MA 112/227 PROBABILITY
LECTURE NOTES

Cathal Seoighe Room C204 (Áras De Brún), Cathal.Seoighe@nuigalway.ie

GENERAL INFORMATION

SYLLABUS

- The role of probability theory in modelling random phenomena and in statistical decision making
- sample spaces and events
- some basic probability formulae
- conditional probability and independence
- Bayes' formula
- counting techniques
- discrete and continuous random variables
- hypergeometric and binomial distributions
- Poisson distributions
- normal distributions
- the distribution of the sample mean when sampling from a normal distribution
- the Central Limit Theorem with applications including normal approximations to binomial distributions.

SAMPLE TEXTS (most of these texts include statistics as well as chapters on probability that are relevant to this course)

- Wackerly, Mendenhall, and Scheaffer *Mathematical statistics with applications*, *6th Ed.* Main Library 519.5 MEN
- Freund, John E. *Modern Elementary Statistics* 519.5 FRE
- Wonnacott, Thomas H. *Introductory Statistics* Main Library 519.5 WON
- Lipschutz, Seymour *Schaum's outline of theory and problems of introduction to probability and statistics.* Main Library 519.2 LIP

- Freund, John E. *John E. Freund's mathematical statistics : with applications*. 2004 (and other versions) Main Library 519.5 FRE (also versions with authors Miller, Irwin]
- Pestman, Wiebe R., *Mathematical statistics : an introduction* 1998 Main Library 519.5 PES Mendenhall, William *Mathematical statistics with applications* 1990 Main Library 519.5 MEN

SAMPLE ELECTRONIC RESOURCES:

Click on some of the On-Line texts at
`http://people.hofstra.edu/liora_p_schmelkin/weblink/methods.html`

A textbook is available online at
`http://www.dartmouth.edu/~chance/teaching_aids/books_articles/probability_book/amsbook.mac.pdf`. This is a relatively advanced textbook but it does cover the material in this course.

Lectures: Tuesdays at 11:10 am in AC 201 and Thursdays at 11:10 am in AC202.

Tutorials: TBA. Commencement date will be announced in class.

Assessment: Assessment will consist of homework, class test, and a final exam. Please attend all lectures

These notes contain a synopsis of the theory that will be covered in this course, followed by worked example questions.

Why study probability?

In the real world we often want to predict what will happen (e.g. what is the chance of 6 numbers right in the next draw of the Lotto; many applications in science, communications, finance, gambling, weather forecasting, and just about all fields of human endeavour). Probability is also the basis of all of statistics. It crops up pretty much everywhere from biology to games to psychology.

What is Probability?

There is still some debate about how to think about probability and the issues can become quite philosophical (what is it? how should we interpret it? is it subjective? etc.) Mathematically, the picture is clearer. Probabilities have to satisfy certain properties - e.g. a probabilty can never be a negative number. These properties are provided later, together with a brief overview of some of the different interpretations of probability.

Some definitions

Experiment: In probability, an experiment is anything that gives rise to a defined set of possible outcomes. E.g. toss a con once, roll a die twice, pick a person at random from the cancer unit of a hospital, measure the temperature tomorrow, count the number of cars that arrive on Campus between 9am and 9:01 am on a random day). By definition, an experiment is the process by which we obtain data, and is intended to include scientific experiments and observational studies; note that an experiment implies more control over extraneous variables than happens in a observational study.

Sample space: The sample space, Ω (or S), is the set of all possible outcomes of the experiment.

For example, if the 'experiment' is a single die toss the sample space is {1,2,3,4,5,6}

Event: A subset A of Ω is called an event. A singleton subset is referred to as an *elementary event* or *basic outcome* or *sample point*. An event is said to *occur* if the outcome of the experiment is a member of the set A.

For example, an even number coming up in a single die toss is an event. In this case A = {2,4,6}. An event can contain a single sample point, e.g. A = {1} is the event that a one comes up in a single die toss.

Probability: A probability function assigns a number (its *probability*) to each event A (i.e. it is a mapping from events to numbers on the real line). There

are some technical issues that have to do with what kinds of subsets can have probabilitites assigned to them, but these are beyond the scope of this course.

Back to the definition of probability

***Classical* or *"equally likely"* definition of probability** If Ω has a finite number of equally likely elements, then we compute $P(A)$ from the formula $P(A) = \frac{\#\text{ elements in } A}{\#\text{ elements in } \Omega}$.

***Empirical* or *statistical* or *relative frequency* definition of probability** $P(A)$ is defined as the limit, as $N \longrightarrow \infty$, of the proportion of times A will occur in N repetitions of the experiment.

***Logical* or *Bayesian* or *subjective* definition of probability** : Probability indicates the degree of plausibility of a proposition, given the available evidence. Emphasizes the conditional nature of probability (always depends on what you know). Seen as an extension of Aristotolean logic to accommodate uncertainty.

***Axiomatic* definition of probability**: As seen above, we can think of a probability as a mathematical function that maps the subsets of Ω to the real numbers. In order for this function to be called a *probability* function it has to satisfy certain criteria (given next).

Probability axioms (Kolmogorov)

A probability function on subsets of Ω must satisfy these conditions

1. $P(\Omega) = 1$, where Ω is the sample space.
2. $P(A) \geq 0$ for all A.
3. $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ if A_i, $i = 1, 2, ..., \infty$ are pairwise disjoint, i.e. mutually exclusive (i.e. $A_i \cap A_j = \emptyset$).

Tools for solving probability problems

Most basic problems in probability can be solved using a combination of the rules of probability and combinatorics (i.e. counting). These rules are listed below. In class we will derive most of the probability formulae from the axioms above.

Some useful formulae in probability

1. *[Addition formula for any two events A and B.]* $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

2. **If** A and B are disjoint then the last formula becomes $P(A \cup B) = P(A) + P(B)$.

3. *(Addition formula for any n events $A_1, A_2, ..., A_n$)*
$P(A_1 \cup A_2 \cup \ldots \cup A_n) = \sum_{i=1}^{n} P(A_i) - \sum_{i<j}^{n} P(A_i \cap A_j) + \sum_{i<j<l}^{n} P(A_i \cap A_j \cap A_l) - ... + (-1)^{n-1} P(\cap_{i=1}^{n} A_i)$.

4. **If** $A_1, A_2, ..., A_n$ are mutually disjoint then the last formula reduces to $P(A_1 \cup A_2 \cup ... \cup A_n) = \sum_{i=1}^{n} P(A_i)$, which is included in Formula 3 above (on putting $A_{n+1} = A_{n+2} = ... = \emptyset$.)

5. *[Multiplication formula for any two events A and B.]* $P(A \cap B) = P(A)P(B \mid A) = P(B)P(A \mid B)$.

6. *[Independence of two events A and B.]* A and B are *independent* if and only if $P(A \cap B) = P(A)P(B)$,or equivalently if and only if $P(A \mid B) = P(A)$ or equivalently if and only if $P(B \mid A) = P(B)$

7. The definition of the conditional probability (note the equivalence of this and the multiplication formula above): $P(B \mid A) = \dfrac{P(A \cap B)}{P(A)}$.

8. *[Multiplication formula for any n events $A_1, A_2, ..., A_n$.]*

$$P(\cap_{i=1}^{n} A_i) = P(A_1)P(A_2 \mid A_1) \ ... \ P(A_n \mid A_1 \cap A_2 \cap ... \cap A_{n-1}).$$

9. *[Independence of n events $A_1, A_2, ..., A_n$.]* $A_1, A_2, ..., A_n$ are (mutually) independent if for any combination $1 \leq i_1 < i_2 < ... < i_k \leq n$ ($k = 2, 3, ..., n$), we have $P(A_{i_1} \cap A_{i_2} \cap ... \cap A_{i_k}) = P(A_{i_1})P(A_{i_2})...P(A_{i_k})$. In particular, if $A_1, A_2, ..., A_n$ are independent then $P(\cap_{i=1}^{n} A_i) = P(A_1)P(A_2) \ ... \ P(A_n)$.

10. *[Partitioning/Marginal probability]* Let $A_1, A_2, ..., A_n$ be mutually disjoint, and also collectively exhaustive (i.e. $\cup_{k=1}^{n} A_k = \Omega$). Then for any event B, we have $P(B) = \sum_{k=1}^{n} P(B \cap A_k)$.

11. Let $A_1, A_2, ..., A_n$ be mutually disjoint and collectively exhaustive. Then for any event B satisfying $P(B) > 0$ and any $i = 1, 2, ..., k$, we have $P(A_i \mid B) = \dfrac{P(A_i)P(B \mid A_i)}{\sum_{k=1}^{n} P(A_k)P(B \mid A_k)}$.

Some useful formulae in combinatorics

1. If we have k sets $S_1, S_2, ..., S_k$ containing $n_1, n_2, ..., n_k$ elements respectively, then it is possible to form exactly $n_1 n_2 ... n_k$ ordered $k-$tuplets containing one element from each set. This is known as the *fundamental counting rule.*

2. n distinct objects can be arranged in a row in

$$n! := n(n-1)(n-2) \times ... \times (1) \text{ ways.}$$

3. Given n distinct objects, the number of distinct groups, each of size r, that can be taken when order within each selected group **is** important is $nPr := \dfrac{n!}{(n-r)!} = n(n-1)...(n-r+1)$. (*Applies to permutations of n objects taken r at a time.*)

4. Given n distinct objects, the number of distinct groups, each of size r, that can be taken when order within each selected group **is not** important is $nC_r := \binom{n}{r} = \dfrac{n!}{r!\,(n-r)!}$ (*Combinations of n objects taken r at a time.*)

   There are other interpretations of $\binom{n}{r}$. Indeed it also represents: **(b)** the number of ways of placing n (distinguishable) objects into 2 cells such that the first cell contains r objects and the other cell contains the remaining $n-r$ objects; **(c)** the number of ways of arranging n objects in a row when r of the objects are of one kind and the remaining $n-r$ are of a second kind; **(d)** the coefficient of x^r in the (binomial) expansion of $(1+x)^r$

5. (Multinomial coefficient formula)
   $\dfrac{n!}{r_1! r_2! ... r_k!}$

One interpretation: The # ways of distributing n (distinguishable) objects into k cells such that the ith cell contains r_i objects, $i = 1, 2, ...k$. Here of course $\sum_{i=1}^{k} r_i = n$; the order of the objects within each cell is *not* important; i.e. we do not get a different arrangement by permuting the objects in any given cell.

Another interpretation: The number of ways of arranging n objects in a row when r_1 of the objects are of one kind, r_2 are of a second kind, ..., r_k are of a kth kind. Again of course $\sum_{i=1}^{k} r_i = n$.

It is important to note that 5 is an extension of 4 to the case of more than two categories.

End of formulae

Discrete random variables

Often what we are interested in is some *number* which is *associated* with the outcome of the experiment rather than in the detailed outcome of the experiment. For example, when we flip a coin a number of times, we might not be so much interested in the sides that come up, but rather in the *number* of times each side arises. Similarly, when a person is picked at random, we may well not be interested in the person's name, but rather some *number* associated with him/her, such as height or weight or income. The fact that so often the outcomes in probability experiments are mapped to numbers leads to the concept of the *random variable*

A **random variable** X is a function that associates a number with each element of the sample space S of an experiment. An easier way to think about a this is that a random variable X represents a numerical quantity, the value of which depends on the outcome of the experiment.

Indeed, the outcomes of any experiment can be mapped to numbers (i.e. represented by a random variable). For example, if we pick a person at random and ask that person whether or not he favours a certain political issue. Let $X = 1$ or 0 according to whether the person favours or does not favour the issue. Here the sample space S has the two elements Favour and Oppose, and X can be exhibited as below:

ω		$X(\omega)$
Favour	$\longrightarrow$	1
Oppose	$\longrightarrow$	0.

X is a numerical-valued function and we cannot tell whether it will take the value 1 or the value 0 in any one repetition of the experiment, because we do not know the outcome of the experiment. Often we can forget about the fact that X is a function and just think of it as variable whose value is unknown and depends on the outcome of the experiment. Notice also that the X in our example is **discrete** (a discrete function can take a finite or countably infinite set of values). The underlying experiment here is called *Bernoulli* because it has exactly two possible outcomes (which in general are usually denoted 'success' and 'failure'). The random variable X that takes two values 0 and 1 is called a **Bernoulli random variable**, and is the simplest kind of random variable we encounter.

Definition: A *Bernoulli Random Variable* is a random variable with two possible outcomes (i.e. a binary random variable).

Definition*:* If X is a discrete random variable, the distribution of X, also called the probability mass function or the probability density of X is the function $f(x)$, such that $f(x) = P(X = x)$. I.e. it gives the probability with which each value occurs.

Note that in any discrete distribution, the probabilities are non-negative and must add to one; this is another definition of a distribution.

Mean and Variance

The *mean* and *variance* are often used to summarize a distribution.

Definition: Let X be a discrete random variable. The *mean value* of X or *expected value* of X,or *population mean* (sometimes loosely called the average value of X) is the centre of gravity of the distribution of X and is defined as

$$\mu := E(X) = \sum_x xP(X = x)$$

(that is, multiply each value x of X by its probability $P(X = x)$ and add).

The E here stands for Expected. Another word for the mean of the distribution is the expected value.

Definition: The *variance* of X, also called the *population variance*, is defined by:

$$Var(X) = \sum_x (x - \mu)^2 P(X = x)$$

Finally, the *standard deviation* of X, which is traditionally given the symbol σ, is simply $\sigma = \sqrt{Var(X)}$. The standard deviation rather than the variance of a distribution is often quoted and for this reason the variance is sometimes referred to as σ^2.

Interpretation of μ and σ
$E(X)$ (the mean of the distribution) is the 'long-term average' value of the random variable. For example, suppose you are involved in a work sweepstake. Each week you get €20 if you win and have a probability of 0.04 of winning and probability 0.96 of not winning. After a large number of weeks you would expect to have won about 4% of the time. The formula for $E(X)$ is logical: if you play the game many many times, you will win about 4% of the time and get €20 and lose about 96% of the time and get €0. If this went on for 1000 weeks you would expect to come away with about $1000 \times (0.04\times$ €20 $+0.96\times$ €0 $= 1000 \times 0.8$

Your long-term average (i.e. $E(X)$) winnings would be $\frac{1000\times 0.8}{1000} = 0.8$

In a real sense it is worth about 80 cents to you per week to play this sweepstake. Of course if you are paying €1 per week to play you can expect to make a net

loss in the long term.

Somewhat similarly, σ^2 tells you how large, on average, $(x-\mu)^2$ is (the formula for the variance just multiplies each value of $(x-\mu)^2$ by how likely that value is to occur and adds them all up). This tells you something about how far, on average, you are from the mean of the distribution.

For example, consider two students who take a course with continuous assessment. One student is eratic and only hands in some of the assignments, but the assignmetns completed are done well. The other student completes all the assignment but does a mediocre job each time. Suppose both students get a grade of 50% for the course. The marks of the first student would be expected to show a higher variance from the mean, reflecting his eratic behaviour.

Two important discrete distributions

The **binomial** and **hypergeometric** distributions both occur in problems involving the number of successes or failures in n Bernoulli trials (recall that a Bernoulli trial is an experiment that can have one of two possible outcomes).

Binomial distribution

Consider a random variable, X, which represents the number of successes we will obtain in n independent Bernoulli trials. If p denotes the probability of a success on any one trial then

$$P(X=x) = \binom{n}{x} p^x (1-p)^{n-x},\ x = 0, 1, 2, ..., n. \quad .$$

Formal definition: A random variable X has a **binomial** distribution with parameters n and p [abbreviated $X \sim Bin(n,p)$] if its mass function, $f(x)$ is

$$f(x) := P(X=x) = \binom{n}{x} p^x (1-p)^{n-x},\ x = 0, 1, 2, ..., n. \quad .$$

[Some texts use π or θ for p. Note also that many texts denote $1-p$ by the letter q, denoting the probability of failure on any one trial.]

Mean and variance of the binomial distribution

If X is a binomially distributed random variable, then the mean and variance of X are

$E(X) := \mu_X := \mu = np$, and $Var(X) := \sigma_X^2 := \sigma^2 = np(1-p)$.

Hypergeometric distribution

Consider now the number of successes we obtain when n items are taken *without replacement* from a population that consists of N items, of which a are of one kind (successes) and the remaining $N-a$ are of a second kind (failures). Unlike in the case of the binomial random variable the number of successes you have obtained so far influences whether the next trial is a success or not, because the number of successes in the population is reduced by one every time you sample a success. I.e. p, the probability of sampling a success, changes in the hypergeometric case but remains fixed for the binomial

Formal definition: A random variable X has a **hypergeometric** distribution if its mass function is

$$f(x) := P(X = x) = \frac{\binom{a}{x}\binom{N-a}{n-x}}{\binom{N}{n}},\ x = 0, 1, 2, ..., \min(a, n).$$

Mean and variance of the hypergeometric distribution

If X is a hypergeometric random variable, then the mean and variance of X are

$E(X) := \mu_X := \mu = n\frac{a}{N}$, and $Var(X) := \sigma_X^2 := \sigma^2 = n\frac{a}{N}(1 - \frac{a}{N})\frac{N-n}{N-1}$.

How do I know if I should use the binomial or hypergeometric?

Consider an experiment that consists of n identical trials, where each trial is Bernoulli (i.e. has only two possible outcomes, 'success' and 'failure'), e.g. pick n people at random from a room and observe if the person is male or female, pick n companies at random and observe if each individual company made a profit or loss last year. To find the distribution of X, the number of successes in the n trials, use the hypergeometric below if the trials involve sampling without replacement from a finite population and use the binomial formula below if the trials are independent and the probability of a success is the same from trial to trial (as in choosing people with replacement from a room, or tossing a coin n times).

Note: In situations where the hypergeometric is appropriate, you can approximate it using the (easier) binomial if $n/N \leq 0.05$.

The Poisson process

The Poisson process arises when we want to model the number of *arrivals* in an interval of time or space. Here we use the word *arrivals* in a broad sense to mean occurrences of something in space or events in time).

Examples of processes that produce arrivals are:

- the observation of disintegration of atoms of a radioactive substance,
- the occurrence of goals in a football game, the occurrence of defects in a sheet of carpet,
- the occurrence of phone calls to a telephone exchange throughout the day,
- the occurrence of typographical errors in a manuscript, etc.

In each of the above, the following three assumptions seem reasonable

1. In a small time interval or amount of space, the probability of occurrence of the event of interest is proportional to the size of the interval (e.g., it is twice as likely to find a defect in the next two metres of carpet as in the next metre, it is five times as likely that a goal will be scored in the next five minutes of a soccer game as in the next minute);
2. The probability of two or more occurrences in small time or space interval is negligible compared with the probability of one occurrence (e.g. the chance of two or more goals in the next minute of a soccer game is much less likely than the probability of one goal, the chance of two or more people arriving at the bank in the next second is very small compared with the probability of one person arriving, etc.);
3. Arrivals in non-overlapping intervals are independent (e.g. the number of arrivals at the bank between 10 and 10:01 am should not affect the number of arrivals between 11:00 and 11:01 am).

Using these assumptions it is possible to show that $X :=$ the number of arrivals/occurrences in any time period of length t (or region of size t) has a distribution of the following form, for some positive number α :

$$f(x) = P(X = x) = e^{-\alpha t}\frac{(\alpha t)^x}{x!},\ x = 0, 1, 2, ..., \infty \qquad (*)$$

(Question: Why is (*) a probability mass function?)
The number α represents the arrival rate, i.e. the mean number of arrivals per unit time. It can be shown that $E(X) = \alpha t$ and that in fact $Var(X)$ is also αt.

Poisson approximation to the Binomial

Suppose that the Binomial distribution with parameters n and p is appropriate model for a random variable X. If $n \geq 20$ and $p \leq 0.05$, it can be shown that the Poisson distribution (with $\alpha t = np$) can be used instead, as a good approximation, to work out probabilities about X. The approximation is excellent if $n \geq 100$ and $np \leq 10$.

Continuous random variables

A random variable is called *continuous* if it can take values anywhere in an interval. For a continuous random variable X, we shall consider a curve, called the density function $f(x)$ of X such that the area under this curve between any two points a and b gives the probability that X lies between a and b. (Such a function $f(x)$ will be non-negative for every real number x, and the total area under it will be one.)

Definition: The probability density function, $f(x)$, of a random variable, X, is the function such that $\int_a^b f(x) = P(a \leq X \leq b)$

An alternative way to describe a continuous random variable is through the *cumulative distribution function*. The cumulative distribution function, $F(x)$, gives the probability that the random variable takes on a value less than or equal to x.

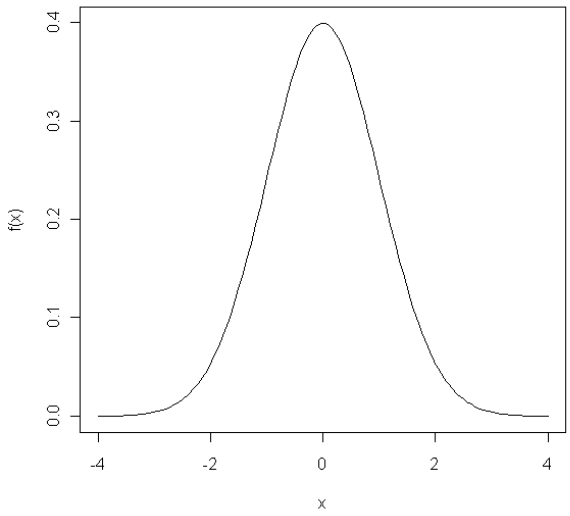
Definition: The cumulative distribution function, F, of a random variable is the function such that $F(x) = P(X \leq x)$

If X is a continuous random variable, the definitions of the population mean $\mu = E(X)$, the population variance $\sigma^2 = Var(X)$, and the population standard deviation $\sigma = \sqrt{Var(X)}$ require calculus but the interpretation is relatively straightforward. If we took a large sample of a the random variable and calculated the mean and variance of the sample (called the sample mean and sample variance, respectively), then the mean and variance of the continuous random variable would be approximated by the sample mean and sample variance. Of course, this also holds true for discrete random variables.

An important continuous random variable: The normal distribution

Continuous random variables that have a normal distribution are the most frequently encountered, as a consequence of the Central Limit Theorem (see next). The height of a random person, the sales of a company on a random day, the mark of a random student on an exam, etc., are all often modelled by a normal distribution. The density function of a normal random variable X that has mean μ and variance σ^2 is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{\frac{-1}{2\sigma^2}(x-\mu)^2\right\}, -\infty < x < \infty.$$



The normal curve above has mean, $\mu = 0$, and variance $\sigma^2 = 1$. Notice that this distribution is symmetric around μ and has its maximum at μ. As $x \longrightarrow \infty$ or $-\infty$, $f(x) \longrightarrow 0$. As you can see from the diagram the areas of highest probability density (i.e. where $f(x)$ is highest) are around the mean. If X has a normal distribution with mean μ and variance σ^2, we write this for short as $X \backsim N(\mu, \sigma^2)$,that is, X is distributed as normal with mean μ and variance σ^2.

To work out the probability that a normally distributed random variable lies in some interval (say between two points a and b on the x-axis) we need tables (or these days a computer) that will tell us the area under this curve between the two points. However, the tables are available only for a normal variable Z that has a mean of 0 and a variance of 1. (Such a normal variable is called a standard normal random variable.) To calculate probabilities about any normal random variable, we use a process called standardization, that is, we convert the original variable into one that has a mean of 0 and a standard deviation of 1, and the normality is maintained. The formula for this is

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

[formula for standardizing to a $N(0,1)$ normal random variable, Z]

This standardization can be applied to any variable, but when it is applied to one that has a normal distribution, the resulting variable Z will have the standard normal distribution $N(0,1)$. You will encounter this standardization proceedure in most of the example questions that involve a normal random variable.

The Central Limit Theorem

Definition A sequence of random variables $X_1, X_2, ..., X_n$ constitutes a *random sample* if they are independent and identically distributed (iid).

Independence means that $f_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) = \Pi_{i=1}^{n} f_{X_i}(x_i)$, for all values x_i of X_i, $i = 1, 2, ..., n$. This is analogous to the definition of independence we had for events, extended to continuous random variables. It is a mathematical way of specifying that the next value to appear in the sample is not influenced by the values that have been sampled so far.

Identically distributed means that the X_i have the same distribution, i.e. $f_{X_i}(x)$ is the same for all $i = 1, 2, ..., n$.

For example, do you think that the heights $X_1, X_2, ..., X_{10}$ of the next 10 people you observe at random constitute a random sample? (if you think the answer is no, then you are either implying that the heights are not independent of each other and/or that some one of the observations is more likely to be bigger than some other.)

Do you think that the sales $X_1, X_2, ..., X_{10}$ over the next 10 years of a randomly selected retail company are a random sample? (Answer: Probably not – there is surely a correlation over time.)

Theorem 1 *Suppose we take a random sample $X_1, X_2, ..., X_n$ of size n from a population that has mean μ and variance σ^2. Let $\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$ be the mean of the random sample. Then*

(a) if the population has a normal distribution (that is, if $X \backsim N(\mu, \sigma^2)$, where X denotes one random member of the population) then the distribution of $\overline{X}$ is normal with mean μ and variance σ^2/n, that is,

$$\overline{X} \backsim N(\mu, \tfrac{\sigma^2}{n}).$$

*(b) [**The Central Limit Theorem**] No matter what distribution the population has, provided n is large, the distribution of $\overline{X}$ is approximately normal with mean μ and variance σ^2/n. That is,*

$$\overline{X} \approx N(\mu, \tfrac{\sigma^2}{n}).$$

Often "n large" is taken as meaning that "$n > 30$". It is well worth your while understanding the last theorem, because it is the basis of many applications in statistical inference.

The square root of the variance of $\overline{X}$ is, by definition, the standard deviation of $\overline{X}$. Its formula is $\frac{\sigma}{\sqrt{n}}$ and this is often called the *standard error of the mean.*

Sampling methods

Suppose that we have a finite population (e.g. the population could represent all students at NUI, Galway, or all voters in Ireland, or all cancer patients in a given hospital) of size N and that we wish to take a sample of size n from this population . The objective of *survey designs* is (like all branches of statistical inference) to maximize the amount of information (about the population) contained in the sample for a fixed cost, or to minimize cost for a fixed sample size. *Methods of data collection* include personal interviews, telephone interviews, self-administered questionnaires, direct observation and examination of records.

Some problems that can arise in attempting to obtain a representative sample include systematic *bias* caused by e.g. non-response, sensitive questions, etc. We briefly describe some sample designs that incorporate randomness (which will permit reduction of bias and enable probability statements to be given to justify inferences made).

Simple random sampling

By definition, this design gives each sample of size n an equal chance, $1/\binom{N}{n}$, of being selected. Each individual in the population will have a probability of $\frac{\binom{1}{1}\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{\frac{(N-1)!}{(n-1)!(N-n)!}}{\frac{N!}{n!(N-n)!}} = \frac{n}{N}$ of being included in the sample. Inferences can be made about certain parameters based on this design For example, we can use the sample mean daily sales, $\overline{x}$, over 10 business days of a certain retail store as an estimate of the population mean daily sales μ over the year, or the blood pressure of 10 students as an estimate of the population mean blood pressure μ of all students in this class. Also, we can use the sample proportion, $\hat{p}$, of smokers in a random sample of 100 Galwegians as an estimate of the population proportion, p, of smokers in Galway. It is important to note that $\hat{p}$ is a special case of the sample mean $\overline{x}$. To see this, imagine assigning 1 to everyone in the population who smokes, and 0 to each non-smoker. The ith person sampled will then have value $x_i = 1$ or 0. The sample mean is then (as always) $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. But this is clearly the sample proportion, $\hat{p}$, of smokers.

Remark: As long as the population size N is reasonably large, the theory of inference based on the above simple random sampling – where the sampling is assumed to be without replacement – is essentially the same as if we are sampling with replacement from a finite population, or taking a random sample from an infinite population.

Stratified random sampling

Here the population (e.g. students at NUI, Galway) is divided into L subpopulations called strata (e.g. 2 strata consisting of males and females, or five strata representing the different colleges in NUI Galway) and a simple random sample is taken from each stratum. There should be little variability among the units within a given stratum (i.e. the strata should be homogeneous).

Reasons for stratification: The idea is that by dividing the population into subpopulations we can (i) obtain information about these subpopulations, (ii) reduce costs due to administrative convenience, and (most importantly) (iii) usually get improved estimation of population parameters than would be possible from a simple random sample. Note that if we stratify according to sex, for example, we would ensure to obtain some females and some males, whereas a simple random sample might not include enough females, and this would lead to poor estimation if females tend to think differently about certain issues than males.

For this reason stratification generally leads to better estimates than simple random sampling if the items within the strata are fairly homogeneous (and there is heterogeneity between strata).

Sample sizes in stratified random sampling: How many items n_i should we take from the ith stratum, $i = 1, 2, ..., L$? Intuitively, if one stratum is much larger than another, we should (other things being equal) take more units from that population. With *probability proportional to size* (also called *proportional allocation*), the number to take from stratum i is

$$n_i = n\frac{N_i}{N},\ i = 1, 2, ..., L\ \textit{(proportional allocation formula)}$$

where N_i is the number of units in stratum i.

If one stratum has much larger variance than another, then intuitively, we should take more items from that stratum to 'get a handle' on what is happening in that stratum. Assuming that we have the same cost associated with sampling any item, the method of *optimal allocation*, that is, the allocation that *minimizes the variance of the sample mean* is to sample the following number of units from stratum i :

$$n_i = n\frac{N_i\sigma_i}{N_1\sigma_1+N_2\sigma_2+...+N_L\sigma_L},\ i = 1, 2, ..., L\ \textit{(optimal allocation formula)}$$

where σ_i is the standard deviation in the ith stratum.

Cluster sampling
In cluster sampling, we divide the population into groups (clusters) such that within each cluster the units are as heterogeneous (different) as possible, and one cluster should be similar to another. We then take a random sample of

clusters and our sample consists of all the units in these clusters. Clustering is effective when a frame (listing) of the population elements is either unavailable or difficult to obtain and when the cost of obtaining observations increases with the distance between units.

Other Sampling Strategies
Other methods of sampling include *systematic sampling*, *ratio estimation*, *regression estimation*, *quota sampling*, and *multi-stage sampling*. This last is an extension of cluster sampling. In two-stage cluster sampling, for example, we might take a random sample of city blocks, and then a random sample of buildings within these blocks. We then survey all occupants of these buildings.

Sample mean and sample standard deviation

If you take a random sample of size, n, from a population then the **sample mean** is

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

This is the same notion of *average* that is familiar in everyday life. A key property of the sample mean is that it is an unbiased estimate of the population mean. I.e.

$$E(\bar{X}) = \mu$$

Most people would find this intuitively obvious. Essentially, this formula says that if I take a large enough random sample, e.g. of NUI Galway students and measure their heights, I would expect the average of my sample to be the same as the population average (i.e. the average height of all NUI Galway students).

Given that we now know the expected value of $\bar{X}$, can we say something about it's variance? It turns out that

$$Var(\bar{X}) = \frac{\sigma^2}{n}$$

This last formula is incredibly important because it tells you something about how far the sample mean ($\bar{X}$) can stray from its expected value (the population mean). Combining this with the Central Limit Theorm allows us to say with confidence what the probability is that the population mean has any particular value, given the mean of a random sample from the population.
Lastly, the **sample variance** is like the average of the squared difference between my sampled values and the sample mean. Mathematically:

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

The sample variance is an unbiased estimate of the population variance. Given a large enough sample we could estimate the population variance to arbitrary precision via the sample variance.

In the above formula there is a technical reason for dividing the sum by $n-1$ rather than by n that has to do with the fact that the mean, $\bar{X}$, has been estimated from the sample. If you are the kind of person who enjoys worrying about the technical details then you can think about it like this: If you had a sample of size 1 you would have no information at all about the population variance. A sample of size 2 gives you one piece of information about the population variance - i.e. the distance between your two sample points, which comes into the above formula in the form of their distances from their midpoint. A sample of size 3 gives 2 pieces of information about the variance, etc...

Normal approximation to binomial distribution

Examples: Let X = the number of heads we will obtain in n flips of a coin that has probability p of coming up heads; or let X = the the number of people, in a random sample of n voters, who will vote for a particular candidate in an election in Ireland, when the population proportion of people who will vote for the candidate is p.

In the examples above, we know from earlier that $X \backsim \text{Bin}(n,p)$, [i.e. X has the binomial distribution with parameters n and p]. Thus the probability mass function of X is

$$f(x) = P(X = x) = \binom{n}{x}p^x(1-p)^{n-x},\ x = 0, 1, 2, ..., n\ .$$

If n is large (more precisely if np and $np(1-p)$ are both at least 5), it follows from the Central Limit Theorem that:

$$\boxed{\text{the distribution of } X \text{ is approx. } N(\mu, \sigma^2) \text{ where } \mu = np \text{ and } \sigma^2 = np(1-p)} \quad (\otimes)$$

This means that we can calculate binomial probabilities fairly accurately by getting areas under a normal curve that has the same mean and variance as the binomial.

Letting $\hat{p} := \frac{X}{n}$ be the sample proportion of successes, we can write the above result as

$$\boxed{\text{the distribution of } \hat{p} \text{ is approximately } N(p, \tfrac{p(1-p)}{n})}$$

End of theory section

Worked examples

Sample spaces, events and their probabilities

1. Among the 20 students in this class, 10 study Arts, 12 have blue eyes and 4 are both Arts students and have blue eyes.

   (a) Using A for the event that a random student studies Arts and B for the event that a random student has blue eyes, write the above information symbolically.

   (b) Find $P(A|B)$, $P(B|A)$ and $P(A \cup B)$ and $P(\overline{A} \cap \overline{B})$.

   (c) Are the events A and B (i) disjoint (i.e. mutually exclusive), (ii) independent?

   Solution: (a) $\#A = 10$ $\#B = 12$ $\#(A \cap B) = 4$
   (b) $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{4}{20}}{\frac{12}{20}}$
   (c) (i) Not disjoint since $A \cap B \neq \emptyset$ (ii) Not independent since $P(A \cap B) \neq P(A) \times P(B)$

2. A card is picked at random from a pack of 52 cards. Let A_1 be the event that the selected card is a Jack, let A_2 be the event that the card is a King, and let A_3 be the even that the card is a Diamond. Are the events A_1 and A_2 (i) disjoint (i.e. mutually exclusive), (ii) independent?

   Solution: (i) No. It's possible for a card to be both a diamond and a king (i.e. the king of diamonds) (ii) Yes. Knowing that the card is a king does not change the probability that the card is a diamond (and *vice versa*).

3. *[Multiple choice format]* Among the 681 finishers in the marathon, there were 215 women and 466 men. Ten of the women and 87 of the men were over the age of 50. **(a)** What is the probability that a randomly sampled participant is a man or is over the age of 50?

   A) 87/681 B) 302/681 C) 87/466 D) 476/681

   Solution: D. Let A,B and C be the events that the participant is a male, female or over 50, respectively. $P(A \cup C) = P(A) + P(C) - P(A \cap C) = \frac{466+97-87}{681}$

4. Suppose a single participant is randomly chosen from the 681 finishers. What is the conditional probability that the individual is a man, given the information that the individual is over the age of 50?

   A) 87/466 B) 87/97 C) 97/681 D) 466/681

   Solution: $P(A|C) = \frac{P(A \cap C)}{P(C)} = \frac{\frac{87}{681}}{\frac{97}{681}} = \frac{87}{97}$

5. There are seven days in the week. Assume that with 7 randomly selected people in a room, each of these days is equally likely to occur as the day of the week of birth for each person. Find the minimum number of people

such that the probability is at least 0.35 that two or more of the people are born on the same day of the week.

Solution: The probability that no two of n people are born on the same day is $\frac{\frac{7!}{(7-n)!}}{7^n}$. The probability that they are all born on different days is one minus this. By trial and error you can show that this probability is 0.3878 for $n = 3$ (and that 3 is the first value of n for which this probability is greater than or equal to 0.35).

Example 5. A man is accused of having murdered his business partner. The business partner was murdered in their shared office after hours in an apparent burglary. Based on witness statements and forensic evidence obtained at the scene and at the man's home the state probabilist has calculated that the man's probability of guilt is 90%. New evidence in the form of phone records now emerges, showing that the man was in the vicinity of the office after five O'clock. The man normally kept regular office hours and phone records show that over the previous 100 days he was in the vicinity of the office after five on just three occasions. What is the probability that the man is guilty following the analysis of this new evidence?

Example 6. [**Diagnostic Testing**] In medicine, diagnostic testing forms the basis for clinical decision-making, and encompasses patient history, physical examination, laboratory tests, imaging techniques (e.g. CT scans, x-rays) and procedures (e.g. ECG). A new diagnostic test for cancer, for example, might be adopted if it has a high probability of detecting the disease in a person who has cancer, and a low probability of declaring a healthy person as having cancer. The table below shows the possible outcomes of the test as a function of patient condition.

	Patient has disease	**Patient does not have disease**
Test is positive	No error committed. The probability that a diseased individual will have a positive test result is called the *sensitivity*, or *true positive rate*, (TPR) of the test.	Error committed. The probability that a disease-free individual will have a positive test result is called the false positive rate (FPR) of the test.
Test is negative	Error committed. The probability that a diseased individual will have a negative test result is called the *false negative rate* (FNR) of the test.	No error committed. The probability that a disease-free individual will have a negative test result is called the *specificity*, or *true negative rate* (TNR) of the test.

Suppose now that the proportion of Irish people with cancer is 1/200, and that a new cancer screening test has a TPR of 0.95 and an FPR of 0.01.

(a) What proportion of patients who take the test will have a positive result.

(b) Given that a patient has a positive result, **what** is the probability that he/she has cancer?

[*Answer*: Using formula 15) on *Some Useful Formulae in Probability*, we get

(a) $\frac{1}{200} \times 0.95 + \frac{199}{200} \times 0.01$,while for b) we use Bayes' Formula (i.e. 16) on *Some Useful Formulae in Probability*) to get $\frac{\frac{1}{200} \times 0.95}{\text{answer in a)}}$.]

6. Box A contains two gold coins, Box B contains two silver coins, and Box C contains one gold and one silver coin. A box is chosen at random and a coin selected from it. If the selected coin is Gold, what is the probability that the other coin in the selected box is Gold (i.e. what is the probability that the selection was made from Box A?) [*Answer:* Using Bayes' Formula (or use a tree diagram or Venn diagram), we get answer $\frac{\frac{1}{3} \times 1}{\frac{1}{3} \times 1 + \frac{1}{3} \times 0 + \frac{1}{3} \times \frac{1}{2}} = \frac{2}{3}$.

   Note: Many people think that the answer to this problem should be $\frac{1}{2}$. This misses the point that if we selected a gold coin it is in fact more likely that we were drawing from A than from C.

7. **Birthdays** Suppose that there are r randomly selected individuals in a room. What is the probability that at least two of them have the same birthday (i.e. born on the same month and day)?

   [*Note:* Ignore leap years, so assume 365 days in a year, and assume that each person has a probability of $\frac{1}{365}$ of being born on any particular day.]

   Solution: $P(\text{at least two have same birthday}) = 1 - P(\text{no two have same birthday}) =$

$$1 - P\left(\begin{array}{c}\text{1st person}\\ \text{is born on}\\ \text{any day}\end{array} \text{ and } \begin{array}{c}\text{2nd person}\\ \text{is born on a day}\\ \text{different}\\ \text{from the first}\end{array} \text{ and . . . and } \begin{array}{c}r\text{th person}\\ \text{is born on a day}\\ \text{different from}\\ \text{the previous } r-1\end{array}\right)$$

$$\underset{\text{by independence}}{=}$$

$$1 - P\left(\begin{array}{c}\text{1st person}\\ \text{is born on}\\ \text{any day}\end{array}\right) P\left(\begin{array}{c}\text{2nd person}\\ \text{is born on a day}\\ \text{different}\\ \text{from the first}\end{array}\right) \times \ldots \times P\left(\begin{array}{c}r\text{th person}\\ \text{is born on a day}\\ \text{different from}\\ \text{previous } r-1\end{array}\right)$$

$$= 1 - \frac{365}{365} \times \frac{364}{365} \times \ldots \times \frac{365 - r + 1}{365}.$$

That is, if we let p_r be the probability that at least two of r people have the same birthday, then $p_r = 1 - \frac{365}{365} \times \frac{364}{365} \times \ldots \times \frac{365 - r + 1}{365}$ **(*)**

TABLE OF VALUES OF (*) FOR SELECTED VALUES OF r

r	p_r
2	$1 - \frac{365}{365} \times \frac{364}{365} = \frac{1}{365}$
4	$1 - \frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} \times \frac{362}{365} = 0.016$
7	0.056
15	0.253
22	0.475
23	0.507
30	0.7064
40	0.891
50	0.970
60	> 0.99
70	> 0.999

Note: For another solution to the Birthday Problem, see *Example 6* of the *Sample Combinatorics Problems* below. There we show that p_r equals $1 - \frac{^{365}P_r}{(365)^r}$ where for positive integers n and r with $n \geq r$, ^nP_r is defined in the next section below.

Counting problems - combinatorics

1. Consider distributing 3 objects into 2 cells. If the objects are *distinguishable*, say A, B, and C, then letting, e.g., (ABC, none) denote the placement of A,B and C into the first cell and no object in the second cell, the possible distributions are:

   (ABC, none), (none, ABC), (AB, C), (AC, B), (BC, A), (A, BC), (B, AC), (C, AB).

   Note that (see Formula 1 above) the first of these, (ABC, none), represents the $\frac{3!}{3!0!} = 1$ way of placing the objects such that the first cell gets all three objects; the second, (none, ABC), is enumerated by $\frac{3!}{0!3!} = 1$ which is the number of ways of placing the objects such that the first cell gets none of the objects and the second gets all three objects; similarly the third, fourth and fifth entries above are quantified by $\frac{3!}{2!1!} = 3$, and the last three entries are enumerated by $\frac{3!}{1!2!} = 3$.

   The various occupancies and their enumeration is thus as follows:

$$\underbrace{(ABC,\ none)}_{\frac{3!}{3!0!}=1},\ \underbrace{(none,\ ABC)}_{\frac{3!}{0!3!}=1},\ \underbrace{(AB,\ C),(AC,\ B),(BC,\ A)}_{\frac{3!}{2!1!}=3},\ \underbrace{(A,\ BC),(B,\ AC),(C,\ AB)}_{\frac{3!}{1!2!}=3}$$

Of course, the total number of occupancies is 1+1+3+3=8, and this 8 could be

obtained directly from *The Fundamental Counting Rule* above the first object can be placed in any of 2 cells, the second in any of 2 cells independently of the first object, and the third in any of 2 cells also; hence the total number of occupancies is $2 \times 2 \times 2 = 2^3 = 8$.

Note: At the level of this course, we will not study in detail the case in which the balls are *indistinguishable*.

2. **(a)** In how many ways can the letters of the word MARMALADE be arranged in a row?

   (b) If an arrangement is picked at random, what is the probability that it begins with an M? **(c)** that the vowels will be in their correct order?

3. There are four married couple in a room. If these 8 people are arranged in a row, **(a)** how many arrangements are possible? If an arrangement is picked at random, **(b)** what is the probability that each couple will be together?

   (c) that all the males will be next to each other?

   (d) that all the males will be next to each other and all the females will be next to each other?

4. Suppose that 15 students will be distributed at random into 3 classes in such a way that each class will get 8 students. If there are 3 whiz kids among the 16 students, write down the probability that each class gets one.

5. In how many ways can 30 football players be divided into two teams of 15 players each?

6. Refer to Example 8 (Birthdays) in the section *Sample Probability Problems* above. Solve the problem using combinatorial techniques.

   Solution: $P(\text{at least two have same birthday}) = 1 - P(\text{no two have same birthday})$

   (using 5.of *Some Useful Probability Formulae*)

$$= 1 - \frac{\#\text{ of groups of } r \textit{ different} \text{ days that can be chosen from 365 days}}{\#\text{ of groups of } r \text{ days that can be taken from 365 days}} = 1 - \frac{^{365}P_r}{(365)^r}$$

Note that this answer is the same as that given in (*) of the solution to *Example* 8 in the section *Sample Probability Problems above because*

$$1 - \frac{^{365}P_r}{(365)^r} = 1 - \frac{\frac{(365)!}{(365-r)!}}{(365)^r} = 1 - \frac{365 \times 364 \times ... \times (365 - r + 1)}{(365)^r} = 1 - \frac{365}{365} \times \frac{364}{365} \times ... \times \frac{365 - r + 1}{365}$$

Note also that in our derivation above, we used $^{365}P_r$ and not $\binom{365}{r}$ because the order within each group of r birthdays is important (compare 3 and 4 in the *Useful Combinatorial Formulae* section above). For example, if you have a birthday on January 1 and I have birthday on January 2, that is quite a different event than you having birthday on January 2 and me having birthday on January 1.

7. Pick a person at random and ask that person whether or not he favours a certain political issue. Let $X = 1$ or 0 according as the person favours or does not favour. Here the sample space S has the two elements Favour and Oppose, and X can be exhibited as below:

8. Let $X =$ the number of heads we will obtain in $n =$3 flips of a fair coin. Notice that X takes values $0, 1, 2, 3$. Here the sample space is S, the random variable X, and the range of X are as shown below:

ω		$X(\omega)$
HHH	$\longrightarrow$	3
HHT	$\longrightarrow$	2
HTH	$\longrightarrow$	2
THH	$\longrightarrow$	2
TTH	$\longrightarrow$	1
THT	$\longrightarrow$	1
HTT	$\longrightarrow$	1
TTT	$\longrightarrow$	0.

It might help you to think of X as the set of numbers that can result from the experiment - i.e. {0,1,2,3}, but please do not think of X as fixed number! Notice that the values of X do not have equal probability – clearly, for example, the probability that X takes the value 2 is larger than the probability that X will equal 0; that is, $P(X = 2) > P(X = 0)$. The goal of probability is to find a formula that gives the likelihood (long-run relative frequency) of the various outcomes; this remark is identical with out statement above about "mathematical models for..."The X just described is also discrete – it takes just 4 values.

9. Let X = the number of diseased trees in a randomly selected forest. Then we see that X takes values $0, 1, 2, ...\ N$ (where N = # trees in the forest).

ω		$X(\omega)$
Boherbue Forest	$\longrightarrow$	12
Coillte's new forest near Athy	$\longrightarrow$	1
etc.	$\longrightarrow$	etc.
etc.	$\longrightarrow$	etc.

Notice that X is discrete. From now on, we shall not exhibit the sample space Ω, but will think of X as the set of numerical outcomes.

10. Let X = the number of radioactive particles emitted by a Geiger counter in a random second. Here X is discrete (though it can take countably infinite values, in theory anyway).

11. Let X = the number of cars that will arrive on Campus between 9 am and 10 am on a random weekday. Here X is discrete (though it is usually allowed to take not just a finite but countably infinite number of values).

12. A building contractor will make €100,000 if he is successful in a bid, while he will make nothing if he is not successful. Let X = the profit he will make. X is discrete, having values €0 and €100,000.

13. Let X = the height of a randomly selected person. X is treated as continuous (not discrete).

    A **continuous** random variable is one that can take values anywhere in a continuum. (See Continuous Random Variables section later.)

14. Let X = the weight of a randomly selected person.

15. Let X = the temperature on a random summer's day.

16. Let X = the sales of a company on a random day. Then (in theory) X can take values anywhere in the positive real line, so in particular it is a continuous random variable.For now we concentrate on discrete random

    variables. We want to assign probabilities to their values, and summarize these distributions. Accordingly, we make the following important definitions.

Probability mass function

1. The number of birdies you will get in Saturday's 18-hole golf game is either 0,1, 2 according to whether it rains heavily, rains lightly or does not rain at all. The probability with which it rains heavily is 0.4, the probability that it rains lightly is 0.4 and the probability that it will not rain at all is 0.2. Find the distribution of X = the number of birdies you will obtain.

   Solution: Clearly, $P(X = 0) = 0.4, P(X = 1) = 0.4$ and $P(X = 2) = 0.2$.

   We can write this distribution in tabular form if we wish:

x = # birdies	$P(X = x)$ = prob. that the random variable X takes the value x
0	0.4
1	0.4
2	0.2

2. Suppose that in the previous example, you will play a second golf game on Sunday, If you score no birdies on Saturday, you will certainly not get one on Sunday. If you score one birdie on Saturday, then you will with equal probabilities, 0.5, score 0 or 1 birdies on Sunday, while, finally, if you get 2 birdies on Saturday, you will certainly score either 1 or 2 birdies on Sunday with probability 0.5 each. Let Y = the total number of birdies you score on the two days. What is the distribution of Y ?

   SOLUTION: Clearly Y takes values 0,1,2,3, and 4. The distribution of Y is then

# scored Sat.	# scored Sun.	y	$P(Y = y)$
0	0	0	$0.4 \times 1 = 0.4$
1	0	1	$0.4 \times 0.5 = 0.2$
1	1	2	$0.4 \times 0.5 = 0.2$
2	1	3	$0.2 \times 0.5 = 0.1$
2	2	4	$0.2 \times 0.5 = 0.1$

Mean, variance and standard deviations of discrete random variables

1. Refer to the GOLF example. Find the mean and standard deviation of the number of birdies you will score on a random Saturday.

   Solution:
   Here $E(X) = \sum_x xP(X = x) = 0(0.4) + 1(0.4) + 2(0.2) = 0.8$.
   $\sigma^2 = \sum_x (x - \mu)^2 P(X = x) = (0 - 0.8)^2(0.4) + (1 - 0.8)^2(0.4) + (2 -$

$0.8)^2(0.2) = 0.256 + 0.016 + 0.288 = 0.56$. Hence the standard deviation of X is $\sigma = \sqrt{0.56} = 0.7483$.

There is another formula, which is often a more convenient way of calculating the variance:

$$\sum_x x^2 P(X = x) - \mu^2 \text{for} \sigma^2$$

In this case we would obtain $\sigma^2 = (0)^2(0.4)+(1)^2(0.4)+(2)^2(0.2)-(0.8)^2 = 1.2 - 0.64 = 0.56$,
and hence $\sigma = \sqrt{0.56} = 0.7483$ as above.]

Binomial and hypergeometric distributions

It is important that students be able to solve problems involving hypergeometric and binomial distributions – see lectures, problem sheets, your text, etc. One well-known situation in which the hypergeometric can be used is in calculating probabilities associated with the numbers we get right when we purchase one Lotto ticket.

1. *[Application of the binomial distribution formula.]* Suppose that 30% of employees of a large firm take public transport to work and that 20 employees will be taken at random.

   (a) What is the probability that 18 of them take public transport?
   /item What is the probability that at least 18 take public transport?
   /item What is the probability that at most 2 take public transport?
   /item What is the probability that all 20 take public transport given that at least 18 of them take public transport?
   /item What are the expected value and the variance of the number in the sample that take public transport to work.

   (b) Suppose that you are suspect of the assumption that $p = 0.3$ of the employees take public transport to work, and that in fact you believe it to be more. Accordingly, you formulate the following *alternative hypotheses* (which are statements about the *population* of employees [not the sample!]): the *null hypothesis* [the '*status quo*'] $H_0 : p = 0.3$, and the *alternative hypothesis* [your '*research hypothesis*'] $H_1 : p > 0.3$.

2. If in fact the sample of 20 showed that 18 took public transport, **what** is the conclusion of your statistical test? [It is understood that you will justify your answers in examination questions!]

Solution: Let X be the number of the 20 who take public transport. (Note that X is a random variable, and it is reasonable to model its distribution by the binomial distribution formula, even though sampling is without replacement.) We thus have $X \sim Bin(n, p)$ with $n = 20$ and $p := P(\text{a random employee from the company takes public transport}) = 0.3$. Then:

(a) $P(X = 18) = \binom{20}{18}(0.3)^{18}(0.7)^2 = 190 \times \underbrace{0.3 \times 0.3 \times ... \times 0.3}_{\text{18 appearances of 0.3}} \times 0.7 \times 0.7$

$= \underline{0.000000036}$.

(b) $P(X \geq 18) = P(X = 18) + P(X = 19) + P(X = 20) = \quad 0.000000036 + \binom{20}{19}(0.3)^{19}(0.7)^1 + \binom{20}{20}(0.3)^{20}(0.7)^0 = 0.000000036 + 0.0000000001 + 0.000000000035 = \underline{0.000000037}$.

(c) $P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) = \binom{20}{0}(0.3)^0(0.7)^{20} + \binom{20}{1}(0.3)^1(0.7)^{19} + \binom{20}{2}(0.3)^2(0.7)^{18} = 0.000797922 + 0.006839337 + 0.027845872 = 0.035483131$ or approx. $\underline{0.0355}$.

(d) We want $P(X = 20 \mid X \geq 18)$. Using our multiplication formula $P(A \cap B) = P(A)P(B|A)$ in reverse, this is $P(X = 20 \mid X \geq 18) = \frac{P(X=20 \cap X \geq 18)}{P(X \geq 18)}$. Now note that exactly 20 use public transport **and** at least 18 use public transport can only occur only if exactly 20 use public transport. In other words, we have that the events "$X = 20$ " and " $X \geq 18$ " occur simultaneously only if the event " $X = 20$ " occurs. Accordingly, we have the following formula (which you should try to see is intuitive without the above probability arguments), $P(X = 20 \mid X \geq 18) = \dfrac{P(X = 20\ )}{P(X \geq 18)}$.

From the calculations in part (a) above, we then get finally

$$P(X = 20\ |X \geq 18) = \frac{0.000000000035}{0.000000037} = 0.000968551 \text{ or approx. } \underline{0.001}.$$

Note: Many people incorrectly think that the answer to this question should

be $\frac{1}{3}$,ignoring the fact the events that 18, 19 and 20 people use public transport are not equally likely. [The probabilities of 18, 19 and 20 are implied in (a) above, and are not equal.]

(e) From the formulae for the mean and vairnace of a binomial random variable, we get here that $E(X) = np = 20(0.3) = 6$, and $Var(X) = np(1 - p) = 20(0.3)(0.7) = 4.2$.

Important: *Try to see that the formula for the mean of a binomial variable is actually obvious – if 30% of the employees use public transport,*

then in a random sample of 20, you'd expect 20×0.3 = 6 to be using public transport. The interpretation of the mean and variance of any random variable are similar to those given in the context of Lotto winnings in Problem 3 below. Ensure that you understand them.

(f) We would conclude that $p > 0.3$. Why? Well, if the null hypothesis H_0 was true, then only 30% of employees use public transport. For a random sample of 20 employees, we showed in (b) above that under this null hypothesis (i.e. if this null hypothesis is true), the chance of observing *18 or more* taking public transport is *only* 0.000000037, or about 1 chance in 27 million! This is a very small probability, so either a very very rare event occurred, or our H_0 is false.

It is reasonable then to reject the null hypothesis and conclude H_1. [If you do not understand this, note that all that is being said is that if only 30% of the population use public transport, it would be very unlikely to *get a sample result at least as extreme as we got.* Notice that if $p > .3$, e.g. if $p = .9$, then the sample result of 18 would not be unlikely. It makes sense then to conclude that p exceeds 0.3.]

USEFUL NOTE: In *any* significance testing problem **the p-value of the statistical test is the probability of obtaining a sample result at least as extreme as that which we did observe, when the null hypothesis is true**. If this p-value is deemed 'too small', then we say that our sample result (= 18 in the above example) is *statistically significant*, and we reject the null hypothesis.

[*Note:* You should practice other binomial problems. For example, start with something like: suppose that 1% of screws produced in a manufacturing process are defective; suppose that 51% of births are females; suppose that 60% of items sold in a store are priced at more than £20 and so on. Then in a random sample that will be taken, work out answers to questions similar to those in (a) – (e) above.]

3. Let X be a discrete random variable. **Prove** that the two formulas you have seen for $Var(X)$ are identical.

Solution:

We must prove that $\sum_x (x-\mu)^2 P(X=x) = \sum_x x^2 P(X=x) - \mu^2$.

This follows from the following calculation:

$\sum_x (x-\mu)^2 P(X=x) = \sum_x (x^2 + \mu^2 - 2\mu x) P(X=x) \underset{\text{why?}}{=}$

$\sum_x x^2 P(X=x) + \sum_x \mu^2 P(X=x) - \sum_x 2\mu x P(X=x) =$

$\sum_x x^2 P(X=x) + \mu^2 \sum_x P(X=x) - 2\mu \sum_x xP(X=x)$
[since constants can be taken outside summations]
$= \sum_x x^2 P(X=x) + \mu^2 (1) - 2\mu(\mu)$ [since $\sum_x P(X=x) = 1$ (why?)

Poisson process

1. Suppose that the number of goals in a random Celtic football game has a Poisson distribution with, on average, 2 goals per game. Find the probability of

   (a) one goal in the next game,

   (b) at least one goal in the next game,

   (c) one goal in the next two games.

   Solution: In (*), we put $\alpha = 2$ throughout the question. Let $X =$ the number of goals scored in one random game. Then *(a)* $P(X = 1) \underset{\text{put } t=1 \text{ and } x=1 \text{ in (*)}}{=} e^{-2}\frac{(2)^1}{1!} = 2e^{-2}$.

   (This is approximately 0.2707 using a calculator.)

   (b) $P(X \geq 1) \underset{\text{(working with the complement saves much work)}}{=} 1 - P(X=0) \underset{\text{put } t=1 \text{ and } x=0 \text{ in (*)}}{=} 1 - e^{-2}\frac{(2)^0}{0!} = 1 - e^{-2}$. ‘ Now let $Y =$ the number of goals scored in the next *two* games. Then

   (c) $P(Y=1) \underset{\text{put } t=2 \text{ and } x=1 \text{ in (*)}}{=} e^{-4}\frac{(4)^1}{1!} = 4e^{-4}$.

   [Note: This result could have been derived by partitioning as in formula (15) of *Useful Probability Formulae*, as follows. P(one goal in the next two games) = P(one goal in the first game and no goal in the second game) +

   P(no goal in the first game and one goal in the second game)

   $\underset{\text{by independence – see (A3)}}{=} P$(one goal in the first game)P(no goal in the second game) +

   P(no goal in the first game)P(one goal in the second game) $\underset{\text{using (*) with } \alpha=2,\ t=1, \text{ and } x=0,\ 1}{=}$

   $e^{-2}\frac{(2)^1}{1!}e^{-2}\frac{(2)^0}{0!} + e^{-2}\frac{(2)^0}{0!}e^{-2}\frac{(2)^1}{1!} = 4e^{-4}$, as above.

2. In the last example, calculate the probability that two goals will be scored in the next game given that at least one goal will be scored.

   Solution: We require

   $$P(X=2 \mid X \geq 1) \underset{\text{multiplication formula}}{=} \frac{P(X = 2 \cap X \geq 1)}{P(X \geq 1)} \underset{\text{why?}}{=} \frac{P(X = 2)}{P(X \geq 1)} =$$

   $$\frac{P(X = 2)}{1 - P(X = 0)} \underset{\text{using (*)}}{=} \frac{e^{-2}\frac{(2)^2}{2!}}{1 - e^{-2}\frac{(2)^0}{0!}} = \frac{2}{e^2 - 1}.$$

Poisson approximation to the Binomial

1. Suppose that 1% of screws produced in a factory are defective. If 100 screws are selected at random, what is the *(a)* exact and *(b)* approximate probability that 5 of them will be defective.

   Solution: Let X = the number of defectives in the random sample of $n = 100$.

   We require $P(X = 5)$.

   (a) Using the binomial distribution function $P(X = x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$, $x = 0, 1, 2, ..., n$

   with $n = 100$, $\theta = P(\text{a random bolt is defective}) = 0.01$, and $x = 5$, we have $P(X = 5) = \binom{100}{5}(0.01)^5(1 - 0.01)^{95} = 0.0029$.

   (b) Using the Poisson approximation to the binomial) we use the formula (*) with $\lambda = \alpha t =$ expected number of defectives in the sample $= \theta n = (0.01)100 = 1$.

   Then $P(X = 5) \overset{\circ}{=} e^{-1}\frac{(1)^5}{5!} = 0.0031$.

   (Notice how close the two answers are, but how much easier the Poisson is.)

Normal random variables

1. Suppose that heights of people have a normal distribution with mean $\mu = 68$ and standard deviation $\sigma = 3$. **(a)** What proportion of people have heights above 71? **(b)** What proportion of people have heights below 62? **(c)** If 2.5% of people have heights above a, what is a?

   Solution: Let X = height of a random person. Then we are given that $X \backsim N(\mu, \sigma^2)$ with $\mu = 68$ and $\sigma = 3$.

(a) We want $P(X > 71)$. Using the standardization formula (*) above, we find that this is $P(X > 71) = P(\frac{X-\mu}{\sigma} > \frac{71-\mu}{\sigma}) = P(Z > \frac{71-68}{3}) = P(Z > 1) = 0.1587$. Here, and elsewhere unless stated otherwise, Z denotes a $N(0,1)$ random variable, and we have used standard normal tables to get the number 0.1587.

(b) Similarly, $P(X < 62) = P(Z < -2) \underset{\text{by symmetry}}{=} P(Z > 2) = 0.0228$.

(c) We want a so that $P(X > a) = 0.025$. (Notice that this is just a reversal of the type of question in (b) above.

Standardizing, we have $P(Z > \frac{a-68}{3}) = 0.025$. Hence form $N(0,1)$ tables, or the information in the question, we get $\frac{a-68}{3} = 1.96$. Hence $a = 68 + 3 \times 1.96 = 73.88$.

Central Limit Theorem

1. Refer to our "heights" example above. Suppose that 100 people will be selected at random. What is the probability that their mean height $\overline{X}$ will exceed 71?

   Solution: We require $P(\overline{X} > 71)$. From the above theorem, $\overline{X} \backsim N(\mu, \frac{\sigma^2}{n})$, so we just apply our standardization procedure (*). We get $P(\overline{X} > 71) = P(\frac{X-\mu}{\sigma/\sqrt{n}} > \frac{71-\mu}{\sigma/\sqrt{n}}) = P(Z > \frac{71-68}{3/\sqrt{100}}) = P(Z > 10) \doteq 0$. (Note that we cannot find the number 10 in our z-tables, but the tables give the area above 3.09 to be about 0.001, so the area above 10 must be extremely close to 0.)

   Note that by the Central Limit Theorem, it was not necessary to know that the distribution of the population was normal. this is because our calculation relied on knowledge of the distribution of $\overline{X}$, and the Central Limit Theorem says that this distribution is approximately normal anyway, since the sample size n is large. Thus the answer we obtained would be approximately correct anyway.

2. Refer again to our HEIGHTS examples above. Suppose that μ is actually unknown. In estimating μ, would you prefer to use the value x of one random person's height X, or the value $\overline{x}$ of the mean $\overline{X}$ of 100 randomly selected peoples' heights?

   Solution: Intuitively, you'd prefer to use the mean $\overline{x}$ (surely, the more observations you take from a population, the more information we obtain about some unknown aspect of this population!) But let's answer the question rigorously. The distribution of each of X and $\overline{X}$ are normal with the same mean μ (the population mean). But while X has standard deviation $\sigma = 3$, $\overline{X}$ has standard deviation $\frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{100}} = 0.3$. This means

that we are much more likely to get a value of $\overline{X}$ "close" to μ than we are of getting a value of X "close" to μ. Here is another more mathematical way of providing an answer: First note that from standard normal distribution tables, we see that if we have a standard normal variable Z, then $P(-1 < Z < 1) = 0.6826$, $P(-2 < Z < 2) = 0.9544$, $P(-2 < Z < 2) = 0.9987$. These numbers are useful to remember, and state that about 68%, 95%, and practically all, values of a normal variable lie within, $\pm$ one standard deviation, $\pm$two standard deviations, and $\pm$three standard deviations of the mean of the variable. Hence [see the standardization formula (*) above], since $\frac{X-\mu}{\sigma}$ and $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ each have the standard normal distribution, we find for example that $P(-1 < \frac{X-\mu}{\sigma} < 1) = 0.6826$ and $P(-1 < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < 1) = 0.6826$. That is, $P(-\sigma < X - \mu < \sigma) = 0.6826$ and $P(-\sigma/\sqrt{n} < \overline{X} - \mu < \sigma/\sqrt{n}) = 0.6826$, i.e. (since, in our example, $\sigma = 3$ and $n = 100$), $P(-3 < X - \mu < 3) = 0.6826$ and $P(-0.3 < \overline{X} - \mu < 0.3) = 0.6826$. We thus see that with probability approx. 0.68, X will lie within only a distance 3 of μ, while with probability approx. 0.68, $\overline{X}$ will be within 0.3 of μ. This confirms our preference for using $\overline{X}$ rather than one random observation X. Also check that the bigger the sample size, the more likely it is that $\overline{X}$ will lie within any given distance of μ (replace $n = 100$ by, e.g., $n = 1,000$ above).

Sampling methods

1. It is desired to take a sample of 100 students from an university that has $10,000$ students. Suppose we divide the population into two strata, males and females, and that there are $6,000$ male students and $4,000$ female students in the university. **How many** of each sex should be included in the sample of 100 if we use proportional allocation?

   Solution: We would take $100 \times \frac{6000}{10000} = 60$ males and $100 \times \frac{4000}{10000} = 40$ females.

2. Refer to the last example and suppose that the survey of 100 students is to be conducted to estimate the mean annual consumption, μ, of beer by students at NUI, Galway. Suppose that the standard deviation of the number of units of beer consumed by males and females are $\sigma_1 = 100$ and $\sigma_2 = 50$, respectively. [Note: If these standard deviations were unknown, we'd estimate them by the sample standard deviations got from a pilot study, or by some other method.] **How many** students from each stratum should be taken if optimal allocation is used?

   Solution: The number of males we should sample is

   $$100 \times \frac{6000 \times 100}{6000 \times 100 + 4000 \times 50} = 75$$

   and the number of females will equal

$$100 \times \frac{4000 \times 50}{6000 \times 100 + 4000 \times 50} = 25$$

[this 25 could more easily be obtained by subtracting 75 from the total sample size of 100].

3. Refer to the last example. Suppose that the mean number of units of alcohol consumed by the 75 males in the past year turned out to be $\overline{x}_1 = 800$ and that the mean for the 25 females was $\overline{x}_2 = 100$. What is the best estimate of μ?

   Solution: $\dfrac{75 \times 800 + 25 \times 100}{75 + 25} = 625.$

   Note that this is a weighted average, $\frac{n_1}{n}\overline{x}_1 + \frac{n_2}{n}\overline{x}_2 = \frac{75}{100} \times 800 + \frac{25}{100} \times 100 = 625$, of the two sample means 800 and 100, with weights given by the proportions of males ($\frac{75}{100}$) and females ($\frac{25}{100}$) in the sample of 100. If you have difficulty with this, note that we would estimate μ by means of the sample mean number of units drank. This sample mean is of course

$$\frac{\text{total number of units drank by the 100 students in the sample}}{\text{total number of students in the sample}} =$$
$$\frac{\text{total number of units drank by the 75 males + total number of units drank by the 25 females}}{100} =$$
$$\frac{75 \times 800 + 25 \times 100}{100} = 625.$$

   It is important to note that the correct answer is *not* the ordinary average $\frac{800+100}{100} = 450$ of the numbers 800 and 100.

Sampling distributions

The purpose of the following long exercise is to give you an intuitive feeling for how sample means and sample standard deviations relate to the population mean and standard deviation. Some students may find this exercise very helpful, others less so. If you do not find it is helping you to improve your understanding of this topic you can skip it.

Recall that statistics is primarily concerned with making valid inferences about populations based on information contained in samples taken from these populations. In this sheet, we will examine some ideas that are key throughout the course in estimating unknown population parameters from the values of sample statistics. For illustrative purposes, we will take a *known* population, but note that in practise this will not be the case.
Consider the following population consisting of the 3 elements: {1, 2, 3}.
We first calculate the mean and variance of the population.
The mean and variance are got using the following general formulae for the mean and variance of *any finite* population $\{y_1, y_2, ..., y_N\}$:

$$\text{Population mean} := \mu = \frac{1}{N}\sum_{i=1}^{N} y_i \qquad (1)$$

We get $\mu = \frac{1}{3}(1+2+3) = 2$ (2)

Also, we have the following formula for the variance of *any finite* population $\{y_1, y_2, ..., y_N\}$:
Population variance $:= \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu)^2$ (3)

We thus obtain $\sigma^2 = \frac{1}{3}\left[(1-2)^2 + (2-2)^2 + (3-2)^2\right] = \frac{2}{3}$ (4)
Alternatively and equivalently, we can imagine a random variable X taking values 1,2 and 3 with probabilities

$$P(X = x) = \frac{1}{3},\ x = 1, 2, 3.$$

This distribution is shown in table form below.

TABLE 1

x	$P(X = x)$
1	1/3
2	3/3
3	1/3

Then note that the above population is equivalent to this random variable and its distribution. Then using the following formula for the mean (i.e. the expected value) of any *discrete* random variable Y

$$E(Y) = \sum yP(Y = y) \tag{5}$$

We get

$$\mu = E(X) = 1 \times \tfrac{1}{3} + 2 \times \tfrac{1}{3} + 3 \times \tfrac{1}{3} = 2 \tag{6}$$

agreeing with (2).

Also, we obtain the variance of X by using either of the following formulae for the variance of any *discrete* random variable Y :

$$\sigma^2 = Var(Y) = E(Y - E(Y))^2 = \sum (y - E(Y))^2 P(Y = y) \tag{7}$$

We thus get $\sigma^2 = Var(X) = (1-2)^2 \times \frac{1}{3} + (2-2)^2 \times \frac{1}{3} + (3-2)^2 \times \frac{1}{3} = \frac{2}{3}$ (8)
agreeing with (4) above.

Remark 2 *We could have used another equivalent formula for the variance (3); this equivalent formula is.*

$$Var(X) = EX^2 - E^2(X) \tag{9}$$

which since our X is discrete here, becomes

$$\sum x^2 P(Y = x) - E^2(X)$$

(Check that this again gives $\frac{2}{3}$ in our example).

Remark 3 *ASIDE: If X was **continuous** with density $f(x)$, we would have used the following formulae for the mean and variance:*

$$\mu = E(X) = \int xf(x)dx \tag{10}$$

and

$$\sigma^2 = Var(X) = E(X - E(X))^2 = \int (x - E(X))^2 f(x)dx = EX^2 - E^2(X) = \int x^2 \ f(x)dx - \mu^2 \tag{11}$$

Remark 4 *In all cases, the population standard deviation is always defined as*

$$\sigma = +\sqrt{\sigma^2} \tag{12}$$

We now study samples of size $n = 2$ taken with replacement from the above population. **Note that in the real world, only one sample is taken (without replacement), but by examining every possible sample, we can get an idea of how good any one sample would be in giving information about the population.**

The possible samples are: (1,1), (1,2), (2,1), (1,3), (3,1), (2,2,), (2,3), (3,2), (3,3).
We will calculate the mean and variance of each sample and construct two distributions.
Note that the mean and variance of any set of numbers $x_1, x_2, ..., x_n$ are, respectively,

$$\text{Sample Mean:= } \bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i \tag{13}$$

and

$$\text{Sample Variance:= } s^2 := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right] \tag{14}$$

You can try to fill in the following table (a few of the rows are already completed)

TABLE 2

Samples	Sample mean $\bar{x}$	Sample variance s^2	Sample st. dev. $s = \sqrt{s^2}$
$\{1,1\}$	$\frac{1+1}{2} = 1$	$\frac{1}{2-1}\{(1-1)^2 + (1-1)^2\} = 0$	$\sqrt{0} = 0$
$\{1,2\}$	$\frac{1+2}{2} = 1.5$	$\frac{1}{2-1}\{(1-1.5)^2 + (2-1.5)^2\} = \frac{1}{2}$	$\sqrt{\frac{1}{2}}$
$\{2,1\}$			
$\{2,2\}$			
$\{1,3\}$	2	2	$\sqrt{2}$
$\{3,1\}$			
$\{2,3\}$	2.5	$\frac{1}{2}$	$\sqrt{\frac{1}{2}}$
$\{3,2\}$			
$\{3,3\}$			

We next calculate the average of all the sample means, and the average of all the sample variances.
Let $\overline{X}$ be the mean of a random sample, and let S^2 be the variance of a random sample.

We get [using (1)] that the mean of the population of all nine $\overline{x}$ values is

$$E(\overline{X}) = \frac{2+1.5+1.5+2+2+2+2.5+2.5+3}{9} = 2. \tag{15}$$

Alternatively, we can get $E(\overline{X})$ by using the distribution of $\overline{X}$, and then the formula for the mean of a discrete random variable. The distribution of $\overline{X}$ is [from Table 2]

TABLE 3

$\overline{x}$	$P(\overline{X} = \overline{x})$
1	1/9
1.5	2/9
2	3/9
2.5	2/9
3	1/9

Then the mean of the sampling distribution of $\overline{X}$ is [using (5)]

$$E(\overline{X}) = 1 \times 1/9 + 1.5 \times 2/9 + 2 \times 3/9 + 2.5 \times 2/9 + 3 \times 1/9 = 2 \tag{16}$$

(agreeing with (15).

Similarly we see from Table 2 and (3) that the variance of $\overline{X}$ is

$$Var(\overline{X}) = \frac{(1-2)^2+(1.5-2)^2+(2-2)^2+(2.5-2)^2+(3-2)^2}{9} = 1/3 \tag{17}$$

Alternatively but equivalently, we see from Table 3 and (7) that the variance of the sampling distribution of $\overline{X}$ is

$$Var(\overline{X}) = (1-2)^2 \times 1/9 + (1.5-2)^2 \times 2/9 + (2-2)^2 \times 3/9 + (2.5-2)^2 \times 2/9 + (3-2)^2 \times 1/9 = 1/3 \tag{18}$$

(agreeing with (17).
Similarly, we can see from Table 2 that the sampling distribution of S^2 is

TABLE 4

s^2	$P(S^2 = s^2)$
0	3/9
1/2	4/9
2	2/9

From Table 2 and (1) that the mean of the population of nine s^2 values is

$$E(S^2) = \frac{0+0+0+1/2+1/2+1/2+1/2+2+2}{9} = \frac{2}{3} \tag{19}$$

Alternatively, using (5), we can see that the mean of the sampling distribution of S^2 is

$$E(S^2) = 0 \times 3/9 + 1/2 \times 4/9 + 2 \times 2/9 = 6/9 = 2/3 \quad (20)$$

Alternatively,using (5), we can see that the mean of the sampling distribution of S^2 is

$$E(S^2) = 0 \times 3/9 + 1/2 \times 4/9 + 2 \times 2/9 = 6/9 = 2/3 \quad (21)$$

From (15) or (16) and (2) or (6), we see that

$$E(\overline{X}) = \mu \quad (22)$$

that is, the mean of all the sample means is the population mean,
i.e. the mean of all the samples means is the populations mean,
i.e. the expected value of the random sample mean $\overline{X}$ is the
population mean μ.
Also comparing (17) or (18) with (4) or (8) we see that

$$Var(\overline{X}) = \sigma^2/n \quad (23)$$

That is, the variance of all the sample means is the population variance divided by the sample size.
Finally, notice from (4) or (8) with (20) or (21), we see that

$$E(S^2) = \sigma^2 \quad (24)$$

Summarizing, we have the following important properties:

Remark 5 (Important general properties of the sample mean $\overline{X}$ and sample variance S^2) *If a random sample of size n is taken without replacement from an infinite population (or from as finite population when sampling is with replacement) that has mean μ and variance σ^2, then the sample mean $\overline{X}$ and sample variance S^2 satisfy*

$$E(\overline{X}) = \mu \quad (25)$$
$$Var(\overline{X}) = \sigma^2/n \quad (26)$$

and

$$E(S^2) = \sigma^2 \quad (27)$$

In particular then, $\overline{X}$ and S^2 are *unbiased* estimators of μ and σ^2, respectively. This is because any estimator is said to be an unbiased estimator of a parameter if the mean of the distribution of that estimator is the parameter.

Remark 6 *It is helpful to note think of μ and σ as the limit, as the sample size n tends to ∞, of $\bar{x}$ and s, respectively.*

Remark 7 *In practise, we will not know the elements of the population and in particular will not know the parameters μ and σ^2. We are allowed to take one sample from the population. Ask yourself if you think that the sample mean $\bar{x}$ would be a good estimate of μ. From (22) or (25) we see that it is unbiased for μ. From (23) you can see that if n is large, the variance of $\overline{X}$ will decrease. In particular, you can see that you'd prefer to take as big a sample size as is feasible*

(taking cost, time and other considerations into account). Putting this another way, you'd be more likely to get a sample mean that is close to the unknown μ if you take a big sample size than if you take a small sample size. ***These points are very important!***

Normal approximation to the binomial distribution

1. Over-booking]To minimize loss due to no-shows, hotels and airlines often take more bookings than they can accommodate. Suppose that a certain airline has seats for 150 passengers on its Boeing 737-400 but has sold 153 tickets for the next flight. Assuming that 2% of all passengers for its flights do not show to claim their reservation, what is the probability that all passengers who show up will get a seat?

   Solution: Let X = the number in the sample of n =153 who will show up. We require $P(X \leq 150) \underset{\text{using the } \textbf{standardizing formula and } (\otimes)}{=}$
   $P(\frac{X-np}{\sqrt{np(1-p)}} \leq \frac{150-np}{\sqrt{np(1-p)}})$
   $\underset{\text{by the normal approx. to the binomial}}{\doteq} P(Z \leq \frac{150-np}{\sqrt{np(1-p)}}) = P(Z \leq \frac{150-153(0.98)}{\sqrt{153 \times 0.02 \times 0.98}})$
   $= P(Z \leq \frac{0.06}{\sqrt{2.9988}}) = P(Z \leq \frac{0.06}{1.73170}) = P(Z \leq 0.03465)$ =$1 - P(Z > 0.03465) = 1 - 0.4862 = 0.5138$, from standard normal tables.

2. Solve the last example using ($\otimes\otimes$) instead of ($\otimes$), that is, work with the proportion, $\hat{p}$, of people who will show up rather than the number, X, who will show up.

   Solution: We require $P(\hat{p} \leq \frac{150}{153}) \underset{\text{using the } \textbf{standardizing formula and } (\otimes\otimes)}{=}$
   $P(\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \leq \frac{\frac{150}{153}-p}{\sqrt{\frac{p(1-p)}{n}}}) \underset{\text{by the normal approx. to the binomial}}{\doteq} P(Z \leq \frac{\frac{150}{153}-p}{\sqrt{\frac{p(1-p)}{n}}}) =$
   $P(Z \leq \frac{\frac{150}{153}-0.98}{\sqrt{\frac{0.02\times 0.98}{153}}}) = P(Z \leq 0.03465) = P(Z \leq 0.03465) = 1 - P(Z > 0.03465) = 1-0.4862 = 0.5138$=$1-P(Z > 0.03465) = 1-0.4862 = 0.5138$, as above.

3. What is the minimum number of Galwegians that should be sampled at random so that with probability at least 0.95, the proportion of smokers in the sample will not differ from the unknown population of smokers by more than ± 0.03 ? *Note:* If $Z \backsim N(0,1)$, $P(Z > 1.96) = 0.025$.

Solution: Let X = the number of smokers in the random sample of size n. We must find n so that

$$P(|\frac{X}{n} - p| \leq 0.03) = 0.95 \qquad (*)$$

(Here the equals sign really means " $\geq$ " if we can't get exact equality, which we won't be able to do). Now (*) is

$$P(|X - np| \leq 0.03n) = 0.95 \qquad (**)$$

Standardizing in (**) we get

$$P(|\frac{X - np}{\sqrt{np(1-p)}}| \leq \frac{0.03n}{\sqrt{np(1-p)}}) = 0.95$$

so that (by the normal approximation to the binomial)

$$P(|Z| \leq \frac{0.03n}{\sqrt{np(1-p)}}) = 0.95 \qquad (***)$$

Hence from Z tables, or the information in the question, we must have

$$\frac{0.03n}{\sqrt{np(1-p)}} = 1.96.$$

We want to solve this for n, but we do not know p. To maximize the variance of X, we put $p = 0.5$ [this will maximize the denominator of (***) and hence minimize the left hand side of (***), thus ensuring that no matter what value p has, we will have a probability of at least 0.95 that the sample proportion will be within 0.03 of the population proportion]. Solving the equation $\frac{0.03n}{\sqrt{n(0.5)(1-0.5)}} = 1.96$, i.e. $2 \times 0.03\sqrt{n} = 1.96$ for n, we get $\sqrt{n} = \frac{1.96}{0.06} = 32.666667$ or $n = 1067.111$. Now we round up, getting $n = 1068$. [If we rounded down, the probability with which the sample proportion will be within 0.03 of the population proportion could be slightly less than 0.95.]

Remark 8 (Helpful Hint) *In general, if we require a sample proportion to be within $\pm E$ of the (unknown) population proportion p with probability at least 0.95, simply put*

n equal to the smallest integer that is greater than or equal to $\left(\frac{1.96}{2E}\right)^2$