Probability Theory

This Version: July 24, 2015
1 Set Theory

We want to draw conclusions about a collection of objects by conducting experiments.

What are the possible outcomes?

Definition 1 The set $S$, of all possible outcomes of a particular experiment is called the sample space for the experiment.

If we toss a coin, $S = \{H, T\}$
If we roll a dice, $S = \{1, 2, 3, 4, 5, 6\}$
If we measure time (in years) until your country wins a World Cup, $S = (0, \infty)$ (or $S = (4, \infty)$ if you are Colombian or American today).

Sample spaces can be
- **Countable:** if its cardinality (number of elements) equals that of a subset of the natural numbers.
  Example: $S = \{H, T\}, S = \{\text{Grains of salt in South Beach}\}$
- **Uncountable:** if its cardinality number is larger than that of the set of all natural numbers. Example: $S = (0, \infty), S = (0, 1)$
Now we can consider collections of possible outcomes

**Definition 2** An event is any collection of possible outcomes of an experiment, any subset of $S$.

As such, the set $A \subseteq S$, or event $A$, can be the outcome of an experiment.

**Definition 3** Relationships between sets:

**Containment:** $A \subseteq B \iff x \in A \implies x \in B$

**Equality:** $A = B \iff A \subseteq B$ and $B \subseteq A$

### 1.1 Set Operations:

**Union:** $A \cup B = \{x : x \in A \text{ or } x \in B\}$

**Intersection:** $A \cap B = \{x : x \in A \text{ and } x \in B\}$

**Complementation:** $A^c = \{x : x \notin A\}$

Example: Selecting a card at random and noting its suit: clubs(C), diamonds (D), hearts (H) or spades(P).
Proposition 1  Properties of set operations. For any three events $A$, $B$, and $C$, defined on a sample space $S$,

1. Commutativity: $A \cup B = B \cup A$
   $A \cap B = B \cap A$

2. Associativity: $A \cup (B \cup C) = (A \cup B) \cup C$
   $A \cap (B \cap C) = (A \cap B) \cap C$

3. Distributive Laws: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
   $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

4. DeMorgan’s Laws: $(A \cup B)^c = A^c \cap B^c$
   $(A \cap B)^c = A^c \cup B^c$
Unions and intersections can be defined over infinite collections of sets. **Countable case:** Let $A_1, A_2, A_3, \ldots$ be a collection of sets defined each on a sample space $S$, then

\[ \bigcup_{i=1}^{\infty} A_i = \{ x \in S : x \in A_i \text{ for some } i \} \]

\[ \bigcap_{i=1}^{\infty} A_i = \{ x \in S : x \in A_i \text{ for all } i \} \]

Examples: let $S = (0, 1]$ and define $A_i = [1/i, 1]$. Then

\[ \bigcup_{i=1}^{\infty} A_i = S \]

\[ \bigcap_{i=1}^{\infty} A_i = \{1\} \]
Not let us define sets that share no points, sets that do not overlap.

**Definition 4** Two events $A$ and $B$ are disjoint if

$$A \cap B = \emptyset$$

Events $A_1, A_2, \ldots$ are pairwise disjoint (or mutually exclusive) if

$$A_i \cap A_j = \emptyset \text{ for all } i \neq j$$

Example: $A_i = [i, i + 1)$, for $i = 0, 1, \ldots$

Sets can be represented as a collection of disjoint parts:

**Definition 5** Partition. If $A_1, A_2, \ldots$ are pairwise disjoint and

$$\bigcup_{i=1}^{\infty} A_i = S$$

then the collection $A_1, A_2, \ldots$ forms a partition of $S$. 
2 Basic Probability Theory

2.1 Axiomatic Foundations

If we repeat a experiment several times, the notion of “more probable outcomes” seems to emerge naturally.

For each event \( A \subset S \) we would like to assign a number \( P(A) \) that embodies the concept of “how probable \( A \) is.”

Hence it would seem natural to define the domain of \( P \) as all subsets of \( S \). However (for technical reasons), things are not that simple and hence... sigma-algebras appear.
Definition 6  **Sigma Algebra.** A collection of subsets of $S$ is called a sigma algebra (also called sigma field), denoted $\Sigma$, if it satisfies

1. $\emptyset \in \Sigma$  (empty set is an element of $\Sigma$)

2. If $A \in \Sigma$, then $A^c \in \Sigma$  (closed under complements)

3. If $A_1, A_2, \ldots \in \Sigma$, then $\bigcup_{i=1}^{\infty} A_i \in \Sigma$  (closed under countable unions)

$\emptyset$ is a subset of any set so 1 and 2 imply $S$ is in any sigma algebra.

Also since

$$\left( \bigcup_{i=1}^{\infty} A_i^c \right)^c = \bigcap_{i=1}^{\infty} A_i$$

2, 3 and DeMorgan’s Law imply that a sigma algebra is closed under countable intersections.

$$\{\emptyset, S\}$$ is the trivial sigma algebra

We will focus on the smallest sigma algebra that contains all open sets in $S$

Example:

508-B (Statistics Camp, Wash U, Summer 2015)
2.2 Computing Probabilities

Definition 7 **Probability Function**. Given a sample space and an associated sigma algebra $\Sigma$, a probability function is a function $\Pr$ with domain $\Sigma$ that satisfies

1. $P(A) \geq 0$, $\forall A \in \Sigma$
2. $P(S) = 1$
3. $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ for any collection of disjoint sets $A_1, A_2, \ldots \in \Sigma$

Axioms do not tell what specific $P$ we should use, instead, for any function $P$ they tell whether $P$ is a probability function.

The triple $(S, \Sigma, P)$ is called a probability space

Probabilities are a special case of measures. The differences is that the measure ($\mu$) of the entire space, $\mu(S)$, can be any nonnegative number.

Example: Head and Tails.
Proposition 2  If $P$ is a prob. function and $A$ is any set in $\Sigma$, then

1. $P(\emptyset) = 0$
2. $P(A) \leq 1$
3. $P(A^c) = 1 - P(A)$
Proposition 3  If \( P \) is a prob. function and \( A \) and \( B \) are any sets in \( \Sigma \), then

1. \( P (B \cap A^c) = P (B) - P (A \cap B) \)

2. \( P (A \cup B) = P (A) + P (B) - P (A \cap B) \)

3. If \( A \subset B \), then \( P (A) \leq P (B) \)

From 2, we obtain a lower boundary for the prob of simultaneous events as

\[ P (A \cap B) \geq P (A) + P (B) - 1 \]

which can be negative (correct and useless)!

508-B (Statistics Camp, Wash U, Summer 2015)
Proposition 4  If $P$ is a prob. function, then

1. $P(A) = \sum_{i=1}^{\infty} P(A \cap C_i)$ for any partition $C_1, C_2, \ldots$

2. $P(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i)$ for any sets $A_1, A_2, \ldots$
2.3 Some Counting

Counting can be quite a task and sometimes it is quite important: for example in 2012’s Micro prelim!!

It’s useful to define probabilities if $S$ is finite and all outcomes in $S$ are equally likely.

**Proposition 5** Fundamental Theorem of Counting. If a job consists of $k$ separate tasks, the $i$th of which can be done in $n_i$ ways, $i = 1, \ldots, k$, then the entire job can be done in $n_1 \times n_2 \times \ldots \times n_k$ ways.

**Counting with or without replacement**: Lottery numbers from an urn... can I throw the number back in the urn once I pulled it out?

**Counting ordered or unordered**: Lottery numbers from an urn... Do I care in which order I write the numbers on the board?
Number of possible arrangements of size $r$ from $n$ objects

<table>
<thead>
<tr>
<th></th>
<th>w/o replacement</th>
<th>w replacement</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ordered</strong></td>
<td>$\frac{n!}{(n-r)!}$</td>
<td>$n^r$</td>
</tr>
<tr>
<td><strong>Unordered</strong></td>
<td>$\binom{n}{r}$</td>
<td>$\binom{n+r-1}{r}$</td>
</tr>
</tbody>
</table>

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

508-B (Statistics Camp, Wash U, Summer 2015)
3 Conditional Probability and Independence

New information may allow us to update probabilities in the sense of computing them relative to an updated sample space. What’s the probability it’ll rain in a given day? What’s the probability it’ll rain in a given day conditional on day being humid?

Definition 8 Conditional Probability. If $A$ and $B$ are events in $S$, and $P(B) > 0$, then the conditional probability of $A$ given $B$, written $P(A|B)$, is

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

$B$ becomes sample space: $P(B|B) = 1$

If sets are disjoint $P(A|B) = P(B|A) = 0$

Example:
Definition of conditional prob. implies

\[ P(A \cap B) = P(A|B)P(B) = P(B|A)P(A) \]

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

\[ P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_k P(B|A_k)P(A_k)} \]

**Definition 9 Proof.** Follows from expressing intersection in terms of conditional probabilities and expressing \( B \) as a union of disjoint sets using the partition. ■

Example:
It may happen that the occurrence of an event bears no information regarding the occurrence of another.

**Definition 10  Statistical Independence.** Two events $A$ and $B$, are statistically independent if

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

Examples: tosses of die: $P$ (at least 1 six in 4 rolls) =

Does disjoint mean independent?
Proposition 6 If $A$ and $B$ are independent events, then the following pairs are also independent

1. $A$ and $B^c$
2. $A^c$ and $B$
3. $A^c$ and $B^c$

In general, a collection of events $A_1, \ldots A_n$ are mututally independent if for any subcollection $A_{i_1}, \ldots A_{i_k}$ we have

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

In most cases, we are not interested in working with the entire sample space.

Random variables enable us to work with summary variables without having to use the entire sample space.

Definition 11 A random variable is a measurable function $f : S \to \mathbb{R}$

Examples: CD4 count change after taking different amount of an antiretroviral. The sum of two dies tossed.

What is a measurable function?

Why do we need r.v. to be measurable?
A measurable function is a “well-behaved” function between measurable spaces. Ok, but what are measurable spaces? Remember that probability space? Just think of a general measure as opposed to a probability measure.

If $\Sigma$ is a $\sigma$-algebra over a set $W$, and $T$ is a $\sigma$-algebra over a set $Y$, then a function $f: W \to Y$ is measurable if $\forall A \in T \ f^{-1}(A) \in \Sigma$.

In this class, usually $Y =$ real line, $T$ is the Borel $\sigma$-algebra, and instead of $f()$, we use $X()$

Really? What’s the Borel $\sigma$-algebra? It is the smallest $\sigma$-algebra containing all open sets
\[ X^{-1} ((a, b)) \in \Sigma \]

\[ X^{-1} (A) = \{ \omega : \omega \in \Sigma, X (\omega) \in A \} \]

Why does \( X \) have to be measurable?

If \( X^{-1} ((a, b)) \notin \Sigma \) we cannot define \( P (a < X < b) \)!!

Preimage of our random variable must belong to the \( \sigma \)-algebra (\( \Sigma \)) defined over the original sample space \( S \).

508-B (Statistics Camp, Wash U, Summer 2015)
5 Distribution Functions

Random variables have associated functions called distributions. The distribution function attaches a probability to each Borel set, and it enables us to compute probabilities of all sets in the $\sigma$-algebra.

Definition 12 The cumulative distribution function or cdf of a r.v. $X$, denoted $F_X(x)$, is defined by

$$F_X(x) = P(X \leq x) \text{ for all } x$$

Example:

$F_X$ can be continuous or discontinuous and must be defined for all $x$.

By construction it cannot be negative.

It can have jumps but by definition $F_X$ takes value at the top of the jump: this is, $F_X$ is right-continuous.

What is so special about the jump points? Size of jump at any point $x$ is $P(X = x)$.
Proposition 7 **Properties of a cdf.** The function $F(x)$ is a cdf iff the following three conditions hold

1. $\lim_{x \to \infty} F(x) = 1$ and $\lim_{x \to -\infty} F(x) = 0$

2. $F(x)$ is nondecreasing (could be flat, just not decreasing)

3. $F(x)$ is right-continuos: for every $x_0$, $\lim_{x \downarrow x_0} F(x) = F(x_0)$

Example: geometric distribution (discrete), logistic (continuous)
Continuity or lack thereof in the cdf corresponds to the associated r.v.

A r.v. $X$ is continuous if $F_X(x)$ is a continuous function of $x$

A r.v. $X$ is discrete if $F_X(x)$ is a step function of $x$

Of course, a r.v. can be a mixture of both:

**Amount paid to director in a Non For Profit:** $P(\text{Officer\_Pay} = 0) > 0$ and $P(\text{Officer\_Pay} = x) = 0$ \(\forall x > 0\)
We say r.v.'s $X$ and $Y$ are identically distributed if, for every set $A$ in the Borel $\sigma$-algebra

$$P (X \in A) = P (Y \in A)$$

That’s not to say that $X = Y$!!!

Example: tossing a fair coin.

**Proposition 8** If $X$ and $Y$ are identically distributed, then

$$F_X (x) = F_Y (x) \text{ for every } x$$
Associated with a random variable $X$ and its cdf there is another function, called the probability density function or pdf (or the probability mass function or pmf), denoted by $f_X(x)$

**Definition 13 pdf/pmf**

- If $X$ is discrete, pmf is given by:
  \[ f_X(x) = P(X = x) \text{ for all } x \]

- If $X$ is continuous, pdf is given by:
  \[ F_X(x) = \int_{-\infty}^{x} f_X(w) \, dw \text{ for all } x \]

Notice that if $X$ is continuous, $P(X = x) = 0$

**Example:**

508-B (Statistics Camp, Wash U, Summer 2015)
Proposition 9  A function $f_X (x)$ is a pdf (or pmf) of a r.v. $X$ iff

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

**Uniform:** $U[a, b]$ with parameters $a < b \in \mathbb{R}$

\[
f(x) = \frac{1}{b-a} \quad F(x) = \frac{x-a}{b-a} \quad \forall x \in [a, b]
\]

where $f(x)$ is zero elsewhere and $F(x) = 0$ for all $x < a$ and $F(x) = 1$ for all $x < b$.

Used to: valuations in auctions.

**Exponential:** with parameter $\lambda > 0$

\[
f(x) = \lambda e^{-\lambda x} \quad F(x) = 1 - e^{-\lambda x} \quad \forall x > 0
\]

where $f(x)$ and $F(x)$ are zero elsewhere

Used to: length of time until the next innovation
**Normal**: $N(\mu, \sigma^2)$, parameters $\mu$ and $\sigma > 0$

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2} \right\} \quad \forall x \in R$$

$F(x)$ does not have a closed form solution

$N(0, 1)$ (Standard Normal) it satisfies symmetry around zero: (i) $f(x) = f(-x)$ and (ii) $F(x) = 1 - F(-x)$

Notation: denote $F(x)$ by $\Phi(x)$

**Standard Logistic**

$$f(x) = \frac{e^x}{(1+e^x)^2} \quad F(x) = \frac{e^x}{1+e^x} \quad \forall x \in R$$