Conditional Probability and Independence

Reuy-Lin Sheu

Department of Mathematics, National Cheng Kung University,
Tainan, Taiwan

November, 2023

Conditional Probability

Let $(\Omega, \mathcal{F}, P)$ be a probability space.

- Suppose a random experiment is conducted, and $\omega \in \Omega$ is the outcome. Let $A,\ B \in \mathcal{F}$ be two events. The probability that $\omega \in A$ is $P(A)$.
- Suppose ω is not known to you, but you are told a partial information that $\omega \in B$. Conditional on the information that $\omega \in B$, the probability that $\omega \in A$ becomes

$$P(A|B) \triangleq \frac{P(A \cap B)}{P(B)} \in [0, 1].$$

- In general, the information whether B has just occurred may well change our way of betting on the event A. So, $P(A|B) \neq P(A)$.
- When $P(A|B) = P(A)$, in which case we say A is independent of B. The additional information from learning that B happened does not permit any inference about the probability for the occurrence of A. That is, A is independent of B *iff*

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

Independence

Let $(\Omega, \mathcal{F}, P)$ be a probability space.

- Notice that, whether two sets are independent depends on the probability measure P.
- For example, in the two-stage binomial asset pricing model, recall that the probability $p \in [0, 1]$ for a single toss H and $q = 1 - p$ for T with which we define the probability P in $(\Omega, \mathcal{F}, P)$ as

$$P(HH) = p^2, P(HT) = pq, p(TH) = pq, P(TT) = q^2.$$

- Let $A = \{HH, HT\}$ be the event that the stock price moves up in the first trading, whereas $B = \{HT, TH\}$ be the event that "one up and one down" with $P(B) = 2pq$.
- Since $P(A)P(B) = 2p^2q$ and $P(A \cap B) = P(HT) = pq$, the two events A and B are independent if and only if $2p^2q = pq$. That is, $p = 0.5 \vee p = 1(q = 0) \vee p = 0$.
- When $p = 0.5$, the probability of B that "one up and one down" is 0.5. Suppose you are told that in the first day the stock price actually went up (i.e. Event A happened), the probability of $P(B|A)$, which is now the probability of "down" in the second day, is still 0.5. So, events A and B are independent.

Independence

Let $(\Omega, \mathcal{F}, P)$ be a probability space and

$$P(A) = p, P(B) = 2pq, P(A)P(B) = 2p^2q, P(A \cap B) = pq.$$

- Suppose now the probability measure P changes to $p = 0.01$. The chance for the stock price to go up is doomed. It is very likely to end up with "two downs" in a two-days trading section.
- In this case, the probability of B, one up and one down, is still quite small. It is $P(B) = 2pq = 2 \cdot 0.01 \cdot 0.99 = 0.0198$.
- However, if you are told a surprise that in the first day the stock price indeed went up, now you know $B = \{HT, TH\}$ (one up and one down) becomes very highly probable. In fact, $P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{pq}{p} = q = 0.99$.
- With the probability measure specified by $p = 0.01$, not only do we see that $P(B|A) \neq P(B)$, but also $P(B|A) \gg P(B)$ that knowing A greatly change the way we think of the probability of B to happen.

Independence

Let $(\Omega, \mathcal{F}, P)$ be a probability space.

- We first observe that, when $P(A) = 0$, then A is independent of any $B \in \mathcal{F}$ since $B \cap A \subset A \Rightarrow 0 \leq P(B \cap A) \leq P(A) = 0$. Therefore, $P(A \cap B) = 0 = P(A)P(B)$.
- Now assume that $P(A) \neq 0, P(B) \neq 0$. It is obvious that, if A is independent of B, then B is also independent of A. That is,

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

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

- Moreover, if A and B are independent, from

$$P(A|B^c) = \frac{P(AB^C)}{P(B^c)} = \frac{\overbrace{1 - P(A^c)}^{P(A)} - \overbrace{P(AB)}^{P(A)P(B)}}{1 - P(B)} = P(A),$$

we know A is independent of B^c.

- Immediately, if A and B are independent, B is also independent of A^c and also there is B^c independent of A^c.

Independence

Let $(\Omega, \mathcal{F}, P)$ be a probability space.

- Moreover, any event $A \in \mathcal{F}$ must be independent of Ω and also of $\Omega^c = \emptyset$.
- This is because $P(A \cap \Omega) = P(A) = P(A)\underbrace{P(\Omega)}_{=1}$ is always true.
- In other words, suppose that A and B are independent, any pair of events $(A', B'),\ A' \in \sigma(A),\ B' \in \sigma(B)$ is also independent.
- Let $\mathcal{G}$ and $\mathcal{H}$ be sub-σ-algebras of $\mathcal{F}$. We say that $\mathcal{G}$ and $\mathcal{H}$ are *independent* if every set in $\mathcal{G}$ is independent of every set in $\mathcal{H}$. That is, $P(A \cap B) = P(A)P(B),\quad \forall A \in \mathcal{G},\ B \in \mathcal{H}$.
- Therefore, two events $A, B \in \mathcal{F}$ are independent if and only if the two sub-sigma-algebras $\sigma(A)$ and $\sigma(B)$ are independent.
- Notice that, for an event $A \in \mathcal{F}$, it must not be independent of its own complement A^c since $P(A \cap A^c) = P(\emptyset) = 0 \neq P(A)P(A^c)$, unless $P(A) = 0 \vee 1$.

Independence of σ-algebras (Exercise)

- Toss a coin twice with $\Omega_2 = \{HH, HT, TH, TT\}$ and $\mathcal{F}_2$ (with $|\mathcal{F}_2| = 16$) is the σ-algebra consisting of all information up to time 2.
- Let the probability of tossing a H in the first trial is $p_1 \in (0,1)$ and that for the second trial is $p_2 \in (0,1)$. Define the probability P as

  $$P(HH) = p_1 p_2, P(HT) = p_1(1-p_2), p(TH) = (1-p_1)p_2, P(TT) = (1-p_1)(1-p_2).$$

- Then, $(\Omega_2, \mathcal{F}_2, P)$ is a probability space.
- Let $\mathcal{G} = \{\emptyset, \Omega_2, \{HH, HT\}, \{TH, TT\}\} \subset \mathcal{F}_2$ be the sub-σ-algebra consisting of information of the first toss; while $\mathcal{H} = \{\emptyset, \Omega_2, \{HH, TH\}, \{HT, TT\}\} \subset \mathcal{F}_2$ be the sub-σ-algebra consisting of information of the second toss.
- Show that the two σ-algebras $\mathcal{G}, \mathcal{H}$ are independent under the specified probability.
- If we change the probability P of $(\Omega_2, \mathcal{F}_2, P)$ to become $P(HH) = \frac{1}{9}, P(HT) = \frac{2}{9}, p(TH) = \frac{1}{3}, P(TT) =$??. Show that, the two sub-σ-algebras $\mathcal{G}, \mathcal{H}$ become dependent.

Independence of random variables

Let $(\Omega, \mathcal{F}, P)$ be a probability space.

- We say that two random variables X and Y are *independent* if the σ-algebras they generate, $\sigma(X)$ and $\sigma(Y)$, are independent.
- For example, in the probability space of the binomial asset pricing model with three independent tosses of the same coin, the two random variables S_2 and $\frac{S_3}{S_2}$ are independent.
- The σ-algebra generated by S_2 is $\mathcal{F}_2$

$$\sigma(S_2) = \{\emptyset, \Omega, A_{HH}, A_{TT}, A_{HH}^c, A_{TT}^c, A_{HH} \cup A_{TT}, A_{HT} \cup A_{TH}\}.$$

- On the other hand, the random variable S_3 takes on values uS_2 if the third toss is H; and dS_2 otherwise. So, $\frac{S_3}{S_2}$ takes only two values u, d. Then,

$$\sigma\left(\frac{S_3}{S_2}\right) = \{\emptyset, \Omega_3, A_{..H}, A_{..T}\}$$

- It is easy to see the independence by the following expression:

$$P(A_{HH} \cap A_{..H}) = P(\{HHH\}) = p^3 = \underbrace{P(A_{HH})}_{=p^2} \underbrace{P(A_{..H})}_{=p}$$

Independence of random variables

Let $(\Omega, \mathcal{F}, P)$ be a probability space.

- Suppose X and Y are independent random variables. Let g and h are Borel measurable functions from $(\mathbb{R}, \mathcal{B})$ to $(\mathbb{R}, \mathcal{B})$. Then, $g(X)$ and $h(Y)$ are also independent random variables.
- This is so because $\sigma(g(X)) \subset \sigma(X)$ and $\sigma(h(Y)) \subset \sigma(Y)$.
- For example, if X and Y are independent random variables, then X^2 and e^Y are also independent to each other.
- Information gets suppressed after operations and becomes less informative. For example, suppose $X(\omega) = 1$ if $\omega = H$ and $X = -1$ if $\omega = T$. Then, $X^2(\omega) = 1, \forall \omega \in \{H, T\}$.
- In this example, X^2 provides less information than X does:

$$\sigma(X^2) = \{\emptyset, \Omega\} \subset \sigma(X) = \{\emptyset, \Omega, \{H\}, \{T\}\}$$

Independent disjoint grouping theorem

Let $(\Omega, \mathcal{F}, P)$ be a probability space

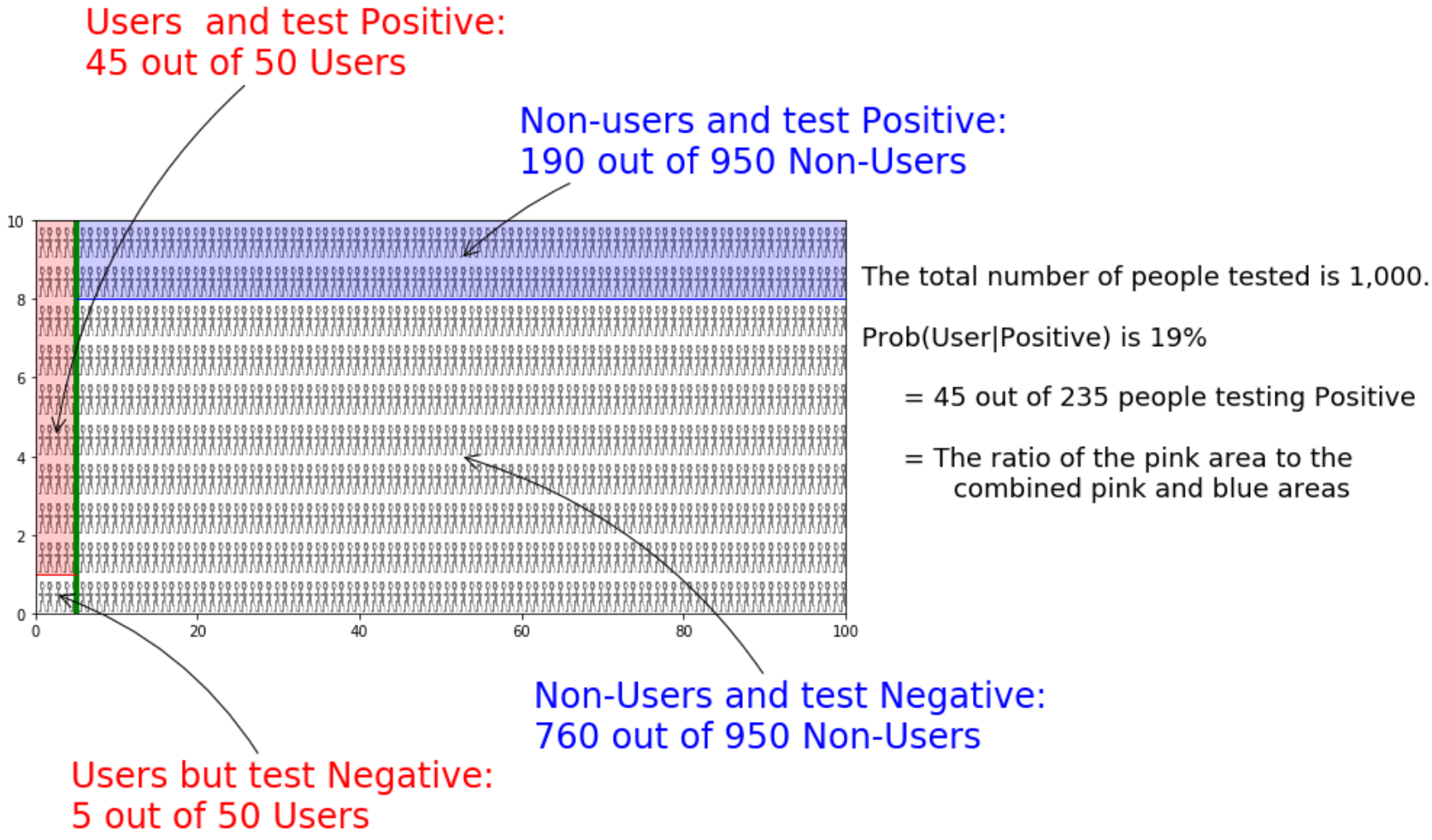
- Suppose $\mathcal{G}_1, \mathcal{G}_2, \cdots, \mathcal{G}_n$ are sub-σ-algebras of $\mathcal{F}$. Then, $\mathcal{G}_1, \mathcal{G}_2, \cdots, \mathcal{G}_n$ are said to be independent, if for all choices of sets $G_i \in \mathcal{G}_i,\ i = 1, 2, \ldots, n$, we have
$$P\left(\bigcap_{i=1}^{n} G_i\right) = \prod_{i=1}^{n} P(G_i)$$
- Moreover, let $X_i,\ i = 1, 2, \ldots, n$ be random variables on $(\Omega, \mathcal{F}, P)$. Then, $X_i,\ i = 1, 2, \ldots, n$ are said to be independent if the sub-σ-algebras $\sigma(X_i),\ i = 1, 2, \ldots, n$ are independent.
- In particular, for each $(x_1, x_2, \ldots, x_n)^T \in \mathbb{R}^n$, $P\{X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n\} = P(X_1 \leq x_1)P(X_2 \leq x_2) \cdots P(X_n \leq x_n)$.
- Let $\{\mathcal{G}_t, t \in T\}$ be an independent family of sub-σ-algebras of $\mathcal{F}$ so that any finite sub-collection of it is independent. Let $\{T_s, s \in S\}$ be a family of disjoint nonempty subsets of T. If $\mathcal{G}_{T_s} = \sigma\{\mathcal{G}_t, t \in T_s\}$, then $\{\mathcal{G}_{T_s}, s \in S\}$ is an independent family.
- For example, if X_1, X_2, X_3, X_4 are independent, then X_2X_3 (a random variable depending only on the outcome of X_2 and X_3) and $X_1 + \cos(2X_4)$ (a random variable depending only on the outcome of X_1 and X_4) are independent.

Formulae based on conditional probability - Bayes' theorem

- Conditional probabilities are sometimes given, or can be easily determined especially in sequential random experiments.
- For example, in Drug testing sample, a test for whether someone has been using cannabis is 90% sensitive, meaning the true positive rate (TPR)$= 0.90 = \frac{45}{50}$. It means that the probability of a positive test result, conditioned on the individual truly being positive, is 0.9.
- In addition, the drug test sample is also 80% specific, meaning true negative rate (TNR) $= 0.80 = \frac{760}{950}$. Therefore, the test correctly identifies 80% of non-use for non-users. The probability of a negative test result, conditioned on the individual truly being negative, is 80%.
- The prevalence of the sample (or the population), the probability of a random person who uses cannabis, is $5\% = \frac{50}{1000}$.
- We would like to know the probability that a random person in the population who tests positive is really a cannabis user? That is, compute the conditional probability $P(\text{User}|\text{Positve})$.

Formulae based on conditional probability - Bayes' theorem

- In this sample, there are $45 + 190 = 235$ persons who are tested positive; among which 45 persons are true users. Then, we infer that, in the population, $P(\text{User}|\text{Positve}) = \frac{45}{235} \approx 19.15\%$.



Formulae based on conditional probability - Bayes' theorem

- Bayes' theorem is stated mathematically as the following equation:

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

- In the example, A is the set for those tested positive, B is the set for true cannabis uses. Then,

$$P(\text{User}|\text{Positve}) = \frac{\overbrace{P(\text{Positve}|\text{User})}^{\text{TPR}=0.9} \overbrace{P(\text{User})}^{\text{Prevalence}=0.05}}{P(\text{Positve})} = \frac{0.045}{P(\text{Positve})}.$$

- As for those tested positive, they may be contributed from either users or non-users. That is,

$$\begin{aligned} P(\text{Positve}) &= P(\text{Positve} \wedge \text{User}) + P(\text{Positve} \wedge \text{NonUser}) \\ &= P(\text{Positve}|\text{User})P(\text{User}) + P(\text{Positve}|\text{NonUser})P(\text{NonUser}) \\ &= 0.9 \cdot 0.05 + (1 - \text{TNR} = 0.2) \cdot 0.95 = 0.235 \end{aligned}$$

- Therefore, $P(\text{User}|\text{Positve}) = \frac{0.045}{0.235} \approx 0.1915$, which is much higher than the prevalence rate 0.05.

Formulae based on conditional probability- Bayes' theorem (Theorem 4.2 at page 28 in the textbook)

- In general, we have the following more general form of Bayes' theorem: Let $(\Omega, \mathcal{F}, P)$ be a probability space. Assume that $F_1, F_2, \ldots, F_n$ are pairwise disjoint events in $\mathcal{F}$ and that $F_1 \cup F_2 \cup \ldots \cup F_n = \Omega$.
- Each event $F_j,\ j = 1, 2, \ldots, n$ is unknown and often called as a hypothesis, while $P(F_j)$ its prior probability can be obtained.
- In the above drug testing example, F_1 is the hypothesis set of drug users with the prevalence 0.05 as its prior probability, while F_2 is the complement of hypothesis F_1.
- Let $A \in \mathcal{F}$ be an evidence event. Then, the conditional probability $P(F_j|A)$ is the posterior probability of the hypothesis F_j given the evidence A.
- In the example, A is the event of people who are tested positive. Given the evidence that a person is tested positive, we want to compute the posterior probability that the person is indeed a cannabis user can be formulated as $P(F_1|A)$.

Formulae based on conditional probability- Bayes' theorem (Theorem 4.2 at page 28 in the textbook)

- Bayes' theorem says that, the posterior (conditional) probability of the hypothesis F_j to be true, given the evidence A can be computed through the set of (assumed) prior probability $P(F_i), i = 1, 2, \ldots, n$ on all the hypotheses; as well as the conditional probabilities for the evidence A to happen, given the various hypotheses.
- Namely,

$$P(F_j|A) = \frac{P(F_j \cap A)}{P(A)} = \frac{P(A|F_j)P(F_j)}{P(A|F_1)P(F_1) + \cdots + P(A|F_n)P(F_n)}$$

- In the example of cannabis testing, $P(A|F_1)$ (TPR) is the rate of actual cannabis users that were correctly identified by the test, while $P(A|F_2)$ (TNR) is the rate of non-cannabis users that were correctly passed by the test.

Formulae based on conditional probability- Bayes' theorem (Theorem 4.2 at page 28 in the textbook)

$$P(F_j|A) = \frac{P(F_j \cap A)}{P(A)} = \frac{P(A|F_j)P(F_j)}{P(A|F_1)P(F_1) + \cdots + P(A|F_n)P(F_n)}$$

- In general, $P(A|F_j)$ means the rate of known samples in class F_j that have been correctly predicted by the underlined classification model.
- They $(P(A|F_j),\ j = 1, 2, \ldots, n)$ can be obtained from a given training set of data.
- In the denominator of the formula, since $\Omega = F_1 \cup F_2 \cup \ldots \cup F_n$ is partitioned into various hypothesis, we have

$$\begin{aligned} P(A) &= P(A \cap \Omega) = P\left(A \cap (\cup_{j=1}^{n} F_j)\right) \\ &= P\left(\cup_{j=1}^{n} (A \cap F_j)\right) \\ &= P(A \cap F_1) + P(A \cap F_2) + \cdots + P(A \cap F_n) \\ &= P(A|F_1)P(F_1) + P(A|F_2)P(F_2) + \cdots + P(A|F_n)P(F_n) \end{aligned}$$

Joint distribution (general setting)

Let $(\Omega, \mathcal{F}, P)$ be a probability space and X and Y be two random variables on it. The induced measure by X and Y are $\mathcal{L}_X$ and $\mathcal{L}_Y$ respectively.

- Consider the random vector

$$(X, Y) : (\Omega, \mathcal{F}, P) \to (\mathbb{R} \times \mathbb{R}, \sigma(\mathcal{B} \times \mathcal{B})).$$

The induced measure by the r. vector (X, Y) is defined by

$$\mathcal{L}_{(X,Y)}(C) = P\{\omega \in \Omega |\ (X(\omega), Y(\omega)) \in C\},\ \forall C \in \sigma(\mathcal{B} \times \mathcal{B}).$$

- In particular, we should consider a "rectangular" type of C in $\mathbb{R}^2$ that $C = A \times B,\ A, B \in \mathcal{B}$. In this case,

$$\{\omega \in \Omega | (X(\omega), Y(\omega)) \in A \times B\} = \{\omega | X(\omega) \in A\} \bigcap \{\omega | Y(\omega) \in B\}.$$

- Not every $2D$ borel set $\sigma(\mathcal{B} \times \mathcal{B})$ is rectangular. For example, the union of two rectangles is usually not a rectangle.

Joint distribution (discrete case) (page 70 in the textbook)

Suppose X and Y be two discrete r.v.'s on $(\Omega, \mathcal{F}, P)$ with discrete range spaces $\mathcal{R}(X)$ and $\mathcal{R}(Y)$ respectively.

- The random vector $(X, Y) : (\Omega, \mathcal{F}, P) \to (\mathbb{R} \times \mathbb{R}, \sigma(\mathcal{B} \times \mathcal{B}))$ has the joint probability mass function defined on $\mathcal{R}(X) \times \mathcal{R}(Y)$, denoted by, $\forall (x, y) \in \mathcal{R}(X) \times \mathcal{R}(Y)$,

$$\begin{aligned} p_{(X,Y)}(x, y) &= P(X = x, Y = y) \\ &= P(\omega \in \Omega | X(\omega) = x, Y(\omega) = y) = \mathcal{L}_{(X,Y)}(\{x\} \times \{y\}) \end{aligned}$$

  where $C = \{x\} \times \{y\}$ is the most important special "rectangle" type of Borel sets in the discrete case.

- Let $A \in \sigma(\mathcal{B} \times \mathcal{B})$ be a joint Borel set. Then,

$$P((X, Y) \in A) = P(\omega \in \Omega | (X(\omega), Y(\omega)) \in A) = \sum_{(x,y) \in A} p_{(X,Y)}(x, y)$$

- The *marginal probability functions* of the r. vector (X, Y) are

$$P(X = x) = \sum_{y \in \mathcal{R}(Y)} P(X = x, Y = y); \; P(Y = y) = \sum_{x \in \mathcal{R}(X)} P(X = x, Y = y)$$

Joint distribution (discrete case) (Example 7.1 at page 70 in the textbook)

An urn has 2 red, 5 white, and 3 green balls. Select 3 balls at random and let X be the number of red balls; and Y be the number of white balls.

- We first notice that the sample space Ω contain all possible $C_3^{10} = 120$ selection of 3 balls from a basket of 10 balls.
- The r.v. X sends every selection $\omega \in \Omega$ to the number of red balls. It is easy to see that the range space of X is $\mathcal{R}(X) = \{0, 1, 2\}$.
- As for the r.v. Y, $Y(\omega)$ is the number of white balls in a selection of 3 balls $\omega \in \Omega$. Thus, $\mathcal{R}(X) = \{0, 1, 2, 3\}$.
- The random vector (X, Y) sends an $\omega \in \Omega$ to a point in $\mathbb{R}^2$. Specifically, $(X, Y)(\omega) = (X(\omega), Y(\omega)) \in \{0, 1, 2\} \times \{0, 1, 2, 3\} \subset \mathbb{R}^2$
- The joint probability mass function is defined on $\forall (x, y) \in \{0, 1, 2\} \times \{0, 1, 2, 3\}$ for the probability that both $X = x$ and $Y = y$ to happen simultaneously.

Joint distribution (discrete case) (Example 7.1 at page 70 in the textbook)

An urn has 2 red, 5 white, and 3 green balls. Select 3 balls at random and let X be the number of red balls; and Y be the number of white balls.

- For example, $(0,0) \in \{0,1,2\} \times \{0,1,2,3\}$. The joint probability for both $X=0$ and $Y=0$ to happen simultaneously is

$$p_{(X,Y)}(0,0) = P(\omega \in \Omega | X(\omega)=0, Y(\omega)=0) = \mathcal{L}_{(X,Y)}(\{0\} \times \{0\}) = \frac{1}{120}$$

since there is only one possible choice: {green, green, green} that has no red nor white balls.

- Another example: $(1,0) \in \{0,1,2\} \times \{0,1,2,3\}$. Then,

$$p_{(X,Y)}(1,0) = P(\omega \in \Omega | X(\omega)=1, Y(\omega)=0) = \mathcal{L}_{(X,Y)}(\{1\} \times \{0\}) = \frac{2 \cdot 3}{120}$$

- For $(X, Y) = (2,0)$, $p_{(X,Y)}(2,0) = \frac{3}{120}$.
- The marginal probability $P(Y=0)$, the probability that a 3-balls selection contains no white ball is

$$P(Y=0) = P(X=0, Y=0) + P(X=1, Y=0) + P(X=2, Y=0) = \frac{10}{120}.$$

Joint distribution (discrete case) (Example 7.1 at page 70 in the textbook)

An urn has 2 red, 5 white, and 3 green balls. Select 3 balls at random and let X be the number of red balls; and Y be the number of white balls.

- The following table shows that joint probability mass function and two marginal probability mass functions.

$y \backslash x$	0	1	2	$P(Y = y)$
0	$1/120$	$2 \cdot 3/120$	$3/120$	$10/120$
1	$5 \cdot 3/120$	$2 \cdot 5 \cdot 3/120$	$5/120$	$50/120$
2	$10 \cdot 3/120$	$10 \cdot 2/120$	0	$50/120$
3	$10/120$	0	0	$10/120$
$P(X = x)$	$56/120$	$56/120$	$8/120$	1

- From this table, we can also compute

$$P(X = 2|X \geq Y) = \frac{P(X = 2, X \geq Y)}{P(X \geq Y)} = \frac{3 + 5}{1 + 6 + 3 + 30 + 5} = \frac{8}{45}.$$

Joint distribution (continuous case) (page 71-72 in the textbook)

Suppose X and Y be two continuous r.v.'s on $(\Omega, \mathcal{F}, P)$ with range spaces $\mathcal{R}(X)$ and $\mathcal{R}(Y)$ respectively.

- The random vector $(X, Y) : (\Omega, \mathcal{F}, P) \to (\mathbb{R} \times \mathbb{R}, \sigma(\mathcal{B} \times \mathcal{B}))$ has the joint probability density function $f_{(X,Y)}(x, y) \geq 0$ defined on $\mathcal{R}(X) \times \mathcal{R}(Y)$, so that, $\forall S \in \sigma(\mathcal{B} \times \mathcal{B})$,

$$\begin{aligned} P((X, Y) \in S) &= P(\omega \in \Omega | X(\omega) = x, Y(\omega) = y, (x, y) \in S) \\ &= \iint_S f(x, y) dx \wedge dy \end{aligned}$$

- The two marginal densities of the random vector (X, Y), which are densities of X and Y, can be computed through the joint probability density $f_{(X,Y)}(x, y)$ as

$$f_X(x) = \int_{y \in \mathcal{R}(Y)} f_{(X,Y)}(x, y) dy; \; f_Y(y) = \int_{x \in \mathcal{R}(X)} f_{(X,Y)}(x, y) dx$$

Joint distribution (continuous case) (Example 7.8 at page 73 in the textbook)

- The simplest joint density function for a pair of continuous r.v.'s X and Y on $(\Omega, \mathcal{F}, P)$ is that (X, Y) is jointly uniform over $S = \mathcal{R}(X) \times \mathcal{R}(Y)$, where S is assumed to be compact.
- It has the following joint probability density function

$$f_{(X,Y)}(x, y) = \begin{cases} \frac{1}{\text{area}(S)}, & \text{if } (x, y) \in S; \\ 0, & \text{otherwise.} \end{cases}$$

Joint distribution (continuous case) (Example 7.8 at page 73 in the textbook)

- For example, $S = (x, y) : \ 0 \leq y \leq x \leq 1$. Then, the random vector (X, Y) jointly uniform over S has the joint p.d.

$$f_{(X,Y)}(x, y) = \begin{cases} 2, & \text{if } 0 \leq y \leq x \leq 1; \\ 0, & \text{otherwise.} \end{cases}$$

- The two marginal density functions are

$$f_X(x) = \int_0^x 2 \cdot dy = 2x, \ \forall x \in [0, 1]; \ \ f_Y(y) = \int_y^1 2 \cdot dx = 2(1-y), \ \forall y \in [0, 1].$$

- Notice that neither X nor Y is uniformly distributed on $[0, 1]$. Moreover, since $f_{(X,Y)}(x, y) \neq f_X(x) f_Y(y)$, the two r.v.'s are not independent[1].

[1] We shall see that, if X and Y are independent continuous r.v.'s, there must be $f_{(X,Y)}(x, y) = f_X(x) f_Y(y)$.

Joint distribution (continuous case) (Example 7.9 at page 74 in the textbook)

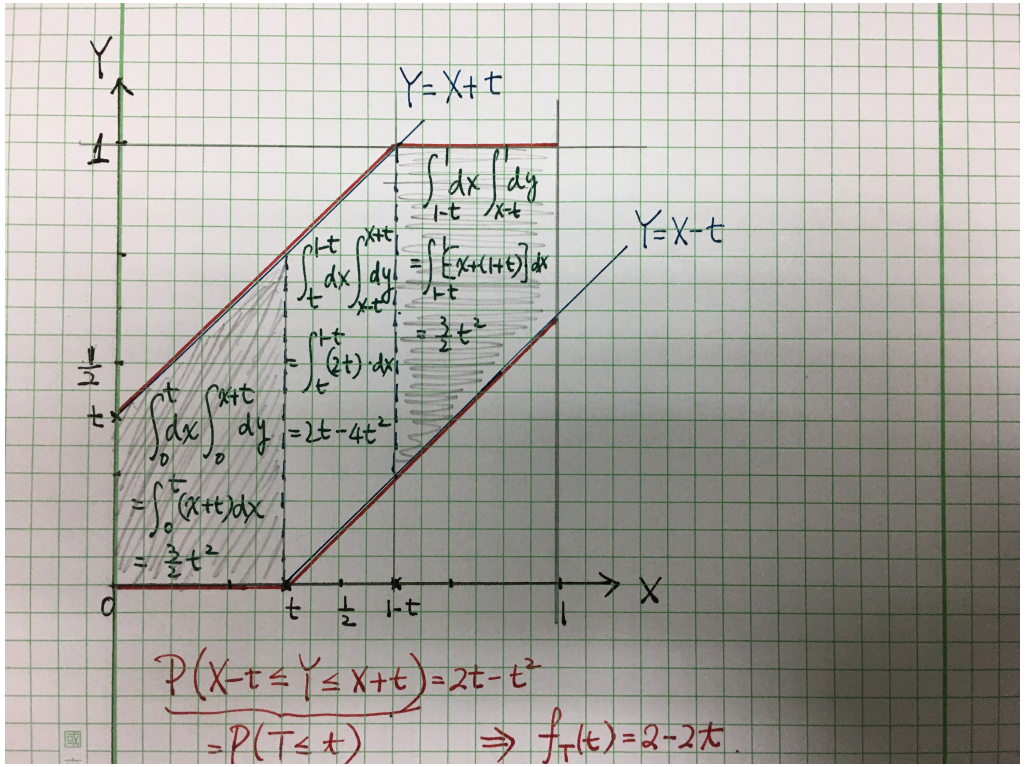
- Mr. and Mrs. Smith agree to meet at a specific location "between 5 and 6 p.m.". Assume that they both arrive there at a random time between 5 and 6 and that their arrivals are independent.
- What is the density function for the time one of them have to wait for the other?
- We may assume that X be the time when Mr. Smith arrives, while Y being the time when Mrs. Smith arrives with the time unit 1 hour. The assumption imply that the random vector (X, Y) is uniform on $S = [0, 1] \times [0, 1]$.
- We would like to find the probability density for $T = |X - Y|$, which has possible values in $[0, 1]$.
- Then, for $t \in [0, 1]$,

$$\begin{aligned} P(T \leq t) &= P(|X - Y| \leq t) \\ &= P(-t \leq Y - X \leq t) \\ &= P(X - t \leq Y \leq X + t) \end{aligned}$$

Joint distribution (continuous case) (Example 7.9 at page 74 in the textbook)

- To compute $P(T \le t) = P(X - t \le Y \le X + t)$ for $t \in [0, 1]$, we divide into two cases: $0 \le t \le \frac{1}{2}$ and $\frac{1}{2} \le t \le 1$.
- When $0 \le t \le \frac{1}{2}$, we have $t \le 1 - t$ and

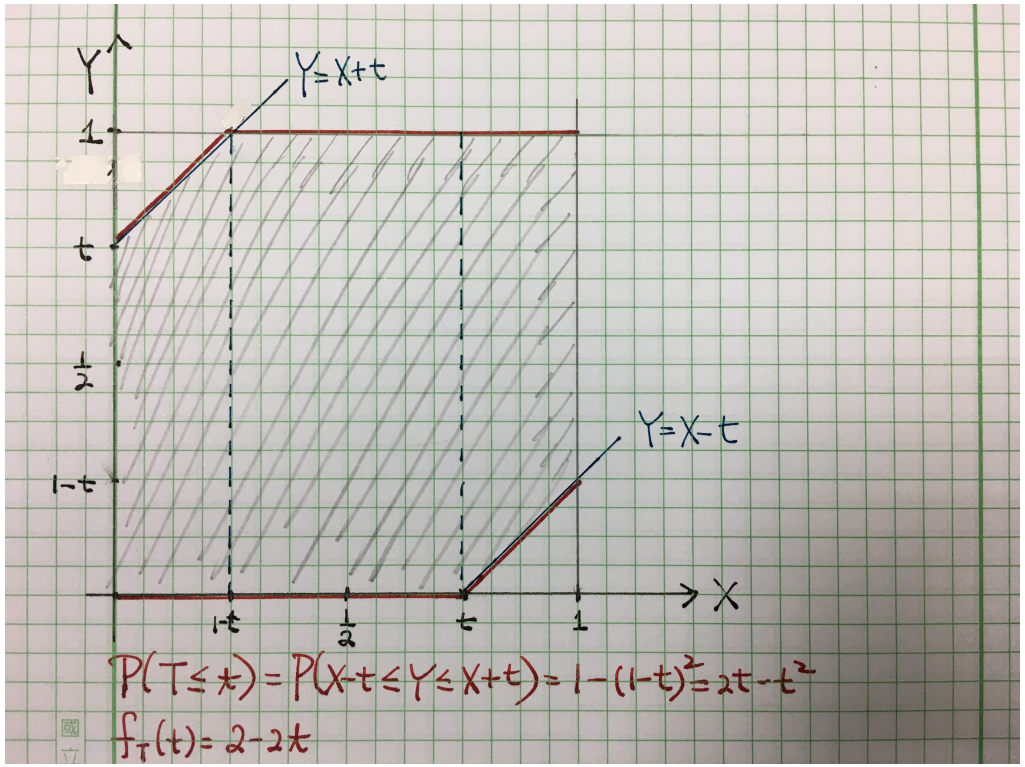
$$P(X - t \le Y \le X + t) = \begin{cases} P(0 \le Y \le X + t), & \text{if } 0 \le X \le t; \\ P(X - t \le Y \le X + t), & \text{if } t \le X \le 1 - t; \\ P(X - t \le Y \le 1), & \text{if } 1 - t \le X \le 1. \end{cases}$$



Joint distribution (continuous case) (Example 7.9 at page 74 in the textbook)

- When $\frac{1}{2} \leq t \leq 1$, we have $1 - t \leq t$ and

$$P(X - t \leq Y \leq X + t) = \begin{cases} P(0 \leq Y \leq X + t), & \text{if } 0 \leq X \leq 1 - t; \\ P(0 \leq Y \leq 1), & \text{if } 1 - t \leq X \leq t; \\ P(X - t \leq Y \leq 1), & \text{if } t \leq X \leq 1. \end{cases}$$



Expectation with Joint distribution (page 87 in the textbook)

- Suppose X and Y are two r.v.'s defined on $(\Omega, \mathcal{F}, P)$ with range spaces $\mathcal{R}(X)$ and $\mathcal{R}(Y)$ respectively, and g is a two-variable real-valued function defined on $\mathcal{R}(X) \times \mathcal{R}(Y)$.
- Then, $g(X, Y) : (\Omega, \mathcal{F}, P) \xrightarrow{(X,Y)} \mathcal{R}(X) \times \mathcal{R}(Y) \xrightarrow{g} \mathbb{R}$ can be calculated for its expectation

$$E(g(X, Y)) = \iint_{\mathcal{R}(X)\times\mathcal{R}(Y)} g(x, y) f_{(X,Y)}(x, y) dx \wedge dy$$

- When (X, Y) is instead a discrete pair with joint probability mass function $p_(X, Y)(x, y)$, then

$$E(g(X, Y)) = \sum_{(x,y)\in\mathcal{R}(X)\times\mathcal{R}(Y)} g(x, y) p_{(X,Y)}(x, y).$$

Expectation with Joint distribution (discrete case) (Example 8.1 at page 87 in the textbook)

Assume that 2 among 5 items are defective. Put the items in a random order and inspect them one by one. Let X be the number of inspections needed to find the first defective item; and Y be the number of additional inspections needed to find the second defective item. Compute $E|X-Y|$.

- We first construct a table for the joint p. m. f. of (X, Y).

joint p. m. f. of (X, Y) $(g(i,j) = \|i-j\|)$				
$X = i \setminus Y = j$	1	2	3	4
1	0.1(0)	0.1(1)	0.1(2)	0.1(3)
2	0.1(1)	0.1(0)	0.1(1)	0
3	0.1(2)	0.1(1)	0	0
4	0.1(3)	0	0	0

- For example, in the first column, $p(X=1, Y=1) = \frac{2}{5} \cdot \frac{1}{4} = 0.1$. $p(X=2, Y=1) = \frac{3}{5} \cdot \frac{2}{4} \cdot \frac{1}{3} = 0.1$. $p(X=3, Y=1) = \frac{3}{5} \cdot \frac{2}{4} \cdot \frac{2}{3} \cdot \frac{1}{2} = 0.1$. $p(X=4, Y=1) = \frac{3}{5} \cdot \frac{2}{4} \cdot \frac{1}{3} = 0.1$.
- $E[g(X, Y)] = E|X-Y| = 14 \cdot 0.1 = 1.4$

Expectation with Joint distribution (continuous case) (Example 8.2 at page 87 in the textbook)

Assume that (X, Y) is a random point in the triangle

$$S = \{(x, y) : x, y \geq 0, x + y \leq 1\}.$$

Compute EX and $E(XY)$.

- The joint probability density on S is

$$f_{(X,Y)}(x, y) = \begin{cases} 2, & \text{if } (x, y) \in S; \\ 0, & \text{otherwise.} \end{cases}$$

- We first compute the density of X through the joint p.d. of (X, Y) by $f_X(x) = \int_{y \in \mathcal{R}(Y)} f(x, y)dy = \int_0^{1-x} 2 \cdot dy = 2(1 - x)$.
- Then, $EX = \int_0^1 x \cdot 2(1 - x) \cdot dx = \frac{1}{3}$.
- Moreover,

$$\begin{aligned} E(g(X, Y)) &= E(XY) \\ &= \int_S 2g(x, y)dx \wedge dy \\ &= \int_0^1 dx \int_0^{1-x} 2xydxdy = \frac{1}{12}. \end{aligned}$$

Joint distribution and independence

Let $(\Omega, \mathcal{F}, P)$ be a probability space and X and Y be two random variables on it. The induced measure by X and Y are $\mathcal{L}_X$ and $\mathcal{L}_Y$ respectively.

- Recall that X and Y are independent if and only if $\sigma(X)$ and $\sigma(Y)$ are independent under the given probability measure P.
- Then, X and Y are independent if and only if, for any "rectangular" type of C in $\mathbb{R}^2$ such that $C = A \times B,\ A, B \in \mathcal{B}$.

$$
\begin{aligned}
& P(\{X \in A\} \cap \{Y \in B\}) \\
= \quad & \underbrace{P(\{\omega \in \Omega | (X(\omega), Y(\omega)) \in A \times B\})}_{=\mathcal{L}_{(X,Y)}(A \times B)} \\
= \quad & P\{X \in A\} P(Y \in B) \\
= \quad & \underbrace{\mathcal{L}_X(A)}_{\text{margianl dist. of } X} \cdot \underbrace{\mathcal{L}_Y(B)}_{\text{margianl dist. of } Y} .
\end{aligned}
$$

- In other words, for independent random variables X and Y, under probability measure P, the *joint distribution* represented by the measure $\mathcal{L}_{(X,Y)}$ factors into the product of the *marginal distributions* represented by the measures $\mathcal{L}_X$ and $\mathcal{L}_Y$.

Independence of random variables

- Let X and Y be two independent r.v.'s on a probability space $(\Omega, \mathcal{F}, P)$.
- Let $\mathbf{1}_A$ and $\mathbf{1}_B$ be indicator functions on two Borel sets A and B respectively.
- Then, $\mathbf{1}_A \circ X$ and $\mathbf{1}_B \circ Y$ are Bernoulli r.v.'s with success probabilities $P(\{X \in A\})$ and $P(\{Y \in B\})$, respectively.
- In fact, $(\mathbf{1}_A \circ X)(\mathbf{1}_B \circ Y)$ is also a Bernoulli r.v. with success probability $P(\{X \in A\} \cap \{Y \in B\})$.
- Then,

$$\begin{aligned} E[(\mathbf{1}_A \circ X)(\mathbf{1}_B \circ Y)] &= P(\{X \in A\} \cap \{Y \in B\}) \\ &= P(\{X \in A\})\, P(\{Y \in B\}) \\ &= E[\mathbf{1}_A \circ X] E[\mathbf{1}_B \circ Y]. \end{aligned}$$

Independence of random variables

- A basic result in real analysis is that any Borel measurable function is an a.e. limit of a step-function sequence[2].
- (Theorem 8.2 at page 91 in the textbook) For independent random variables X and Y; and for any two Borel functions u and v, the formula $E[(\mathbf{1}_A \circ X)(\mathbf{1}_B \circ Y)] = E[\mathbf{1}_A \circ X]E[\mathbf{1}_B \circ Y]$ can be extended to

$$E[u(X)v(Y)] = E[u(X)]E[v(Y)].$$

by the standard procedure taking the limit of all the step functions via MCT(monotone convergence theorem).

[2] A step function is a finite linear combination of indicator functions of the form: $f(x) = \sum_{i=0}^{n} c_i \mathbf{1}_{A_i}$, where A_i are Borel.

Independence of random variables

- Let $f_{(X,Y)}(x,y)$ be the joint density of (X,Y) and $f_X(x), f_Y(y)$ be its two marginal densities. Then, for $g(x,y) = u(x)v(y)$,

$$E[u(X)v(Y)] = \iint u(x)v(y)f_{(X,Y)}(x,y) \cdot dx \wedge dy$$

- On the other hand,

$$\begin{aligned} E[u(X)]E[v(Y)] &= \left(\int u(x)f_X(x) \cdot dx\right) \cdot \left(\int v(y)f_Y(y) \cdot dy\right) \\ &= \iint u(x)v(y)f_X(x)f_Y(y) \cdot dx \wedge dy \end{aligned}$$

- (page 72 in the textbook) Therefore, X and Y are two continuous independent r.v.'s, if and only if, for any two Borel functions u and v, $E[u(X)v(Y)] = E[u(X)]E[v(Y)]$; and, if and only if

$$f_{(X,Y)}(x,y) = f_X(x)f_Y(y)$$

Independence of random variables (Example 7.11 at page 74 in the textbook)

Assume that you are waiting for two phone calls, from Alice and from Bob. The waiting time T_1 for Alice's call has expectation 10 minutes and the waiting time T_2 for Bob's call has expectation 40 minutes. Assume T_1 and T_2 are independent exponential r.v.'s. What is the probability $P(T_1 < T_2)$ that Alice's call will come first?

- Recall that the exponential density function with parameter $\lambda > 0$ is:

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \in [0, \infty); \\ 0, & \text{if } x < 0. \end{cases}$$

- Since $ET_1 = \frac{1}{\lambda_1} = 10$, $ET_2 = \frac{1}{\lambda_2} = 40$ and the two r.v,'s are independent, we thus know the joint p.d. of (T_1, T_2) is

$$f_{(T_1,T_2)}(t_1, t_2) = f_{T_1}(t_1) f_{T_2}(t_2) = \frac{1}{10} e^{-\frac{t_1}{10}} \frac{1}{40} e^{-\frac{t_2}{40}}, \ (t_1, t_2) > 0.$$

- Therefore,

$$\begin{aligned} P(T_1 < T_2) &= \iint_{0<t_1<t_2<\infty} \frac{1}{400} e^{-\frac{t_1}{10}} e^{-\frac{t_2}{40}} dt_1 \wedge dt_2 \\ &= \int_0^\infty \frac{1}{400} e^{-\frac{t_1}{10}} dt_1 \int_{t_1}^\infty e^{-\frac{t_2}{40}} dt_2 = \int_0^\infty \frac{1}{10} e^{-\frac{5}{40} t_1} dt_1 = 0.8 \end{aligned}$$

Independence of random variables (Example 8.8 at P. 91)

Pick up a random point (X, Y) in the simplex $\Delta = \{(x, y) : x, y \geq 0, x + y \leq 1\}$. The area of Δ is 0.5. So, the joint p.d. of (X, Y) is $f_{(X,Y)}(x, y) = 2, \ \forall (x, y) \in \Delta$.

- We can compute the two marginal p.d.'s as

$$f_X(x) = \int_0^{1-x} 2dy = 2(1-x), \forall x \in [0, 1]; \ f_Y(y) = \int_0^{1-y} 2dx = 2(1-y), \forall y \in [0, 1]$$

  Then, $EX = EY = \int_0^1 2x(1-x)dx = \frac{1}{3}$.
- It is obvious that $f_{(X,Y)}(x, y) = 2 \neq f_X(x) f_Y(y), \ \forall (x, y) \in \Delta$.
- Therefore, X and Y are not independent.
- In fact, when $u(x) = x, \ v(y) = y$,

$$\begin{aligned} E[u(X)v(Y)] &= E[XY] = \iint_\Delta xy \cdot 2dxdy = 2 \int_0^1 xdx \int_0^{1-x} ydy = \frac{1}{12} \\ E[u(X)]E[v(Y)] &= EX \cdot EY = \frac{1}{9} \end{aligned}$$

- That is, $E[XY] \neq EX \cdot EY$.

Independence of random variables (Example 8.8 at P. 92)

Pick up a random point (X, Y) in the unit diamond S. That is, S is the square with corners at $(0,1),(1,0),(0,-1)$ and $(-1,0)$.

- We can compute the two marginal p.d. as

$$f_X(x) = \begin{cases} x+1, & \text{if } x \in [-1,0]; \\ 1-x, & \text{if } x \in [0,1] \end{cases} \quad ; \quad f_Y(y) = \begin{cases} y+1, & \text{if } y \in [-1,0]; \\ 1-y, & \text{if } y \in [0,1] \end{cases}$$

  Then, $EX = \int_{-1}^{0} x(x+1)dx + \int_{0}^{1} x(1-x)dx = 0$. Also, $EY = 0$.
- It is obvious that $f_{(X,Y)}(x,y) = \frac{1}{\text{area(S)}} = \frac{1}{2} \neq f_X(x)f_Y(y), \ \forall (x,y) \in S$.
- Therefore, X and Y are not independent.
- However, when $u(x) = x, \ v(y) = y$,

$$\begin{aligned} E[u(X)v(Y)] &= E[XY] = \iint_S \frac{1}{2} \cdot xy \cdot dxdy \\ &= \frac{1}{2}\int_{-1}^{1} xdx \cdot \int_{-1+|x|}^{1-|x|} ydy = \frac{1}{2}\int_{-1}^{1} xdx \cdot 0 = 0. \end{aligned}$$

- Since X and Y are not independent, there must be some Borel functions $u(x)$ and $v(y)$ such that $E[u(X)v(Y)] \neq E[u(X)]E[v(Y)]$.

Covariance of (X, Y) (P. 94 in the textbook)

Let X and Y be two independent r.v.'s on a probability space $(\Omega, \mathcal{F}, P)$. Define the *covariance* between X and Y as

$$\begin{aligned}
\mathrm{Cov}(X, Y) &= E[(X - \mu_X)(Y - \mu_Y)] \\
&= \iint_{\mathcal{R}(X)\times\mathcal{R}(Y)} (x - \mu_X)(y - \mu_Y) f_{(X,Y)}(x, y) dx \wedge dy \\
&= E[XY] - \mu_X \cdot E[Y] - \mu_Y \cdot E[X] + E[\mu_X \mu_Y] \\
&= E[XY] - \mu_X \mu_Y = E[XY] - EX \cdot EY.
\end{aligned}$$

- The algebraic formula for the covariance has the form of summing all corresponding products (with certain weights), and thus the covariance is a kind of inner product.
- When X and Y are independent, $\mathrm{Cov}(X, Y) = 0$ and thus $X - \mu_X$ and $Y - \mu_Y$ are orthogonal vectors.
- The converse, however, is not true. The example to pick up a random point from a unit diamond has 0 covariance between its x and y coordinates, but the two are not independent.
- When $x \in [-1, 1]$ is picked, the y coordinate is then limited to be selected from $[1 - |x|, 1 + |x|]$.

Covariance of (X, Y) (Example 8.14 at P. 96 in the textbook)

Roll a die 10 times. Let X be the number of 6's rolled and Y be the number of 5's rolled. Compute the $\mathrm{Cov}(X, Y)$.

- Observe that $X = \sum_{i=1}^{10} I_i$, where I_i is the indicator r.v. for the i^{th} roll to be 6.
- Similarly, we write $Y = \sum_{i=1}^{10} J_i$, where J_i is the indicator r.v. for the j^{th} roll to be 5.
- Then, both $X,\ Y \sim \mathrm{Binomial}(10, \frac{1}{6})$ with $EX = EY = np = \frac{10}{6} = \frac{5}{3}$.
- Next, we compute $E[XY] = \sum_{i=1}^{10} \sum_{j=1}^{10} E[I_i J_j]$. Notice that, when $i = j$, $E[I_i J_j] = 0$ since both 5 and 6 cannot be rolled on the same toll. Due to the independence on each roll,

$$E[XY] = \sum_{i \neq j} E[I_i J_j] = \sum_{i \neq j} E[I_i] E[J_j] = (10^2 - 10)\frac{1}{6^2} = \frac{5}{2}.$$

- Therefore, $\mathrm{Cov}(X, Y) = \frac{5}{2} - \left(\frac{5}{3}\right)^2 = -\frac{5}{18}$.

Covariance of (X, Y) (Example 8.14 at P. 96 in the textbook)

Roll a die 10 times. Let X be the number of 6's rolled and Y be the number of 5's rolled. We have $\mu_X = \mu_Y = \frac{5}{3}$; $\mathrm{Cov}(X, Y) = -\frac{5}{18}$.

- Covariance of (X, Y) is the sum of black and (red) figures from each block in the following table. There are $32(+); 34(-)$ and many 0's.
- The negative covariance $\mathrm{Cov}(X, Y) = -\frac{5}{18}$ indicates that, for the pair of values $(X = x, Y = y)$ (which appear in random), "on the average," it is more likely that, when you roll a die 10 times, the number of 6's (x) deviate from mean $\frac{5}{3}$ oppositely from how the number of 5's (y) deviate from its mean.

joint p. m. f. of (X, Y) $(g(i,j) = (x-\mu_X)(y-\mu_Y))$						
$X \backslash Y$	0	1	2	3	...	10
0	$(\frac{4}{6})^{10}$ $(\frac{25}{9})$	$c_1^{10}(\frac{1}{6})(\frac{4}{6})^9$ $(\frac{10}{9})$	$c_2^{10}(\frac{1}{6})^2(\frac{4}{6})^8$ $(-\frac{5}{9})$	$c_3^{10}(\frac{1}{6})^3(\frac{4}{6})^7$ $(-\frac{20}{9})$	...	$(\frac{1}{6})^{10}$ $(-\frac{125}{9})$
1	$c_1^{10}(\frac{1}{6})(\frac{4}{6})^9$ $(\frac{10}{9})$	$c_1^{10}c_1^{9}(\frac{1}{6})^2(\frac{4}{6})^8$ $(\frac{4}{9})$	$c_1^{10}c_2^{9}(\frac{1}{6})^3(\frac{4}{6})^7$ $(-\frac{2}{9})$	$c_1^{10}c_3^{9}(\frac{1}{6})^4(\frac{4}{6})^6$ $(-\frac{8}{9})$	...	0
2	$c_2^{10}(\frac{1}{6})^2(\frac{4}{6})^8$ $(-\frac{5}{9})$	$c_2^{10}c_1^{9}(\frac{1}{6})^3(\frac{4}{6})^7$ $(-\frac{2}{9})$	$c_2^{10}c_2^{9}(\frac{1}{6})^4(\frac{4}{6})^6$ $(\frac{1}{9})$	$c_2^{10}c_3^{9}(\frac{1}{6})^3(\frac{4}{6})^3$ $(\frac{4}{9})$	...	0
3	$c_3^{10}(\frac{1}{6})^3(\frac{4}{6})^7$ $(-\frac{20}{9})$	$c_3^{10}c_1^{9}(\frac{1}{6})^4(\frac{4}{6})^6$ $(-\frac{4}{9})$	$c_3^{10}c_2^{9}(\frac{1}{6})^5(\frac{4}{6})^5$ $(\frac{4}{9})$	$c_3^{10}c_3^{9}(\frac{1}{6})^6(\frac{4}{6})^4$ $(\frac{16}{9})$	...	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	0
10	$(\frac{1}{6})^{10}$ $(-\frac{125}{9})$	0	0	0	0	0

Covariance of (X, Y) (P. 94 in the textbook)

Let $(\Omega, \mathcal{F}, P)$ be a probability space and $A, B \in \mathcal{F}$ be two events with $\mathbf{1}_A$, $\mathbf{1}_B$ be their indicator r.v.'s, respectively.

- Then, the covariance of $(\mathbf{1}_A, \mathbf{1}_B)$ is

$$\begin{aligned} \mathrm{Cov}(\mathbf{1}_A, \mathbf{1}_B) &= E(\mathbf{1}_A \mathbf{1}_B) - E\mathbf{1}_A \cdot E\mathbf{1}_B \\ &= P(A \cap B) - P(A)P(B) \\ &= P(A)[P(B|A) - P(B)] \end{aligned}$$

- If $P(B|A) > P(B)$, knowing that A happened increases the probability for the prediction that B happens, we say the two events are *positively correlated* under the probability measure P. In this case, the covariance of their indicator functions $(\mathbf{1}_A, \mathbf{1}_B)$ is positive.
- If $P(B|A) < P(B)$, knowing that A happened decreases the the probability for prediction that B happens, the two events are *negatively correlated* under the probability measure P, and the covariance of their indicator functions $(\mathbf{1}_A, \mathbf{1}_B)$ is negative.
- When $P(B|A) = P(B)$, the two events are independent under P. In this case, the covariance is 0. (The converse may not be true).

Pearson correlation coefficient (Statistics)

- Covariance of (X, X) happens to be variance of X. That is,

$$\mathrm{Cov}(X, X) = E[(X - \mu_X)(X - \mu_X)] = \mathrm{Var}(X).$$

- Pearson correlation coefficient of two r.v.'s (X, Y) is defined as

$$\begin{aligned} \rho_{X,Y} &= \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Cov}(X, X)}\sqrt{\mathrm{Cov}(Y, Y)}} \\ &= \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} \\ &= \frac{< X - \mu_X, Y - \mu_Y >}{\|X - \mu_X\| \|Y - \mu_Y\|} \end{aligned}$$

- Thus, by Cauchy–Schwarz inequality, $-1 \leq \rho_{X,Y} \leq 1$ and it is the "abstract angle" between $X - \mu_X$ and $Y - \mu_Y$.
- Pearson correlation coefficient is essentially a normalized measurement of the covariance (still a kind of covariance), such that, regardless what X and Y are to be compared with, the value is always between -1 and 1.

Linear Data has Pearson correlation coefficient ± 1

- Assume that a set of 2D data $(X, Y) = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ fall exactly on the line $y = ax + b$. That is,

$$y_i = ax_i + b, \ \forall i \in [1 : n].$$

- It implies that the average $(\bar{x}, \bar{y})$ of the data (X, Y) also fall on the same line. That is,

$$\sum_{i=1}^{n} y_i = a \sum_{i=1}^{n} x_i + nb \ \Rightarrow \ \bar{y} = a\bar{x} + b$$

- The data covariance is somehow *not* the same as the covariance of two random variables. But here we pretend they share the same type of formula.

Linear Data has Pearson correlation coefficient ± 1

- We first form two n vectors:

  $$X-\bar{x} = (x_1-\bar{x}, x_2-\bar{x}, \ldots, x_n-\bar{x}); \text{ and } Y-\bar{y} = (y_1-\bar{y}, y_2-\bar{y}, \ldots, y_n-\bar{y}).$$

- Since the data (X, Y) and their average $(\bar{x}, \bar{y})$ satisfy

  $$y_i = ax_i + b; \text{ and } \bar{y} = a\bar{x} + b,$$

  there is $y_i - \bar{y} = a(x_i - \bar{x}), \ \forall i \in [1 : n]$.

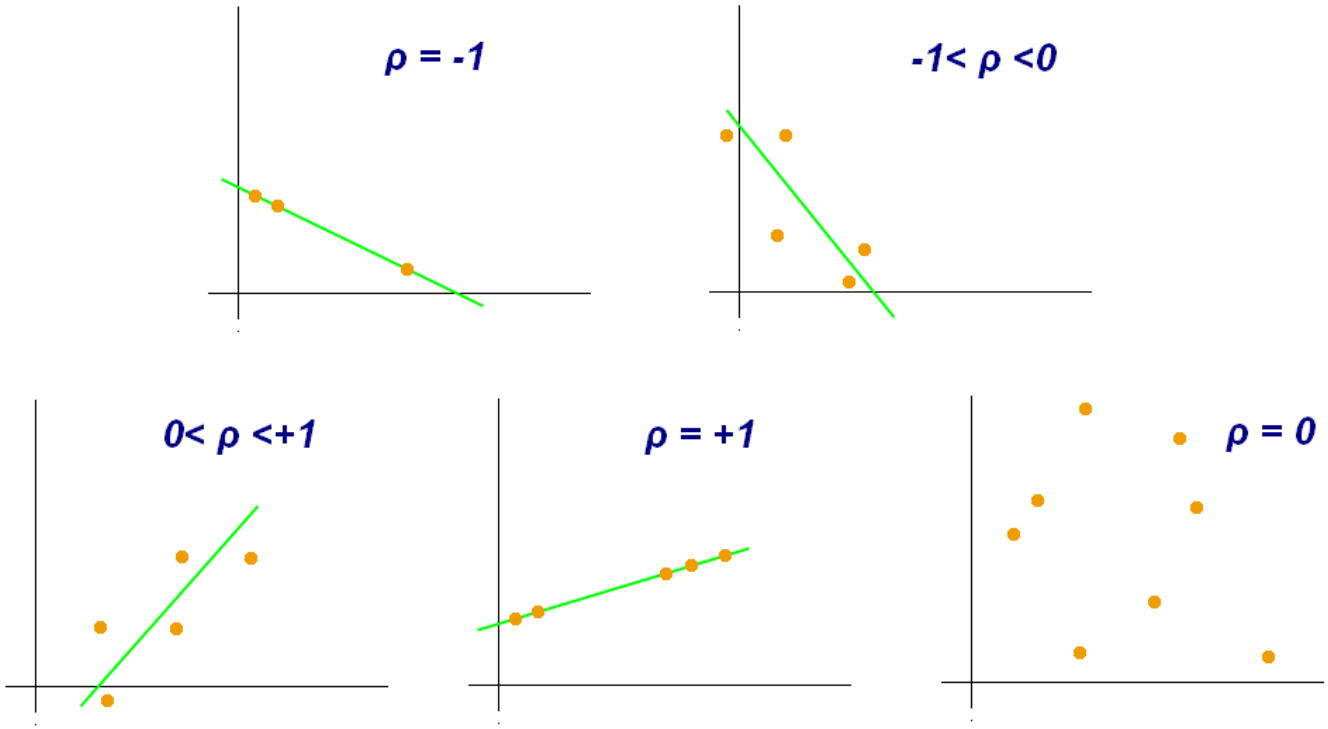
- The two n vectors thus become

  $$X-\bar{x} = (x_1-\bar{x}, x_2-\bar{x}, \ldots, x_n-\bar{x}); \text{ and } Y-\bar{y} = a(x_1-\bar{x}, x_2-\bar{x}, \ldots, x_n-\bar{x}).$$

- The Pearson correlation coefficient is now computed by

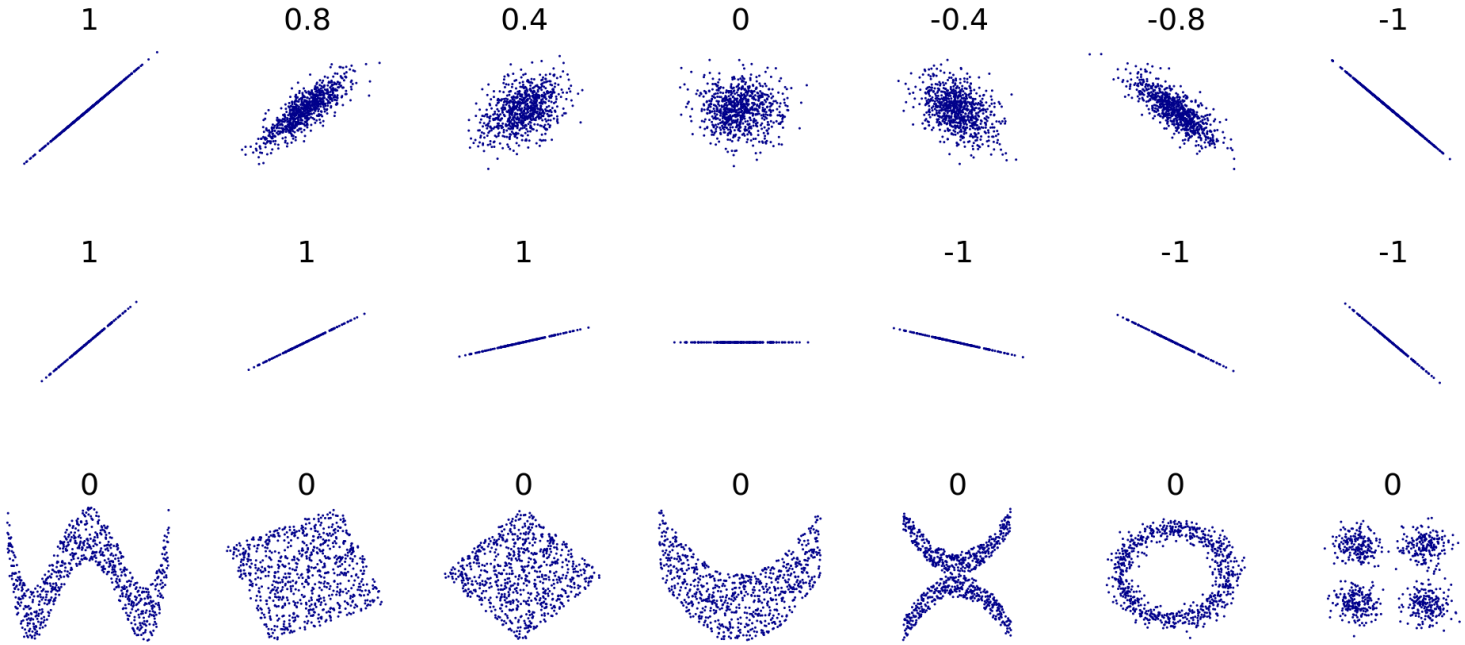
$$\begin{aligned}
\rho_{X,Y} &= \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Cov}(X, X)}\sqrt{\mathrm{Cov}(Y, Y)}} \\
&= \frac{< X - \bar{x}, Y - \bar{y} >}{\|X - \bar{x}\| \, \|Y - \bar{y}\|} \\
&= \frac{a(x_1 - \bar{x})^2 + a(x_2 - \bar{x})^2 + \cdots + a(x_n - \bar{x})^2}{\sqrt{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}\sqrt{a^2(x_1 - \bar{x})^2 + a^2(x_2 - \bar{x})^2 + \cdots + a^2(x_n - \bar{x})^2}} \\
&= \frac{a}{|a|} = \pm 1.
\end{aligned}$$

Examples of scatter diagrams with different values of correlation coefficient (From Wikipedia)



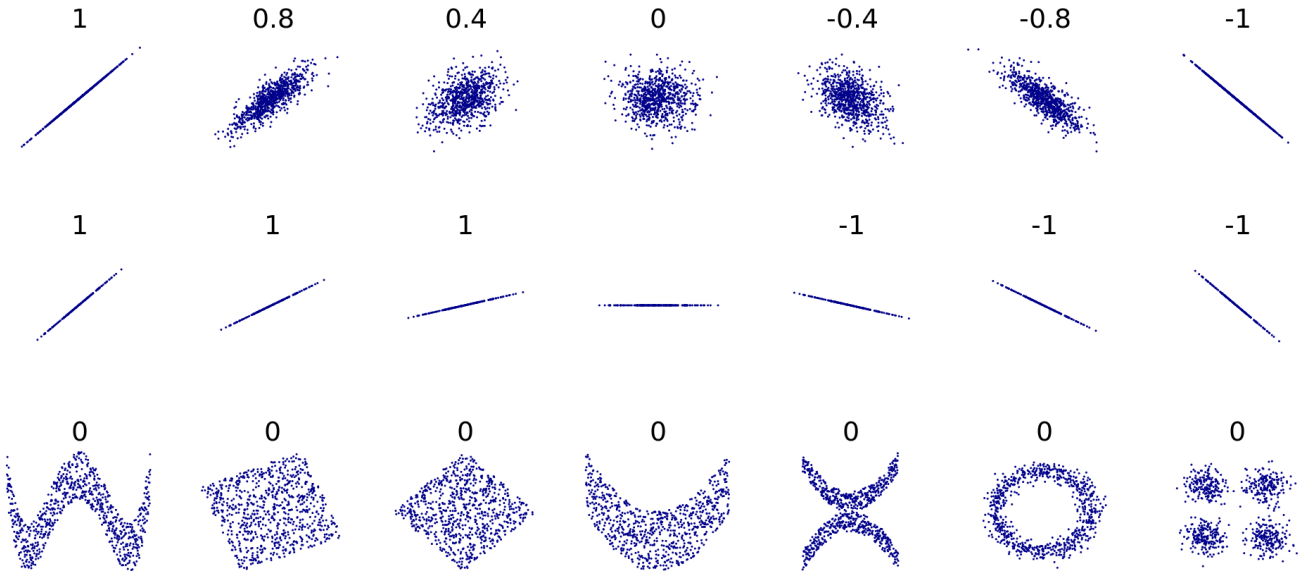
- Pearson correlation coefficient measures linear correlation between two sets of data, and ignores many other types of relationships or correlations.

Examples of scatter diagrams with different values of correlation coefficient (From Wikipedia)



- (Top row) The correlation reflects the strength and direction of a linear relationship.
- The smaller the absolute value of the correlation coefficient (i.e. the closer it is to 0), the linear relationship between the two sets of data is dispersed, from a straight line to an increasingly blurred one.

Examples of scatter diagrams with different values of correlation coefficient (From Wikipedia)



- (Middle row) No matter the slope of the relationship of (X, Y) is, as long as it stays well close to a line, the strength of the linear relation is strong.
- (Bottom row) Many aspects of nonlinear relationships have 0 correlation coefficient.
- (Center one) N.B.: the figure in the center has a slope of 0 but in that case the correlation coefficient is undefined because the variance of Y is zero.

Variance of sums of random variables (page 94 in the textbook)

- Theorem 8.4: Variance-covariance formula:

$$
\begin{aligned}
E\left[\sum_{i=1}^{n} X_i\right]^2 &= \sum_{i=1}^{n} E\left[X_i^2\right] + \sum_{i \neq j} E(X_i X_j) \\
\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) &= E\left[\sum_{i=1}^{n} X_i - E(\sum_{i=1}^{n} X_i)\right]^2 \\
&= E\left[\sum_{i=1}^{n} (X_i - EX_i)\right]^2 \\
&= \sum_{i=1}^{n} \operatorname{Var}(X_i) + \sum_{i \neq j} \operatorname{Cov}(X_i, X_j).
\end{aligned}
$$

- If $X_1, X_2, \ldots, X_n$ are independent, then

$$\operatorname{Var}(X_1 + X_2 + \cdots + X_n) = \operatorname{Var}(X_1) + \operatorname{Var}(X_2) + \cdots + \operatorname{Var}(X_n).$$

- For example, toss a coin independently for n times. Let $S_n = X_1 + X_2 + \cdots + X_n \sim \text{Binomial}(n, p)$. Then, $\operatorname{Var} S_n = n \cdot \operatorname{Var} X_1 = npq$.

Conditional Expectation

Let $(\Omega, \mathcal{F}, P(\cdot|H)))$ be a conditional probability space given that the event $H \in \mathcal{F}$ has occurred with $P(H) > 0$.

- Let Y be a discrete random variable and $b \in \mathcal{R}(Y)$. Then the conditional probability distribution on $\{Y = b\},\ b \in \mathcal{R}(Y)$ under the hypothesis H is

$$\mathcal{L}_{Y,P(\cdot|H))}(\{b\}) = P\{Y = b|H\} = \frac{P(\{Y = b\} \cap H)}{P(H)}.$$

- We then define the conditional expectation of Y given H as

$$E[Y|H] = \sum_b bP\{Y = b|H\}.$$

- In particular, let X be a discrete random variable taking values in a state space $\mathcal{R}(X) \subset \mathbb{R}$ and $a \in \mathcal{R}(X)$. Then, the conditional expectation of Y given $H = \{X = a\}$ is

$$E[Y|X = a] = \sum_b bP\{Y = b|X = a\}.$$

Conditional Expectation

$$E[Y|X=a] = \sum_b bP\{Y=b|X=a\},\ a \in \mathcal{R}(X) \tag{1}$$

- As the number a varies in $\mathcal{R}(X)$, (1) defines a function f on the state space $\mathcal{R}(X)$ of X, whose values are various conditional expectations of Y given different $\{X=a\}$. That is,

$$f(a) = E[Y|X=a]. \tag{2}$$

- By the conditional expectation of a random variable Y given another random variable X, written as $E[Y|X]$, we mean a random variable $E[Y|X](\omega) = f(X(\omega))$ where f is defined by (2)
- When a is fixed, every $\omega \in \Omega$ such that $X(\omega) = a$ shares the same conditional expectation $E[Y|X](\omega) = f(a)$.
- The conditional expectation of Y given X is a random variable on Ω, and also a function on the state space $\mathcal{R}(X)$ of X:

$$E[Y|X] = f(X) : \omega \in \Omega \xrightarrow{X} S = \mathcal{R}(X) \xrightarrow{f} \mathbb{R}.$$

An Example on Conditional Expectation

Consider the three stage binomial asset pricing model by tossing a coin with $P(H) = p$ independently for three times. The sample space is $\Omega_3 = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$

- We first compute the expectation of S_1.

  $E(S_1) = uS_0P(A_H)+dS_0P(A_T) = uS_0 \cdot p(p^2+2pq+q^2)+dS_0 \cdot q = (pu+qd)S_0.$

- Secondly, we would like to compute $E[S_1|S_2]$, the conditional expectation of S_1 given the information of S_2.

- We first notice that $f(S_2) = E[S_1|S_2]$ is a random variable on Ω_3. For each $\omega \in \Omega_3$, we first observe the value $S_2(\omega)$. Then compute the conditional expectation $f(S_2(\omega)) = E[S_1|S_2 = S_2(\omega)]$.

- Although there are 8 sample points in Ω_3, there are only 3 different S_2 values:

$$S_2(\omega) = \begin{cases} u^2S_0, & \text{if } \omega = HHH,\ HHT \\ udS_0, & \text{if } \omega = HTH,\ HTT,\ THH,\ THT \\ d^2S_0, & \text{if } \omega = TTH,\ TTT \end{cases}$$

An Example on Conditional Expectation

- If $\omega = HHH \vee HHT$, then $S_2(\omega) = u^2 S_0$. So,

$$\begin{aligned} E[S_1|S_2](\omega) &= \sum_b bP\{S_1 = b|S_2 = u^2 S_0\} \\ &= (uS_0)P\{S_1 = uS_0|S_2 = u^2 S_0\} + (dS_0)P\{S_1 = dS_0|S_2 = u^2 S_0\} \\ &= (uS_0)P\{S_1 = uS_0|S_2 = u^2 S_0\} \\ &= (uS_0)\frac{P(\{A_H\} \cap \{A_{HH}\})}{P\{A_{HH}\}} \\ &= uS_0. \end{aligned}$$

- Similarly, when $\omega = TTH \vee TTT$, $S_2(\omega) = d^2 S_0$.

$$E[S_1|S_2](TTH) = E[S_1|S_2](TTT) = dS_0.$$

An Example on Conditional Expectation

- Finally, if $\omega \in \{HTH, HTT, THH, THT\}$, then $S_2(\omega) = udS_0$. So,

$$\begin{aligned} E[S_1|S_2](\omega) &= \sum_b bP\{S_1 = b|S_2 = udS_0\} \\ &= (uS_0)P\{S_1 = uS_0|S_2 = udS_0\} + (dS_0)P\{S_1 = dS_0|S_2 = udS_0\} \\ &= (uS_0)\frac{P(\{A_{HT}\})}{P(\{A_{HT}\} \cup \{A_{TH}\})} + (dS_0)\frac{P(\{A_{TH}\})}{P(\{A_{HT}\} \cup \{A_{TH}\})} \\ &= uS_0\frac{pq}{2pq} + dS_0\frac{pq}{2pq} = \frac{1}{2}(u+d)S_0. \end{aligned}$$

- In conclusion, we can write $E[S_1|S_2](\omega) = f(S_2(\omega))$ where

$$f(x) = \begin{cases} uS_0, & \text{if } x = u^2S_0, \\ \frac{1}{2}(u+d)S_0, & \text{if } x = udS_0, \\ dS_0, & \text{if } x = d^2S_0. \end{cases}$$

- It is easy to understand the value of $E[S_1|S_2](\omega)$.
  - (i) If you were told that $S_2 = u^2S_0$, it must be $\omega \in A_{HH}$. You should predict $S_1 = uS_0$ without any doubt.
  - (i) If $S_2 = d^2S_0$ has been observed, it is sure that $S_1 = dS_0$.
  - (iii) If $S_2 = udS_0$, the chance that S_1 be uS_0 is half, while the other half chance of it be dS_0. We predict $S_1 = \frac{1}{2}(u+d)S_0$ by taking the average.

An Example on Conditional Expectation

$$E[S_1|S_2](\omega) = f(S_2(\omega)) = \begin{cases} uS_0, & \text{if } x = u^2 S_0, \\ \frac{1}{2}(u+d)S_0, & \text{if } x = udS_0, \\ dS_0, & \text{if } x = d^2 S_0. \end{cases}$$

- In the sample space Ω_3, there are 8 sample points which take only two S_1 values (uS_0, dS_0). Given the information of S_2, $E[S_1|S_2](\omega)$ predicts S_1 by classifying the 8 sample points into three categories based on the possible values of S_2, and thus $E[S_1|S_2](\omega) = f(S_2(\omega))$ is $\sigma(S_2)$-measurable[3].
- Since the conditional expectation $E[S_1|S_2](\omega)$ is a random variable on Ω_3, and since $E[S_1|S_2]^{-1}(\{uS_0\}) = A_{HH}$, $E[S_1|S_2]^{-1}(\{\frac{1}{2}(u+d)S_0\}) = A_{HT} \cup A_{TH}$, and $E[S_1|S_2]^{-1}(\{uS_0\}) = A_{TT}$, the sigma-algebra generated by $E[S_1|S_2]$ is the same as $\sigma(S_2)$:

$$\sigma(E[S_1|S_2]) = \sigma(S_2) = \{\emptyset, \Omega, A_{HH}, A_{TT}, A_{HT} \cup A_{TH}, A_{HH} \cup A_{TT}, A_{HH}^c, A_{TT}^c\}).$$

[3]Recall that a random variable X is $\mathcal{G}$-measurable if $\sigma(X) \subset \mathcal{G}$

An Example on Conditional Expectation

$$E[S_1|S_2](\omega) = f(S_2(\omega)) = \begin{cases} uS_0, & \text{if } x = u^2 S_0, \\ \frac{1}{2}(u+d)S_0, & \text{if } x = udS_0, \\ dS_0, & \text{if } x = d^2 S_0. \end{cases}$$

- Since the conditional expectation $E[S_1|S_2]$ is a random variable on Ω_3, we can also compute its expectation:

$$\begin{aligned} E(E[S_1|S_2]) &= \sum_{\omega \in \Omega_3} E[S_1|S_2](\omega) P(\omega) \\ &= \sum_{x_k = uS_0, \frac{1}{2}(u+d)S_0, dS_0} x_k \mathcal{L}_{E[S_1|S_2]}(\{x_k\}) \\ &= uS_0 \cdot p^2 + \frac{1}{2}(u+d)S_0 \cdot 2pq + dS_0 q^2 \\ &= S_0(up + dq) \\ &= E(S_1) \end{aligned}$$

- Originally, $E[S_1|S_2](\omega)$ predicts the value of S_1 based on the three possible values of S_2. However, after one more average over the three categories, $E(E[S_1|S_2])$ now predicts S_1 using only the value $S_0(up + dq)$ for all 8 sample points in Ω_3.
- Taking more times of average results in worse prediction.

An Example on Conditional Expectation

$E[S_1|S_2](\omega) = f(S_2(\omega))$ where

$$f(x) = \begin{cases} uS_0, & \text{if } x = u^2 S_0, \\ \frac{1}{2}(u+d)S_0, & \text{if } x = udS_0, \\ dS_0, & \text{if } x = d^2 S_0. \end{cases}$$

- Finally, recall that the conditional expectation $E[S_1|S_2](\omega)$ predicts S_1 on all 8 sample points using the information of S_2.
- In the following, we see that, S_1 and its prediction $E[S_1|S_2](\omega)$ have the same average value on all sets $A \in \sigma(S_2)$, that is,

$$\int_A E[S_1|S_2]dP = \int_A S_1 dP.$$

- Take $A_{HH}^c \in \sigma(S_2) = \{\emptyset, \Omega, A_{HH}, A_{TT}, A_{HT} \cup A_{TH}, A_{HH} \cup A_{TT}, A_{HH}^c, A_{TT}^c\}$.

$$\begin{aligned} \int_{A_{HH}^c} E[S_1|S_2]dP &= \int_{A_{TT}} E[S_1|S_2]dP + \int_{A_{HT} \cup A_{TH}} E[S_1|S_2]dP \\ &= dS_0 \cdot q^2 + \frac{1}{2}(u+d)S_0 \cdot 2pq = qdS_0 + uS_0pq. \end{aligned}$$

$$\int_{A_{HH}^c} S_1 dP = \int_{A_T} S_1 dP + \int_{A_{HT}} S_1 dP = qdS_0 + uS_0pq.$$