Conditional Expectation

Contents

1 Introduction **2**

2 Measures and Random Variables **2**
2.1 Cumulative Distribution function 3
2.2 Radon- Nikodym Theorem and Probability density functions 3

3 Differentiation **5**

4 Fundamental Theorem of Calculus **5**

5 Radon- Nikodym Theorem **5**

6 Conditional Expectation **6**

7 Conditional Expectation Notes-1 **7**
7.1 General Case 10

8 Conditional Expectations Notes 2 (Jun Shao) **10**

1 Introduction

The existence of conditional probability and conditional expectation comes from the Radon-Nikodym theorem. Radon-Nikodym theorem itself is a generalization of the fundamental theorem of calculus for general measures. Therefore we first discuss differentiation and fundamental theorem of calculus as a precursor to Radon-Nikodym theorem. We first review the basics of measure, integration and differentiation.

2 Measures and Random Variables

Definition (σ-algebra): Let $\mathcal{F}$ be a collection of subsets of a sample space Ω. $\mathcal{F}$ is called a σ-field or σ-algebra if and only if it has the following properties:

1. The empty set $\emptyset \in \mathcal{F}$.
2. If $A \in \mathcal{F}$, then the complement $A^c \in \mathcal{F}$.
3. If $A_i \in \mathcal{F}, i = 1, 2, \cdots$, and their union $\cup A_i \in \mathcal{F}$.

A pair $(\Omega, \mathcal{F})$ consisting of a set Ω and a σ-field $\mathcal{F}$ of subsets of Ω is called a **measurable space**. The elements of $\mathcal{F}$ are called measurable sets in measure theory or events in probability and statistics.

Definition (Measure): Let $(\Omega, \mathcal{F})$ be a measurable space. A set function ν defined on $\mathcal{F}$ is called a **measure** if and only if it has the following properties:

1. $0 \leq \nu(A) \leq \infty$ for any $A \in \mathcal{F}$
2. $\nu(\emptyset) = 0$.
3. If $A_i \in \mathcal{F}, i = 1, 2, \cdots$, and A_i's are disjoint, then,

$$\nu\Big(\cup_{i=1}^{\infty} A_i \Big) = \Sigma_{i=1}^{\infty} \nu(A_i) \tag{1}$$

The triple $(\Omega, \mathcal{F}, \nu)$ is called a **measure space**. If $\nu(\Omega) = 1$, then ν is called a probability measure and we usually deonte it by P instead of ν, in which case $(\Omega, \mathcal{F}, P)$ is called a **probability space.**

Although measure is an extension of length, area, or volume, sometimes it can be quite abstract. For example, the following set function is a measure:

$$\nu(A) = \begin{cases} \infty & A \in \mathcal{F}, A \neq 0 \\ 0 & A = \emptyset \end{cases} \tag{2}$$

The following examples provided two very important measures in probability and statistics.

Example (Counting Measure): Let Ω be a sample space, $\mathcal{F}$ the collection of all subsets and $\nu(A)$ the number of elements in $A \in \mathcal{F}$ ($\nu(A) = \infty$ if A has infinitely many elements).

Then ν is a measure on $\mathcal{F}$ and is called the **counting measure**.
Example (Lebesgue Measure): There is a unique measure m on $(\mathbb{R}, \mathcal{B})$ that satisfies

$$m([a,b]) = b - a \tag{3}$$

for every finite interval $[a,b], -\infty < a \leq b < \infty$. This is called the **Lebesgue measure**.
If we restrict m to be the measurable space $([0,1], \mathcal{B}_{[0,1]})$, then m is a probability measure.

If Ω is countable in the sense that there is one-to-one correspondence between Ω and the set of all integers, then one can usually consider the trivial σ-field that contains all subsets of Ω and a measure that assigns a value to every subset of Ω. When Ω is uncountable it is not possible to define a reasonable measure for every subset of $\mathcal{R}$. This is why it is necessary to introduce σ-fields that are smaller than the power set.

2.1 Cumulative Distribution function

There is a one-to-one correspondence between the set of all probability measures on $(\mathbb{R}, \mathcal{B})$ and a set of functions on $\mathbb{R}$. Let P be a probability measure. The **cumulative distribution function**(c.d.f) of P is defined to be

$$F(x) = P((-\infty, x]), \qquad x \in \mathbb{R} \tag{4}$$

Proposition: Let F be a c.d.f on $\mathcal{R}$. Then

1. $F(-\infty) = \lim_{x \to -\infty} F(x) = 0$
2. $F(\infty) = \lim_{x \to -\infty} F(x) = 0$
3. F is nondecreasing, that is, $F(x) \leq F(y)$ if $x \leq y$.
4. F is right continuous, that is, $\lim_{y \to x, y > x} F(y) = F(x)$

Suppose that a real-vlaued function F on $\mathbb{R}$ satisfies $(1) - (4)$ above, then F is the c.d.f. of unique probability measure on $(\mathbb{R}, \mathcal{B})$.

Note that the starting point is the availability of probability measure. Once a measure is available, one can define c.d.f and also p.d.f (Radon Nikodym derivative) with respect to either counting measure or Lebesgue measure depending on with respect to which measure the P is absolutely continuous. This will be important in the case of conditional probability.

2.2 Radon- Nikodym Theorem and Probability density functions

Definition (σ-finite): A measure ν on $(\Omega, \mathcal{F})$ is said to be σ-finite if and only if there exists a sequence $\{A_1, A_2, \cdots\}$ such that $\cup A_i = \Omega$ and $\nu(A_i) < \infty$ for all i.

- Any finite measure , such as probability measure, is clearly σ-finite.
- The Lebesgue measure is σ-finite, since $\mathbb{R} = \cup A_n$ with $A_n = (-n, n), n = 1, 2, \cdots$

- The counting measure is σ-fintie if and only if Ω is countable.

Let $(\Omega, \mathcal{F}, \nu)$ be a measure space and f be a non-negative Borel function. One can show that the set function

$$\lambda(A) = \int_A f d\nu, \quad A \in \mathcal{F} \tag{5}$$

is a measure on $(\Omega, \mathcal{F}$. Note that

$$\nu(A) = 0 \Rightarrow \lambda(A) = 0 \tag{6}$$

If (6) holds for two measures λ and ν defined on the same measurable space, then we say that λ is **absolutely continuous** w.r.t ν and write $\lambda << \nu$.

Theorem (Radon-Nikodym Theorem): Let ν and λ be two measures on $(\Omega, \mathcal{F})$ and ν be σ-finite. If $\lambda << \nu$, then there exists a nonnegative Borel function f on Ω such that (5) holds. Furthermore, f is unique a.e. ν, that is, if $\lambda(A) = \int_A g d\nu$ for any $A \in \mathcal{F}$, then $f = g$ a.e. ν.

If (5) holds, then the function f is called the **Radon-Nikodym derivative** or density of λ w.r.t ν and is denoted by $d\lambda/d\nu$.

- If $\int_\Omega f d\nu = 1$ for an $f \geq 0$ a.e ν, then λ given by (5) is a probability measure and f is called its **probability density function** (p.d.f.) w.r.t. ν.

- For any probability measure P on $(\mathbb{R}^k, \mathcal{B}^k)$ corresponding to a c.d.f. F or a random vector X, if P has a p.d.f. f w.r.t. a measure ν, then f is also called the p.d.f. of F or X w.r.t. ν.

Example: Let Ω be a countable set, $\mathcal{F}$ be all subsets of Ω and ν be the counting measure. For any Borel function f, it can be shown that

$$\int f d\nu = \sum_{\omega \in \Omega} f(\omega) \tag{7}$$

Example (Discrete c.d.f. and p.d.f.): Let $a_1 < a_2 < \cdots$ be a sequence of real numbers and let $p_n, n = 1, 2, \cdot$, be a sequence of positive numbers such that $\Sigma_{n=1}^{\infty} p_n = 1$. Define

$$F(x) = \begin{cases} \Sigma_{i=1}^{n} p_n & a_n \leq x < a_{n+1}, n = 1, 2, \cdots \\ 0 & -\infty < x < a_1 \end{cases} \tag{8}$$

Then F is a stepwise c.d.f. It has a jump of size p_n at each a_n and is flat between a_n and $a_{n+1}, n = 1, 2, \cdots$. Such a c.d.f. is called a **discrete c.d.f.** and the corresponding random variable is called a discrete random variable. Let $\Omega = \{a_1, a_2, \cdots\}$, $\mathcal{F}$ be the collection of all subsets of Ω,

$$P(A) = \sum_{i: a_i \in A} p_i, \quad A \in \mathcal{F} \tag{9}$$

We can show that P is a probability measure and F is the c.d.f. of P.
Let ν be the counting measure on the power set of Ω. From the previous example we have,

$$P(A) = \int_A f d\nu = \sum_{a_i \in A} f(a_i), \quad A \subset \Omega \tag{10}$$

where $f(a_i) = p_i, i = 1, 2, \cdots$. That is f is the p.d.f. of P or F w.r.t. ν. Hence, any discrete c.d.f. has a p.d.f. w.r.t. counting measure. A p.d.f. w.r.t counting measure is called a discrete p.d.f.
Example (Discrete c.d.f. and p.d.f.):
Example (Mixed c.d.f and p.d.f):

After observing the above 3 examples, we would like to ask the following question:

Given a probability measure P on $(\mathbb{R}, \mathcal{B})$, we will be able to define the c.d.f F associated with it as :

$$F(x) = P((-\infty, x]), \qquad x \in \mathbb{R} \tag{11}$$

Does there always exists some measure ν such that we will be able to define p.d.f. of P w.r.t. ν?

3 Differentiation

We first prove the Lebesgue theorem that every monotonic increasing function is differentiable almost everywhere. We prove this theorem by using the Rising sun Lemma, which is discussed first.
Lemma:(Rising Sun Lemma)

4 Fundamental Theorem of Calculus

5 Radon- Nikodym Theorem

Let $(X, \mathcal{M})$ be a measurable space. For μ a measure on $(X, \mathcal{M})$ and f a nonnegative function on X that is measurable with respect to $\mathcal{M}$, define the set function ν on $\mathcal{M}$ by

$$\nu(E) = \int_E f d\mu \quad \text{for all } E \in \mathcal{M} \tag{12}$$

Exercise: Show that ν is a measure on measurable space $(X, \mathcal{M})$. We can show that the set function ν is a measure on the measurable space $(X, \mathcal{M})$ and it has the property that

$$\text{if } E \in \mathcal{M} \text{ and } \mu(E) = 0, \text{ then } \nu(E) = 0 \tag{13}$$

The Radon-Nikodym theorem asserts that if μ is σ-finite, then every σ-finite measure ν on $(X, \mathcal{M})$ that posseses the property (13) is given by (12) for some nonnegative function f on

X that is measurable with respect to $\mathcal{M}$.
A measure ν is said to be **absolutely continuous** with respect to the measure μ provided (13) holds. We use the symbolism $\nu << \mu$ for ν absolutely continuous with respect to ν.

The following example shows that the hypothesis in the Radon-Nikodym Theorem that μ is σ-finite cannot be omitted. Let $X = [0, 1]$, $\mathcal{B}$ the class of Lebesgue measurable subsets of $[0, 1]$, and take ν to be Lebesgue measure and μ to be the counting measure on $\mathcal{B}$. Then ν is finite and absolutely continuous with respect to μ, but there is no function f such that

$$\nu(E) = \int_E f d\mu \quad \text{for all } E \in \mathcal{B} \tag{14}$$

6 Conditional Expectation

Given a probability space $(\Omega, \mathcal{F}, P)$ and two events A and B in $\mathcal{F}$ with $P(B) > 0$, the conditional probability of A given B is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{15}$$

The conditional expectation of a random variable X given the event B is defined (when it exists) as

$$\mathbb{E}(X|B) = \frac{\int_B X dP}{P(B)} \tag{16}$$

Example: Consider the experiment consisting of 3 tosses of a fair coin.

$$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\} \tag{17}$$

Let A be the event consisting of the outcomes which have 2 heads, $A = \{HHT, HTH, THH\}$.
Let B be the event consisting of those outcomes with first head, $B = \{HHH, HHT, HTH, HTT\}$.
Let us form the σ-algebra $\mathcal{F}_1$ consisting of $\{A, B, A^c, B^c, A \cup B, A \cup B^c, A^c \cup B, A^c \cup B^c, A \cap B, A \cap B^c, A^c \cap B, A^c \cap B^c, \Omega, \emptyset\}$. In order to assign probabilities let us assume that p is the probability of getting a head.

Set	Elements of the set	Probability	p=0.4
A	$\{HHT, HTH, THH\}$	$3p^2(1-p)$	0.2880
B	$\{HHH, HHT, HTH, HTT\}$	$p^3 + 2p^2(1-p) + p(1-p)^2$	0.4
A^c	$\{HHH, HTT, THT, TTH, TTT\}$	$p^3 + 3p(1-p)^2 + (1-p)^3$	0.7120
B^c	$\{THH, THT, TTH, TTT\}$	$p^2(1-p) + 2p(1-p)^2 + (1-p)^3$	0.6
$A \cup B$	$\{THH, HHH, HHT, HTH, HTT\}$		0.4960
$A \cup B^c$	$\{HHT, HTH, THH, THT, TTH,$ $TTT\}$		
$A^c \cup B$	$\{HHH, HTT, THT, TTH, TTT\}$		
$A^c \cup B^c$	$\{HHH, HTT, THT, TTH,$ $TTT, THH\}$		
$A \cap B$	$\{HHT, HTH\}$	$p^2(1-p) + (1-p)p^2 = 2p^2(1-p)$	0.1920
$A \cap B^c$	$\{THH\}$	$p^2(1-p)$	0.0960
$A^c \cap B$	$\{HHH, HTT\}$		
$A^c \cap B^c$	$\{THT, TTH, TTT\}$	$2p(1-p)^2 + (1-p)^3$	
Ω	$\{HHH, HHT, HTH, HTT, THH,$ $THT, TTH, TTT\}$	1	

We will consider 3 cases. In each case, the way in which information is given changes:

- **Case 1:** It is informed that some set $B \in \mathcal{F}$ has occured. This changes the measure assigned to all the sets in $\mathcal{F}$.
- **Case 2:** In this case we are given a σ-algebra$\mathcal{G}$ instead of a set B. We are informed that the outcome belongs to one of the sets in the $\mathcal{G}$.
- **Case 3:** We define a random variable, say Y. We are informed that the random variable has occured.

Let us now discuss each of the cases in detail:

- **Case 1:** It is informed that some set $B \in \mathcal{F}$ has occured. This changes the measure assigned to all the sets in $\mathcal{F}$.
  This case is simple. When you are informed that the set B has occured, we can define a new probability measure P' on $\mathcal{F}$.
- **Case 2:** In this case we are given a σ-algebra $\mathcal{G}$ instead of a set B. We are informed that the outcome belongs to one of the sets in the $\mathcal{G}$.
  Suppose you are given the σ-field $\{\emptyset, \Omega, B, B^c\}$.
- **Case 3:** We define a random variable, say Y. We are informed that the random variable has occured.

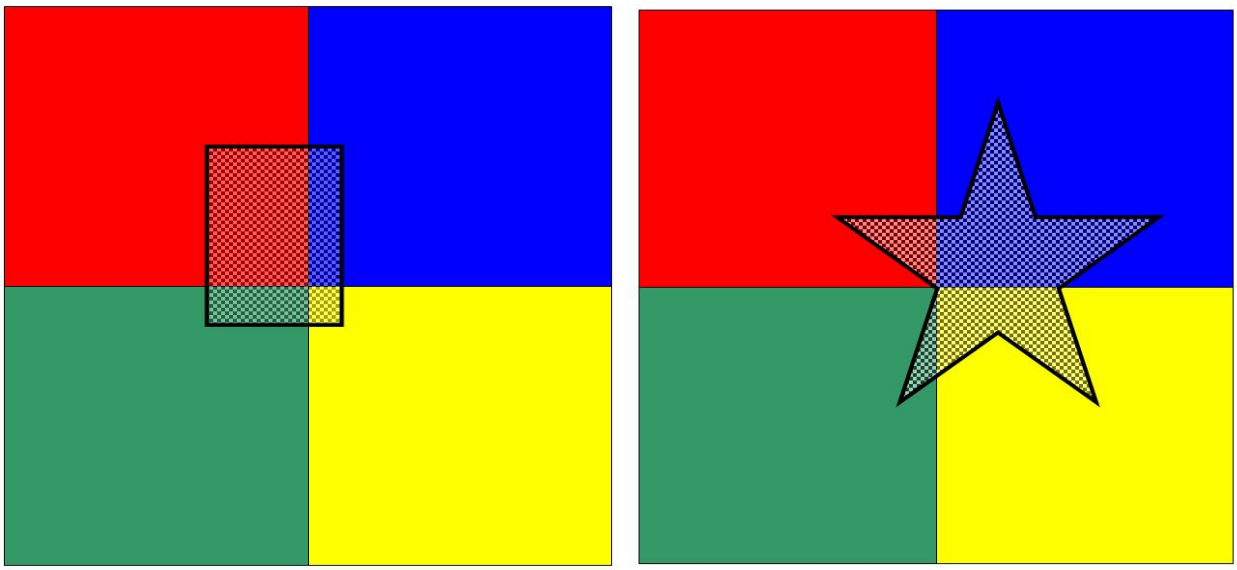
7 Conditional Expectation Notes-1

Conditional expectation had been defined as

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

in terms of the probability of the events. It was Kolmogrov, in his Foundations of Theory of Probability, gave the interpretation of conditional probabilities as random variables, based on the axiomatic approach to probability. Let us understand the concept of conditional probability with the help of the following example. Consider a slightly modified dart game with the following modification:

- the dart board is square and has a checkered region as shown in the figure below.The board is divided into 4 regions, indicated by the 4 colours.
- The player should throw the dart blind-folded
- The player should hit the dart in the checkered region to win. The score is based on the distance from the checkered region. Closer the dart is to the checkered region, higher will be the score.
- One game consists of 3 chances to hit the board.
- At the begining of each game, the shape of the checkered region is changed and the player sees the region and is then blinded folded.
- When the player throws the dart, there is a game arbitrator who tell s the colour of the region where the dart hit. This acts a guidance to the player in his next throw.

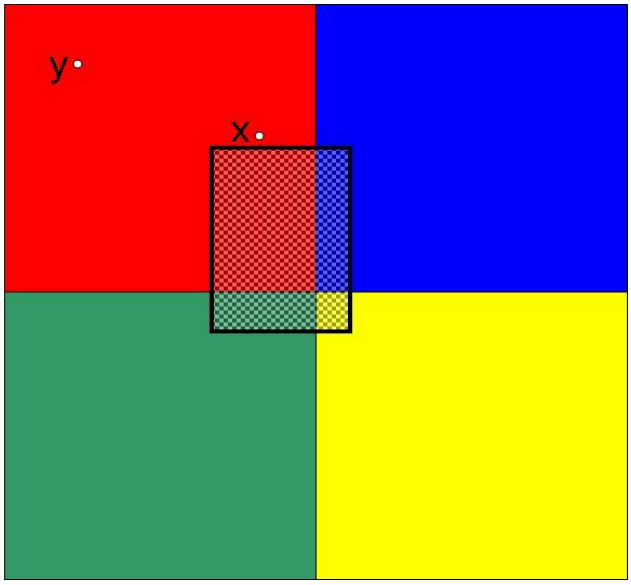


Let us analyse this game from the mathematical point of view. The square board forms the outcome of each throw, in other words, the sample space Ω. Let $\{\Omega, \mathcal{F}, P\}$ be the measure space associated with probability measure P. Let us name the four disjoint coloured regions, A_R, A_B, A_G, A_Y. These four regions are such that $\Omega = A_R \cup A_B \cup A_G \cup A_Y$. Let us consider the sub-σ-algebra $\mathcal{G} \subset \mathcal{F}$, formed by the finite union of the sets A_R, A_B, A_G, A_Y.
Notice the following:

- The σ-algebra $\mathcal{G}$ consists only of those sets (events) about which the arbitrator can give information to the player. For example,"Did the dart hit red?","Did the dart hit yellow?", "Did the dart hit green", "Did the dart hit blue?", "Did the dart hit yellow or red?" ,"Did the dart hit top half of the board?", this question is answer if the arbitrator calls out the colour, the player would know whether the dart hit top half or bottom half. Essentially, $\mathcal{G}$ models the information that the player will receive from the arbitrator. This shows how σ-algebra models the information, which justifies the statement that we often come across in statistics and probability that the information is contained in the σ-algebra.

- Notice that the arbitrator does not give any information about what is the distance of the current hit from the checkered region. All the arbitrator does is call out the colour of the region where the dart has hit.

So the player proceeds is, at the begining of the games, he sees the shape of the checkered region and calculates beforehand the probability of winning associated with each colour. Let the checkered region be denoted by C. Even though the arbitrator does not give the position of the dart, but just the colour, the best choice for the player is to throw the dart into the region with the highest probability of winning. With each point on the board, the player can associate what is the probability of winning better based on arbitrator's information, if the dart hits that particular point.



Even though a point x is closer to the checkered region than point y, since the player does not get that proximity information from the arbitrator, for player hitting anypoint in red is equivalent and so he associates the same probability of winning to any point in red.

So to each point $\omega \in \Omega$, the player associated a value as given below:

$$f(\omega) = \begin{cases} P(C|A_R) \text{ if } \omega \in A_R \\ P(C|A_Y) \text{ if } \omega \in A_Y \\ P(C|A_B) \text{ if } \omega \in A_B \\ P(C|A_G) \text{ if } \omega \in A_G \end{cases} \tag{18}$$

This $f(\omega)$ is the conditional probability of C given $\mathcal{G}$.

7.1 General Case

In the general case, we do not expect that the σ-algebra $\mathcal{G}$ comes from such a partition. But what is important here is to understand that $\mathcal{G}$ models the information that is available. Now instead of defining the conditional probability based on the partitions of Ω, we define a new measure ν. To do so let us fix a set $C \in \mathcal{F}$ and define a finite measure ν on $\mathcal{G}$.

$$\nu(G) = P(C \cap G) \text{ for all } G \in \mathcal{G} \tag{19}$$

Radon Nikodym..
Definition
Gambling Interpretation

8 Conditional Expectations Notes 2 (Jun Shao)

Let $\mathcal{C}$ be a collection of subsets of Λ. We define

$$f^{-1}(\mathcal{C}) = \{f^{-1}(C) : C \in \mathcal{C}\} \tag{20}$$

Definition (Measurable function): Let $(\Omega, \mathcal{F})$ and $(\Lambda, \mathcal{G})$ be measurable spaces and f a function from Ω to Λ. The function f is called a measurable function from $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$ if and only if $f^{-1}(\mathcal{G}) \subset \mathcal{F}$.

If f is measurable from $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$, then $f^{-1}(\mathcal{G})$ is a sub-σ-field of $\mathcal{F}$. It is called the σ-field generated by f and is denoted by $\sigma(f)$.

Definition (Conditional Expectation): Let X be an integrable random variable on $(\Omega, \mathcal{F}, P)$.

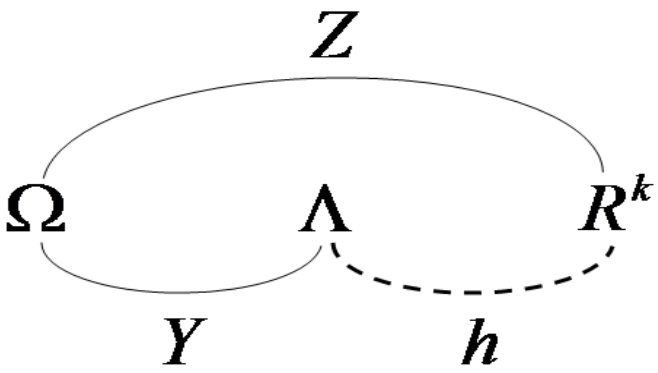
1. Let $\mathcal{A}$ be a sub-σ-field of $\mathcal{F}$. The conditional expectation of X given $\mathcal{A}$, denoted by $E(X|\mathcal{A})$, is the a.s.-unique random variable satisfying the following two conditions:
   (a) $E(X|\mathcal{A})$ is measurable from $(\Omega, \mathcal{A})$ to $(\mathbb{R}, \mathcal{B})$
   (b) $\int_A E(X|\mathcal{A})dP = \int_A XdP$ for any $A \in \mathcal{A}$.

2. Let $B \in \mathcal{F}$. The conditional probability of B given $\mathcal{A}$ is defined to be $P(B|\mathcal{A}) = E(I_B|\mathcal{A})$

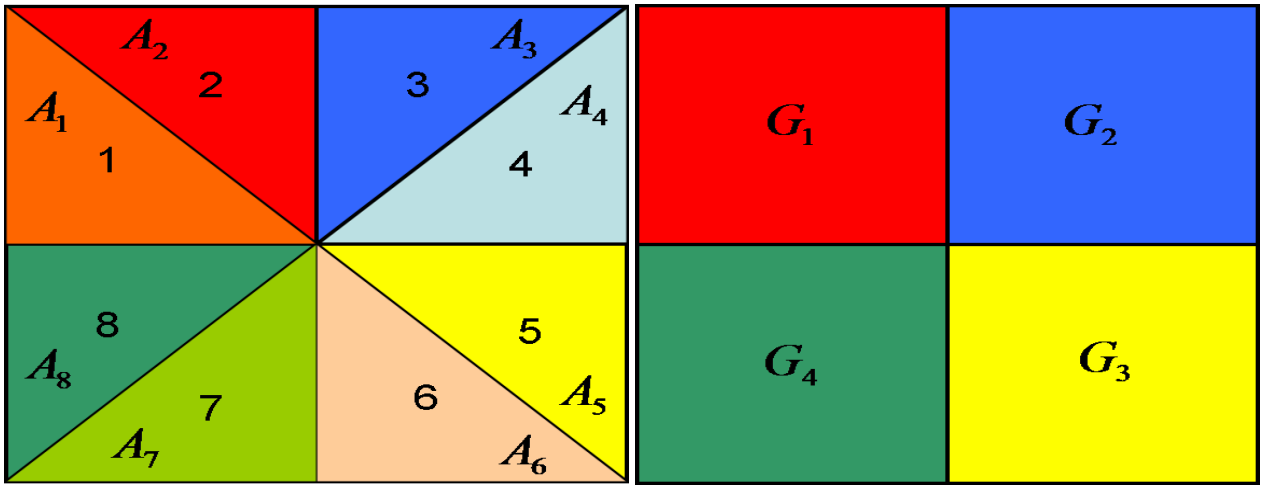
3. Let Y be measurable from $(\Omega, \mathcal{F}, P)$ to $(\Lambda, \mathcal{G})$. The conditional expectation of X given Y is defined to be $E(X|Y) = E[X|\sigma(Y)]$

Example: Let us continue with the dart board example to understand the definition of conditional expectation. Let us assign numbers to each of the colours. Let $Red = 10, Blue = 20, Green = 30, Yellow = 40$. Let Y be the random variable defined on $(\Omega, \mathcal{F}, P)$ to $(\mathbb{R}, \mathcal{G})$. Here $\mathcal{F}$ consists of the Borel sets of $\mathbb{R}^2$. $\mathcal{G}$ consists of the finite union of the sets A_R, A_B, A_G, A_Y, as shown in the example above.

Lemma 1.2: let Y be a measurable from $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$ and Z a function from $(\Omega, \mathcal{F})$ to $\mathcal{R}^k$. Then Z is measurable from$(\Omega, \sigma(Y))$ to $(\mathcal{R}^k, \mathcal{B}^k)$ if and only if there is a measurable function h from $(\Lambda, \mathcal{G})$ to $(\mathcal{R}^k, \mathcal{B}^k)$ such that $Z = h \circ Y$.
Proof: Given that $Y : \Omega \to \Lambda$ is measurable from $\langle \mathcal{F}, \mathcal{G} \rangle$.
Only If: If Z is measurable w.r.t $\langle \mathcal{F}, \mathcal{R}^k \rangle$, then show that there exists a function h such that $Z = h \circ Y$ and h is measurable w.r.t $\langle \mathcal{G}, \mathcal{R}^k \rangle$.
Define h such that $h(\nu) = Z(\omega)$, if $\nu = Y(\omega)$, $\nu \in \Lambda$. Now we have to show that h is measurable w.r.t $\langle \mathcal{G}, \mathcal{R}^k \rangle$. That is, we have to show that, $h^{-1}(B) \in \mathcal{G}$ for all $B \in \mathcal{B}^k$.



Example: Let us consider the dart game example again. The board space, the square, is the Ω. The σ-algebra $\mathcal{F}$ consists of finite union of the set $A_i, i = 1, \cdots, 8$, as shown in the figure below. Let $\mathcal{G}$ be the σ-algebra of finite union of $G_i, i = 1, \cdots, 4$. Let $Y : \Omega \to \Lambda$ be a measurable function $\langle \mathcal{F}, \mathcal{G} \rangle$. Of course in this case $\Lambda = \Omega$, which is say equal to $[-a, a] \times [-a, a]$.



References

[1] Patrick Billingsley,"Probability and Measure ", 2nd Edition, Chapter 6.

[2] M. Loeve," Probability Theory", Chapter: Conditioning.