parameters are

n = number of trials
p = probability of success

while

X counts the number of (possible) successes

We say that X follows a binomial distribution and write $X \sim B(n,p)$.

Since n is the number of trials, X can take on the values

$$0, 1, 2, 3, 4, \ldots, n$$

The probabilities P(X=0), P(X=1), P(X=2), etc can be obtained by the GDC.

(and also by the formula mentioned in the introduction, but as we have said this formula is not in the syllabus).

Example 1:

We toss a die 5 times. The success is to get a six. Create a probability distribution table for this situation.

Now, find the probability of getting:

a. exactly 3 sixes

b. at most 3 sixes

c. less than 3 sixes

d. more than 3 sixes

d. at least 3 sixes

What is the expected value and variance of X?

Example 2:

A box contains 5 balls, 1 BLACK and 4 WHITE. We win if we select a BLACK ball. We play this game 10 times.

Find:

(a) The probability to win exactly 4 times

(b) The probability to win at most 4 times

(c) The probability to win at least once

(d) The expected number of winning games.

(e) The variance of the number of winning games.

Example 3:

Let p = 0.2 and n unknown. It is given that P(X = 1) = 0.268. Find n.

Example 4:

Determine the mode of each of the following situations.

a. n = 20, p = $\frac{1}{6}$

b. n = 5, p = $\frac{1}{6}$

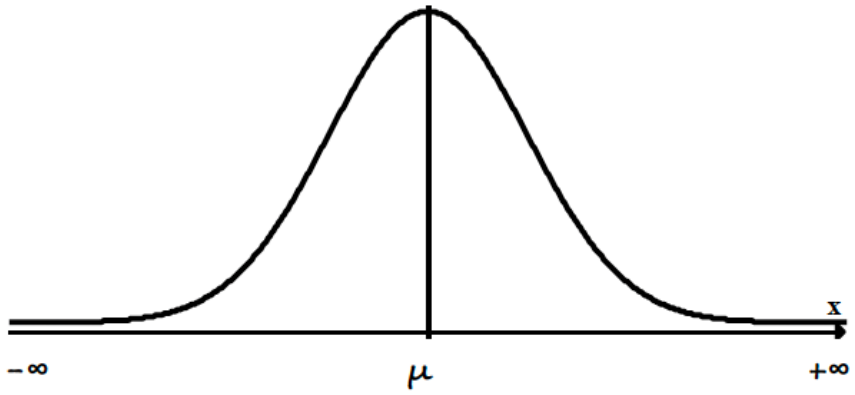
4.11 Normal Distribution

It is the distribution of a **continuous random variable** X with values form $-\infty$ to $+\infty$. The parameters of this distribution are

μ = mean

σ = standard deviation.

The "behavior" of the probability is described by a function which looks like



Roughly speaking, there is a highly likely mean value μ and all the other values of X spread out symmetrically about the mean. As we move away from the mean (either to the left or to the right of the mean) the probability decreases dramatically!

We say that X follows a normal distribution with mean μ and standard deviation σ (or variance σ^2) and we write $X \sim N(\mu, \sigma^2)$.

NOTICE

- The whole area under the curve is 1 (i.e. 100%). The area before the mean as well as the area after the mean is 0.5 (i.e. 50%)
- Theoretically, the distribution of X ranges between $-\infty$ to $+\infty$. In practice, we may assume that almost the whole population (in fact 99,7%) ranges between $\mu-3\sigma$ and $\mu+3\sigma$.
- The standard deviation σ indicates the spread of the population.

Example 1:

The mass of packets for a certain type of coffee is normally distributed with a mean of 500 g and standard deviation of 15 g.

(a) Find the probability that a packet weighs more than 520 g.

(b) The lightest 4% of the packets weigh less than a. The heaviest 5% of the packets weigh more than b. Find a and b.

The packs in question (b) are rejected from the market.

(c) In a daily production of 1600 packs how many of them are expected to be rejected?

(d) We select 2 packs. Find the probability that both are rejected.

(e) We select 5 packs. Find the probability that at least one is rejected.

(f) Find Q1 and Q3, the lower and upper quartiles of the weights

Example 2:

For a random variable X we know that:

35% is less than 60

25% is more than 90

That is

$$P(X \leq 60) = 0.35, P(X \geq 90) = 0.25.$$

Find μ and σ.

4.12 Continuous Distributions in General (HL only)

In general, for **a continuous random variable X** with

probability density function (or **pdf**) $f(x)$

it holds

(i) $f(x) \geq 0$, i.e. the function is non-negative

(ii) $\int_{-\infty}^{+\infty} f(x)dx = 1$, i.e the total area under the curve is 1

while the probability that X takes values between a and b is

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

Notice that

$P(a \leq X \leq b)$ and $P(a < X < b)$

are exactly the same as the probability that X takes a particular value, say P(X=a) is zero!

- THE EXPECTED VALUE $\mu = E(X)$

The **mean** μ, or otherwise the **expected value E(X)** is defined by

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

- THE VARIANCE Var(X)

It is defined by

$$Var(X) = E(X-\mu)^2$$

that is

$$Var(X) = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)dx$$

An equivalent (and more practical) definition is

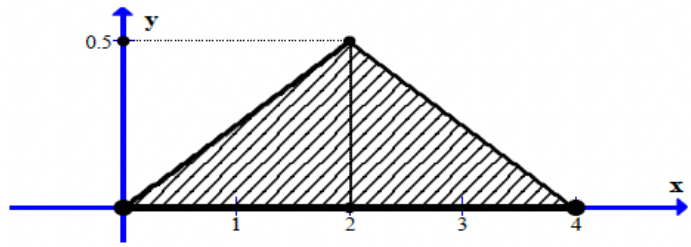
$$Var(X) = E(X^2) - \mu^2$$

where

$$E(X^2) = \int_{-\infty}^{+\infty} x^2 f(x)dx$$

Example 1:

Let X be a continuous random variable in [0,4] with pdf:



$$f(x)=\begin{cases}\frac{x}{4}, & 0\le x\le 2\\ 1-\frac{x}{4}, & 2\le x\le 4\end{cases}$$

a. Verify that $f(x)$ is a pdf.

b. Find:

i. expected value

ii. variance

iii. median

4.13 Counting – Permutations – Combinations (HL only)

Example 1:

Two dice are tossed. How many results are there?

Example 2:

The Latin alphabet has 26 letters. How many combinations of two letters are there if:

a) repetition of letters is allowed

b) no repetition is allowed (i.e. different letters)

Example 3:

Find the total number of words with 4 letters if our alphabet is

a) the alphabet of the 26 Lattin letters

b) only the letters A, B, C, D, E

c) the ten digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9

d) only the digits 0, 1 (known as binary alphabet)

Example 4:

A password has the form XXYYY where:

X is one of the 26 Latin letters

Y is one of the digits 0,1,2,3,4,5,6,7,8,9

What is the total number of possible passwords?

Example 5:

Three people, Alex, Bill, Chris must sit in three chairs in a row!

How many ways are there for them to be arranged?

COMBINATIONS	PERMUTATIONS
We do not mind about the order (The r objects are seen as a group)	We mind about the order (AB is different than BA)
nCr	**nPr**

Example 6:

Consider $n = 5$ objects, A, B, C, D, E. We select $r = 2$ out of them.

Find the number of:

a. combinations

b. permutations

Example 7:

There are 10 people in a room. We choose 3 people out of them. Find the number possible if we consider:

a) the 3 people as a group

b) If we arrange the 3 people in order

Example 8:

We choose 6 numbers out of 49. Determine the number of possible ways.

Example 9:

There are 26 Latin letters (A, B, C, …, Z).

a. Find the total number of 3 letter words that are possible.

b. Find:

i. how many of them consist of different letters.

ii. how many of them begin with A.

iii. how many of them do not begin with A.

iv. how many of them do not contain the letter F.

v. how many of them contain the letter F (at least once).

Example 10:

A school class consists of 30 students, 10 boys and 20 girls. We select a committee of 5 students.

Find the number of ways to select a committee of 5 students if:

a) the committee consists of 2 boys and 3 girls

b) the committee consists of boys only

c) the committee consists of students of the same gender

d) there are at most two boys in the committee