PART 2

Probability Theory

2.1

SET THEORY

Definition 2.1.1. The set, S, of all possible outcomes of a particular experiment is called the *sample space* for the experiment.

Definition 2.1.2. An *event* is any collection of possible outcomes of an experiment, that is, any subset of S (including S itself).

Example 2.1.3 (Event operations)**.** Selecting cards...

Theorem 2.1.4. *For any three events, A, B, C, defined on a sample space S,*

a. *Commutativity*

$$A \cup B = B \cup A,$$
$$A \cap B = B \cap A;$$

b. *Associativity*

$$A \cup (B \cup C) = (A \cup B) \cup C,$$
$$A \cap (B \cap C) = (A \cap B) \cap C;$$

c. *Distributive Laws*

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)\,,$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)\,;$$

d. *DeMorgan's Laws*

$$(A \cup B)^c = A^c \cap B^c,$$
$$(A \cap B)^c = A^c \cup B^c.$$

Proof can be done as an exercise

Definition 2.1.5. Two events A and B are disjoint (or *mutually exclusive*) if $A \cap B = \emptyset$. The events $A_1, A_2, ...$ are *pairwise disojoint* (or *mutually exclusive*) if $A_i \cap A_j = \emptyset$ for all $i \neq j$.

Definition 2.1.6. If $A_1, A_2, ...$ are pairwise disjoint and $\bigcup_{i=0}^{\infty} A_i = S$, then the collection $A_1, A_2, ...$ forms a *partition* of S.

Basics of probability Theory

Theorem 2.1.7. *If P is a probability function and A is any set in S, then*

a. $P\left(\emptyset\right) = 0$, *where $\emptyset$ is the empty set;*
b. $P\left(A\right) \leq 1$*;*
c. $P\left(A^c\right) = 1 - P\left(A\right)$*.*

Theorem 2.1.8. *If P is a probability function and A and B are any sets in $\mathcal{B}$, then*

a. $P\left(B \cap A^c\right) = P\left(B\right) - P\left(A \cap B\right)$*;*
b. $P\left(A \cup B\right) = P\left(A\right) + P\left(B\right) - P\left(A \cap B\right)$*;*
c. *If $A \subset B$, then $P\left(A\right) \leq P\left(B\right)$.*

Theorem 2.1.9. *If a job consists of k separate tasks, the ith of which can be done in n_i ways, $i = 1, ..., k$, then the entire job can be done in $n_1 \times n_2 \times ... \times n_k$ ways.*

This is sometimes known as the Fundamental Theorem of Counting. The proof for this can be done as an exercise.

Example 2.1.10 (Lottery - II)**.** Although the Fundamental Theorem of Counting is a reasonable place to start...

Definition 2.1.11. For a positive integer n, $n!$ (read n factorial) is the product of all of the positive integers less than or equal to n. That is,

$$n! = n \times (n-1) \times (n-2) \times ... \times 3 \times 2 \times 1.$$

Furthermore, we define $0! = 1$.

Definition 2.1.12. For nonnegative integers n and r, where $n \geq r$, we define the symbol $\binom{n}{r}$, read n *choose* r, as

$$\binom{n}{r} = \frac{n!}{r!\left(n-r\right)!}$$

	Without replacement	With replacement
Ordered	$\frac{n!}{(n-r)!}$	n^r
Unordered	$\binom{n}{r}$	$\binom{n+r-1}{r}$

Example 2.1.13 (Poker)**.** Consider choosing a five-card poker hand...

Example 2.1.14 (Sampling with replacement)**.** Consider sampling $r = 2$ items from $n = 3$ items...

Some authors argue that it is appropriate to assign equal probabilities to the unordered outcomes when "randomly distributing r indistinguishable balls into n distinguishable urns." That is, an urn is chosen at random and a ball placed in it, and this is repeated r times. The order in which the balls are placed is not recorded so, in the end, an outcome such as $\{1, 3\}$ means one ball is in urn

1 and one ball is in urn 3.

But here is the problem with this interpretation. Suppose two people observe this process, and Observer 1 will assign probability $\frac{2}{9}$ to the event $\{1,3\}$. Observer 2, who is observing exactly the same process, should also assign probability $\frac{2}{9}$ to this event. But if the six unordered outcomes are written on identical pieces of paper and one is randomly chosen to determine the placement of the balls, then the unordered outcomes each have probability $\frac{1}{6}$. So Observer 2 will assign probability $\frac{1}{6}$ to the event $\{1,3\}$.

The confusion arises because the phrase "with replacement" will typically be interpreted with the sequential kinda of sampling we described above, leading to assigning a probability $\frac{2}{9}$ to the event $\{1,3\}$. This is the correct way to proceed, as probabilities should be determined by the sampling mechanism, not whether the balls are distinguishable or indistinguishable.

Example 2.1.15 (Caldulating an average)**.** An illustration of the distinguishable/indistinguishable approach...

If there are k places and we have m different numbers repeated $k_1, k_2, ..., k_m$ times, then the number of ordered samples is $\frac{k!}{k_1!k_2!...k_m!}$. This type of counting is related to the *multinomial distribution*, which we will see in Section 4.6.

Figure 1.2.2 (p. 19 of book) is an elementary version of a very important statistical technique known as the *bootstrap* (Efron and Tibshirani 1993). We will return to the bootstrap in Section 10.1.4

Conditional Probability and Independence

Definition 2.1.16. If A and B are events in S and $P(B) > 0$, then the *conditional probability* of *A given B*, written $P\left(A|B\right)$, is

$$P\left(A|B\right) = \frac{P(A \cap B)}{P(B)}$$

Note that what happens in the conditional probability calculation is that B becomes the sample space: $P\left(B|B\right) = 1$. The intuition is that our original sample space, S, has been updated to B. All further occurrences are then calibrated with respect to their relation to B. In particular, note what happens to conditional probabilities of disjoint sets. Suppose A and B are disjoint, so $P(A \cap B) = 0$. It then follows that $P\left(A|B\right) = P\left(B|A\right) = 0$.

Theorem 2.1.17 (Bayes' Rule)**.** *Let $A_1, A_2, ...$ be a partition of the sample space, and let B be any set. Then for each $i = 1, 2, ...,$*

$$P\left(A_i|B\right) = \frac{P\left(B|A_i\right) P(A_i)}{\sum_{j=1}^{\infty} P\left(B|A_j\right) P\left(A_j\right)}.$$

Definition 2.1.18. Two events, A and B, are *statistically independent* if

$$P(A \cap B) = P(A)\, P(B).$$

Note that independence could have been equivalently defined by either $P\left(A|B\right) = P(A)$ or $P\left(B|A\right) = \frac{P\left(A|B\right)P(B)}{P(A)} = \frac{P(A)P(B)}{P(A)} = P(B)$. The advantage of defining it this way is that it treats the events symmetrically and will be easier to generalise to more than two events.

Theorem 2.1.19. *If A and B are independent events, then the following pairs are also independent:*

a. *A and B^c*
b. *A^c and B*
c. *A^c and B^c*

All three statements can be proved as an exercise.

Definition 2.1.20. A collection of events $A_1, ..., A_n$ are *mutually independent* if for any subcollection $A_{i_1}, ..., A_{i_k}$, we have

$$P\left(\bigcap_{j=1}^{k} A_{i_j}\right) = \prod_{j=1}^{k} P\left(A_{i_j}\right).$$

Random Variables

Definition 2.1.21. A *random variable* is a function from a sample space S into the real numbers.

Distribution Functions

Definition 2.1.22. The *cumulative distribution function* or *cdf* of a random variable X, denoted by $F_X(x)$, is defined by

$$F_X(x) = P_X(X \leq x), \ \forall x.$$

Theorem 2.1.23. *The function $F(x)$ is a cdf if and only if the following three conditions hold:*

a. $\lim_{x\to-\infty} F(x) = 0$ *and* $\lim_{x\to\infty} F(x) = 1$.
b. *$F(x)$ is a nondecreasing function of x.*
c. *$F(x)$ is right-continuous; that is, for every number x_0, $\lim_{x\downarrow x_0} F(x) = F(x_0)$.*

Whether a cdf is continuous or has jumps corresponds to the associated random variable being continuous or not. In fact, the association is such that it is convenient to define continuous random variables in this way.

Definition 2.1.24. A random variable X is *continuous* if $F_X(x)$ is a continuous function of x. A random variable is *discrete* if $F_X(x)$ is a step function of x.

Definition 2.1.25. The random variables X and Y are *identically distributed* if, for every set $A \in \mathcal{B}$, $P(X \in A) = P(Y \in A)$.

Theorem 2.1.26. *The following two statements are equivalent:*

a. *The random variables X and Y are identically distributed.*
b. *$F_X(x) = F_Y(x)$ for every x.*

Sufficiency can be proven as an exercise, necessity is very hard. See page 34.

Density and Mass Functions

Associated with a random variable X and its cdf F_X is another function, called either the probability density function (pdf) or probability mass function (pmf). The terms pdf and pmf refer, respectively, to the continuous and discrete cases. Both pdfs and pmfs are concerned with "point probabilities" of random variables.

Definition 2.1.27. The *probability mass function (pmf)* of a discrete random variable X is given by

$$f_X(x) = P(X = x) \ \forall x.$$

Definition 2.1.28. The *probability density function* or *pdf*, $f_X(x)$, of a continuous random variable X is the function that satisfies

$$F_X(x) = \int_{-\infty}^{x} f_X(t) \ dt \ \forall x.$$

Note (notation)*.* The expression "X has a distribution given by $F_X(x)$" is abbreviated symbolically by "$X \sim F_X(x)$", where we read the symbol "$\sim$" as "is distributed as". We can similarly write $X \sim f_X(x)$ or, if X and Y have the same distribution, $X \sim Y$.

Since $P(X = x) = 0$ if X is a continuous random variable,

$$P(a < X < b) = P(a < X \leq b) = P(a \leq X < b) = P(a \leq X \leq b).$$

It should be clear that the pdf (or pmf) contains the same information as the cdf and we should choose whichever one makes the problem simpler.

Theorem 2.1.29. *A function $f_X(x)$ is a pdf (or pmf) of a random variable X if and only if*

a. *$f_X(x) \geq 0$ for all x.*
b. *$\sum_x f_X(x) = 1$ (pmf) or $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$ (pdf).*

2.2

EXPECTATION

Expectation of a Random Variable

Definition 2.2.1. The expected value, or mean, or first moment, of X is defined to be

$$\mathbb{E}\left[X\right] = \int x dF\left(x\right) = \begin{cases} \sum_x x f\left(x\right) & \text{If } X \text{ is discrete} \\ \int x f\left(x\right) dx & \text{If } X \text{ is continuous} \end{cases} \tag{1}$$

assuming that the sum (or integral) is well defined. Use the following notation to denote the expected value of X

$$\mathbb{E}\left[X\right] = \mathbb{E}X = \int x dF\left(x\right) = \mu = \mu_X. \tag{2}$$

$\int x dF\left(X\right)$ is used as a convenient unifying notation so we do not have to write $\sum_x x f\left(x\right)$ for discrete RVs and $\int x f\left(x\right) dx$ for continuous ones but be aware that $\int x dF\left(X\right)$ has a precise meaning that is discussed in real analysis course.

To ensure that $\mathbb{E}\left[X\right]$ is well defined, we say that $\mathbb{E}\left[X\right]$ exists if $\int_x \left|x\right| dF_X\left(x\right) < \infty$. Otherwise, we say that the expectation does not exist.

To compute $\mathbb{E}\left[Y\right]$ when $Y = r\left(X\right)$, one way is to find $f_Y\left(y\right)$ then compute $\mathbb{E}\left[Y\right] = \int y f_Y\left(y\right) dy$.

Theorem 2.2.2 (The Rule of the Lazy Statistician)**.** *Let $Y = r\left(X\right)$. Then*

$$\mathbb{E}\left[Y\right] = \mathbb{E}\left[r\left(X\right)\right] = \int r\left(x\right) dF_X\left(x\right). \tag{3}$$

$$Z = r\left(X, Y\right) \Rightarrow \mathbb{E}\left[Z\right] = \mathbb{E}\left[r\left(X, Y\right)\right] = \int \int r\left(x, y\right) dF\left(x, y\right). \tag{4}$$

The k^{th} moment of X is defined to be $\mathbb{E}\left[X^k\right]$ assuming that $\mathbb{E}\left[\left|X\right|^k\right] < \infty$.

Theorem 2.2.3. *If the k^{th} moment exists and if $j < k$ then the j^{th} moment exists.*

Proof.

$$\begin{aligned}
\mathbb{E}\left[|X|^j\right] &= \int_{-\infty}^{\infty} |x|^j f_X(x)\ dx \\
&= \int_{|x|\leq 1} |x|^j f_X(x)\ dx + \int_{|x|>1} |x|^j f_X(x)\ dx \\
&\leq \int_{|x|\leq 1} f_X(x)\ dx + \int_{|x|>1} |x|^k f_X(x)\ dx \\
&\leq 1 - \mathbb{E}\left[|X|^k\right] < \infty.
\end{aligned}$$

■

The k^{th} central moment is defined to be $\mathbb{E}\left[(X-\mu)^k\right]$.

Properties of Expectations

Theorem 2.2.4. *If $X_1, ..., X_n$ are RVs and $a_1, ..., a_n$ are constants, then*

$$\mathbb{E}\left[\sum_i a_i X_i\right] = \sum_i a_i \mathbb{E}\left[X_i\right]. \tag{5}$$

Theorem 2.2.5. *Let $X_1, ..., X_n$ be independent RVs. Then*

$$\mathbb{E}\left[\prod_{i=1}^{n} X_i\right] = \prod_i \mathbb{E}\left[X_i\right]. \tag{6}$$

Note the summation rule does not require independence but the product rule does.