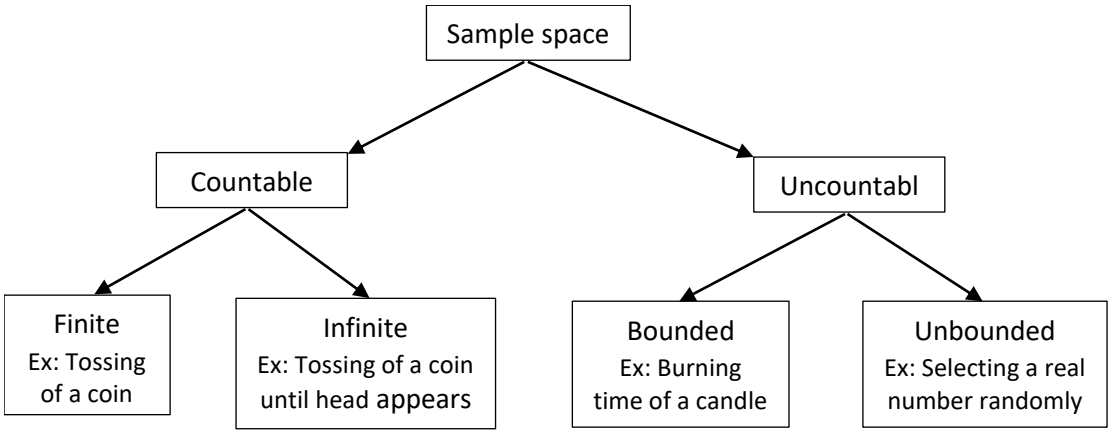
Module 0

Review of probability

Topics: Events and probability, Random variables, Random vectors, Stochastic process

Events and probability

Random experiment, sample space and events: An experiment (in a broad sense) whose outcome cannot be predicted with certainty is called a random experiment. The set of all outcomes of a random experiment constitutes its sample space. Depending on countability and finiteness/boundedness, sample spaces can be classified as shown below. Depending on the type of sample space, the way we measure probability changes.



Events are subsets of sample space, i.e., a collection of outcomes constitutes an event. We say that an event occurs whenever an outcome in the event occurs. We define probability for events (not for outcomes). Events are sets; thus, set operations (like complementation, union, intersection, difference), their properties (like associative and distributive properties) and other relations (like De Morgan's law) are applicable for events.

Two events are called disjoint or mutually exclusive if they have no common outcome. A group of events are called disjoint if all pairs of events from the group are disjoint. We need not go beyond pair because if A and B are disjoint, then there is no common element in a group of events that contains A and B. A group of events are collectively exhaustive if their union is the sample space. A partition of the sample space is a collection of mutually exclusive and collectively exhaustive events.

Axioms of probability: Over time different ways of measuring probability were developed, starting with the classical approach, followed by the frequentist's approach, and then the subjective approach (with updating). We needed a common framework so that a coherent theory of probability can be developed. Kolmogorov formulated this framework.

Probability is a real-valued function defined on the set of events satisfying three axioms:

(i) $P(A) \geq 0$ for all events A

(ii) $P(\Omega) = 1$ (Ω denotes the sample space)
(iii) $P(A \cup B) = P(A) + P(B)$ whenever A and B are disjoint events

Observe that the third axiom implies $P(E_1 \cup E_2 \cup \cdots \cup E_n) = P(E_1) + P(E_2) + \cdots + P(E_n)$ whenever $E_1, E_2, \ldots, E_n$ are disjoint for all $n \geq 2$. Note that the above axiomatic definition of probability does not tell us the value of probability of an event. It merely imposes some reasonable restrictions on how probability should be measured. The whole of probability theory rests on these axioms. If a probability measure satisfies these axioms, then all results of probability theory are applicable to it.

Some immediate implications of probability axioms are:
(i) $P(A^c) = 1 - P(A)$ for all events A
(ii) $P(A) \leq P(B)$ whenever $A \subset B$
(iii) $P(\bigcup_{i=1}^{n} E_i) \leq \sum_{i=1}^{n} P(E_i)$ for every collection $E_1, E_2, \ldots, E_n$
(iv) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(AB) - P(BC) - P(AC) + P(ABC)$
In general, $P(\bigcup_{i=1}^{n} E_i) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \leq i_1 < i_2 < \cdots < i_k \leq n} P\left(E_{i_1} E_{i_2} \cdots E_{i_k}\right)$

Probability measures: Classical measure of probability was the first to be developed. It is applicable when all outcomes of the random experiment are equally likely. Here, probability of an event is measured as the ratio of number of outcomes in the event to the total number of outcomes, i.e., $P(A) = |A|/|\Omega|$. For uncountable sample spaces, it's the ratio of 'size' of the event to the 'size' of the sample space. If $\Omega \subset \mathbb{R}$, then length is the size, if $\Omega \subset \mathbb{R}^2$, then area is the size, and so on. It's obvious that the classical measure satisfies 1st and 2nd axioms. Since $|A \cup B| = |A| + |B|$ for disjoint events A and B, third axiom is also satisfied.

Classical measure of probability is not applicable when all outcomes of a random experiment are not equally likely. Outcomes of an infinite/unbounded sample space are never equally likely. It can happen for finite/bounded case as well. Frequentist's measure of probability overcomes these shortcomings of the classical measure. Here, probability of an event is measured as the proportion of times an event occurs in an infinite repetition of the random experiment, i.e., $P(A) = \lim_{n \to \infty} n_A/n$ where n_A denotes the number of times A occurs out of n repetitions. Again, it's obvious that the frequentist's measure satisfies 1st and 2nd axioms. Since $n_{A \cup B} = n_A + n_B$ for disjoint events A and B, 3rd axiom is also satisfied.

While frequentist's measure is applicable for all types of random experiments, there were two concerns. First, how can we be sure that $\lim_{n \to \infty} n_A/n$ will converge? The law of large numbers settled this doubt. The second concern is about implementation of the formula in some settings, e.g., probability of failure of a new rocket. It would require many trial runs of the rocket, which is too expensive to implement. In these situations, expert opinion (which is an amalgamation of knowledge, experience, simulation, etc.) is sought, and the resultant probability measure is called subjective probability. As we gather more data, experts update their opinion and probability measures are refined.

Counting techniques: In the classical measure for countable sample spaces, we need to count the number of outcomes in an event. We mention some counting techniques here.

(i) If we can perform X in m different ways, Y in n different ways, and X and Y do not influence one another, then we can perform X and Y in mn different ways.

(ii) Number of permutations of r objects from a set of n distinct objects without repetition is: $n!/(n-r)!$. Permutation refers to selection and arrangement.

(iii) Number of permutations of r objects from a set of n distinct objects with unlimited repetition is: n^r.

(iv) Number of combinations of r objects from a set of n distinct objects without repetition is: $n!/r!\,(n-r)!$. Combination refers to selection alone.

(v) Number of combinations of r objects from a set of n distinct objects with unlimited repetition is: $(n-1+r)!/r!\,(n-1)!$.

Conditional probability: Consider tossing of two fair coins. $\Omega = \{(HH), (HT), (TH), (TT)\}$ and classical measure applies to Ω. We are interested in the event of at least one head, i.e., in the event $A = \{(HH), (HT), (TH)\}$. Clearly, $P(A) = 3/4$. Now, if we have the information that the coins show the same face, then the outcomes (HT) and (TH) are impossible, and this changes $P(A)$. Event $B = \{(HH), (TT)\}$ captures the given fact that the coins show the same face. The new sample space is B (not Ω), and we need to define a probability measure on B in order to calculate probability of A given that B has happened, denoted by $P(A|B)$.

Instead of starting from scratch, we can use the probability measure defined on Ω to construct a probability measure on B as: $P(A|B) \coloneqq P(A \cap B)/P(B)$. Note that $A \cap B$ is the part (of A) through which A can occur given that B has occurred. This probability measure on B is reasonable and satisfies the first two axioms. For disjoint events A_1 and A_2, $P(A_1 \cup A_2|B) = P\big((A_1 \cup A_2) \cap B\big)/P(B) = P\big((A_1 \cap B) \cup (A_2 \cap B)\big)/P(B)$. Since $A_1 \cap B$ and $A_2 \cap B$ are disjoint and the probability measure on Ω satisfies all three axioms, $P(A_1 \cup A_2|B) = \big(P(A_1 \cap B) + P(A_2 \cap B)\big)/P(B) = P(A_1 \cap B)/P(B) + P(A_2 \cap B)/P(B) = P(A_1|B) + P(A_2|B)$. So, the third axiom is satisfied by the probability measure on B. With this measure, $P(A|B) = P(A \cap B)/P(B) = P(HH)/P(B) = 0.25/0.5 = 1/2$.

Two useful formulas: $P(A|B) = P(A \cap B)/P(B) \Rightarrow P(A \cap B) = P(A|B)P(B)$. Similarly, $P(ABC) = P(A|BC)P(B|C)P(C)$. In general, $P(E_nE_{n-1}\cdots E_1) = P(E_n|E_{n-1}E_{n-2}\cdots E_1) \cdot P(E_{n-1}|E_{n-2}E_{n-3}\cdots E_1)\cdots P(E_2|E_1)\cdot P(E_1)$. This formula is known as the chain rule of probability. It can be useful in calculating probabilities of certain events. For example, let us calculate the probability of n people having different birthdays. Let us define $E_1, E_2, \ldots, E_n$ as: E_k is the event that k-th person has a different birthday from the first $k-1$ persons for $k = 1,2,\ldots,n$. Then we are interested in finding $P(E_nE_{n-1}\cdots E_1) = P(E_n|E_{n-1}E_{n-2}\cdots E_1) \cdot P(E_{n-1}|E_{n-2}E_{n-3}\cdots E_1)\cdots P(E_2|E_1)\cdot P(E_1) = \left(1-\frac{n-1}{365}\right)\cdot\left(1-\frac{n-2}{365}\right)\cdots\left(1-\frac{1}{365}\right)\cdot 1$.

The second formula is known as the law of total probability. Consider students in IME625 can be classified as sincere as well as smart (with respect to the course), sincere but unsmart, insincere but smart, and insincere as well as unsmart. Let B_1, B_2, B_3, B_4 denote these types. Past experiences suggest that 40% students are of type B_1, 30% are of type B_2, 20% are of type B_3, and the remaining 10% are of type B_4. Let A denote the event that a student correctly answers a question asked during class. Again, past experiences tell that $P(A|B_1) = 0.9$, $P(A|B_2) = 0.6$, $P(A|B_3) = 0.5$, and $P(A|B_4) = 0.2$. I am interested in the probability that a randomly selected student answers a question correctly, i.e., $P(A)$. Observe that B_1, B_2, B_3, B_4 is a partition. Then $A = A \cap \Omega = A \cap (\cup_i B_i) = \cup_i (A \cap B_i)$. Since $A \cap B_i$ for different i are disjoint, $P(A) = \sum_i P(A \cap B_i) = \sum_i P(A|B_i)P(B_i)$. This formula is known as the law of total probability. Here, $P(A) = \sum_{i=1}^{4} P(A|B_i)P(B_i) = 0.9 \times 0.4 + 0.6 \times 0.3 + 0.5 \times 0.2 + 0.2 \times 0.1 = 0.66$.

Bayes theorem: In the context of the previous paragraph, consider that a randomly selected student answers the question correctly, i.e., event A occurs. Given this, I am interested in finding the probability that the student is sincere and smart, i.e., $P(B_1|A)$. By conditional probability, $P(B_i|A) = P(B_i \cap A)/P(A) = P(A|B_i)P(B_i)/\sum_i P(A|B_i)P(B_i)$. This formula is known as the Bayes theorem. Here, $P(B_1|A) = P(A|B_1)P(B_1)/\sum_{i=1}^{4} P(A|B_i)P(B_i) = 0.9 \times 0.4/0.66 = 0.545$. A partition may consist of B and B^c. Then Bayes theorem takes the form: $P(B|A) = P(A|B)P(B)/\{P(A|B)P(B) + P(A|B^c)P(B^c)\}$.

Before any question were asked, my belief about a student being sincere and smart is simply $P(B_1) = 0.4$. With a question correctly answered, my belief changes to $P(B_1|A) = 0.545$. If the question were answered incorrectly, then belief would have changed to $P(B_1|A^c) = P(A^c|B_1)P(B_1)/\sum_{i=1}^{4} P(A^c|B_i)P(B_i) = 0.1 \times 0.4/(0.1 \times 0.4 + 0.4 \times 0.3 + 0.5 \times 0.2 + 0.8 \times 0.1) = 0.04/0.34 = 0.118$. With another question answered correctly/incorrectly, my belief would change further. This process is known as Bayesian updating.

Independence of events: Two events A and B are said to be independent if occurrence/non-occurrence of one does not influence occurrence/non-occurrence of the other, i.e. $P(A|B) = P(A)$, $P(A^c|B) = P(A^c)$, and six more conditions, all of which are equivalent to $P(A \cap B) = P(A) \cdot P(B)$. Three events A, B, C are said to be independent if occurrence/non-occurrence of some of these events does not influence occurrence/non-occurrence of some other of these events, which is equivalent to four conditions: $P(AB) = P(A)P(B)$, $P(BC) = P(B)P(C)$, $P(AC) = P(A)P(C)$, $P(ABC) = P(A)P(B)P(C)$. Independent of $E_1, E_2, \ldots, E_n$ is defined in a similar manner and it requires $\binom{n}{2} + \binom{n}{3} + \cdots + \binom{n}{n}$ conditions to be checked. It shall be noted that disjoint events cannot be independent and vice versa.

Random variables

The idea of a random variable: Consider tossing of three fair coins. Here, $\Omega = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$ and the classical measure applies. This much is enough to

answer any question of interest. However, if our interest is only about the number of heads, then it makes sense to consider $\Omega' = \{0,1,2,3\}$ as the sample space. Classical measure does not apply to Ω', but the classical measure of Ω can be used to define a probability measure on Ω' as follows: $P_{\Omega'}(\{0\}) = P_{\Omega}(\{TTT\}) = 1/8$, $P_{\Omega'}(\{1\}) = P_{\Omega}(\{HTT, THT, TTH\}) = 3/8$, $P_{\Omega'}(\{2\}) = P_{\Omega}(\{HHT, HTH, THH\}) = 3/8$, $P_{\Omega'}(\{3\}) = P_{\Omega}(\{HHH\}) = 1/8$.

In the above illustration, we transformed the original sample space into a new one and used the original probability measure to construct a new one. If the outcomes are transformed into numbers, then we can use many tools including calculus to better understand the random experiment. In the above example, mean and variance have no meaning in Ω, but those are meaningful in $\Omega' \subset \mathbb{R}$. Random variables perform this transformation. Random variable is a real-valued function defined on the sample space, i.e., $X: \Omega \to \mathbb{R}$. Let $X(\Omega) \subseteq \mathbb{R}$ denote the new sample space. Then the new probability measure $P_{X(\Omega)}(A) \coloneqq P_{\Omega}(\{\omega \in \Omega: X(\omega) \in A\})$ for $A \subseteq X(\Omega)$. One can verify consistency of this measure with probability axioms.

Distribution function: Distribution function of a random variable captures probabilities of events of the form $\{X \leq x\}$ for all $x \in \mathbb{R}$. It is defined as: $F: \mathbb{R} \to [0,1]$ such that $F(x) = P_{X(\Omega)}(\{X \leq x\}) = P_{\Omega}(\{\omega \in \Omega: X(\omega) \leq x\})$. It satisfies the following properties:

(i) $F(-\infty) = 0$ and $F(\infty) = 1$
(ii) $F(x)$ is non-decreasing in $x \in \mathbb{R}$
(iii) $F(x)$ is right-continuous in $x \in \mathbb{R}$

The first two properties can be proved easily. The third one is difficult to prove, but one can verify it by plotting distribution functions familiar random variables. We can use distribution function to obtain probabilities of different types of events. For example,

(i) $P(\{X = x\}) = P(\{X \leq x\}) - P(\{X < x\}) = F(x) - F(x^-)$
(ii) $P(\{a \leq X \leq b\}) = P(\{X \leq b\}) - P(\{X < a\}) = F(b) - F(a^-)$
(iii) $P(\{a \leq X < b\}) = P(\{X < b\}) - P(\{X < a\}) = F(b^-) - F(a^-)$
(iv) $P(\{a < X \leq b\}) = P(\{X \leq b\}) - P(\{X \leq a\}) = F(b) - F(a)$
(v) $P(\{a < X < b\}) = P(\{X < b\}) - P(\{X \leq a\}) = F(b^-) - F(a)$

Note: $F(x^-)$ is the left-limit of $F(\cdot)$, which exists irrespective of $F(\cdot)$ is left-continuous or not. If it's left-continuous, then $F(x^-) = F(x)$, else $F(x^-) < F(x)$. Due to right-continuity, $F(x^+) = F(x)$.

Any event about a random variable, i.e., any $A \subseteq X(\Omega) \subseteq \mathbb{R}$ can be represented as the union of some disjoint intervals and singletons. Then probability of the event is sum of probabilities of the constituent intervals and singletons, which we know how to calculate. So, distribution function is enough to calculate probability of any event about a random variable. It's like the signature of a random variable.

Mass and density functions: For all random variables, distribution function $F(x)$ increases from 0 to 1 as x increases from $-\infty$ to ∞. If this increase happens in a continuous manner, i.e., if $F(\cdot)$ is continuous (both left and right continuous), then the random variable is called a continuous random variable. If the increase happens only through jumps, i.e., if $F(\cdot)$ is a step

function, then the random variable is called a discrete random variable. If the increase is a result of both continuous and jump increase, then the random variable is called a mixed random variable. You are familiar with continuous and discrete random variables. Example of a mixed random variable is exam completion time of students. For a 3-hour exam, if the time limit is removed, assume that a student completes the exam anytime between 2 and 4 hours with all outcomes equally likely. Let τ denote this time. Now with the time limit, the exam completion time $T = \min(\tau, 3)$. If you plot distribution function of T, you will see a linear increase of $F(t)$ from 0 to 0.5 as t increases from 2 to 3^-, and then there is a jump of 0.5 in $F(t)$ as t goes from 3^- to 3.

For discrete random variables, we define mass function as follows. $p: X(\Omega) \to [0,1]$ such that $p(x) = P_{X(\Omega)}(\{X = x\}) = P_{\Omega}(\{\omega \in \Omega: X(\omega) = x\})$ for $x \in X(\Omega)$, the set of all values that X can take. Observe that $p(x) = F(x) - F(x^-)$ and $F(a) = \sum_{x \le a, x \in X(\Omega)} p(x)$. If we encounter a discrete random variable, we can either obtain its distribution function or its mass function, and the other one can be calculated straightforwardly.

For continuous random variables, we define density function as follows. $f: \mathbb{R} \to [0, \infty)$ is a density function of X if $F(a) = \int_{-\infty}^{a} f(x)dx$ for all $a \in \mathbb{R}$. Unlike mass function, density function does not measure probability and its definition is indirect. If $F'(\cdot)$ exists, then it's a density function, because (i) $F'(\cdot)$ is defined on $\mathbb{R}$, (ii) $F'(\cdot)$ is non-negative as $F(\cdot)$ is non-decreasing, and (iii) $F(a) = \int_{-\infty}^{a} F'(x)dx$ for all $a \in \mathbb{R}$ by the fundamental theorem of calculus. Unlike mass function for discrete random variables, density function may not exist for certain continuous random variables, but this does not happen commonly. If we encounter a continuous random variable, we obtain its distribution function first, and then get its density function. Only in rare cases, we go for density function first.

Common distributions: We encounter some probability distributions more often than others. Three of the most common discrete random variables are binomial, geometric, and Poisson random variables. The first two are associated with Bernoulli trials, where outcomes of a random experiment are classified as success and failure. A binomial random variable with parameters n and p is the number of successes in n independent repetitions of a Bernoulli trial with success probability p. Its mass function is: $p(k) = \binom{n}{k} p^k (1-p)^{n-k}$ for $k = 0,1, \dots, n$. A geometric random variable with parameter p is the number of independent repetitions of a Bernoulli trial with success probability p required to get the first success. Its mass function is: $p(k) = (1-p)^{k-1} p$ for $k = 1,2,3, \dots$.

Three of the most common continuous random variables are uniform, normal, and exponential random variables. A uniform random variable is one that is equally likely to take all values in a bounded interval. Following the classical measure, distribution function of a uniform random variable with lower limit a and upper limits b can be obtained as: $F(x) = 0$ for $x < a$, $F(x) = (x-a)/(b-a)$ for $a \le x \le b$, and $F(x) = 1$ for $x > b$. Differentiating $F(\cdot)$, we get the density function as: $f(x) = 1/(b-a)$ for $a \le x \le b$ and 0 otherwise. A

normal random variable is encountered when average of 'many' repetitions of a random variable (both discrete and continuous) is taken. A normal random variable with mean μ and variance σ^2 has density function: $f(x) = \left(1/\sqrt{2\pi}\,\sigma\right) \cdot \exp(-(x-\mu)^2/2\sigma^2)$ for all $x \in \mathbb{R}$. Its distribution function does not have a closed form expression. Normal random variable with $\mu = 0$ and $\sigma = 1$ is called the standard normal random variable. Its density function is $\phi(z) = \left(1/\sqrt{2\pi}\right) \cdot \exp(-z^2/2)$ for $z \in \mathbb{R}$ and distribution function $\Phi(z) = \int_{-\infty}^{z} \phi(t)\,dt$ for $z \in \mathbb{R}$ is calculated numerically. If $X \sim N(\mu, \sigma^2)$, then $(X-\mu)/\sigma$ follows standard normal distribution. Then $F(x) = P(X \le x) = P((X-\mu)/\sigma \le (x-\mu)/\sigma) = \Phi((x-\mu)/\sigma)$. This saves numerical calculations for individual normal random variables.

Poisson and exponential random variables are related to one another and are encountered in naturally occurring phenomena. For example, number of earthquakes in a certain duration is a Poisson random variable and the time between two successive earthquakes is an exponential random variable. A Poisson random variable with intensity λ per unit time and duration t has mass function: $p(k) = e^{-\lambda t}\,(\lambda t)^k/k!$ for $k = 0,1,2,\ldots$. The associated exponential random variable has distribution function $F(x) = 1 - e^{-\lambda x}$ for $x > 0$ and 0 otherwise and density function $f(x) = \lambda e^{-\lambda x}$ for $x > 0$ and 0 otherwise.

Function of a random variables: Sometimes, we are interested in a real-valued function of a random variable. Then the function itself can be viewed as a random variable. Let $X: \Omega \to \mathbb{R}$ and $Y: X(\Omega) \to \mathbb{R}$. We know how to obtain distribution function $F_X(\cdot)$ of X from $P_\Omega(\cdot)$. In a similar manner, $F_Y(y) = P_{Y(X(\Omega))}(\{Y \le y\}) = P_{X(\Omega)}(\{x \in X(\Omega): Y(x) \le y\})$, which can be calculated using $F_X(\cdot)$, as explained earlier. Depending on whether Y is discrete/continuous, its mass/density function can be obtained.

Evaluation of $F_Y(\cdot)$ becomes simpler when $Y(\cdot)$ is a strictly monotonic function, because such functions are invertible. Let $Y(\cdot)$ be a strictly increasing function, e.g., $X \ge 0$ and $Y = X^2$. Then $Y \ge 0$ and $F_Y(y) = P_{X(\Omega)}(\{x \in X(\Omega): x^2 \le y\}) = P_{X(\Omega)}\left(\{x \in X(\Omega): x \le \sqrt{y}\}\right) = P_{X(\Omega)}\left(\{X \le \sqrt{y}\}\right) = F_X\left(\sqrt{y}\right)$ for all $y \ge 0$. In general, for a strictly increasing $Y(\cdot)$, $F_Y(y) = P_{X(\Omega)}(\{x \in X(\Omega): Y(x) \le y\}) = P_{X(\Omega)}\left(X \le Y^{-1}(y)\right) = F_X\left(Y^{-1}(y)\right)$. On the other hand, if $Y(\cdot)$ is a strictly decreasing function, e.g., $X > 0$ and $Y = 1/X$, then $Y > 0$ and $F_Y(y) = P_{X(\Omega)}(\{x \in X(\Omega): 1/x \le y\}) = P_{X(\Omega)}(\{x \in X(\Omega): x \ge 1/y\}) = 1 - P_{X(\Omega)}(X < 1/y) = 1 - F_X((1/y)^-)$ for all $y > 0$. In general, for a strictly decreasing $Y(\cdot)$, $F_Y(y) = P_{X(\Omega)}(\{x \in X(\Omega): Y(x) \le y\}) = P_{X(\Omega)}(\{x \in X(\Omega): x \ge Y^{-1}(y)\}) = 1 - P_{X(\Omega)}\left(X < Y^{-1}(y)\right) = 1 - F_X\left(\left(Y^{-1}(y)\right)^-\right)$. If X is continuous, then $F_Y(y) = 1 - F_X\left(Y^{-1}(y)\right)$.

Earlier, we mentioned that if $X \sim N(\mu, \sigma^2)$, then $Y = (X-\mu)/\sigma$ follows standard normal distribution. Now it can be verified. Here, Y is a strictly increasing function. Then $F_Y(y) = F_X\left(Y^{-1}(y)\right) = F_X(\mu + \sigma y) = \int_{-\infty}^{\mu+\sigma y} f_X(x)dx$ for all $y \in \mathbb{R}$, where $f_X(x) = \left(1/\sqrt{2\pi}\,\sigma\right) \cdot \exp(-(x-\mu)^2/2\sigma^2)$. By Leibniz's rule for differentiation under the integral sign, $f_Y(y) =$

$\frac{dF_Y(y)}{dy} = f_X(\mu+\sigma y)\cdot\frac{d(\mu+\sigma y)}{dy} - f_X(-\infty)\cdot\frac{d(-\infty)}{dy} + \int_{-\infty}^{\mu+\sigma y}\frac{df_X(x)}{dy}dx = \sigma f_X(\mu+\sigma y) - 0 + 0 = \left(1/\sqrt{2\pi}\right)\cdot\exp(-y^2/2)$ for all $y \in \mathbb{R}$, as required.

Mean, variance and quantiles: While distribution function 'carries' all information about a random variable, it's not easily comprehensible. Some of its features, most notably the mean, variance, and quantiles are easily understood. Mean or expected value of a random variable measures its average value when the underlying random experiment is repeated infinitely, i.e., $E[X] = \lim_{n\to\infty}\sum_{i=1}^{n} x_i/n$, where x_i is the observed value of X in the i-th repetition. It can be shown that $E[X] = \sum_{x\in X(\Omega)} x\cdot p(x)$ for discrete random variables and $\int_{-\infty}^{\infty} xf(x)dx$ for continuous random variables, provided that the sum/integral exists 'absolutely' (and if it does not exist, then the mean is undefined). Variance of a random variable measures its average deviation around its mean in an infinite repetition of the random experiment, i.e., $Var(X) = \lim_{n\to\infty}\sum_{i=1}^{n}(x_i - E[X])^2/n$. Here deviation is measured as the squared distance. Let $Y = (X - E[X])^2$. Then $E[Y] = \lim_{n\to\infty}\sum_{i=1}^{n}(x_i - E[X])^2/n$, as $E[X]$ is a constant. Then $Var(Y) = E[(X - E[X])^2] = E\left[X^2 - 2E[X]X + E^2[X]\right] = E[X^2] - E^2[X]$. Mean and variance of common random variables are listed below.

Dist.	Mean	Variance	Moment generating function	Median
$Bin(n,p)$	np	$npq; q = 1-p$	$(q+pe^s)^n$ for all $s \in \mathbb{R}$	$\lfloor np \rfloor$ or $\lceil np \rceil$
$Geo(p)$	$1/p$	q/p^2	$pe^s/(1-qe^s)$ for all $s < -\ln q$	$\lceil -1/\log_2 q \rceil$
$U(a,b)$	$(a+b)/2$	$(b-a)^2/12$	$\left(e^{sb}-e^{sa}\right)/s(b-a)$ for $s \neq 0$, else 1	$(a+b)/2$
$N(\mu,\sigma^2)$	μ	σ^2	$e^{(\mu s+\sigma^2 s^2/2)}$ for all $s \in \mathbb{R}$	μ
$Pois(\lambda,t)$	λt	λt	$e^{\lambda(e^s-1)}$ for all $s \in \mathbb{R}$	---
$Exp(\lambda)$	$1/\lambda$	$1/\lambda^2$	$\lambda/(\lambda-s)$ for all $s < \lambda$	$\ln 2/\lambda$

Note: Information on moment generating function and median is more of a reference.

More features of a distribution can be captured by taking higher order moments, i.e., $E[X^k]$ for $k = 1,2,3,\ldots$. Note that mean is the first moment and variance is second relative moment about the mean. $E[X^k] = \sum_{x\in X(\Omega)} x^k\cdot p(x)$ for discrete random variables and $\int_{-\infty}^{\infty} x^k f(x)dx$ for continuous random variables, provided that the sum/integral exists 'absolutely'. We can use moment generating function to obtain all the moments. Moment generating function of random variable X with respect to parameter s is defined as: $m_X(s) = E[e^{sX}]$, where s is suitably chosen so that the expectation exists. $m_X(s) = E[1 + sX + (sX)^2/2 + (sX)^3/3! + (sX)^4/4! + \cdots] = 1 + sE[X] + s^2E[X^2]/2 + s^3E[X^3]/3! + \cdots$. By repeated differentiation with respect to s and then setting $s = 0$, we get $E[X] = m_X'(0)$, $E[X^2] = m_X''(0)$, $E[X^3] = m_X'''(0)$, and so on. Like distribution function, moment generating function is unique for a random variable. It too carries all information about a random variable.

Quantiles are thresholds that put certain fraction of observations to the left of the threshold and the remaining to its right in an infinite repetition of the underlying random experiment. More precisely, q-th quantile for $0 < q < 1$ (generally represented in percentage), $x_q :=$

$\min\{x: \lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{1}(x_i \le x)/n \ge q\}$, where $\mathbb{1}(\cdot)$ is the indicator function. It can be shown that $x_q = \min\{x: F(x) \ge q\}$. Most important of the quantiles is the median. Quantiles are used in simulating a random variable using uniform random number generators.

Earlier, we have discussed how to obtain distribution function of a real-valued function of a random variable. From the distribution function, we can obtain mass/density function, and then calculate mean, variance, etc. However, there is a much simpler way of calculating these quantities that does not require the knowledge of mass/density function of the new random variable. Let $Y: X(\Omega) \to \mathbb{R}$ and we are interested in obtaining $E[Y]$. It can be proved that $E[Y] = \sum_{x\in X(\Omega)} Y(x) \cdot p(x)$ if X is discrete and $\int_{-\infty}^{\infty} Y(x)f(x)dx$ if X is continuous. This result is known as the law of the unconscious statistician. Following the same principle, $Var(Y) = E[Y^2] - E^2[Y] = \sum_{x\in X(\Omega)}\{Y(x)\}^2 \cdot p(x) - \left\{\sum_{x\in X(\Omega)} Y(x) \cdot p(x)\right\}^2$ for discrete X and $\int_{-\infty}^{\infty}\{Y(x)\}^2 f(x)dx - \left\{\int_{-\infty}^{\infty} Y(x)f(x)dx\right\}^2$ for continuous X. The same logic leads to the following useful formulas: $E[aX + b] = aE[X] + b$ and $Var(aX + b) = a^2 Var(X)$, where a, b are constants. We already have used these formulas in this section.

Random Vectors

The idea of a random vector: Consider rolling of two fair dice. $\Omega = \{1,2,3,4,5,6\} \times \{1,2,3,4,5,6\}$ and classical measure applies. We are interested in maximum and (absolute) difference of the outcomes, denoted by X and Y respectively. We can consider X and Y as two different random variables and obtain their mass functions and answer any question of interest about them separately. However, if we consider them together, then we can answer questions that involve both, e.g., probability of the maximum being 6 and the difference not exceeding 1 (i.e., $X = 6$ and $Y \le 1$). Formally, jointly distributed random variables or random vectors are vector valued functions defined on the sample space of a random experiment. Here, we consider only two-dimensional vectors, i.e., $(X, Y): \Omega \to \mathbb{R}^2$.

Joint distribution function of (X, Y) is defined as: $F_{X,Y}: \mathbb{R}^2 \to [0,1]$ such that $F_{X,Y}(x, y) = P_{(X,Y)(\Omega)}(\{X \le x, Y \le y\}) = P_\Omega(\{\omega \in \Omega: (X, Y)(\omega) \le (x, y)\})$ for all $(x, y) \in \mathbb{R}^2$. Important properties of the joint distribution function are:

(i) $F_{X,Y}(-\infty, y) = F_{X,Y}(x, -\infty) = 0$ for all $x, y \in \mathbb{R}$ and $F_{X,Y}(\infty, \infty) = 1$
(ii) $F_{X,Y}(x, y)$ is non-decreasing in both $x, y \in \mathbb{R}$
(iii) $F_{X,Y}(x, y)$ is right-continuous in both $x, y \in \mathbb{R}$
(iv) $F_X(x) = F_{X,Y}(x, \infty)$ and $F_Y(y) = F_{X,Y}(\infty, y)$ for all $x, y \in \mathbb{R}$

The last property connects joint distribution and individual/marginal distributions. This can be verified as: $F_X(x) = P_{X(\Omega)}(\{X \le x\}) = P_\Omega(\{\omega \in \Omega: X(\omega) \le x\}) = P_\Omega(\{\omega \in \Omega: X(\omega) \le x, Y(\omega) \in \mathbb{R}\}) = P_{(X,Y)(\Omega)}(\{X \le x, Y \in \mathbb{R}\}) = F_{X,Y}(x, \infty)$ for all $x \in \mathbb{R}$.

If X and Y both are discrete random variables, then we define joint mass function of (X, Y) as: $p_{X,Y}: (X, Y)(\Omega) \to [0,1]$ such that $p_{X,Y}(x, y) = P_{(X,Y)(\Omega)}(\{X = x, Y = y\}) = P_\Omega(\{\omega \in$

$\Omega: X(\omega) = x, Y(\Omega) = y\})$ for all $(x, y) \in (X, Y)(\Omega)$. Note that $p_X(x) = P_{X(\Omega)}(\{X = x\}) = P_\Omega(\{\omega \in \Omega: X(\omega) = x\}) = P_\Omega(\{\omega \in \Omega: X(\omega) = x, Y(\omega) \in Y(\Omega)\}) = P_{(X,Y)(\Omega)}(\{X = x, Y \in Y(\Omega)\}) = P_{(X,Y)(\Omega)}\left(\bigcup_{y \in Y(\Omega)}\{X = x, Y = y\}\right) = \sum_{y \in Y(\Omega)} P_{(X,Y)(\Omega)}(\{X = x, Y = y\})$, by third axiom of probability; so, $p_X(x) = \sum_{y \in Y(\Omega)} p_{X,Y}(x, y)$ for all $x \in X(\Omega)$. Similarly, $p_Y(y) = \sum_{x \in X(\Omega)} p_{X,Y}(x, y)$ for all $y \in Y(\Omega)$.

If X and Y both are continuous, then we define join density function of (X, Y) as follows. $f_{X,Y}: \mathbb{R}^2 \to [0, \infty)$ is a density function of (X, Y) if $F_{X,Y}(a, b) = \int_{-\infty}^{a} \int_{-\infty}^{b} f_{X,Y}(x, y) dy\, dx$ for all $(a, b) \in \mathbb{R}^2$. If $\partial^2 F_{(X,Y)}(x, y)/\partial x \partial y$ exists, then it's a joint density function of (X, Y), as it satisfies all the conditions. Consider $g(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy$. It's defined for all $x \in \mathbb{R}$, it's non-negative, and $\int_{-\infty}^{a} g(x) dx = \int_{-\infty}^{a} \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy\, dx = F_{X,Y}(a, \infty) = F_X(a)$ for all $a \in \mathbb{R}$. Therefore, $g(x)$ is a density function of X. Hence, $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy$ for all $x \in \mathbb{R}$, and in a similar manner, $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx$ for all $y \in \mathbb{R}$.

Back to the example … Joint mass function of (X, Y) is given below.

$p_{X,Y}$	$Y = 0$	$Y = 1$	$Y = 2$	$Y = 3$	$Y = 4$	$Y = 5$	p_X
$X = 1$	$P_\Omega(1,1) = 1/36$	0	0	0	0	0	$1/36$
$X = 2$	$P_\Omega(2,2) = 1/36$	$2/36$	0	0	0	0	$3/36$
$X = 3$	$P_\Omega(3,3) = 1/36$	$2/36$	$2/36$	0	0	0	$5/36$
$X = 4$	$P_\Omega(4,4) = 1/36$	$2/36$	$2/36$	$2/36$	0	0	$7/36$
$X = 5$	$P_\Omega(5,5) = 1/36$	$2/36$	$2/36$	$2/36$	$2/36$	0	$9/36$
$X = 6$	$P_\Omega(6,6) = 1/36$	$2/36$	$2/36$	$2/36$	$2/36$	$2/36$	$11/36$
p_Y	$6/36$	$10/36$	$8/36$	$6/36$	$4/36$	$2/36$	1

Now, $P_{(X,Y)(\Omega)}(X = 6, Y \leq 1) = p_{X,Y}(6,0) + p_{X,Y}(6,1) = 1/36 + 2/36 = 1/12$.

Conditional distribution and expectation: In the context of random vectors, sometimes we know value of one of the random variables, and that changes probability distribution of the other random variable. In the above example, X takes values $1,2, \ldots, 6$ with certain probabilities. However, if $Y = 1$, then X can no longer take the value 1 and its mass function changes. We capture this change by conditional mass function. $p_{X|Y=y}(x) = P(X = x|Y = y) = P\,(X = x, Y = y)/P(Y = y) = p_{X,Y}(x, y)/p_Y(y)$ for all $x \in X(\Omega)$. In a similar manner, $p_{Y|X=x}(y) = p_{X,Y}(x, y)/p_X(x)$ for all $y \in Y(\Omega)$.

For the continuous case, $f_{X|Y=y}(x) \coloneqq f_{X,Y}(x, y)/f_Y(y)$ for all $x \in \mathbb{R}$ is the conditional density function of $X|Y = y$. It's defined for all $x \in \mathbb{R}$, it's non-negative, and finally $\int_{-\infty}^{a} f_{X|Y=y}(x) dx = \frac{\int_{-\infty}^{a} f_{X,Y}(x,y) dx}{f_Y(y)} = \lim_{\delta \to 0^+} \frac{\int_{-\infty}^{a} f_{X,Y}(x,y) dx \cdot 2\delta}{f_Y(y) \cdot 2\delta} = \lim_{\delta \to 0^+} \frac{\int_{-\infty}^{a} \int_{y-\delta}^{y+\delta} f_{X,Y}(x,t) dt dx}{\int_{y-\delta}^{y+\delta} f_Y(t) dt} =$

$\lim_{\delta \to 0^+} \frac{F_{X,Y}(a,y+\delta) - F_{X,Y}(a,y-\delta)}{F_Y(y+\delta) - F_Y(y-\delta)} = \lim_{\delta \to 0^+} \frac{P(X \leq a, y-\delta \leq Y \leq y+\delta)}{P(y-\delta \leq Y \leq y+\delta)} = \frac{P(X \leq a, Y = y)}{P(Y = y)} = P(X \leq a|Y = y) =$

$F_{X|Y=y}(a)$ for all $a \in \mathbb{R}$. In a similar manner, $f_{Y|X=x}(y) \coloneqq f_{X,Y}(x,y)/f_X(x)$ for all $x \in \mathbb{R}$ is the conditional density function of $Y|X = x$.

Using the conditional mass and density functions, we can define conditional mean, variance, etc. $E[X|Y = y] = \sum_{x \in X(\Omega)} x \cdot p_{X|Y=y}(x)$ for the discrete case and $\int_{-\infty}^{\infty} x f_{X|Y=y}(x)dx$ for the continuous case. Unlike $E[X]$, $E[X|Y = y]$ is a random variable, as its value changes with Y. If we 'uncondition' the conditional mean, i.e., average out the uncertainty of Y, we get the unconditional mean. For the discrete case, $E_Y\big[E[X|Y = y]\big] = \sum_{y \in Y(\Omega)} E[X|Y = y] \cdot p_Y(y) = \sum_{y \in Y(\Omega)} \sum_{x \in X(\Omega)} x \cdot p_{X|Y=y}(x) \cdot p_Y(y) = \sum_{y \in Y(\Omega)} \sum_{x \in X(\Omega)} x \cdot p_{X,Y}(x,y) = \sum_{x \in X(\Omega)} x \cdot \sum_{y \in Y(\Omega)} p_{X,Y}(x,y) = \sum_{x \in X(\Omega)} x \cdot p_X(x) = E[X]$, as claimed. One can show the same for the continuous case as well. We can define conditional higher order moments, i.e., $E[X^k|Y = y]$ for $k = 1,2,3,$ ... in a similar manner, and $E_Y\Big[E[X^k|Y = y]\Big] = E[X^k]$.

Conditional variance can be calculated as: $Var(X|Y = y) = E[X^2|Y = y] - E^2[X|Y = y]$. However, $E_Y[Var(X|Y = y)] \neq Var(X)$. The reason is: $E_Y[Var(X|Y = y)] = E_Y\big[E[X^2|Y = y]\big] - E_Y\big[E^2[X|Y = y]\big] = E[X^2] - E_Y\big[E^2[X|Y = y]\big]$ and $E_Y\big[E^2[X|Y = y]\big] \neq E^2[X]$. Note that $E_Y^2\big[E[X|Y = y]\big] = E^2[X]$. Then $E_Y[Var(X|Y = y)] = E[X^2] - E^2[X] + E_Y^2\big[E[X|Y = y]\big] - E_Y\big[E^2[X|Y = y]\big] = Var(X) - \big\{E_Y\big[E^2[X|Y = y]\big] - E_Y^2\big[E[X|Y = y]\big]\big\} = Var(X) - Var_Y(E[X|Y = y])$. Thus, $Var(X) = E_Y[Var(X|Y = y)] + Var_Y(E[X|Y = y])$. The second term appears because variance is a relative moment.

Covariation and independence: In the previous section, we defined conditional mean and variance of a random variable given that the other random variable is known. When both the random variables are unknown, then the concept of mean and variance are meaningless for the jointly distributed random variables, though they are meaningful for the random variables individually. There is one concept, known as covariation, which is meaningful for jointly distributed random variables. It measures 'joint variation' of two random variables around their means. It is measured as: $Cov(X,Y) = E[(X - E[X]) \cdot (Y - E[Y])] = E[XY - XE[Y] - E[X]Y + E[X]E[Y]\big] = E[XY] - E[X]E[Y]$. Unlike variance, covariance can be negative. A positive covariance indicates that increase/decrease of one random variable with respect to its mean induces similar change in the other random variable. A negative covariance indicates opposing changes in the random variables about their means.

Covariance can be used to measure the degree of association between two random variables. However, $Cov(aX,bY) = E[abXY] - E[aX][bY] = abCov(X,Y)$ for constant a,b tells that choice of measuring units for X and Y influences covariance. To overcome this problem, we define correlation coefficient as: $\rho_{X,Y} = Cov(X,Y)/\sqrt{Var(X)Var(Y)}$. Since $Var(aX) = a^2Var(X)$ and $Var(bY) = b^2Var(Y)$, $\rho_{aX,bY} = abCov(X,Y)/ab\sqrt{Var(X)Var(Y)} = \rho_{X,Y}$, i.e., correlation coefficient is unit-invariant. Furthermore, it can be shown that $-1 \leq \rho_{X,Y} \leq 1$. These two properties make correlation coefficient a good measure of the degree of association between two random variables.

The concept of independence can be extended from events to random variables (and later to stochastic processes). Two events are independent if occurrence/non-occurrence of one event does not influence occurrence/non-occurrence of the other. We say that two random variables are independent if every event about one random variable is independent of each event about the other. Like independence of events A and B is equivalent to $P(AB) = P(A) \cdot P(B)$, independence of random variables X and Y is equivalent to $F_{X,Y}(x,y) = F_X(x) \cdot F_Y(y)$ for all $(x,y) \in \mathbb{R}^2$. If X, Y are discrete, then another equivalent condition is: $p_{X,Y}(x,y) = p_X(x) \cdot p_Y(y)$ for all $(x,y) \in (X,Y)(\Omega)$. If X, Y are continuous, then the alternate condition is: $f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y)$ for all $(x,y) \in \mathbb{R}^2$.

We can extend the idea of independence to multiple random variables, in a manner similar to the independence of multiple events. For three random variables X, Y, Z to be independent, we need (i) $F_{X,Y}(x,y) = F_X(x) \cdot F_Y(y)$ for all $(x,y) \in \mathbb{R}^2$, (ii) $F_{Y,Z}(y,z) = F_Y(y) \cdot F_Z(z)$ for all $(y,z) \in \mathbb{R}^2$, (iii) $F_{X,Z}(x,z) = F_X(x) \cdot F_Z(z)$ for all $(x,z) \in \mathbb{R}^2$, and (iv) $F_{X,Y,Z}(x,y,z) = F_X(x) \cdot F_Y(y) \cdot F_Z(z)$ for all $(x,y,z) \in \mathbb{R}^3$. However, setting $x \to \infty$ in (iv) gives us (ii), $y \to \infty$ gives (iii), and $z \to \infty$ gives (i). Therefor, we need only one condition for independence of X, Y, Z: $F_{X,Y,Z}(x,y,z) = F_X(x) \cdot F_Y(y) \cdot F_Z(z)$ for all $(x,y,z) \in \mathbb{R}^3$. In general, $X_1, X_2, \dots, X_n$ are independent if $F_{X_1,X_2,\dots,X_n}(x_1, x_2, \dots, x_n) = \prod_{i=1}^{n} F_{X_i}(x_i)$ for all $(x_1, x_2, \dots, x_n) \in \mathbb{R}^n$. The idea of independence extends to real-valued functions of random variables. If $X_1, X_2, \dots, X_n$ are independent and $g_1, g_2, \dots, g_n$ are real-valued functions, then $g_1(X_1), g_2(X_2), \dots, g_n(X_n)$ too are independent random variables.

If X, Y are independent, $p_{X|Y=y}(x) = p_{X,Y}(x,y)/p_Y(y) = p_X(x)p_Y(y)/p_Y(y) = p_X(x)$ for all $x \in X(\Omega)$ for the discrete case and $f_{X|Y=y}(x) = f_{X,Y}(x,y)/f_Y(y) = f_X(x)f_Y(y)/f_Y(y) = f_X(x)$ for all $x \in \mathbb{R}$ for the continuous case. Then $E[X|Y=y] = E[X]$ and $Var(X|Y=y) = Var(X)$. Moreover, $E[XY] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_{X,Y}(x,y)dydx = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_X(x)f_Y(y)dydx = \int_{-\infty}^{\infty} x f_X(x)dx \int_{-\infty}^{\infty} y f_Y(y)dy = E[X]E[Y]$, implying $Cov(X,Y) = 0$ as well as $\rho_{X,Y} = 0$. It shall be noted that $Cov(X,Y)$ can be zero even when X and Y are not independent.

Sum of random variables: Consider $Z = X + Y$ or XY or $\max(X,Y)$. These are examples of real-valued functions of jointly distributed random variables, and thus, can be regarded as a new random variable defined on $(X,Y)(\Omega)$. We already have encountered this in the above sections. Most important of these functions is the sum of random variables, which is consider here. One can also consider vector-valued functions of jointly distributed random variables, e.g., $(U,V): (X,Y)(\Omega) \to \mathbb{R}^2$ such that $U = X + Y$ and $V = XY$. Here, we need to obtain joint distribution and mass/density functions of (U,V).

We can calculate mean and variance of the sum of random variables using the law of the unconscious statistician (i.e., without calculating mass/density function of the sum). For the continuous case, $E[X+Y] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)f_{X,Y}(x,y)dydx = \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f_{X,Y}(x,y)dydx +$

$\int_{-\infty}^{\infty} y \int_{-\infty}^{\infty} f_{X,Y}(x,y)dxdy = \int_{-\infty}^{\infty} x f_X(x)dx + \int_{-\infty}^{\infty} y f_Y(y)dy = E[X] + E[Y]$, irrespective of whether X and Y are independent or not. One can verify this for the discrete case as well. In general, $E[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} E[X_i]$. Now we calculate variance of the sum. $Var(X+Y) = E[(X+Y)^2] - E^2[X+Y] = E[X^2 + Y^2 + 2XY] - (E[X] + E[Y])^2 = (E[X^2] - E^2[X]) + (E[Y^2] - E^2[Y]) + 2(E[XY] - E[X]E[Y]) = Var(X) + Var(Y) + 2Cov(X,Y)$. In general, $Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} Cov(X_i, X_j)$. Alternately, $Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n}\sum_{j=1}^{n} Cov(X_i, X_j)$, because $Cov(X_i, X_i) = Var(X_i)$ and $Cov(X_i, X_j) = Cov(X_j, X_i)$. If $X_1, X_2, \dots, X_n$ are independent, then $Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i)$.

Now, we try to calculate mass/distribution/density functions of the sum of two random variables. Let X and Y be discrete and $Z = X + Y$. Then Z is a discrete random variable with mass function: $p_Z(z) = P(\{Z = z\}) = P(\{X + Y = z\}) = P\left(\bigcup_{x \in X(\Omega)} \{X = x, Y = z - x\}\right) = \sum_{x \in X(\Omega)} P(\{X = x, Y = z - x\}) = \sum_{x \in X(\Omega)} p_{X,Y}(x, z - x)$. If X and Y are independent, then $p_Z(z) = \sum_{x \in X(\Omega)} p_X(x) p_Y(z - x)$. With this formula, one can show that

a) If $X \sim Bin(n_1, p)$ and $Y \sim Bin(n_2, p)$ are independent, then $X + Y \sim Bin(n_1 + n_2, p)$

b) If $X \sim Pois(\mu_1)$ and $Y \sim Pois(\mu_2)$ are independent, then $X + Y \sim Pois(\mu_1 + \mu_2)$

The above results can also be obtained using uniqueness of moment generating functions. For the continuous case, $F_Z(z) = P(\{Z \le z\}) = P(\{X + Y \le z\}) = P\left(\bigcup_{x \in \mathbb{R}} \left\{\bigcup_{y \le z - x} \{X = x, Y = y\}\right\}\right) = \int_{-\infty}^{\infty} \int_{-\infty}^{z-x} f_{X,Y}(x,y)dydx$. For independent X, Y, $F_Z(z) = \int_{-\infty}^{\infty} \int_{-\infty}^{z-x} f_X(x) f_Y(y)dydx = \int_{-\infty}^{\infty} f_X(x) F_Y(z - x)dx$, and then using Leibniz's rule for differentiation under the integral sign, $f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x)dx$. With this formula, one can show that

c) If $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ are independent, then $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

Limit theorems: In statistics we encounter average of repetitions of random variables, i.e., average of independent and identically distributed (*iid*) random variables. Let $X_1, X_2, \dots, X_n$ denote *iid* random variables with $E[X_1] = \mu$ and $Var(X_1) = \sigma^2$. Let $\bar{X}_n = \sum_{i=1}^{n} X_i/n$ denote the average of n such random variables. Then $E[\bar{X}_n] = E[\sum_{i=1}^{n} X_i]/n = \mu$ and $Var(\bar{X}_n) = Var[\sum_{i=1}^{n} X_i]/n^2 = \sigma^2/n$. Observe that $E[\bar{X}_n]$ is constant and $Var(\bar{X}_n)$ is a decreasing in n and vanishes as $n \to \infty$. Thus, if we consider average of a sufficiently large repetition of a random variable, the average converges to the mean of the random variable. This observation is formalized in the law of large numbers: $\lim_{n \to \infty} P(|\bar{X}_n - \mu| \le \epsilon) = 1$ for all $\epsilon > 0$. A 'stronger' result is: $P(\lim_{n \to \infty} \bar{X}_n = \mu) = 1$.

The law of large numbers is one of the most important results of probability theory. It is behind many results that we have encountered in this review, including the convergence of the frequentist's measure $\lim_{n \to \infty} n_A/n = P(A)$. There is another limit theorem that talks about the distribution of $\bar{X}_n$. The central limit theorem says that $\bar{X}_n$ converges to $N(\mu, \sigma^2/n)$ for sufficiently large n. Formally, $\lim_{n \to \infty} P\left((\bar{X}_n - \mu)/\left(\sigma/\sqrt{n}\right) \le z\right) = \Phi(z)$ for all $z \in \mathbb{R}$, where $\Phi(\cdot)$ is the distribution function of the standard normal random variable. Central limit

theorem allows us to approximate distribution of the average of a large number of *iid* random variables using normal distribution.

Stochastic process

Introduction to stochastic processes: With basic probability theory, we can model and analyze stochastic systems that can be described using one or a few random variables. Sometimes, we need many random variables to describe a system, and often, such random variables evolve over time, for example, daily closing indices of BSE, population size of a country over time, etc. In such cases, we model the system using stochastic processes and study how the system evolves with time. Formally, stochastic process is a sequence of random variables $\{X_t : t \in T\}$, where T is called the index set. Typically, T represents time and X_t represents state of the system at time $t \in T$. Depending on the natures of T and X_t (discrete vs. continuous), we classify stochastic processes into four types.

	X_t is discrete	X_t is continuous
T is discrete	Discrete-time discrete-state stochastic process Example: Markov chain	Discrete-time continuous-state stochastic process Example: Martingale
T is continuous	Continuous-time discrete-state stochastic process Example: Poisson process	Continuous-time continuous-state stochastic process Example: Brownian motion

In this course, we will consider discrete-state stochastic processes such as Markov chain, Branching chain, Poisson process, Renewal process, and Continuous-time Markov chain. The first two are discrete-time process and the last three are continuous-time process.

Practice problems

Book-1: Introduction to Probability Models by Sheldon Ross [10th edition]

Events and probability
Book-1, Chapter-1, Exercise No. 22, 25, 30, 33, 38

Random Variables
Book-1, Chapter-2, Exercise No. 28, 36, 40, 46, 63

Random Vectors
Book-1, Chapter-2, Exercise No. 55, 70, 74
Book-1, Chapter-3, Exercise No. 11, 14