**MAT2013: Probability and Statistics [1] [2]** **Spring 2017**

# Lecture 1: Introduction, Basics of Probability

*Lecturer: Jae Guk, Kim* *Scribes: Jisung Lim*

**Scribes**: Jisung Lim, *B.S. Candidate of Industrial Engineering in Yonsei University, South Korea.*



## 1.1 Why do we have to study probabilty?

***To model the random phenomena in the mathematical framework with consistency.***

## 1.2 How to model the random experiment?

### 1.2.1 Identifying Sample Space

**Definition 1.2.1. Random Experiment**
A **random experiment** is one whose outcome is determined by chance. The result is unknown, until we observe the outcome.

(ex) Tossing a coin, rolling a dice.

**Definition 1.2.2. Sample Space**
For a random experiment, the set of all possible outcomes of it is called **the sample space** $S$ of **the random experiment**.

ex. Coin-tossing
Sample space: $S = \{\text{HEAD}, \text{TAIL}\}$

**Definition 1.2.3. Event**
An event $A$ is a subset of the sample space $S$.

ex. Rolling a dice
Sample space: $S = \{1, 2, 3, 4, 5, 6\}$, Event: $A_0 = \emptyset, \ldots, \{1, 3\}$

**Definition 1.2.4. Mutually Exclusive**
We say that a bunch of events $A_1, A_2, \ldots$ are **mutually exclusive** if

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

ex. Coin tossing
Sample space: $A_1 = \{\text{HEAD}\}, A_2 = \{\text{TAIL}\}$

### 1.2.2 Assigning a probability to each elements in sample space

#### 1.2.2.1 What do we mean when we say "the probability of Heads," $P$(Head)?

**Answer 1. Relative Frequency Approach (Bernoulli trial)**

- This approach states that the way to determine $P(\text{Head})$ is to flip the coin repeatedly in exactly the same way each time.
  Then, a good approximation to $P(\text{Head})$ will be

$$P(\text{Head}) \simeq \frac{(\text{number of observed heads})}{(\text{total \# of flips})}$$

- A justification for this approach comes from **the law of large numbers**.

**Theorem 1.2.5. The law of large number (LLN).**

$$\frac{S_n}{n} \to P(A) \quad as \quad n \to \infty \tag{1.1}$$

where

$$\begin{aligned} A &: \text{event} \\ n &: \text{the \# of conducted experiments} \\ S_n &: \text{the \# of times that A occurred in the n experiments} \end{aligned} \tag{1.2}$$

**So, is there any problem in using Relative Frequency Approach to calculate the probability?**

- It is just a long-run approximation: It requires the large number of experiments!
- Is it possible to compute the probability of you getting A in this class?

**Answer 2. The Subjective Approach (Bayesian)**

- This approach interprets probability as the experimenter's **degree of belief** that an event of interest will occur.
- Ex) What is the probability that she/he likes me?
  To answer this questions, we refer to the use of intuition, personal belief, and other direct/indirect informations. this is the subjective definition of probability.

**# Brief insight**
In most cases of random phenomena of science and engineering, the relative frequency interpretation of probability is the operative one. (More info, see appendix.)

**Anyway**
In consequence, modeling a random phenomena is a process of identifying a sample space $S$ and assigning probabilities to all the events $A_1, A_2, \ldots$ of the sample space $S$.

#### 1.2.2.2 Probability Model

The mathematical description of a random phenomenon consisting of two parts:

- a sample space S
- a way of assigning probabilities.

It will be tricky as you expect but there is very obvious case in giving probabilities.

We used coin-tossing example to explain the relative frequency interpretation of probability. However, it is obvious to select a model for it,

$$P(\text{Head}) = \frac{1}{2}, P(\text{Tail}) = \frac{1}{2} \tag{1.3}$$

provided that the coin is well-balanced.

In general, it is good to assign a probability to an event in Equally Likely Model through

$$P(A_i) = \frac{1}{N} \quad \forall i \in \{1, \ldots, N\} \tag{1.4}$$

## 1.3 The viewpoint: Probability Measure (function)

### 1.3.1 Proability Function and its Properties

Now, we consider probability as a function from a collection of events $\mathcal{P}(S)$ to the real-valued interval $\mathbb{R} \in [0, 1]$, which is called probability measure. A probability function is a rule that associates with each event $A$ of the sample space $S$. A unique quantity $P(A) = a$, called the probability of $A$.

**Definition 1.3.1.** Any probability function $P$ satisfies the following Kolmogorov Axioms:

1. $P(A) \geq 0$ for any event $A \subseteq S$
2. $P(S) = 1$
3. If the countable events $A_1, A_2, A_3, \ldots$ are mutually exclusive, then

$$P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

*Properties* 1.3.1.1. Let $P$ be the probability function which satisfies the Kolmogorv Axioms:

1. $P(\emptyset) = 0$

   *Proof.* Hint: Let assume $A_i = \emptyset \quad \forall i \in \mathbb{N}$ □

2. For every finite $A_1, A_2, A_3, \ldots, A_m$ of mutually exclusive events $A_i \subseteq S$, we have

$$P\left(\cup_{i=1}^{m} A_i\right) = \sum_{i=1}^{m} P(A_i) \tag{1.5}$$

*Proof.* Hint: $B_i = A_i \forall i = \{1, \dots, m\}$ and $B_j = \emptyset \forall i = \{m+1, m+2, \dots\}$ $\square$

3. $P(A^C) = 1 - P(A)$

   *Proof.* Hint: $1 = P(S) = P(A \cup A^C) = P(A) + P(A^C)$ $\square$

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

   *Proof.* Hint: $B = A \cup (B \cap A^C)$ $\square$

5. $\forall A \subseteq S$, $0 \leq P(A) \leq 1$

   *Proof.* (a) $0 \leq P(A)$
   Hint: Axiom 1.

   (b) $P(A) \leq 1$
   Hint: Axiom 2: $P(S) = 1$, $A \subset S$, and property 4.

   By (a) and (b), the proposition is true. $\square$

6. (Additive rule) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
   **Sketch:** Sketch diagram to grasp intution for solving the problem.
   **Proof:** Hint: $(A \cap B^C) \cup (A \cap B) \cup (A^C \cap B)$ $\blacksquare$

7. (Total probability) Let $B_1, B_2, \dots, B_n$ be mutually exclusive and exhaustive, then $P(A) = P(A \cap B_1) + P(A \cap B_2) + \cdots + P(A \cap B_n)$.
   **Proof:**
   (Mutually exclusive): We say that a bunch of events $B_1, B_2, \dots$ are mutually exclusive if

$$B_i \cap B_j = \emptyset \quad \text{for} \quad i \neq j$$

   (Collectively exhaustive): We say that a bunch of events $B_1, B_2, \dots$ are collectively exhaustive if

$$\cup_{i=1}^{\infty} B_i = S$$

   Hint: $P(A) = P(A \cap S) = P(A \cap (\cup_{i=1}^{\infty} B_i)) = P(\cup_{i=1}^{\infty} (A \cap B_i)) = \sum_{i=1}^{\infty} P(A \cap B_i)$ $\blacksquare$

### 1.3.2 Conditional Probability

Now, we want to deal with a situation like, what is the probability of $A$ occurring given knowledge that $B$ has occured?

*Example* 1.1. Consider a full deck of 52 cards. (1, 2, 3, 4, ..., 13). Select two cards successively.

$$A = \{\text{first card drawn is an 1}\}$$

$$B = \{\text{second card drawn is an 1}\}$$

**Solution:** By definition, our sample space is given by,

$$S = \{(1A, 1B), (1A, 1C), (1A, 1D), \dots, (1A, 13D), \dots, (13B, 13D), (13C, 13D)\}$$

which has the number of cases $n(S) = 52 \times 51$. Before we evaluate some probabilities, let's introduce a basic counting principle called additive rule and product rule.

*Definition* 1.3.2. **The rule of sum**
If a first task can be performed in $m$ ways, while a second task can be performed in $n$ ways, and the **two tasks cannot be performed simultaneously**, then performing either task can be accomplished in any one of $m + n$ ways.

*Definition* 1.3.3. **The rule of product**
If a procedure can be broken down into first and second stages, and if there are $m$ possible outcomes for the first stage and if, for each of these outcomes, there are $n$ possible outcomes for the second stage, then the total procedure can be carried out, in the designated order, in $m \cdot n$ ways.

Then we can calculate the probability of $A$ and $B$, respectively, which is given by

$$P(A) = \frac{4}{52}\frac{51}{51} = \frac{4}{52}$$

and

$$P(B) = \frac{4}{52}\frac{3}{51} + \frac{48}{52}\frac{4}{51}.$$

Also we can find joint probability of $A$ and $B$, which is given by

$$P(A \cap B) = \frac{4 \times 3}{52 \times 51}.$$

Now, let us consider the conditional probability, the probability of $A$ occrring given knowledge that $B$ has occured, which is given by

$$P(B|A) = \frac{3}{51}.$$

■

In the latest example, our evalution process of the probabilities of various events seems to strongly depend on the knoweledge of the specific situation. Now, we want to generalize this evaluation process into systematic framework to solve similar problems without exploiting any situation-based knoweledge.

**Sketch:** First, we may consider roughly (or intuitively) the reduced sample space from $S$ to $B$.

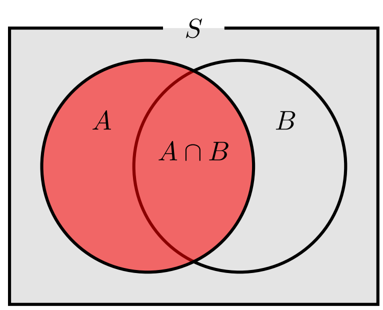


Figure 1.1: conditional distribution

Then, we can evaluate the conditional distribution

$$P[B|A] = \frac{12/(52 \times 51)}{4/52} = \frac{3}{51}.$$

Now, we will define the conditional probablity more precisely, and rigorously as follows:

**Definition 1.3.4.** Let $A$, $B$ be events, $P(A) > 0$. The conditional probability of $B$ given $A$ is

$$P[B|A] = \frac{P(A \cap B)}{P(A)} \tag{1.6}$$

**Theorem 1.3.5.** For any fixed event $A$ with $P(A) > 0$

1. $P[B|A] \geq 0 \quad \forall B \subset S$
2. $P[S|A] = 1$
3. If $B_1, B_2, B_3, \ldots$ are disjoint events, then

$$P(\bigcup_{i=1}^{\infty} B_i|A) = \sum_{i=1}^{\infty} P(B_i|A)$$

That is, the conditional probability $P(\cdot|A)$ is also a probability function.

*Properties* 1.3.5.1. Then the following are immediate, for any events $A, B, C$ with $P(A) > 0$,

1. $P(B^C|A) = 1 - P(B|A)$
2. if $B \subset C$, then $P(B|A) \leq P(C|A)$
3. $P((B \cup C)|A) = P(B|A) + P(C|A) - P((B \cap C)|A)$
4. (The mulitplication Rule) For any two events $A$ and $B$,

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

   And more generally, for events $A_1, A_2, \ldots, A_n$

$$P(A_1 \cap A_2 \cap A_3 \cap \cdots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots P(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1})$$

*Example* 1.2. **Card selection**

Let's calculate the probability of two sequential outcomes being both 1. First, we can count every cases using our knoweledge about the situation, which is given by

$$P(\text{both } 1) = P(A \cap B) = \frac{4 \times 3}{52 \times 51}$$

But, we want to evaluate the probability by exploiting mathematical framework, not using our situation-based knoweledge and intuition.

$$P(B \cap A) = P(A)P(B|A) = \frac{4}{52} \cdot \frac{12/(52 \times 51)}{4/52}$$

*Example* 1.3. **Red-Green Ball**

There is an urn with 10 balls inside: 7 red balls and 3 green balls. We will draw three balls sequentially from the urn, and our event of interest is the case where all three balls are red. Then, we can divide the event into three sub event as follows:

- $A_1 = \{\text{1st ball is red}\}$
- $A_2 = \{\text{2nd ball is red}\}$
- $A_3 = \{\text{3rd ball is red}\}$

And we can evaluate the probability of our event of interest systematically:

$$P(\text{All 3 balls are red}) = P(A \cap B \cap C) = P(A)P(B|A)P(C|B \cap A)$$

### 1.3.3 Independence

Given events $A$ and $B$, the situation where knowledge that $B$ has occurred in no way changes the probability that $A$ will occur. A proper way to represent it would be

$$P(A|B) = P(A) \quad \text{where} \quad P(B) > 0 \tag{1.7}$$

More specifically, consider you toss a coin twice, then sample space $S$ is given by

$$S = \{HH, HT, TH, TT\}$$

and we assume the model has probability with equally likely event,

$$P(HH) = P(HT) = P(TH) = P(TT) = 1/4$$

Now, lets evaluate the following probabilities

$$P(\text{2nd toss is H}|\text{1st toss is H}) = \frac{P(\text{both H})}{P(\text{1st toss is H})} = \frac{1/4}{1/2} = \frac{2}{4} = \frac{1}{2}$$

$$P(\text{2nd toss is H}) = \frac{P(\text{2nd toss is H})}{P(S)} = \frac{2}{4} = \frac{1}{2}$$

$$P(\text{2nd toss is H}|\text{1st toss is H}) = P(\text{2nd toss is H})$$

This means that the information that "the first toss is H" has nothing to do with the probability that "the second toss is H". That is, the coin deos not remenber the result of the first toss.

*Question in Lecture* 1.1. Why coin tossing example can be used in numerous examples?
*ANS.* Symmetric property of coin tossing.

**Definition 1.3.6.** Events A and B are said to be independent if

$$P(A \cap B) = P(A)P(B) \tag{1.8}$$

Otherwise, the events are said to be dependent.

**Theorem 1.3.7.** If the events $A$ and $B$ are independent, then

1. $A$ and $B^C$ are independent
   **Sketch:** w.t.s: $P(A^C \cap B) = P(A^C)P(B) \quad \cdots \quad (1)$

   **Proof:**
   Let $P(B) > 0$, by multiplication rule, (1) becomes

   $$P(A^C \cap B) = P(B)P(A^C|B) \quad \cdots \quad (2)$$

   Since the conditional probability $P(\cdot|B)$ is also the probability function, by property of complement set,

   $$P(B)P(A^C|B) = P(B)(1 - P(A|B)) \quad \cdots \quad (3)$$

   Since $A$ and $B$ are independent, then

   $$P(B)(1 - P(A))$$

   By property of probability function

   $$P(B)(A^C)$$

   ■

2. $A^C$ and $B$ are independent
3. $A^C$ and $B^C$ are independent

What if there are more than two events of interest?

**Definition 1.3.8.** Any collection of events $\{A_i\}_{i\in I}$ is independent if for every finite subset $J \subset I$, one has

$$P(\cap_{i\in J} A_i) = \prod_{i\in J} P(A_i) \tag{1.9}$$

It is often called to be mutually independent.

**Note.** if events are $\{A_i\}_{i\in I}$ is independent, they pairwise independent but the converse is false. ($\{A_i\}_{i\in I}$ are pairwise independent if $A_i$ and $A_j$ are independent for all $i, j$ with $i \neq j$)

*Example* 1.4. Let the sample space $S = \{1, 2, 3, 4\}$ and each single event is equally likely probable $P(i) = 1/4 \quad \forall i \in \{1, 2, 3, 4\}$. Then we can construct a collection of events $\mathbb{C} = \{A, B, C\}$, in which each event is given by:

$$A = \{1, 2\}, B = \{1, 3\}, C = \{2, 3\}$$

Let us evaluate the probability of pair of each event, which is given by

- $P(A \cap B) = P(\{1\}) = 1/4 = (1/2)(1/2) = P(A)P(B)$
- $P(B \cap C) = P(\{3\}) = 1/4 = (1/2)(1/2) = P(B)P(C)$
- $P(A \cap C) = P(\{2\}) = 1/4 = (1/2)(1/2) = P(A)P(C)$

which implies pairwise independent.

But, if we evaluate the probability of $A \cap B \cap C$, then

$$P(A \cap B \cap C) = P(\emptyset) = 0.$$

and since $P(A)P(B)P(C) = 1/8$, then

$$P(A \cap B \cap C) \neq P(A)P(B)P(C),$$

which implies the collection of events $\{A, B, C\}$ is not mutually independent.

#### 1.3.3.1 Independence vs. exclusive, pairwisely vs. mutually

So far, we've discussed about difference between *pairwisely* and *mutually*, in the context of **independence**. As we've seen before, the notion of *pairwisely* and *mutually* was introduced when we had discussed about **exclusive**. You may be confused between two different concepts. Let's take an example of exclusive to separate two different concept—independence and exclusive, in the perspective of *pairwisely* and *mutually*.

First, for our example, let us define the *partition*, which satisfies two properties, mutually exclusive and collectively exhaustive.

**Definition 1.3.9.** A collection of events $\{E_n\}_{n\in\mathbb{N}}$ is called a partition of $S$ if they are pairwise disjoint,

$$P(E_n) > 0, \cup_{n\in\mathbb{N}}(E_n) = S$$

*Example* 1.5. Does partitions have mutually exclusive property? Prove your claim.

**Proof:** *Hint.* A collection of sets are pairwise disjoint implies the collection is mutually exclusive. ∎

**Theorem 1.3.10. Bayes' Theorem**
Let $E_n$ be a finite or countable partition of $S$, and suppose $P(A) > 0$, then

$$P(E_n|A) = \frac{P(A|E_n)P(E_n)}{\sum_m P(A|E_m)P(E_m)} \tag{1.10}$$

**Sketch:**

$$P(E_n|A) = \frac{P(E_n \cap A)}{P(A)} \frac{P(E_n)}{P(E_n)} = \frac{P(A|E_n)P(E_n)}{P(A)} = \frac{P(A|E_n)P(E_n)}{\sum_m P(A \cap E_m)}$$

**Proof:** *D.I.Y.* ∎

*Example* 1.6. p.103 (This solution may be *WRONG* because the scriber refers to additional material. There are some concepts and contents that are not mentioned in the lecture, so I've included references to that contents.)

**Sketch: Modeling the random phenomena**
**1. Specify our random experiment.**
Select one of the documents from the cabinet. We may find two kinds of information in the experiment.

- "Who filed this document?"
- "Is document misplaced?"

The possible outcomes for first question is one of Moe, Larry, and Curly and the possible outcomes for second question is that the document is misfiled or not. Like this case, if we have more than one kind of 'outcomes of interest,' we may construct multiple sample space for each kind of outcomes [3]. Hence, we can construct two sample space for each question, which is given by

$$S_1 = \{m, l, c\},\ S_2 = \{a, \bar{a}\}$$

where $m$, $l$ and $c$ stands for Moe, Larry, and Curly for the first question, respectively, and $a$ implies that the document is misplaced, $\bar{a}$ implies not.

**2. Construct our sample space**
From here, we can define our sample space as the Cartesian product of the two sample spaces noted above:

$$S = S_1 \times S_2 = \{(m, a), (l, a), (c, a), (m, \bar{a}), (l, \bar{a}), (c, \bar{a})\}$$

**3. Determine our events of interest**
Our events of interest can be determined based on our previous two questions and their associated answers. Hence our events are determined as follows

- $M$ = Moe filed the document = $\{(m, a), (m, \bar{a})\}$
- $L$ = Larry filed the document = $\{(l, a), (l, \bar{a})\}$
- $C$ = Curly filed the document = $\{(c, a), (c, \bar{a})\}$

and

- $A$ = The file is miss classified = $\{(m, a), (l, a), (c, a)\}$

From this setting, we can sketch the diagram as follows:

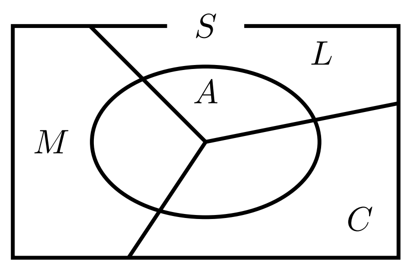


Figure 1.2: sample space diagram

**4. Assign probabilities to each event**
First, picking up any file from the cabinet is assumed to occur equally likely. From this assumption, we can evaluate the probability of each event based on counting. (e.g. $P(M) = n(M)/n(S)$.) Because of the assumption of same labor efficiency, we can say that each probability is proportional to the workload on each assistants. From the given information about the workload, we can assign probabilities to each event, given by

$$P(M) = \frac{6}{10}, P(L) = \frac{3}{10}, P(C) = \frac{1}{10}$$

And also we have another infomration called 'Misfile Rate', the rate of misfiling of each assistant when the documents was filed by that assistant. By the definition, Misfile Rate can be viewed as a conditional probability as follows

$$P(A|M) = 0.003, P(A|L) = 0.007, P(A|L) = 0.010$$

which implies the probability of the file is miss classified, when the knowledge that the file is classified by Moe, Larry, or Curly, respectively, is given.

**5. Properties of events**
Since any two person cannot classify same file simultaneously, each event $M$, $L$ and $C$ is **mutually exclusive**. Since no one classifies files except for those three assistants: Moe, Larry, and Curly, then each event $M$, $L$, and $C$ is **collectively exhaustive**. Hence, by definition, a collection of those three events is a partition of sample space $S$.

**Solution:** Use the Bayes' rule. Find $P(M|A), P(L|A), P(C|A)$. *D.I.Y.* ∎

### 1.3.4 Counting Methods

In the equally likely model,

$$P(A) = \frac{(\#\text{ of A})}{(\#\text{ of S})}$$

Thus, it is important to count outcomes in the event of interest.

**Definition 1.3.11. Multiplication Principle**
If an experiment is composed of $k$ successive steps which may be performed in $n_1, n_2, \ldots, n_k$ distinct ways, respectively, then the experiment may be performed in $n_1 \cdot n_2 \cdots n_k$ distinct ways.

**Theorem 1.3.12.** Every counting problem can be substituted with $k$ balls $n$ bins problem.

- **Ordered sample**
  The number of ways in which one may select an ordered sample of $k$ subjects from a population that has $n$ distinguishable members is
  1. $n(n-1)(n-2)\cdots(n-k+1)$ if sampling is done without replacement.
  2. $n^k$ if sampling is done with replacement.
- **Unordered sample**
  The number of ways in which one may select an unordered sample of $k$ subjects from a population that has $n$ distinguishable members is
  1. $\binom{n}{k}$ if sampling is done without replacement.
  2. $\binom{n+k-1}{n-1} = \binom{n+k-1}{k}$ if sampling is done without replacement.

*Example* 1.7. Take a coin and flip it 7 times. How many sequences of Heads and Tails are possible?
*ANS:* 7 balls to 2 bins.

*Example* 1.8. We rent five videos, we want to see three movies in specific order.
*ANS:* 3 balls to 5 0–1 bins.

## 1.4 Appendix

### 1.4.1 Diffrence between Relative Frequency Approach and Subjective Approach

- Relative Frequency Approach (Frequentist)
  1. "Probabilities represent long run frequencies of events."
  2. Computationally inexpensive relatively.
  3. Do not use subjective information.
  4. But there exist some limitations in performance.
- The Subjective Approach (Bayesian)
  1. "Probability is used to quatify our uncertainty about something or precisely degree of belief."
  2. Computationally expensive, since we need to compute all the distribution.
  3. Use subjective information.
  4. But the performance is better especially in high dimension.

### 1.4.2 Unordered Sample with Replacement

Consider "unordered sample with replacement" case and let's develope the logic. Let's consider it with intuition. Assume $k = 2$, then there are two possible cases,

- If two balls are in distinct bins, then $\frac{n(n-1)}{2!}$.
- If two balls are in the same bin, then $n$.

So,

$$\frac{n(n-1)}{2} + n = \frac{n(n+1)}{2}$$

What if $k = 3$?

- If three balls are in three distinct bins, then ....
- If three balls are in two distinct bins, then ....
- If three balls are in the same bin, then ....

What if $k = k_0$?

- If $k_0$ balls are in $k_0$ distinct bins, then ....
- If $k_0$ balls are in $k_0 - 1$ distinct bins, then ....
- If $k_0$ balls are in $k_0 - 2$ distinct bins, then ....
  ...
- If $k_0$ balls are in two distnct bins, then ....
- If $k_0$ balls are in the same bin, then ....

What if $k$ is too large to consider every each case? it seems hopelessly complicated. So let's do it again with a smart idea. Represent each of the balls by a 0's and the separations between bins by 1's; that is we have k 0's and (n-1) 1's.

Then, each possibility for placing the indistinguishable balls in the boxes can be thought of as a vector of length $n + k - 1$ which is madeup of $n - 1$ entities which are 1's (separations) and k entities which are 0's (balls). This case is represented by unordered without replacement case.

$$\binom{n+k-1}{n-1} = \binom{n+k-1}{k}$$

For example, the sample $k = 3$ and $n = 5$ which is represented by the vector $(0, 0, 2, 0, 1)$ can be thought of as the vector $(w, w, b, b, w, w, b)$

## 1.4.3 The subtlety of probability measure $\mathbb{P}$ and its domain, event space $\mathcal{A}$

### 1.4.3.1 What should a collection of events $\mathcal{A}$ look like?

1. **Motivation.**

   Let $\Omega = [0,1]$ be sample space, then from geometrical intuition we may construct probability function as follows:
   $$P\left(\,(a,b]\,\right) = \frac{|b-a|}{|1-0|} = b-a \quad \text{where} \quad 0 \le a \le b \le 1$$
   Suppose we want this $\mathbb{P}$ to extend to a probability function $\mathbb{P}$ on $2^{[0,1]}$ (power set of $[0,1]$) satisfying axioms. At some cases, however, we CANNOT satisfy those axioms, exactly the third one when the sample space $\Omega$ has uncountably many elements.

2. **Then what we have to do? From "Any subset" to $\sigma$-*algebra*.**

   Because our function $\mathbb{P}$ is intuitively obvious, we might be inclined to take action on our event sapce $\mathcal{A}$. As we've seen so far, when the sample space $\Omega$ has uncountably many elements, the idea of defining the probability of a power set of $\Omega$ in terms of the probabilities of elementary outcomes is unacceptable. Hence, we want to model the event space $\mathcal{A}$ as $\sigma$-*algebra* of sample space $\Omega$ which is a collection of subsets of $\Omega$ with the following definition:

   **Definition 1.4.1.** Given a sample space $\Omega$, a $\sigma$-*algebra* is a collection $A$ of subsets of $\Omega$, with the following properties:

   - (Closed under complement) $A \in \mathcal{A} \Rightarrow E^C \in \mathcal{A}$
   - $\emptyset \in \mathcal{A}$ $(\Omega \in \mathcal{A})$
   - (Closed under countable union) If $E_1, E_2, E_3, \ldots \in \mathcal{A}$, then $\bigcup_{i=1}^{\infty} E_i \in \mathcal{A}$.

   A set $E$ that belongs to $\mathcal{A}$ is called an event, an $\mathcal{A}$-measurable set, or simply a measurable set. The pair $(\Omega, \mathcal{A})$ is called a measurable space.

3. **Construct probability measure**

   And then, we can say that the probability measure $\mathbb{P}$ is given by
   $$\mathbb{P} : \mathcal{A} \to [0, \infty]$$
   which satisfies following two conditions,

   (a) $\mathbb{P}(E^C) = 1 - \mathbb{P}(E)$
   (b) $\mathbb{P}(\Omega) = 1$
   (c) (Countable additivity) If $E_i$ is a sequence of disjoint sets that belong to $\mathcal{A}$, then
   $$\mathbb{P}(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} \mathbb{P}(E_i)$$

   Since we are free from misunderstanding the word "any subset" in $\Omega$, we can accept $\mathbb{P}$ is defined on "any subset" of $\Omega$ through this course.

# References


[1] Youngstown State University G. Jay Kerns. *Introduction to Probability and Statistics Using R*. `http://ipsur.org/index.html`. G. Jay Kerns, 2010.

[2] Babatunde A. Ogunnaike. *Random Phenomena: Fundamentals of Probability and Statistics for Engineers*. CRC Press, 2009.

[3] Wikipedia. *Sample space*. [Online; accessed 21-March-2017]. 2017. URL: `https://en.wikipedia.org/wiki/Sample_space#Multiple_sample_spaces`.