Part (A): Review of Probability [Statistics I revision]

1 Definition of Probability

1.1 Experiment

An experiment is any procedure whose outcome is uncertain

- toss a coin
- throw a die
- buy a lottery ticket
- measure an individual's height h

1.2 Sample Space

A sample space, S, is the set of all possible outcomes of a random experiment

- $S = \{H, T\}$
- $S = \{1, 2, 3, 4, 5, 6\}$
- $S = \{\textit{WIN, LOSE}\}$
- $S = \{h : h \geq 0\}$

1.3 Events

An event, A, is an element, or appropriate subset of S
e.g. roll a die, $S = \{1, 2, 3, 4, 5, 6\}$

- event A, that an even number is obtained is given by $A = \{2, 4, 6\}$

- event B, that a number no greater than 4 is obtained is described by $B = \{1, 2, 3, 4\}$

S can be discrete (e.g. coin, die, lottery), or continuous (e.g. height).

In the first part of this course we will deal with the discrete case (as in Stats 1).

Set operations

If A, B are events, according to Set Theory, we define the operations:

- $A \cup B$: union, 'A or B happening'
- $A \cap B$: intersection, 'A and B happening'
- A^c (or A', $\bar{A}$): complement of A wrt S, 'not A'

  Notice that: $A \cup A^c = S$, $A \cap A^c = \emptyset$

1.4 Probability mass function

A probability function, $P(s)$, is a real-valued function of a collection of events from a sample space, S, attaching a value in $[0, 1]$ to each event, satisfying the axioms:

A1. $P(A) \geq 0$ for any event A

A2. $P(S) = 1$

A3. If $A_1, A_2, A_3, \ldots$ is a countable collection of mutually exclusive events ($A_i \cap A_j = \emptyset$ for $i \neq j$, i.e. they have no common elements), then $P(\cup_i A_i) = \sum_i P(A_i)$

Important properties of $P(s)$

- $P(\emptyset) = 0$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

  This can be generalised for n events, e.g. for $n = 3$:

  $P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3) + P(A_1 \cap A_2 \cap A_3)$.

  Notice: $P(A \cup B) = P(A) + P(B)$ not always true!

  It only holds for disjoint (mutually exclusive) events.

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

  Proof: A, A^c are disjoint, and $A \cup A^c = S$

  $\Rightarrow P(A \cup A^c) = P(S) = 1$

  $\Rightarrow P(A) + P(A^c) = 1$

  In many practical situations it is easier to calculate $P(A^c)$ first and use this property to calculate $P(A)$.

1.5 Independence and Conditional Probability

For any event A with $P(A) > 0$ we can define

$$P(B/A) = \frac{P(A \cap B)}{P(A)} \qquad (1)$$

If A and B are independent then by definition $P(A \cap B) = P(A) \times P(B)$ and therefore

$$P(B/A) = \frac{P(A)P(B)}{P(A)} = P(B).$$

Note that rearrangement of (1) gives the chain (multiplication) rule:

$$P(A \cap B) = P(A)P(B/A) = P(B)P(A/B). \qquad (2)$$

This can also be generalised for n events.

1.6 Total Probability Rule

If $A_1, A_2, \ldots, A_k$ are mutually exclusive events and form a partition of a sample space S (i.e. $\cup_{i=1}^{k} A_i = S$), with $P(A_i) > 0 \qquad \forall i$,

then for an event $B \in S$:

$$P(B) = \sum_{i=1}^{k} P(B \cap A_i) = \sum_{i=1}^{k} P(A_i) P(B/A_i) \qquad (3)$$

Example:

An insurance company covers claims from 4 different motor portfolios, A_1, A_2, A_3 *and* A_4. *Portfolio* A_1 *covers* 4000 *policy holders;* A_2 *covers* 7000*;* A_3 *covers* 13000*; and* A_4 *covers* 6000 *of the total* 30000 *policy holders insured by the company. It is estimated that the proportions of policies that will result in a claim in the following year in each of the portfolios are* $8\%, 5\%, 2\%$ *and* 4% *respectively. What is the probability that a policy chosen randomly from one of the portfolios will result in a claim in the following year?*

Solution ...

1.7 Bayes Theorem

If $P(A) > 0$ and $P(B) > 0$, from (1)

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B/A)}{P(B)} \qquad (4)$$

or, if $A_1, A_2, \ldots, A_k$ form a partition of a sample space S with $P(A_i) > 0 \quad \forall i$, and $P(B) > 0$, then from (3)

$$P(A_j/B) = \frac{P(A_j)P(B/A_j)}{\sum_{i=1}^{k} P(A_i)P(B/A_i)}, \qquad j = 1, 2, \ldots, k. \qquad (5)$$

Example: *(previous cont.)*

If a claim rises from a policy in the year concerned, what is the probability that the policy belongs to portfolio A_3?

Solution ...

1.8 Random variables

Until now we have associated probabilities with events in a sample space. We will now review the concept of random variables.

Definition:

A random variable (r.v.) X is a function from a sample space S to $\mathbb{R}$ (the real numbers).

That is, to any outcome, s, we associate a real number $X(s)$.

The range of a r.v. X, denoted as S_x, is the set $\{r \in \mathbb{R} : r = X(s)$ for some $s \in S\}$.

[i.e. the set of all real numbers r which are equal to $X(s)$ for some outcome s.]

We will usually denote random variables using capital letters, e.g. X, and their realisations (values) using small letters, e.g. x.

If S is finite or countable then the range of X is a set of discrete numbers. We say X is a discrete random variable.

Some simple examples:

(i) Experiment: Toss a balanced coin 3 times.

Sample space S: All possible sequences of T, H of length 3.

Let $X(s)$ = no. of Heads in outcome (sequence) s. Then

X is a discrete random variable.

(ii) Experiment: Toss a balanced coin until a 'H' is thrown.

Sample space: $S = \{H, TH, TTH, \ldots\}$.

Let $X(s)$ = no. of throws (i.e. length of s). Then

X is a discrete random variable.

1.9 Probability function of a discrete r.v.

Any probability function on S leads naturally to a probability function on the range of X, defined by

$$f_X(x) = P(X = x) = \sum_{s:X(s)=x} P(s).$$

In **Example (i)** before:

s :	HHH	HHT	HTH	THH	TTT	TTH	THT	HTT
Prob:	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$
X :	3	2	2	2	0	1	1	1

Probability of 2 Hs is
$P(X = 2) = \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8}$.

Usually we consider r.v.'s which describe some useful summary of the outcome of an experiment, e.g. the total score in a multiple choice test.

Axioms of probability

For a probability function of a discrete r.v. to satisfy the axioms of probability, it is sufficient that

i. $f_X(x) \geq 0$ for all $x \in S_x$

ii. $\sum_{\text{all } x} f_X(x) = 1$

1.10 S infinite

Until now in the examples we have only considered cases where the sample space S consists of a countable and finite number of events. However, there may be also an infinite number of possibilities involved with the experiment under consideration.

Some results on infinite series ...

Example:

Two (TV show) players (A and B) play a game in which each wins a prize if s/he is the first to choose the correct screen out of the 6-screen display panel shown to them. Player A goes first, and the prize is randomly re-allocated to one of the screens after each player's choice (so that the probability of either player winning at a single draw is independently $1/6$). What is the probability that A wins?

Solution ...

2 Some common discrete distributions [Statistics I]

2.1 The binomial distribution: Bin(n, p)

Consider an experiment consisting of n independent trials where the probability of success is $p, 0 < p < 1$.

Let X denote the number of successes (S's). Then

$$f_X(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, 2, \ldots, n. \quad (6)$$

Why? The sample space S consists of all sequences of S and F of length n. For any such sequence, s, with x S's and (therefore) $(n - x)$ F's the probability is
$P(s) = p^x(1-p)^{n-x}$
(since the trials are independent so probabilities multiply).

There are $\binom{n}{x}$ such sequences, and therefore (6) follows.

2.2 The geometric distribution: Geometric(p) [Statistics I revision]

Again consider a sequence of independent trials with probability of ‘success’ $P(S) = p$. Let X be the number of trials required to achieve the 1st success [e.g. number of attempts required to pass the driving test].

$S = S, FS, FFS, FFFS, \ldots$
$X = 1, 2, 3, 4, \ldots$

Therefore for $x = 1, 2, 3, \ldots$

$$f_X(x) = P(X = x) = P(\overbrace{FF \ldots F}^{x-1} S) = (1-p)^{x-1} p. \qquad (7)$$

The memoryless property

The geometric distribution has the memoryless property. That is, given that there have already been n trials without success, the probability that x additional trials will be required for the first success is independent of n:

$$P(X > x + n / X > n) = P(X > x) \qquad \text{[Show]}$$

[Is the number of attempts required to pass the driving test well described by the geometric distribution?]

2.3 [NEW!] The negative binomial distribution: NBin(r, p)

This is a generalisation of the geometric distribution.

Consider the same sequence of independent trials as in (2.2) and let X denote the number of trials required to achieve r successes, where r can be any positive integer. Clearly the range of X is $r, r+1, r+2, \ldots$.

What is $f_X(x) = P(X = x), \qquad x \geq r$?

Let s be an outcome for which $X(s) = x$. Then s is a sequence of S's and F's such that:

(i) s has length x

(ii) s has r S's and $x - r$ F's

(iii) the last entry is S

From (ii), $P(s) = (1-p)^{x-r} p^r$ (independence of trials!!).

How many such outcomes are there?

Trial:	1	2	3	$\ldots$	$x-1$	x
Result:	$?$	$?$	$?$	$\ldots$	$?$	S

$$(r-1)S's \ \& \ (x-r)F's$$

There are $\binom{x-1}{r-1}$ ways of assigning the S's to position $1, 2, \ldots, x-1$.

Therefore

$$f_X(x) = P(X = x) = \binom{x-1}{r-1}(1-p)^{x-r}p^r, \qquad x = r, r+1, \ldots \quad (8)$$

Relationship to geometric distribution

1. NBin$(1, p) \equiv$ Geometric(p)

2. If $X_1, X_2, \ldots, X_r$ are independent Geometric(p) r.v.'s, then
   $Y = X_1 + X_2 + \ldots + X_r \sim$ NBin(r, p).

To summarise:

In a sequence of independent 'trials', each with probability of 'success' p:

- the number of successes in n trials follows a bin(n, p) distribution
- the number of trials until the first success follows a geometric(p) distribution
- the number of trials until the rth success follows a NBin(r, p) distribution.

2.4 [NEW!] The Poisson distribution: Poisson(λ)

This is a distribution which is often used to describe the outcomes of experiments that involve counting objects or events (e.g. the number of road accidents occurring on a stretch of road in a 1-month period).

Its range is $\{0, 1, 2, 3, \ldots\}$, and its probability function is given by

$$f_X(x) = P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \qquad x = 0, 1, 2, 3, \ldots, \qquad \lambda > 0. \tag{9}$$

It has connections with the Bin(n, p) distribution, and also the Exponential distribution (see later).

3 Expectation, variance and independence [Statistics I]

3.1 Mathematical Expectation of a discrete r.v. X

(Also referred to as expected value or mean of X.)

The expectation of a r.v. X is a weighted average of all possible values of X, with weights determined by the probability distribution of X. It measures where the centre of the distribution lies.

Definition:

Let X be a discrete r.v. with probability mass function $f_X(x)$. Then its expected value is

$$E(X) = \sum_{x \in S_x} x f_X(x) = \mu_x \qquad \text{(mean of X)} \tag{10}$$

Notice that if the above sum does not converge for an infinite set S_x, then $E(X)$ is not defined.

Properties of expectation:

If X, Y are r.v.'s and $a, b \in \mathbb{R}$ are constants:

(i) $E(X + Y) = E(X) + E(Y)$

(ii) $E(a) = a, \qquad E(aX + b) = aE(X) + b$

(iii) If $X \geq 0$, then $E(X) \geq 0$

3.2 Expectation of a function of a discrete r.v.

Let X be a r.v. with probability function $f_X(x)$, and $Y = g(X)$ a real-valued function of it. Then Y is also a r.v., and its expectation is given by

$$E(Y) = \sum_{x \in S_x} g(x) f_X(x) = \sum_y y f_Y(y).$$

Example:

An individual invests an amount of £$105k$ on a financial product. The (total) return of his investment at the end of the following year is given in the table:

Return (in £1000's) x :	118	113	110	107
Probability $f_X(x)$:	$\frac{3}{20}$	$\frac{7}{20}$	$\frac{7}{20}$	$\frac{3}{20}$

If the profit of the investment, Y, is given (as a function of the total return X) by the formula $Y = X^{0.99} - 105$, find the expected profit of the investor.

Solution:

$$
\begin{aligned}
E(Y) &= \sum_x g(x) f_X(x) \\
&= (118^{0.99} - 105) \times \frac{3}{20} + (113^{0.99} - 105) \times \frac{7}{20} \\
&\quad + (110^{0.99} - 105) \times \frac{7}{20} + (107^{0.99} - 105) \times \frac{3}{20} \\
&= 7.5 \times \frac{3}{20} + 2.8 \times \frac{7}{20} - 0.1 \times \frac{7}{20} - 2.9 \times \frac{3}{20} = 1.6
\end{aligned}
$$

[An expected profit of $£1600$. However, notice that with probability $1/2$ the investor will lose money! Is this a ‘good’ investment?]

3.3 Variance of a discrete r.v. X

The variance of a random variable x (var(X), also σ_x^2) is defined as the expectation of the deviation (squared) of X from its mean (expected value), i.e.

$$\text{var}(X) = E\{X - E(X)\}^2 \tag{11}$$

and for a discrete r.v. is calculated as

$$\text{var}(X) = \sum_{x \in S_x} \{x - E(X)\}^2 f_X(x) \tag{12}$$

It measures the 'spread' of the r.v. X, i.e. the extent to which the distribution of X is dispersed around its mean μ_x.

It can also be calculated as

$$\text{var}(X) = E(X^2) - E\{X\}^2. \tag{13}$$

The positive square root of the variance is called the standard deviation (s.d.) of X,

$$\sigma_x = \sqrt{\text{var}(X)}$$

Properties of variance

(i) $\text{var}(X) \geq 0$

with equality if and only if X is a constant.

(ii) If $a, b \in \mathbb{R}$ are constants, then

$\text{var}(a + bX) = b^2 \, \text{var}(X)$.

3.4 Independence of r.v.'s

Definition

The discrete r.v.'s X, Y are said to be independent of each other iff

$$P(X = x \cap Y = y) = P(X = x) \times P(Y = y)$$

$\forall x \in S_x, \forall y \in S_y$.

Properties

If X, Y are independent r.v.'s, then

(i) $E(XY) = E(X)E(Y)$

(ii) $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y)$ [Prove]

Counter **example** for property (ii):

If $X = Y$ (NOT independent), then

$$\begin{aligned}\text{var}(X+Y) &= \text{var}(2X) \\ &= 4\text{var}(X) \neq \text{var}(X) + \text{var}(Y)\end{aligned}$$