EE 507 LECTURE PACKET, FALL 2008

Lecture 1: An Axiomatic Approach to Probability

Reading: Papoulis, Chapters 1-3

August 21-September 4, 2006

Please think about what you have done over the last couple days. Each activity that you have taken part in can be viewed as an *experiment*, that is, a process that has an outcome of interest. Some of these experiments have very predictable outcomes, for instance buying a book for $60 will result in your bank account balance decreasing by $60, brushing your teeth will make your mouth feel clean, and attending EE 507 class will make you sleepy.

However, other experiments have outcomes that cannot be predicted (by you or perhaps by anyone), or that are not worth predicting exactly. For instance, none of us will be able to predict exactly the temperature tomorrow (or at least it is too tedious to take into account all factors), and we also won't be able to predict your grade in this class. You also probably won't be able to predict correctly the time I will go to sleep tonight, though I may have a better idea. In fact, philosophically, we might expect that some quantities are fundamentally unpredictable because our observation process changes these quantities.

Nevertheless, you may need to make decisions or take actions using the outcomes of these experiments (for example, deciding what clothing you will wear tomorrow using a prediction of the temperature). This need for using incompletely-known information has motivated the development of a field called *probability theory*, which allows us to systematically represent and process uncertain quantities. In fact, in many settings of interest, the uncertain quantities are in fact uncertain *signals* (time functions); such uncertain signals are known as *random processes*. The purpose of this course is to review probability theory and introduce you to the exciting field of random processes.

In this first of three lectures on probability theory, we will motivate and introduce the most basic notions regarding probability, and start learning how to process uncertain information.

1 What are probabilities?

We are concerned with the unpredictable outcomes of experiments. At the crudest level, we can think of experiments as "always" having a certain outcome, "never" have this outcome, or "sometimes" have the outcome. If you think for a moment, though, I think you will realize that humans in fact have more refined notions for unpredictable outcomes (e.g., most likely it will rain tomorrow), and use these refined notions to make decisions (e.g., carry an umbrella).

A sensible way to describe such unpredictable experiments is as follows. *Given the information that we know about the experiment*, we expect an outcome or event of interest to occur with some *chance* or probability. For instance, if we roll a fair die, we expect to get the outcome '2' with a 16.7% chance. Also, the event that an odd number is rolled will occur with a 50% chance (which we can deduce by noticing that three of the six equally-likely outcomes are odd). Alternately, we can interpret a probability as a *frequency*: if we were to repeat the die-rolling experiment many times, we would expect the event that an odd number is rolled to occur about half of the time.

This natural intuition for probability motivates a formal framework for studying uncertain experiments. We shall introduce this framework in the next section.

2 An Axiomatic Framework for Uncertain Experiments

An **uncertain experiment** is viewed as having a structure and following a set of rules or **axioms**, which are sensible given our above interpretations for probabilities.

Specifically, an uncertain experiment is said to generate or have **outcomes** (sometimes called **elementary outcomes**). These represent the possible results of interest of the experiment. Any **set**, or collection, of possible outcomes is called an **event**[1]. An event is said to occur if the experiment generates one of the outcomes in the set.

There is a space (set) of events[2] that are interesting to us. For each such *event* A, we assign a number $P(A)$ that is called a **probability of event** A. These probabilities are chosen in such a way that the following three axioms are satisfied:

- $P(A) \geq 0$ for all events A. (Interpretation: this axiom captures the idea that chances or frequencies must be non-negative.)
- The event Ω containing all outcomes has probability 1, i.e. $P(\Omega) = 1$. (Interpretation: this axiom simply captures the idea that the experiment, when defined properly, has some outcome.)

[1] We note that the set of no outcomes is also considered an event.

[2] The space of events ought to be *closed under complementation, intersection, and unionization* (see definitions below), in order for probabilities to be well defined; that is, if two events are in the space, then the event containing the common outcomes of the two should be in the space, as should be the event containing all outcomes in both; also, if an event is in the space, another event containing all outcomes that are not in the first event should be in the space.

- For two events A and B that are **disjoint** (that have no outcomes in common), the probability that either A or B occurs (denoted $P(A+B)$, see Section 2.2) equals the probability that A occurs plus the probability that B occurs $(P(A)+P(B))$. Interpretation: this axiom captures the idea that the chance of having one among a many outcomes (or disjoint events) is equal to the sum of the chances of each outcome (or each disjoint event, respectively).

These three axioms are all we need to develop a meaningful representation of uncertain experiments! All the insights that we gain about such experiments can be derived from this simple framework. One important consideration in taking this axiomatic approach is appropriately defining the outcomes for a given experiment. There are many ways to define the outcomes of an experiment; an analyst is wise to choose a definition that *has the appropriate granularity (level of detail)*, and *permits her/him to easily assign probabilities to events.* This will become clearer as we pursue examples.

It is worth noting that this axiomatic approach to probability was first proposed by Kolmogorov. The genius of his approach is that it exactly captures our natural concept of chance (as self-evident truths), and imposes no further unnecessary structure.

Let's conclude this formal discussion of probability with one bit of terminology. We shall refer to the space of possible outcomes of an experiment (which is equivalently the set of all outcomes Ω) as the **sample space** of the experiment.

2.1 Example 1

The following game is played at the WSU casino: a customer flips two fair coins. If an odd number of the coins (i.e., one coin) show Heads, then the casino pays the customer \$1. If an even number of the coins (zero or two coins) show Heads, then the customer has to pay the casino \$10. We would like to study the probability that the casino makes money in a given game.

One sensible way to define outcomes for this experiment is as the number of Heads showing on the two coins:

<u>Outcomes</u>
0 Heads showing
1 Heads showing
2 Heads showing

Since there are three elementary outcomes, we can define an event space with $2^3 = 8$ events. Here are the events, with sensibly-assigned probabilities:

Events
Φ, the **null event** (empty set); *prob=0*
{ 0 Heads }, *prob=0.25*
{ 1 Heads }, *prob=0.5*
{ 2 Heads }, *prob=0.25*
{ 0 Heads, 1 Heads }, *prob=0.75*
{ 0 Heads, 2 Heads }, *prob=0.5*
{ 1 Heads, 2 Heads }, *prob=0.75*
The full set Ω, *prob=1*

Notice that the three axioms of probability are satisfied.

Concepts Question 1: What is another good way to define outcomes for this experiment?

Concepts Question 2: What are some bad (useless) ways to define outcomes for this experiment?

2.2 Example 2

Consider the following uncertain experiment. You close your eyes and throw a dart at a map of Washington State, and you are interested in features of the place where it lands.

Here's one way to define outcomes for this experiment: each point on the map is an outcome of the experiment. One more outcome is that the dart misses the map of Washington altogether. We notice that this experiment has an infinite (and in fact *uncountable*) number of outcomes.

When we define outcomes as above, we can define each event as a region on the map, with or without the outcome that the dart misses. We notice that this experiment has LOTS and LOTS of events[3] (many more than there are outcomes). We can assign probabilities to these events as we wish, as long as they satisfy the three axioms. Conceptually, what does it mean for these axioms to be satisfied?

Concepts Question: What is another good way to define outcomes for this experiment?

2.3 Some Thought-Provoking Questions

1. Why are probabilities assigned to events rather than outcomes? Wouldn't it be simpler to assign probabilities to outcomes, and then think of event probabilities as being derived from outcome probabilities?

2. Seemingly, one concern with this axiomatic approach is that we can assign probabilities to events as we wish, and so there is not connection to the physical reality of the frequency of the event. Should we be concerned?

[3] Notice that many of these events may not be interesting to us; in many settings, such as this one, we may only need to be concerned with some of the events in the space. This is important to remember for the sake of sanity.

2.4 Some Derived Properties of Uncertain Experiments

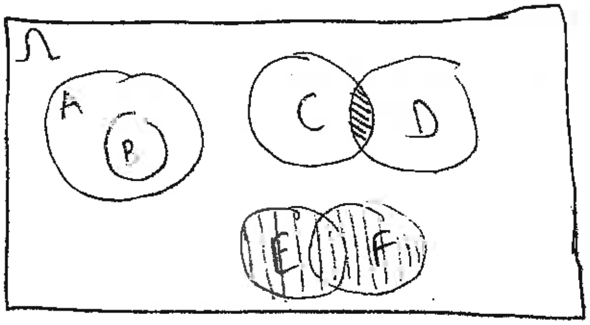
The axioms of probability lead us to several other interesting conclusions about uncertain experiments. To derive these conclusions, we will need to refresh ourselves about the algebra of sets.

2.4.1 Aside: Set Operations

In general, a **set** A is a collection of objects or items or **elements.** A subset B of A is another set, all of whose elements are contained in A. We use the notation $B \subset A$ when B is a subset of A. For us, the sample space Ω of an experiment will often be a set of interest. Events, which are subsets of Ω, are also sets of interest. We notice that the elements of these sets are outcome.

A set and its interesting subsets are often illustrated using a Venn diagram, as shown below. Venn diagrams are often useful for deducing set equivalences, so please review their use.

Venn Diagram



-Notice $B \subset A$

- ▨ CD

-

Often, set theorists are concerned with all the elements that lie in either of two sets (for us, all the possible outcomes within two events). This is known as the **union** of the two sets, and is itself a set. A bit more formally, the union of two sets A and B, which we denote $A+B$, is the set that contains all elements that are in either A or B (including those that are in both). The union of two sets is depicted on the Venn diagram above. Note that the union operation is *commutative* and *associative*: $A + B = B + A$, and $(A+B)+C = A+(B+C)$.

Given two sets, mathematicians are also interested in elements that lie within both sets. (for us, outcomes that are members of both events). That is, the **intersection** of the sets A and B (denoted AB) are the elements that are contained in both A and B. The intersection of two sets is also depicted in the Venn Diagram above. The intersection operation is commutative, associative, and *distributive over the union operation*: $AB = BA$, $(AB)C = A(BC)$, and $A(B+C) = AB + AC$.

Notice that the following relations are true: $AB \subset A \subset A+B$, and $AB \subset B \subset A+B$.

In addition to unions and intersections, we will also sometimes be interested in the **complements** of sets, i.e. the outcomes that are outside a given set. In particular, consider a full set Ω (in our case, the set of all possible outcomes), and consider a set A within this full set. Then the complement of A, denoted $\overline{A}$, contains all elements of Ω that are not in A. The complement of a set is also depicted above.

Algebraic relations between sets can be derived from these basic definitions. Typically, one of three approaches is taken to prove these equivalences: *1)* the definitions are used directly, *2)* elements are counted in Venn diagrams, or *3)* other derived equivalences are used to prove the result. I am sure you have had some experience with proving set relations, so I won't spend much time on this; let's do a couple examples to become fluent again, though:

1. $A\Omega = A$

2. $A+\bar{A}=\Omega$

3. $\overline{AB} = \bar{A}+\bar{B}$

4. $A\bar{B}+\bar{A}B = (A+B)(\overline{AB})$

2.5 Basic Properties of Uncertain Experiments

Now that we have become fluent with the algebra of sets, we are ready to derive properties of uncertain experiments from the three basic axioms.

One property that is "obvious" but nevertheless needs to be derived is that the empty event has zero probability:
$P(\Phi) = 0$.

Proof:

Another relation that is often used connects the probability of a union of two events with the probabilities of the events and of their intersection:
$P(A+B) = P(A) + P(B) - P(AB) \leq P(A) + P(B)$.

Proof:

We also note that the third axiom can be applied repeatedly to obtain the following: if events $A_1, \ldots, A_N$ are mutually exclusive (i.e., all their pairwise intersections are the null set), then $P(A_1 + \ldots + A_N) = P(A_1) + \ldots + P(A_N)$. Thinking about the above result, we might wonder if such a result holds even in the limit as N gets large. In fact, it doesn't, and this shortcoming motivates us to modify the definition of an uncertain experiment just a little bit:

Let's conclude our introduction of probability with an example:

3 Enhancing the Model for Uncertain Experiments: Conditional Probabilities

Often, when we are analyzing an uncertain-experiment model, we are able to obtain further information about the outcome of the experiment, and wish to use this information to appropriately modify the probabilities of events. That is, given that a particular event A which is a subset of the sample space has occurred, we would like to appropriately describe probabilities of other events. From our intuition about probabilities (in particular, their interpretation as chances or frequencies), we will give a plausible description of these probabilities; this will in turn lead us to define the probability of an event *conditioned* on the event A.

3.1 Intuition

To motivate this idea of conditional probabilities, let's consider an example. Let's say that you receive important information through a binary channel. In particular, you receive one of the eight binary numbers $000, 001, \ldots, 111$, each with equal probability. You need to figure out if the received binary number is greater than 4. Unfortunately, the channel doesn't always work perfectly, so you may not be certain whether or not the number sent is greater than 4 (let's call this the event A). Consider the following scenarios:

One day, you receive no information at all. What is the probability that the number that was sent is greater than 4 (i.e., the probability that the event A occurred)?

- All outcomes are equally probable, i.e. have probability 1/8.
- Three of the eight outcomes are in A, hence $P(A) = \frac{3}{8}$

Another day, you receive only the final bit in the sequence. Let us denote by B the event that this final bit is '1'. Given that the final bit is '1' (i.e. given or conditioned on event B), what is probability that event A occurred?

If B happened, then only four outcomes might have happened: 001, 011, 101, 111. Two of these four outcomes are greater than 4, hence the probability is $\frac{2}{4} = \frac{1}{2}$

Notice how we computed this probability: we found the *number* of outcomes in B

that are also in A, and divided this by the total number of outcomes in B. Another way of saying this is that we found the *total probability* of the outcomes in B that are also in A—which is the probability $P(AB)$—and divided this by the the probability of event B $(P(B))$[4]. This probability-based interpretation is more powerful, in that it is sensible even when outcomes do not have equal probability. Let's revisit the example above in the case where the different sequences are sent with different probabilities:

footnote, not exponent!

3.2 Definition and Properties of Conditional Probabilities

Consider a sample space, and two events A and B in that sample space. We use the notation $P(A \mid B)$ for the probability of A **conditioned** on B (or **given** B). We define this **conditional probability** as $P(A \mid B) = \frac{P(AB)}{P(B)}$. As motivated above, conditional probabilities are of interest in many settings, and the expression above can directly be used to find conditional probabilities from a description of the full sample space.

Many other experiments are actually naturally described using conditional probabilities, and our goal is to find probabilities of events (or of their intersections/unions) or other conditional probabilities from this description. Thus, we are motivated to derive expressions that allow us to find such probabilities from conditional probabilities. Let us now derive several such expressions, using an example throughout for motivation.

Example: During the summer in Pullman, 90% of the days are sunny, and 10% of the days are cloudy. On sunny days, the temperature is greater than 100^o with probability 0.3, between 90^o and 100^o with probability 0.5, and between 80^o and 90^o with probability 0.2. On cloudy days, the temperature is between 80^o and 90^o with probability 0.4, and less than 80^o with probability 0.6. Let us answer the following three questions:

1. If we choose a summer day at random, what is the probability that the day is sunny and the temperature is between 80^o and 90^o?

2. If we choose a summer day at random, what is the probability that the temperature is between 80^o and 90^o?

[4] If $P(B) = 0$, then $\frac{P(AB)}{P(B)}$ is undefined. This is actually sensible, since it is not meaningful to ask what the probability of one event is given another event that never occurs.

3. Given that the temperature on a randomly-chosen summer day is between 80^o and 90^o, what is the probability that the day is sunny?

We can model the above scenario as an uncertain experiment with six possible outcomes, each a weather condition (sunny, cloudy) coupled with a temperature range. Let us call the event that the weather is sunny B, and the event that the temperature is between 80^o and 90^o A. Now consider answering the three questions above:

1. The probability that the weather is sunny and the temperature is in the given range is $P(AB)$. Since we have available the conditional probability for A given B, and the probability of B, we notice that we can find $P(AB)$ as $P(A \mid B)P(B)$ (by rearranging the definition of conditional probabilities). Notice that this property that $P(AB) = P(A \mid B)P(B)$ makes very good sense: the probability that A and B both occur is equal to the probability that B occurs times the probability that A occurs given that B occurred. In this example, we thus find that $P(AB) = (0.9)(0.2) = 0.18$.

2. The probability that the temperature is between 80^o and 90^o is $P(A)$. We need to somehow find this probability in terms of conditional probabilities like $P(A \mid B)$ and $P(A \mid \overline{B})$ and probabilities for the events B and $\overline{B}$. To do so, notice that $A = A\overline{B} + AB$, since $A\overline{B} + AB = A(B + \overline{B}) = A\Omega = A$. Noting further that $A\overline{B}$ and AB are mutually exclusive (i.e., don't have any outcomes in common), we see that $P(A) = P(AB) + P(A\overline{B})$. Using the definition of conditioning, we get $P(A) = P(A \mid B)P(B) + P(A \mid \overline{B})P(\overline{B})$. We call this important relationship the **law of total property**: it shows that the probability of an event can be found by conditioning on a set of events that partition the sample space. Thinking about our example, we can convince ourselves that the law of total probability is sensible: the total probability that the temperature is between 80^o and 90^o is the probability the temperature is in this range given that it is sunny times the probability it is sunny, plus the probability the temperature is in this range given that it is not sunny (cloudy) times the probability it is not sunny (cloudy). Substituting the probabilities, we find that $P(A) = (0.9)(0.2) + (0.1)(0.4) = 0.22$. It is worth noting that, in general, we can condition on more than two events when finding total probabilities; you will study this natural generalization of the law of total probability in your homework.

3. Finally, we would like to find the probability that it is sunny given that the temperature is between 80^o and 90^o. This is the probability $P(B \mid A)$. From the definition of conditional probability, we find that $P(B \mid A) = \frac{P(AB)}{P(A)}$. We have already computed $P(AB)$ and $P(B)$ from Tasks 1 and 2, so we can immediately find that $P(B \mid A) = \frac{0.18}{0.22} \approx 0.82$. In case we wish to work directly from the problem formulation, some further simplification of $P(B \mid A)$ is worthwhile. First, by applying the

definition of conditioning, we get $P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)}$. This equality is the celebrated **Bayes' rule**, which shows how the probability of one event B conditioned on another event A can be found from $P(A \mid B)$. In practice, the denominator $P(A)$ in the Baye's rule expression is often found from the law of total probability, whose substitution yields the following: $P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A \mid B)P(B)+P(A \mid \overline{B})P(\overline{B})}$. You will find yourself using this form of Bayes' rule often. Finally, let us take a moment to interpret the answer to the example problem. We notice the probability it is sunny given that the temperature is between 80^o and 90^o is smaller than the (unconditioned) probability that it is sunny; this is sensible since cloudy days are more likely to have temperatures between 80^o and 90^o.

Let us conclude our introduction to conditional probabilities with another example:

3.3 Independent Events

Conditional probabilities are important because they allow us to study how new information about an experiment impact the probabilities of events of interest. Often, we are particularly interested in whether or not new information tells us anything at all about an event of interest. This motivates us to define the notion of **independence**. In particular, we say that event B is **independent** of the event A, if knowledge that A occurred does not change the probability of B, i.e. if $P(B \mid A) = P(B)$.

Let us first make sure our definition of independence is sensible, by thinking about an example. Consider throwing a dart at a circular dartboard, and say that we are equally likely to hit any point on the dartboard. Let A be the event that we hit the left half of the dartboard, and let B be the event that we hit the top half of the dartboard. Conceptually, we would expect the event B to be independent of the event A: from symmetry, the fact that we hit the left half of the dartboard should not give any further insight into whether or not we hit the top or bottom of the dartboard. From the model of this uncertain experiment, $P(B) = \frac{1}{2}$ and $P(B \,|\, A) = \frac{P(AB)}{P(A)} = \frac{0.25}{0.5} = \frac{1}{2}$. Thus, our model yields that $P(B \,|\, A) = P(B)$, as expected.

Several properties of independent events are worth discussing. Perhaps most significantly, we note that independence is a transitive property. That is, if $P(B \,|\, A) = P(B)$, then $P(A \,|\, B) = P(A)$. Let's take a moment to prove this, using Bayes' rule:

$$P(A|B) \underset{\substack{\uparrow \\ \text{Baye's} \\ \text{Rule}}}{=} \frac{P(B|A)P(A)}{P(B)} \underset{\substack{\uparrow \\ \text{from} \\ \text{def. of independence}}}{=} \frac{P(B)P(A)}{P(B)} = P(A).$$

This transitivity property is quite useful (and reassuring), since it permits us to describe two events as independent without concern about directionality. In other words, if two events are independent, then A does not provide information about B, *and* B does not provide information about A. From the proof of transitivity, we also immediately obtain the following condition for independence: *two events A and B are independent if and only if* $P(AB) = P(A)P(B)$.

Let us go through one more (rather surprising) example to gain more insight into independence:.

3.4 Conditioning on Multiple Events

In many cases, we may have multiple pieces of information about an experiment, and may wish to compute probabilities of events given this information. We use the notation $P(A \,|\, B_1, \ldots, B_n)$ for the probability of event A given that events $B_1, \ldots, B_n$ have all occurred. We notice, from the definition of conditioning, $P(A \,|\, B_1, \ldots, B_n) = \frac{P(A, B_1, \ldots, B_n)}{P(B_1, \ldots, B_n)}$.

When we consider the probability of an event conditioned on multiple other ones, the following interpretation of conditioning is very useful: conditioning on an event C can be viewed as redefining an experiment's sample space to be C, with the probabilities of outcomes/events in this sample space scaled up so that the total probability is 1. This idea that we are simply re-defining probabilities for a smaller sample space naturally allows us to generalize the properties/definitions that we have developed. For instance, we see that $P(A+B\,|\,C) = P(A\,|\,C) + P(B\,|\,C) - P(AB\,|\,C)$. Similarly, we obtain that $P(A\,|\,B, C) = \frac{P(A,B\,|\,C)}{P(B\,|\,C)}$. Also, we can generalize the notion of independence. Two events A and B are independent given another event C, if $P(A\,|\,B, C) = P(A\,|\,C)$.

You will have several chances to analyze probabilities of events conditioned on multiple other ones, in your homework.

4 Combined Experiments and Repeated Trials

Let us conclude this introduction to probability theory, by studying how experiments can be combined, and in particular how repeated trials of an experiment can be analyzed. There are several reasons that we study combined experiments and repeated trials:

- Just as we naturally consider re-defining an experiment's sample space given further information, it is natural for us to consider the sample space of outcomes when many experiments are considered at once.
- This class is concerned with random processes, which are sequences/signals generated by uncertain experiments. A natural first step toward studying random sequences/signals is to study sequences of uncertain experiments.
- By analyzing repeated trials, we will be able to show how an uncertain experiment's axiomatic model can be evaluated.
- Studying repeated trials is a sleazy way for us to introduce the art of *counting*, a commonly-used tool for computing probabilities.

4.1 Combined Experiments

Consider the following pair of experiments. In one experiment, we roll a fair die. This experiment has a sample space with six outcomes $1, \ldots, 6$, each that occur with probability $\frac{1}{6}$. In the second experiment, we toss a fair coin. This experiment has two outcomes, H and T, each occurring with probability $\frac{1}{2}$. If I ask for the probability that the die shows 4 and the coin shows H, you will be tempted to answer that this probability is $\frac{1}{6}\frac{1}{2} = \frac{1}{12}$. However, based on our development so far, we do not technically have a way to study this combined experiment. Problems such as these motivate us to carefully define the combination of two (or more) experiments into a single experiment.

Let's say we are combining two experiments with sample spaces Ω_1 and Ω_2. A sensible way to define outcomes of the combined experiment are as ordered pairs of outcomes from the two individual experiments; all such pairs are possible outcomes. For instance, a combination of the die-tossing and coin-tossing experiments has outcomes such as $(4, H)$; there are twelve outcomes in total in this case. Similarly, if two experiments

8

have outcomes that are points on the real line, the combined experiment has outcomes that are pairs of points, or in other words points in a two-dimensional space. One bit of terminology is worth discussing: a set which contains all pairs from two other sets is often referred to as the **Cartesian product** of the other two sets. Thus, the new sample space is the Cartesian product of Ω_1 and $Omega_2$, denoted $\Omega_1 \times \Omega_2$.

Based on the definition of outcomes above, we see that events in events are sets of these outcome pairs. Another way to view the set of interesting events is as the Cartesian product of the sets of interesting events of the two original experiments, together with the intersections/unions of these events.

How we assign probabilities to the combined experiments depends to some extent on the specifics of the two experiments. At the very least, we would expect these probabilities to be *consistent*, i.e. the probability of the event $A_1 \times \Omega_2$ of the combined experiment (where A_1 is an event of experiment 1 and Ω_2 is the full event of experiment 2) to equal the probability of event A_1 in experiment 1. Similarly, $P(\Omega_1 \otimes A_2)$ should equal $P(A_2)$. In many cases, the two experiments can be assumed to be independent, in the sense that $P(A_1 \times A_2) = P(A_1) \times P(A_2)$, where $A_1 \times A_2$ is an event of the combined experiment, and A_1 and A_2 are events of the original experiments.

$$\begin{array}{c|cc} & 1 & 2 \\ \hline a & & \\ b & & \\ c & & \end{array}$$

This formalism for combining experiments allows us to sensibly combine the die-tossing and coin-tossing experiments. I'll leave it as an exercise for you to study this example in detail.

Of course, we can combine more than two experiments with this approach, through a sequential process.

$$\begin{pmatrix} 0.4 & 0 \\ 0.1 & 0.5 \end{pmatrix}$$

4.2 Repeated Trials

$$\begin{array}{c|cc} & 1 & 2 \\ \hline a & x & 0.4-x \\ b & 0.5-x & [illegible] \end{array}$$

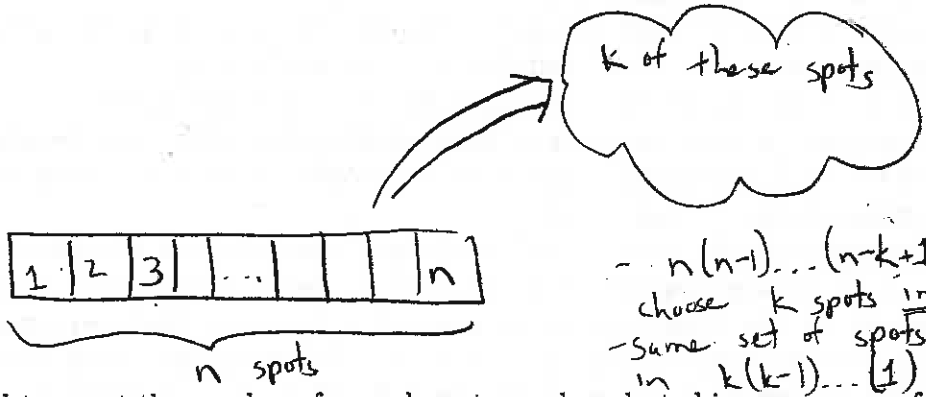
An especially interesting class of combined experiments are those in which the same experiment is repeated many times, independently. Such repeated trials are interesting because they often happen in practice, and because they give us a way to give our axiomatic approach a physical interpretation. Studying repeated trials also is a good way for us to get fluent with counting.

Specifically, we will address the following question concerning repeated trials. For a particular experiment, let's say the probability that an event A occurs is p. We repeat this experiment n times. What is the probability that the event A occurs on exactly k of the n trials.

$0.5 - x + \frac{x}{2} = 0.6$

[illegible]

To find this probability, let us consider the combined experiment. Events of this combined experiment are sequences of n events of the original experiment. Let us, in particular, consider events in which A occurs in a specified set of k trials, while $\overline{A}$ occurs in the remaining $n-k$ trials. Each event of this sort, where A occurs in a specified set of k trials and $\overline{A}$ occurs in the remaining trials (see diagram below), occurs with probability $p^k(1-p)^{n-k}$. Further, we notice that the union of all such events is the event that A occurs in exactly k trials, and further that the intersection of any two of these events is the null set. Thus, the probability the A occurs in exactly k trials is $(C)(p^k(1-p)^{n-k})$, where C is the number of ways k spots can be selected in a sequence of n spots (see diagram below).



- n(n-1)...(n-k+1) ways to choose k spots in order.
- Same set of spots can be ordered in k(k-1)...(1) ways.

We thus need to count the number of ways k spots can be selected in a sequence of n spots. To find this number C, we can envision sequentially pulling out k spots from the sequence of n spots. In the first step, we can pull out any of the n spots. At the next step, we can pull out any of the remaining $n-1$ spots, and so on. Thus, in total, there are $(n)(n-1)...(n-k+1)$ to pull out k spots in order, out of n spots. We can also count this number in another way. By assumption, C is the number of ways that k (unordered) spots can be selected from n spots. For each of these selections, we can order the selected spots in $(k)(k-1)...(2)(1)$ ways. Thus, $(C)(k)(k-1)...(2)(1) = (n)(n-1)...(n-k+1)$, and so $C = \frac{(n)(n-1)...(n-k+1)}{(k)(k-1)...(2)(1)}$. We can conveniently write C in factorial notation, as $C = \frac{n!}{k!(n-k)!}$.

We thus find that the probability of the event A occurring in exactly k of the n trials is $\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$. This computation of the probability that an event occurs on exactly k of n trials is a fundamental one in probability. Repeated trials of this form (where one event is either occurs or does not occur during each trial, independently) are often referred to as **Bernoulli trials**. To get a little more insight into Bernoulli trials, let us actually plot the probability of an event occurring k times as a function of k, for fixed n and p:

Using a very similar analysis, we can find the probability that, given events $A_1, \ldots, A_n$ that partition the sample space, the event A_1 occurs on k_1 trials, the event A_2 occurs k_2

trials, and so forth. I'll let you explore this problem on your next homework set.

4.3 Asymptotic Probabilities and Connection to Reality

Often, we are interested in experiments that are repeated many, many times. For such experiments, the plot of probabilities versus k takes on a special shape. To get an idea of what this special shape is, let us draw the probabilities plot for several n and fixed k, in the context of an example:

(The example will be handed out later.)

$q = 1-p$

By drawing these probability functions on a semi-log plot and fitting them with quadratic polynomials, we can guess the following:

$$P(k\,successes) \approx \frac{1}{\sqrt{2\pi npq}\,(1-p)} e^{-(k-np)^2/2np(1-p)},$$

as long as $n >> \frac{1}{p(1-p)}$. This standard shape taken by the probability function is the celebrated *Bell Curve* or **Gaussian function.** We will use the Gaussian function extensively in this course; I'll give a thorough introduction in Lecture 2.

The approximation above can be proved using *Stirling's formula*, which shows how factorials can be approximated using exponentials. This proof does not give much conceptual insight, so let us instead give a rough conceptual justification for the result:

We also often are concerned about the case where n becomes large (there are many trials), but the probability of an event A occurring on any particular trial scales inversely with n. In this case, the shape of the probability function also becomes distinctive. First, let us motivate why this extreme case is of interest with an example:

Pullman has grown into a metropolis, and now has a 24-hour bus service. We are interested in the times at which buses are at a particular shelter. In particular, let's say that on average five buses arrive at the shelter per hour. Then a reasonable model for bus arrival times over a long interval T hours is that $5T$ buses each independently arrive randomly at a time between 0 and T. Given this model, we might wish to compute the probability that k buses arrive during a 1-hour interval. This computation is especially interesting in the case that T becomes arbitrarily large, since in this limit we are finding the probability that k buses arrival during a "typical" one-hour interval. However, for a given T, notice that the probability of having k buses is $\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$, where $n = 5T$ and $p = \frac{1}{T}$ (see Figure below). Thus, we are motivated to study repeated-trial experiments in the limit that the number of trials gets large, while the probability of the event occurring proportionally gets small.

Let us sketch the probabilities as a function of k to identify the shape of the distribution in the limiting case.

(Please see handout)

From these sketches, we can guess the following approximation for the probability of k occurrences:

$$\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k} \approx e^{-np}\frac{(np)^k}{k!},$$

for large n and np of order 1. We refer to this approximation as the **Poisson approximation.**

Let us take a moment to prove this result:

Applying this result to the the bus-flow example above, we notice that the probability of getting k buses during one hour is given by $e^{-5}\frac{5^k}{k!}$.

Let us also pursue one more example in which the Poisson approximation is used:

Finally, we use the asymptotic results developed in this section to justify one of our interpretations of probabilities! ~~Recall that~~

Lecture 2a: Introduction to Random Variables

Reading: Papoulis, Chapter 4

September 6-13, 2006

So far, we have introduced a framework for analyzing uncertain experiments, and have shown how this framework can be used to deduce probabilities of random events. In this section, we will enhance our framework by defining a notion known as a *random variable* and the associated notion of a *probability distribution*. We will begin by motivating this new notion, and then giving a formal definition. Once we have carefully defined random variables and probability distributions, we will recognize that they are amenable to all sorts of interesting analyses, and will pursue these analyses in detail. In the process, we will give examples that highlight common random variables and probability distributions, and shed some light on the history of these notions.

This handout is the first of three on random variables.

1 Motivation

The concept of a random variable is motivated by the observation that experiments with numerical interpretations for outcomes are amenable to some analyses that other experiments are not. To crystallize this motivation, let me ask you some questions:

1. **Question:** If you roll two dice, what is the probability that Die 1 shows a larger value than Die 2? **Answer:** $\frac{15}{36}$.

2. **Question:** Let's say you pick two letters from the English alphabet. What is the probability that Letter 1 is bigger than Letter 2? **Answer:** Sandip, that question doesn't make any sense, you idiot. What does it mean for one letter to be bigger than the other??

3. **Question:** If we choose a student at random in this class, on average what will her/his GPA be? **Answer:** 3.2.

4. **Question:** If we choose at student a random in this class, on average what color is her/his hair? **Answer:** Sandip, this is a DUMB QUESTION[1]. There's no way to

[1] Actually, there is no such thing as a dumb question, so please keep asking me questions.

average colors. Perhaps one could average the frequencies of light associated with these colors, but averaging the colors themselves doesn't make any sense.

These examples highlight that we can ask and solve some special questions (e.g., questions about averages or comparisons), when experimental outcomes have numerical interpretations. However, so far, our framework for modeling experiments does not differentiate between numerical and non-numerical outcomes: we simply have arbitrary outcomes and events that are sets of outcomes. A random variable is a way of associating *numbers* with outcomes, in which case certain important events can expressed in terms of these numbers, and interesting operations on these numbers can be defined.

2 What is a Random Variable?

Formally, a random variable is a mapping from the sample space to the real numbers ($\mathcal{R}$), the integers ($\mathcal{Z}$), or to any vector space. That is, a random variable is an association of a number with each outcome with an experiment. A bit more formally, given a sample space Ω, the random variable $X(\omega)$ is a numerical function of outcomes $\omega \in \Omega$, i.e. $X(\omega)$ gives a numerical value to each outcome in Ω.

Let us introduce several examples, to clarify what a random variable is.

Example 1: Consider an experiment where we toss a fair coin two times. This experiment has four outcomes: $\Omega = HH, HT, TH, TT$. Let's say we care specifically about the *number* of heads that are tossed in two trials. Notice that each outcome ω in the sample space has a certain number of heads, so we can associate a number $X(\omega)$ which equals the number of heads with each ω. $X(\omega)$ is a random variable. Here's a full definition of the random variable:
$\omega = HH \rightarrow X(\omega) = 2$
$\omega = HT \rightarrow X(\omega) = 1$
$\omega = TH \rightarrow X(\omega) = 1$
$\omega = TT \rightarrow X(\omega) = 0$

Example 2: Consider an experiment where a point on the map of Washington State is selected at random (with all points on the map equally likely to be selected). The sample space Ω of this experiment thus contains all points on the map (perhaps including points that represent the ocean). We might be interested in the elevations of such randomly selected locations in Washington. Thus, it is natural for us to define a random variable $X = X(\omega)$ that maps points ω in the sample space Ω to elevations of the corresponding locations. We notice that the random variable $X(\omega)$ takes on values between $0\,m$ and $4392\,m$ $(14,410$ ft), which is the highest point in the state.

3 Cumulative Distributions: Probabilities of Important Random Variable-Based Events

A primary advantage of using random variables is that events can be defined in terms of the random variable. Since the random variable associates numbers with outcomes, we can sensibly compare outcomes in defining an event (and we can also give "scores" or "average values" to events, as we will see in a later section).

So, what is a useful way to define events in terms of a random variable X associated with the sample space? Let's brainstorm a little bit...

We could consider events comprising outcomes for which the random variable X has a particular value, say x. We shall use the short notation $\{X = x\}$ to describe such events, and refer to the probability that this event occurs (P(X=x)) as the probability that the random variable X equals x. Events defined in this way are quite useful in studying some experiments. For instance, in the coin-tossing experiment, we may well be interested in the event $\{X = 1\}$, i.e. the event comprising all outcomes with one H showing. This event has probability $\frac{1}{2}$ (or in other words we can say that $X = 1$ with probability $\frac{1}{2}$). *Unfortunately, however, defining events in this way has a fatal flaw: in many experiments (specifically, those with a continuum of outcomes), the probabilities of all events $X = x$ are 0 and hence such events are not worthwhile to analyze.* For instance, for the experiment in which X is the elevation of a randomly chosen point in Washington State, the probability of the event $X = 800\,m$ is zero: the elevation is not *exactly* $800\,m$ except along a line on the map.

A neat approach for avoiding the flaw above is to define events as comprising outcomes for which the random variable has at most a certain value x, i.e. outcomes such that $X(\omega) \leq x$. We use the notation $\{X \leq\}$ for such events, and use the notation $P(X \leq x)$ for the probability of this event. By defining an event as a range of outcomes (specifically, those satisfying an inequality constraint), we can obtain a meaningful probability for the event even when outcomes form a continuum. For instance, for the experiment in which X is the elevation of a randomly chosen point in Washington State, the probability of the event $X \leq 800\,m$ is equal to the total fraction of the state that is below $800\,m$ in elevation, which is a non-zero and meaningful amount.

Now consider a plot of the probabilities $P(X \leq x)$ as x changes from $-\infty$ to ∞. This plot gives us the total or *cumulative* probabilities that the random variable X takes on a value less than x, for all x. Thus, we might expect that this plot in some sense tells us everything we might want to know about a random variable, and hence may be valuable for probabilistic analysis of random variables. In fact, this plot and the corresponding function $P(X \leq x)$ are indeed powerful constructs in the study of random variables, as we shall demonstrate throughout the course. Since this function is so important, it has a name: **the cumulative distribution function** (CDF) or **probability distribution function.** You will sometimes see the notation $F_X(x)$ used instead of $P(X \leq x)$ for the cumulative distribution function.

In the remainder of this section, let us practice finding/drawing cumulative distribution functions from descriptions of the uncertain experiment, and derive some instructive properties of the CDF.

3.1 Examples

Let us first find and sketch the cumulative distribution functions for the two random variables defined in Section 2:

Also, consider an experiment where a fair coin is tossed n times, and let the random variable Y be the number of trials showing Heads. Let us plot the CDF for Y. Also, let's say that the random variable Z is the *fraction* of the trials showing Heads. How does the CDF of Z relate to that of Y?

3.2 Properties of the Cumulative Distribution Function

Also, let us derive some common properties of and deductions from cumulative distribution functions. These properties/deductions are helpful for computing probabilities from CDFs and manipulating CDFs in proving further results. They also give some further insight into what a CDF is.

Property 1: *A cumulative distribution function is right-continuous.*

Property 2: *The cumulative distribution function approaches* 1 *as* x *becomes large, and approaches* 0 *as* x *becomes small (highly negative).*

Property (Really, Deduction) 3: *The probability that a random variable* X *is within the interval* $[a, b]$ *can be found as* $F_X(b) - F_X(a)$.

4 Probability Density Functions

Let's think some more about the experiment in which we randomly choose a point on the map of Washington State. In particular, let's think about the following question: are you more likely to choose a point where the elevation X is $800\,m$ or a point where the elevation is $3000\,m$? Well, from our earlier discussion, we know that the probability $P(X = 800)$

and $P(X = 2000)$ are both 0, so this question is moot. However, our intuition suggests that, in some sense, we are more likely to hit the lower elevation because there are many more points with elevation near 800 m than with elevation near 2500 m. That is, the probability of $\{800 \le X \le 801\}$ is greater than the probability $\{2500 \le X \le 2501\}$, and hence we think of elevations *near* 800 m as being more likely than elevations *near* 2500 m. In fact, no matter how small we make the interval around 800 m, we would expect the probability of an elevation in that interval to be larger by a scale factor than for a correspondingly-sized interval at 2500 m (though both probabilities would be smaller). Thus, we realize that there is a difference in the probability per unit elevation at 800 m and 2500 m, or in other words in the *density of probability* at these two elevations.

The above discussion motivates a useful definition, of the probability density of a random variable at a point. In particular, for a random variable X, we define the **probability density** of X at x as

$$f_X(x) = \lim_{\epsilon \to 0} \frac{P(x \le X \le x + \epsilon)}{\epsilon}. \qquad (1)$$

Notice that the probability density of X at x is not the probability that $\{X = x\}$ but rather captures the probability of X being near x with other probabilities.

As with cumulative distributions, we find it convenient to plot the probability density as a function of x. We refer to this function and plot as the **probability density function**. Let us revisit the examples above, to get some practice in plotting probability density functions.

4.1 Examples

Example 1, Two Coin Tosses:

Example 2, Throwing a Dart at the Washington Map:

Example 3, n Coin Tosses:

4.2 Probability Mass Functions

From the coin toss examples, we realize that it is rather cumbersome to work with the probability density function when.random variables take on particular values with non-zero probabilities: in these cases, the PDF includes impulse (delta) functions. In the case where the random variable is actually of **discrete-type**—i.e., it only takes on values in a discrete set—we find it convenient to plot to actually plot the probabilities that the random variable takes on certain values rather than the probability density. That is, for a random variable X of discrete type, we often find it convenient to simply plot $P(X = x)$ as a function of x. We refer to such a plot as the **probability mass function** (PMF) of the random variable X. We notice that the PMF represents actual probabilities rather than probability densities.

Let us find the PMFs for the example random variables of discrete type introduced above:

4.3 Properties of the PDF/PMF

Let us also derive some properties of PDFs and PMFs, with an special focus on how they can be used to compute probabilities (and hence the CDF):

5 More about Random Variables

We have defined the notion of a *random variable*, as well as the associated *cumulative distribution function* and *probability distribution function*, and have identified properties of these constructs. We have already shown that random variables permit comparison of the outcomes of experiments. In this section, we will enhance our study of random variables, in two senses: *1)* we will highlight that the notion of conditioning can be extended to the random variable theory, and *2)* we will highlight one further tractability of random variables, namely the ability to find their *expectations* or averages.

5.1 Conditional Distributions

Earlier, we naturally defined the probability of one event *conditioned on* another event (given that another event occurred). We note that the set of outcomes of an experiment for which an associated random variable X satisfies $X \leq x$ or $\{X \leq x\}$ is in fact an event. We can thus naturally consider the probability of this event $\{X \leq x\}$ conditioned on another event, say A. From the definition of conditional probabilities, this probability is given by $P(\{X \leq x\} \,|\, A) = \frac{P(\{X \leq x\}, A)}{P(A)}$. Thinking of this probability expression as a function of x, we can interpret these probabilities as a *cumulative distribution function (CDF) for X given A*, which we alternately denote by $F_{X\,|\,A}(x)$. Notice that the conditional CDF for X given A identifies the probabilities that X takes on values in any specified range, given that A has occurred. We notice that a conditional CDF has very similar properties as a CDF; we leave a proof of this analogy as an exercise.

From the conditional CDF, we can naturally define the notion of a conditional probability distribution function. In particular, we refer to the function $f_{X\,|\,A} = \frac{dF_{X\,|\,A}(x)}{dx}$ as the *probability distribution function (PDF) for X given A*. The conditional PDF admits the same interpretation as any PDF, as a *density* of the chance of the random variable falling in an interval (in this case, given A).

Let's do an example to gain insight into conditional CDFs and PDFs:

Since the conditional CDF evaluated at each argument value is simply a conditional probability, we should be able to apply the law of total probability and Bayes' rule to

CDFs. Using the manipulations is helpful, because it allows us to compute CDFs and/or event probabilities of interest from the experiment's description. Here are several results:

- Often, the CDFs/PDFs of a random variable *given* certain events can be found easily, but the unconditioned CDF/PDF of the random variable is desired. In this case, the following equalities are handy: $F_X(x) = F_{X\,|\,A_1}(x)P(A_1) + \ldots + F_{X\,|\,A_n}(x)P(A_n)$, $f_X(x) = f_{X\,|\,A_1}(x)P(A_1) + \ldots + f_{X\,|\,A_n}(x)P(A_n)$, where $A_1, \ldots, A_n$ partition the sample space.

- We can in turn also find the conditional probability of and event A given $X \leq x$, from Bayes' rule: $P(A\,|\,X \leq x) = \frac{F_{X\,|\,A}(x)P(A)}{F(x)}$. In a similar fashion, we can find an expression for $P(A\,|\,X = x) \triangleq lim_{\delta \to 0} P(A\,|\,x < X < x + \delta)$. In particular, we find that this probability is $P(A\,|\,X = x) = \frac{f(x\,|\,A)P(A)}{f(x)}$.

Let us take a minute to justify these results:

- In other experiments, the probability for an event A given values of a random variable X may be known, and the unconditioned probability of A may be desired. In such cases, we can find $P(A)$ as follows: $P(A) = \int_{-\infty}^{\infty} P(A\,|\,X = x) f(x)\,dx$.

- In turn, we can find the probability distribution for X given A: $f_{X\,|\,A}(x) = \frac{P(A\,|\,X=x)f(x)}{P(A)} = \frac{P(A\,|\,X=x)f(x)}{\int_{-\infty}^{\infty} P(A\,|\,X=x)f(x)\,dx}$.

Let us again take a moment to prove these relationships:

Let us also do a couple examples:

Finally, we note that it is especially interesting to consider the conditional distribution for X, given the event that X is in a particular range, i.e. $a \leq X \leq b$. Let us study this case a bit further:

5.2 The Expectation of A Random Variable

Because a random variable associates numbers with experimental outcomes, we can naturally consider the *average value* or *expectation* of a random variable. That is, if we were to run the uncertain experiment many times, we could average the random variable values obtained on each trial, and we might expect this average to reach a constant. Formally, we find it convenient simply to define the expectation of a random variable in terms of the PDF of that random variable. Before presenting this definition, however, let us (somewhat informally) motivate the expression for a random variable's expectation.

To motivate the definition, let us consider repeating the experiment z times, and consider the chance that the random variable X lies on the interval $n\delta$ and $(n+1)\delta$, for all integer n and for arbitrarily small δ (see Figure above). As z becomes large, we know from the central limit theorem that this probability approaches $f_X(n\delta)\delta$, and hence that approximately $zf_X(n\delta)\delta$ trials will yield a random variable in this interval. Thus, the total average of the random variables obtained on these trials should be approximately $\frac{1}{z}\sum_{n=-\infty}^{infty} zf_X(n\delta)\delta(n\delta) = \sum_{n=-\infty}^{\infty}(n\delta)f_X(n\delta)\delta \approx \int_{-\infty}^{infty} xf_X(x)\,dx$. That is, we weight each possible value taken on by X by the probability density of this value, and add (integrate) over possible values.

With this motivation in mind, we define the **expectation** or **mean** or **average** of a random variable X as $E[X] = \int_{-\infty}^{\infty} xf_X(x)\,dx$. For discrete-valued random variables, this definition specializes to $E[X] = \sum_x xp_X(x)$.

Let us practice computing the expectations of some typical random variables:

The above computations crystallize that, interestingly, the expectation of a random variable only depends on the PDF of the random variable; the specifics of the mapping from the underlying probability space are irrelevant.

5.2.1 Calculating the Mean of a Random Variable from the CDF

Interesting, one can compute the mean of a random variable directly from the CDF. Let's take a moment to do this computation:

6 Some Interesting Random Variables

When we are modeling uncertainties in the world around us, we find that many uncertain numerical quantities (random variables) are well described by a few common CDFs/PDFs. Thus, it is worth our while to study (and learn how to manipulate) these common CDFs/PDFs. In fact, we are often so concerned with such uncertain quantities, rather than their underlying probabilistic experiment, that we find it convenient to study these common CDFs/PDFs without specifying the underlying experiment. Let us take a moment to convince ourselves that thinking about PDFs/CDFs on their own is actually sensible:

The above concept is known as the **existence theorem.**

Now we believe that studying random variables and hence PDFs/CDFs on their own is couth, so let us describe some common PDFs/CDFs for random variables. I feel that your text's (Papoulis') descriptions of these common distibutions is thorough and instructive, so I've copied the descriptions directly from the text.. Let us, however, also take a moment to describe common settings for these uncertainties:

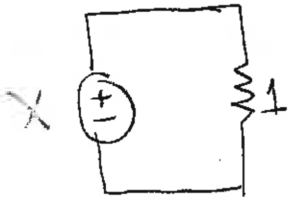
Lecture 2b: Functions of a Single Random Variable

Reading: Papoulis, Chapter 4

September 6-13, 2006

1 Finding PDFs/CDFs for Functions of Random Variables

Often, we are interested in characterizing a function $g(X)$ of a random variable X. For instance, consider the following circuit example, where the constant source voltage X is an exponentially-distributed random variable:



$\text{Power} = \frac{V^2}{R} = X^2$

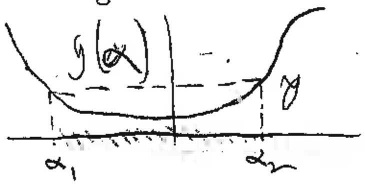
We might want to study the power dissipation in the resistor, which is given by $g(X) = \frac{X^2}{1} = X^2$.

We notice that a (deterministic) function $Y = g(X)$ of a random variable X is itself another random variable defined on the same sample space. This is because each outcome of the experiment maps to a particular number X and in turn X maps to a single $Y = g(X)$, so there is a mapping from each outcome to a value Y.

Since the random variable Y is generated deterministically from the random variable X, we might expect that the CDF and PDF of Y can be found from the CDF and PDF of X. This turns out to be true, and in fact there is a systematic procedure for deciding the CDF/PDF of Y from the CDF/PDF of X. In particular, let us consider the CDF of Y:

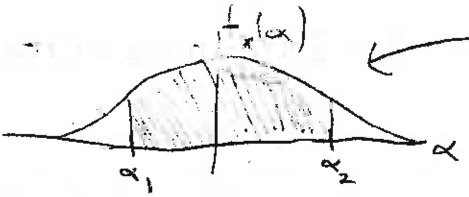
$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(g(X) \leq y). \end{aligned} \tag{1}$$

That is, $F_Y(y)$ is equal to the probability of the event $\{g(X) \leq y\}$. This is the event comprising the outcomes of the experiment for which $g(X)$ is less than y. Thus, to find this probability, all we need to do is to figure out the outcomes that satisfy $g(X) \leq y$, and find the probability of this event. Notice that $g(X) \leq y$ is in general satisfied for ranges of the random variable X (i.e., for outcomes whose corresponding values for X fall in certain ranges), and hence we can find the probability of $\{g(X) \leq y\}$ from the CDF or PDF of X. Once the CDF of Y has been found, the PDF of Y can be determined by taking the derivative of the CDF. This technique for finding the CDF/PDF of Y is diagrammed below:



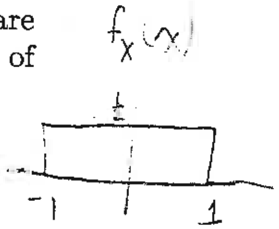
$P(Y \leq y) = P(X$ in marked range$)$ ← this can be found from PDF or CDF

$F_X(\alpha_2) - F_X(\alpha_1)$



Next, let us pursue several examples to illustrate our strategy for finding the CDF/PDF of Y. Several more examples can be found in your text (Papoulis).

1) Say that X is uniformly distributed in the interval $[-1, 1]$, and $Y = X^2$. What are the CDF/PDF of Y? More generally, if X has CDF $F_X(x)$, what are the CDF/PDF of Y in terms of the CDF of X?



$$F_Y(y) = P(Y \leq y) = P(X^2 \leq y)$$
$$= P(-\sqrt{y} \leq X \leq y)$$
$$= \begin{cases} 0, & y < 0 \\ \int_{-\sqrt{y}}^{\sqrt{y}} \frac{1}{2}\,dx, & 0 \leq y < 1 \\ 1, & y \geq 1 \end{cases}$$

$$F_Y(y) = \begin{cases} 0, & y < 0 \\ \sqrt{y}, & 0 \leq y < 1 \\ 1, & y \geq 1 \end{cases}$$

$$\boxed{f_Y(y) = \frac{1}{2\sqrt{y}}, \quad 0 \leq y < 1}$$

- Note that $F_Y(y)$ can also be found as $F_X(\sqrt{y}) - F_X(-\sqrt{y})$

Let's take a moment to interpret the result of this computation. In particular, we notice that Y is more likely to take values near 0 than values that are far away. This makes sense because $g(X)$ is flat near $X = 0$, so that many values X and hence many outcomes map to values of Y near 0.

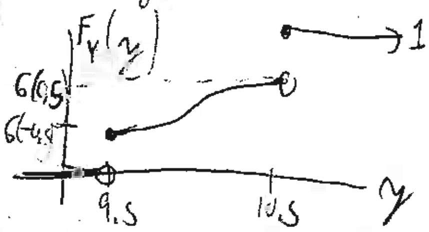
2) Consider a particular mixture containing strands of DNA. If we were to choose one strand at random from this mixture, the mass of this strand is well-described a Gaussian random with mean 10 and standard deviation 1. When we isolate a strand, we can actually measure its mass, but our measuring device saturates: masses less than 9.5 are measured as 9.5, and masses greater than 10.5 are measured as 10.5 (see figure below). Let Y be the measured mass of a randomly-chosen strand. What are the CDF/PDF of Y?

$F_Y(y) = P(Y \leq y)$

For $y < 9.5$, $P(Y \leq y) = 0$

For $9.5 \leq y < 10.5$,
$P(Y \leq y) = P(X \leq y)$
$= G\left(\frac{y-10}{1}\right)$

For $y \geq 10.5$
$P(Y \leq y) = 1$



$f_Y(y)$
$= G(-0.5)\delta(y-9.5) + g(y-10) + (1 - G(0.5))\delta(y-10.5)$
for $-9.5^- < y < 10.5^+$,
and $f_Y(y) = 0$ otherwise

3) Let's say X is a random variable with given distribution $F_X(x)$. In terms of $F_X(x)$, what are the CDF/PDF of $Y = sign(X)$? What about of $Z = aX + b$? And of $A = F_X(X)$?

$$F_Y(y) = P(Y \leq y) = \begin{cases} 0, & y < -1 \\ F_X(0), & -1 \leq y < 1 \\ 1, & y \geq 0 \end{cases}$$

$$f_Y(y) = F_X(0)\delta(y+1) + (1 - F_X(0))\delta(y-1)$$

$$F_Z(z) = P(Z \leq z) = P(aX + b \leq z) = P\left(X \leq \frac{z-b}{a}\right).$$

$$\boxed{F_Z(z) = F_X\left(\frac{z-b}{a}\right)}$$

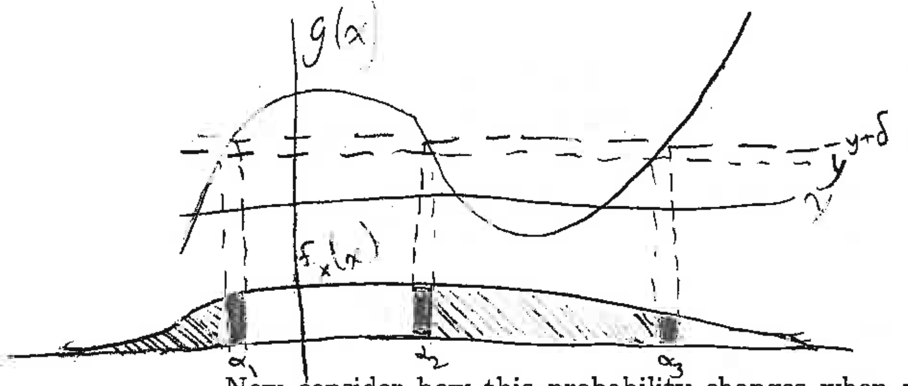
$$f_Z(z) = \frac{d}{dz} F_Z(z) = \frac{d}{dz} F_X\left(\frac{z-b}{a}\right)$$

$$\boxed{f_Z(z) = \frac{1}{a} f_X\left(\frac{z-b}{a}\right)}$$

1.1 A Clever Formula for the PDF of Y

Above, we have developed a general strategy for finding the CDF/PDF of Y from the CDF or PDF of X. This strategy is essentially based on finding the CDF of Y from that of X, and in turn computing the PDF. Since many random variables are standardly described by their PDFs, it would be helpful (in the sense of reducing computational effort and making the computation more systematic) to have a simple formula for computing the PDF of Y in terms of that of X. In fact, for most mappings, a clever and simple formula can be developed.

To develop this simple formula, let us take the general approach considered above. In particular, we note that $F_Y(y) = P(Y \leq y) = P(g(X) \leq y)$. For most mappings, there are a countable set of points $\alpha_1, \ldots, \alpha_n$ such that $g(\alpha_i) = y$, for each y. In such cases, there is a simple explicit formula for $f_Y(y)$. For simplicity, let us study the case where there are three values α for which $g(\alpha) = y$, i.e. $n = 3$. Our analysis for this case will generalize readily to the case of arbitrary n. The probability $F_Y(y)$ can be found readily from the PDF of X in this case, as indicated by the figure and expression below:

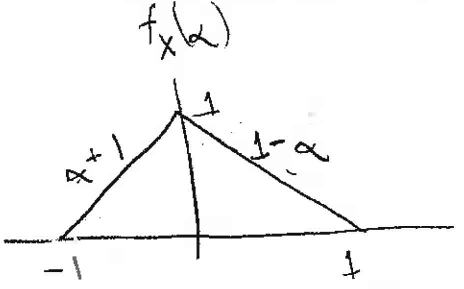


▨ : Area is probability that $Y \le y$

■ : Area is $P(Y \le y+\delta) - P(Y \le y)$

Now consider how this probability changes when we increment y slightly, i.e. let us find $F_Y(y+\delta) - F_Y(y)$. From the figure above, we see that $F_Y(y+\delta) - F_Y(y) \approx \sum_{i=1}^{3} \frac{\delta}{|\frac{dg(\alpha_i)}{dx}|} f_X(\alpha_i)$, with the approximation becoming an equality in the limit of small δ. Thus, $f_Y(y) = \lim_{\delta \to 0} \frac{F_Y(y+\delta)-F_Y(y)}{\delta} = \sum_{i=1}^{3} \frac{f_X(\alpha_i)}{|\frac{dg(\alpha_i)}{dx}|}$. For a general n, we thus notice that $f_Y(y) = \sum_{i=1}^{n} \frac{f_X(\alpha_i)}{|\frac{dg(\alpha_i)}{dx}|}$. Thus, we have given an explicit formula of $f_Y(y)$ in terms of $f_X(x)$ and the mapping $g()$.

Let us apply this formula to an example:



$Y = X^2$: $f_Y(y) = ?$

$g(\alpha) = y$ has roots $-\sqrt{y}, \sqrt{y}$

Thus for $0 \le y < 1$,

$$f_Y(y) = \frac{f_X(\sqrt{y})}{|\frac{d}{dx}(x^2)|_{\sqrt{y}}|} + \frac{f_X(-\sqrt{y})}{|\frac{d}{dx}(x^2)|_{-\sqrt{y}}|} = \frac{1-\sqrt{y}}{2\sqrt{y}} + \frac{-\sqrt{y}+1}{2\sqrt{y}}$$

$$= \frac{1-\sqrt{y}}{\sqrt{y}} = \frac{1}{\sqrt{y}} - 1$$

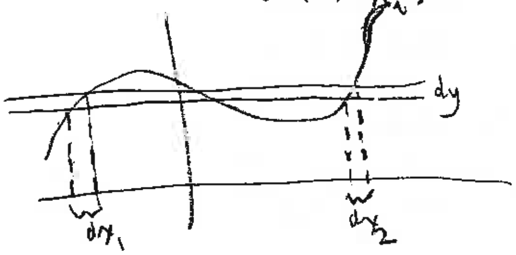
Thus, $\boxed{f_Y(y) = \frac{1}{\sqrt{y}} - 1, \ 0 \le y \le 1}$

2 Expectations of Functions of RVs

Here, we study the expectation (average) of a function of a random variable. In particular, we first show that the expected value for $Y = g(X)$ can be computed simply from the PDF of X. We then study the expectations of some particular functions of X, which give some special insight into the random variable X. We also compute such expectations for random variables with common probability distributions, so that we have a table of *statistics* ready for use.

2.1 Computing the Expectation of $Y = g(X)$

Noting that $Y = g(X)$ is itself a random variable defined on the same sample space as X, we can compute the expected value of Y as $E[Y] = \int_{-\infty}^{\infty} y\, f_Y(y)\, dy$. From this expression, it seems that we must find the the PDF for Y in order to compute its expectation. In fact, however, we can compute the expectation of Y directly from the PDF of X. To write an expression in terms of $f_X(x)$, we notice that $f_Y(y)\,dy$ can be written as sum of differentials of the form $f_X(x_i)\,dx_i$.



Notice that $f_Y(y)\,dy = f_X(x_1)\,dx_1 + f_X(x_2)\,dx_2$

$$\Downarrow$$

$$\int_{-\infty}^{\infty} y\, f_Y(y)\,dy = \int_{-\infty}^{\infty} g(x) f_X(x)\,dx$$

We thus recover that $E[Y] = E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx$.

Let's do an example, to practice applying the computation above. In particular, consider a random variable X that has P.D.F. $f_X(\alpha) = 3\alpha^{-4}$, for $\alpha \geq 1$. Let us find the expectation of X^a for each $a > 0$:

$$E[X^a] = \int_1^{\infty} x^a (3x^{-4})\, dx$$

$$E[X^a] = 3\int_1^{\infty} x^{a-4}\, dx = \frac{x^{a-3}}{a-3}\Big]_1^{\infty} = 0 - \left(\frac{1}{a-3}\right), a<3$$

$$\boxed{E[X^a] = \frac{1}{3-a}, a<3}$$

$$\boxed{E[X^\alpha] \text{ D.N.E, for } \alpha \geq 3}$$

The above example gets us to think about a couple interesting questions: em 1) Does the random variable X always have an expectation? *2)* If X has an expectation, is it necessarily true that the expectation of $g(X)$ will exist? We can easily see that both statements are untrue: if $f_X(x)$ falls off slowly with x, then $E[X]$ may not exist. Also, unless $g()$ grows linearly with respect to its argument, then the expectation of $g(X)$ may not exist even if the expectation of X does.

For a random variable of discrete time, we note that the expectation of a function of a random variable can be computed as follows: $E[g(X)] = \sum_i g(x_i) P(X = x_i)$, where the sum is taken over all values x_i that the random variable X can take on.

2.2 An Important Expectation: the Variance

For simplicity, let us refer to the expectation of a random variable X as u in the following.

We are often interested in the expectation $E[(X - E[X])^2] = E[(X - u)^2]$, because this expectation indicates the *spread* of the random variable around its average value.

This expectation is known as the **variance** of the random variable X. We can directly compute the variance from the PDF of X, as $\int_{-\infty}^{\infty}(x-u)^2 f_X(x)\,dx$.

We can also develop another convenient expression for the variance with a little bit of algebra: we note that $E[(X-u)^2] = E[X^2 - 2uX + u^2] = E[X^2] - 2uE[X] + u^2$, where the last equality follows from the linearity of the expectation operator. We thus recover that $E[(X-u)^2] = E[X^2] - u^2 = E[X^2] - (E[X])^2$. This expression is often a convenient way to find the variance, because the expectation $E[X^2]$ can be found with (comparatively) little effort, and the expectation $E[X]$ is already known.

For practice, let us find the variances of a couple random variables:

Example 1: Variance of a Gaussian Random Variable

$$E(X^2) = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} x^2 e^{-\frac{(x-\mu)^2}{2\sigma^2}} \doteq ?$$

- We can evaluate this integral by parts.

- Instead, let us play a trick:

Note that

$$\int_{-\infty}^{\infty} e^{-\frac{(x-u)^2}{2\sigma^2}}\,dx = \sigma\sqrt{2\pi}$$

Taking derivatives of each side w/ respect to σ, we get

$$\int_{-\infty}^{\infty} \frac{(x-u)^2}{\sigma^3} e^{-\frac{(x-u)^2}{2\sigma^2}} = \sqrt{2\pi}$$

$$\Rightarrow \int_{-\infty}^{\infty} \frac{(x-u)^2}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-u)^2}{2\sigma^2}} = \sigma^2$$

i.e. $E[(X-u)^2] = \sigma^2$

Example 2: Mean and Variance of a Poisson Random Variable

$$P(X=k) = (e^{-a})\frac{a^k}{k!},\ k=0,1,\ldots$$

$$E[X] = \sum_{k=0}^{\infty}(e^{-a})\frac{a^k}{k!}\cdot k = e^{-a}\sum_{k=1}^{\infty}\frac{a^k}{(k-1)!} = (e^{-a})(a)\sum_{k=1}^{\infty}\frac{a^{k-1}}{(k-1)!} = e^{-a}\cdot a\cdot\sum_{l=0}^{\infty}\frac{a^l}{l!} = e^{-a}\cdot a\cdot e^{a} = \boxed{a}$$

$$E[X^2] = a^2 + a$$

Prove this in a very similar way to the expression for $E[X]$

Thus,

$$\text{Variance} = E[X^2] - (E[X])^2 = a^2 + a - a^2 = \boxed{a}$$

2.3 Other Important Statistics (Expectations): Moments

The mean indicates the average value of a random variable, while the variance roughly indicates the spread of the random variable around the mean. These quantities can thus be viewed as *statistics*, i.e. scalars that condense the information contained in a random variable's PDF, in such a way that one important feature of the random variable is highlighted. It turns out that expectations of powers of a random variable, which are known as *moments*, are particularly important statistics of the random variable. In particular, statisticians commonly consider the moments $m_n = E[X^n]$ and central moments $u_n = E[(X-u)^n]$. The higher central moments, in particular, give indications about the shape of the PDF of X. There are several expressions relating moments and central moments, as well as other statistics of a random variable. In the interest of brevity, we will not pursue these relationships further. Your text contains a good summary of these relationships. I'll also get you to think further about moments in a homework problem on the Gaussian random variable.

2.4 Another Important Statistic: the Moment-Generating Function

Please think back for a minute to your undergraduate courses on circuits and systems. In these courses, you often used *transforms* (e.g., the Laplace transform or Fourier transform) to solve differential equations. These transforms were equivalent representations for a time function, that were generated by computing a function of another variable from the time function; the transforms were useful because they allowed us to solve differential equations using algebraic techniques.

Based on this background in linear systems, we might wonder whether transforms of probability density functions can be found, and whether these transforms might simplify computations of moments and other probability distributions. In fact, transformations of PDFs can indeed be used to simplify the computations of moments, as well as PDFs/CDFs of functions of the corresponding random variable.

The transform that turns out to be useful is in fact very similar to the one we use regularly in linear systems. Specifically, we consider the transform

$$G_X(s) \triangleq \int_{-\infty}^{\infty} f_X(\alpha)e^{s\alpha}\,d\alpha = E[e^{sX}],$$

where s is in general a complex number. It's worth noting that the transform, which looks an awful lot like a Laplace transform (but with e^{sx} rather than e^{-sx}), can in fact be viewed as the expectation of a function of X, namely e^{sX}. This transform is called the **moment-generating function** for X (for a reason that will become clear shortly).

We also find it convenient to evaluate the moment-generating function along the complex axis, i.e. for $s = jw$. Doing so, we obtain the transform

$$G_X(jw) = \int_{-\infty}^{\infty} f_X(\alpha)e^{jw\alpha}\,d\alpha = E[e^{jwX}] \triangleq G_X(w). \tag{2}$$

This transform, which is analogous to Fourier transform in linear systems analysis, is known as the **characteristic function**.

In a moment, we will learn how the moment-generating function and characteristic function can be used to characterize and relate random variables. Before pursuing these applications, let us first practice finding the moment-generating function and characteristic function, with an example:

$$f(x) = \frac{1}{x_2 - x_1},\ x_1 \le x \le x_2$$

$$G_X(s) = E[e^{sX}] = \int_{x_1}^{x_2} \frac{1}{x_2 - x_1} e^{s\alpha}\,d\alpha$$

$$= \frac{1}{x_2 - x_1} \frac{e^{s\alpha}}{s}\Bigg]_{x_1}^{x_2} = \boxed{\frac{1}{x_2 - x_1}\frac{\left[e^{sx_2} - e^{sx_1}\right]}{s}}$$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$$

$$G_X(s) = \int_{-\infty}^{\infty} e^{sx} e^{-\frac{x^2}{2\sigma^2}} \cdot \frac{1}{\sigma\sqrt{2\pi}}\,dx$$

$$= \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{x}{\sigma} - \sigma s\right)^2} e^{\frac{s^2\sigma^2}{2}} \cdot \frac{1}{\sigma\sqrt{2\pi}}\,dx$$

$$= \boxed{e^{\frac{s^2\sigma^2}{2}}}$$

Let us now pursue several uses of the moment-generating function and characteristic function:

2.4.1 Application: Finding Moments from the Moment-Generating Function

The moment-generating function for a random variable X can straightforwardly be used to find the moments of X. To see how, notice that

$$\frac{d^n F(s)}{ds^n} = \frac{d^n\, E[e^{sX}]}{ds^n} = E[X^n e^{sX}]$$

Thus, we see that

$$\frac{d^n F(s)}{ds^n}|_{s=0} = E[X^n]. \tag{3}$$

That is, we can find the nth moment of X by computing the nth derivative of the moment-generating function with respect to s, and plugging in $s = 0$.

Let's do an example that illustrates how the moment-generating function can be used to find the expectation of a random variable:

Exponential R.V. : $f_X(x) = \lambda e^{-\lambda x},\ x \geq 0$

$$G_X(s) = \frac{\lambda}{\lambda - s}$$

$$E[X] = \frac{dG_X(s)}{ds}\Big|_{s=0} = \frac{\lambda}{(\lambda-s)^2}\Big|_{s=0} = \frac{\lambda}{\lambda^2} = \frac{1}{\lambda}$$

$$E[X^2] = \frac{d^2}{ds^2} G_X(s)\Big|_{s=0} = \frac{2\lambda}{(\lambda-s)^3}\Big|_{s=0} = \frac{2}{\lambda^2}$$

$$\Rightarrow Var(X) = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \boxed{\frac{1}{\lambda^2}}$$

It's worth noting that, in finding statistics, the moment-generating function is serving the same purpose as the Laplace transform in linear systems analysis: the derivative of a function is one domain is being equivalenced with the multiplication of the function by a power of the argument in the other domain. In linear systems analysis, derivatives are typically transformed algebraic expressions, so as to simply compute the solution; the opposite methodology is used in probability theory.

2.5 Application: Understanding the Connection between Moments and PDF

The following is a question of both academic and practical interest: if we know all the moments of a random variable, can we come up with the PDF of the random variable? The moment-generating function helps us to answer this question, because we know that the PDF of a random variable can be obtained if the moment-generating function can be found (by taking a "inverse Laplace transform"'). Thus, by noting that the moment generating function can be written as $F_X(s) = \sum_{n=0}^{\infty} \frac{E[X^n]}{n!} s^n$, we see that the moment-generating function and hence the PDF can be computed from the moments as long as the moments are in fact finite, and further the series for $F_X(s)$ converges absolutely for s near 0. Thus, we have obtained a broad set of sufficient conditions (on the moments) given which the PDF can be computed from all the moments.

2.5.1 Application: Easily Finding the PDF for $g(X)$

In some cases, the characteristic function of X can also be used to easily find the distribution of a function of X. Let us expose how this is done through an example:

- X is uniform in $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$
- What is $f_Y(y)$, where $Y = \sin(X)$?

$$G_Y(jw) = \int_{-\infty}^{\infty} e^{jw\cdot\sin(x)} f_X(x)\, dx = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} e^{jw\cdot\sin(x)} \cdot \frac{1}{\pi}\, dx$$

- Let $y = \sin(x)$
- Then $dy = \cos(x)\, dx = \sqrt{1-y^2}\, dx$,

and so $G_Y(jw) = \frac{1}{\pi} \int_{-\infty}^{\infty} e^{jwy} \frac{1}{\sqrt{1-y^2}}\, dy$

$\boxed{\text{Thus, } f_Y(y) = \frac{1}{\pi\sqrt{1-y^2}}}$

2.5.2 Characteristic Function for R.V.s of Discrete Type

Characteristic Functions can analogously be determined for random variables of discrete type.

Characteristic function: $\sum_i P(X=x_i) e^{jwx_i}$

If R.V. takes on only discrete values, the following is preferable

Moment Gen. Funct. $= \Gamma(z) = E[z^X] = \sum_{n=-\infty}^{\infty} P(X=n) z^n$

Char. Function $= \Gamma(e^{jw}) = \sum_{n=-\infty}^{\infty} P(X=n) e^{jwn}$

Moment Theorem: $E[X(X-1)\cdots(X-k+1)] = \left.\frac{d^k \Gamma(z)}{dz^k}\right|_{z=1}$

LECTURE 2C: PAIRS OF RANDOM VARIABLES

Section 1

In this third part of the lecture on random variables, we will consider how two random variables defined on a sample space can be characterized. This study of pairs of random variables will allow us to better understand how multiple uncertain quantities in an experiment are related. Although we focus our discussion on pairs of random variables, the notions that we develop can naturally be extended to three or more random variables. Thus, this study of pairs of random variables is a natural first step toward understanding random signals (random processes).

The remainder of the section is organized as follows:

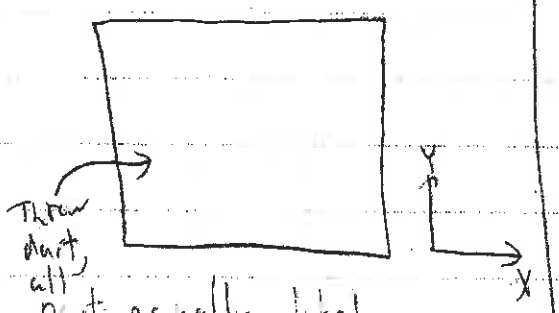
- Section 2: We introduce the notions of joint PDFs/CDFs, conditional PDFs/PMFs, and independence.
- Section 3: We discuss functions of two random variables; in particular, we show how PDFs/CDFs of these functions can be found. In turn, we explore how statistics of a pair of random variables can be computed, and introduce some particularly useful statistics.

Section 2: CDFs/PDFs of pairs of R.V.s

Often, we wish to understand the joint characteristics of two uncertain quantities (random variables), or to characterize one given the other. Such problems motivate us to try to find the probability that two random variables are in specified sets (e.g., the probability that one random variable is smaller than the other), as well as aggregate statistics of random-variable pairs. For single random variables, the CDF or PDF provides us with enough information to permit computation of the probability that the R.V. is in any given set. We might wonder if a similar notion of a PDF/CDF can be defined for a pair of random variables, such that probabilities of the R.V.s falling in specified sets can be easily obtained. Our goal here is to appropriately define CDFs/PDFs for pairs of random variables, and understand how these CDFs/PDFs can be used to compute probabilities of interest.

Consider two random variables X and Y. A first query of interest is whether the individual CDFs/PDFs of X and Y permit us to compute the probabilities that the RVs X and Y are in given sets. In fact, the individual PDFs/CDFs are not enough, as shown in the following example:

Consider the following two experiments:



X: horiz. position of dart location

Y: vertical position of dart location

⇓

For this experiment, X and Y are uniform on (0,1)

$P(X>Y)=\frac{1}{2}$

X is uniform on [0,1]

$Y=X$

⇓

For this experiment, X and Y are also uniform on [0,1].

However, $P(X>Y)=0$.

- Thus, we find the PDFs of X and Y are not sufficient to compute the probability i.e. the probability that (X,Y) is in a given set (in this case, $X>Y$).

This example indicates to us that an appropriate CDF/PDF for a pair of random variables must codify probabilities for intersections/unions of events defined from the two random variables. If our notion for a CDF/PDF captures probabilities of intersections (or equivalently unions, since we can go back and forth), we should be able to find the probability that the random variables are in any set in $\mathbb{R}^2$.

With this motivation in mind, we define the CDF of a pair of random variables as $F_{X,Y}(x,y) = P(X \leq x, Y \leq y) = P(X \leq x \cap Y \leq y)$.

as before, it's easiest to start by thinking about CDFs

That is, we call the probability that X is in the range $[-\infty, x]$ AND Y is in the range $[-\infty, y]$, as the CDF of X and Y, evaluated at the pair (x, y).

<u>Properties of the CDF</u>

Let us list some properties of the CDF, which make computation of probabilities for events associated with the random variables easier.

1. Drawn with respect to the plane, the CDF is continuous from the right and from above

2a. $P(X\le x)=F_X(x) = F_{X,Y}(x,\infty)$.

2b. $P(Y\le y)=F_Y(y)=F_{X,Y}(\infty,y)$.

3. $P(X\le x \cup Y\le y)=F_X(x)+F_Y(y)-F_{X,Y}(x,y)$.

4. $P(x_1\le X\le x_2, y_1\le Y\le y_2)$
$= F_{X,Y}(x_2,y_2)-F_{X,Y}(x_1,y_2)-F_{X,Y}(x_2,y_1)+F_{X,Y}(x_1,y_1)$.

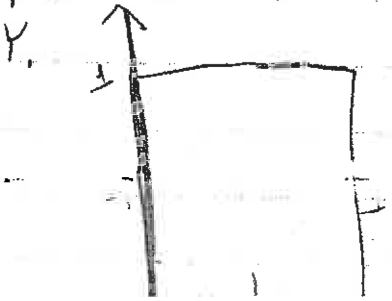
5. $F(-\infty,y)=F(x,-\infty)=0,\quad F(\infty,\infty)=1$

These results are easy to prove, so I haven't included the details here. Using the above, notice that we can easily find the probabilities that the random variable falls within rectangular regions (and their union). However, we can't easily find $P(X<Y)$, for instance.

Let us do a couple examples to learn to draw joint CDFs, and find probabilities from them.

Example 1

We throw a dart, which is equally likely to hit any point on the square dartboard shown below. We define the horizontal position of the dart by X, and the vertical position by Y.



Let us find the CDF for this example:

- If $x_1 < 0$ or $y_1 < 0$, $F_{X,Y}(x_1, y_1) = 0$
- If $0 \le x_1 < 1$, $0 \le y_1 < 1$, $F_{X,Y}(x_1, y_1) = x_1 y_1$
- If $0 \le x_1 < 1$, $y_1 \ge 1$, $F_{X,Y}(x_1, y_1) = x_1$
- If $x_1 \ge 1$, $0 \le y_1 < 1$, $F_{X,Y}(x_1, y_1) = y_1$

If $x_1 \ge 1$, $y_1 \ge 1$, $F_{X,Y}(x_1, y_1) = 1$

Using properties of the CDF, we can find the probability that the random variable falls in a rectangular region, e.g. $P(0.2 \le X \le 0.8,\ 0.3 \le Y \le 0.4)$

This probability equals

$$F(0.8, 0.4) - F(0.8, 0.3) - F(0.2, 0.4) + F(0.2, 0.3)$$

$$= 0.32 - 0.24 - 0.08 + 0.06 = \boxed{0.06}$$

As with single random variables, pairs of random variables are often most naturally specified in terms of probability densities instead of cumulative probabilities. This motivates us to define a joint probability density function (PDF) of a pair of random variables. The PDF is actually further important for the two-R.V. case, because probabilities of the random variable falling in non-rectangular regions are difficult to find from the CDF.

It is natural for us to define the probability density as the probability of the random variables falling in a small square region $(dx \times dy)$, divided by the size of the region. This leads to the following definition for the joint PDF:

$$f_{X,Y}(x_1, y_1) = \frac{\partial^2}{\partial x_1 \partial y_1}\left[F(x_1, y_1)\right]$$

Let us present some properties of the joint PDF:

1. $F_{X,Y}(x_1, y_1) = \int_{-\infty}^{x_1} \int_{-\infty}^{y_1} f_{X,Y}(\alpha, \beta)\, d\beta\, d\alpha.$

2. The probability that the pair (X, Y) is in a particular set D is
$$\iint_D f_{X,Y}(\alpha, \beta)\, d\beta\, d\alpha.$$

3. $f_{X,Y}(x_1, y_1) \geq 0$ for all x_1, y_1, and $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x_1, y_1)\, dy_1\, dx_1 = 1.$

4. $$\left.\begin{aligned} f_X(x_1) &= \int_{-\infty}^{\infty} f_{X,Y}(x_1, y_1)\, dy_1 \\ f_Y(y_1) &= \int_{-\infty}^{\infty} f_{X,Y}(x_1, y_1)\, dx_1 \end{aligned}\right\} \text{Their individual P.D.F.s of } X \text{ and } Y \text{ are called marginal PDFs.}$$

It's worth thinking a bit further about the case that (X, Y) take on particular values with non-zero probability, i.e. $P(X=1, Y=2) = 0.3$. In this case, the CDF will have a staircase shape, and the PDF will have the term $0.3\,\delta(x_1-1)\,\delta(y_1-2)$

Also of interest, given a value of X, the R.V. Y may have a particular value with some non-zero probability (e.g., Y may be a function of X) In such cases, the PDF will be impulsive along a line, i.e. the PDF will have a term of the form $\delta(y_1 - g(x_1))$.

Let's return to the dartboard example:

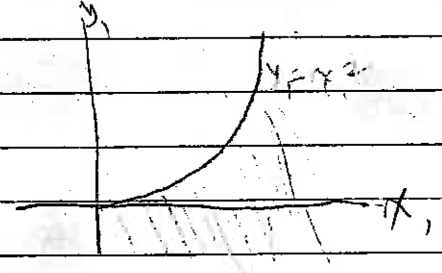
From the problem description, it is evident that X and Y are uniformly distributed in the square region $(0 \le x_1 \le 1,\ 0 \le y_1 \le 1)$. That is,

$$f_{X,Y}(x_1, y_1) = \begin{cases} c, & 0 \le x_1 \le 1,\ 0 \le y_1 \le 1 \\ 0, & \text{otherwise} \end{cases}$$

Since $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x_1, y_1)\,dy_1\,dx_1 = 1$, we see that $C = 1$

Thus, $$f_{X,Y}(x_1, y_1) = \begin{cases} 1, & 0 \le x_1 \le 1,\ 0 \le y_1 \le 1 \\ 0, & \text{otherwise} \end{cases}$$

Using the PDF, we can find the probability that (X, Y) is in any specified region. For instance, let us compute the probability that $Y < X^2$



$$P(Y \le X^2) = \int_0^1 \int_0^{x_1^2} 1 \; dy_1\, dx_1$$
$$= \int_0^1 y_1 \Big]_0^{x_1^2} dx_1$$
$$= \int_0^1 x_1^2\, dx_1 = \frac{x_1^3}{3} = \boxed{\frac{1}{3}}$$

Example 2: Jointly Gaussian R.V.s

Two random variables X and Y are said to be jointly Gaussian if

$$f_{X,Y}(x_1, y_1) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} \exp\left\{-\frac{1}{2(1-r^2)}\left[\frac{(x_1-\eta_1)^2}{\sigma_1^2} + \frac{(y_1-\eta_2)^2}{\sigma_2^2} - 2r\frac{(x_1-\eta_1)(y_1-\eta_2)}{\sigma_1\sigma_2}\right]\right\},$$

where η_1 and η_2 are real numbers, σ_1 and σ_2 are positive numbers, and r is between 0 and 1.

Claim: The marginal P.D.F.s for X and Y are as follows:

$$X \sim \mathcal{N}(\eta_1, \sigma_1^2), \quad Y \sim \mathcal{N}(\eta_2, \sigma_2^2)$$

Proof

$$f_Y(y_1) = \int_{-\infty}^{\infty} f_{X,Y}(x_1, y_1)\,dx_1$$

To continue, notice that the term in the exponent of $f_{X,Y}(x_1, y_1)$ can be written (through completing the square) as

$$\frac{1}{2(1-r^2)}\left(\left[\frac{(x_1-\eta_1)}{\sigma_1} - r\frac{y_1-\eta_2}{\sigma_2}\right]^2 + (1-r^2)\frac{(y_1-\eta_2)^2}{\sigma_2^2}\right)$$

Thus, $\int_{-\infty}^{\infty} f_{X,Y}(x_1, y_1)\,dx_1 = \exp\left[\frac{-(1-r^2)(y_1-\eta_2)^2}{2(1-r^2)\sigma_2^2}\right] \cdot \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}}\int_{-\infty}^{\infty} \exp\left[-\left(\frac{x_1-\eta_1}{\sigma_1} - r\frac{y_1-\eta_2}{\sigma_2}\right)^2 \frac{1}{2(1-r^2)}\right] dx_1$

The integral in the previous expression is a constant (doesn't depend on y_1)... do you see why??

Thus $f_Y(y_1) = C \exp\left[\frac{(y-\eta_2)^2}{2\sigma_2^2}\right]$

9

Notice that, in general, it seems difficult to compute the CDF of (X, Y) in a closed form or even in a way that allows us to read off the CDF from a table. We will return to this dilemma a little later in the course.

At least, let's think a bit more about the PDF of jointly Gaussian random variables. In particular, let us sketch contours in the plane along which the PDF is a constant. It is our contention that these contours are ellipses. For instance, consider the Gaussian density with $\mu_1 = \mu_2 = 0$, $\sigma_1 = 1$, $\sigma_2 = 2$, $r = \frac{1}{2}$

Notice that a contour of constant probability can be found by finding x and y such that

$$\frac{x^2}{1} + \frac{y^2}{4} - \frac{xy}{2} = C, \text{ for each positive constant } C.$$

For instance, we can find x, y such that

$$x^2 + \frac{y^2}{4} - \frac{xy}{2} = 1$$

We notice that this is the equation for an ellipse. Do you remember how to plot the ellipse?

$$\frac{y^2}{4} - \frac{y}{2} = 0 \qquad \left(x - \frac{y}{4}\right)^2$$

$$y\left(\frac{y}{4} - \frac{1}{2}\right) = 0$$

$$y = 0$$

Conditional P.D.F's and Independence

Now that we are considering two random variables, it is sensible to consider the PDF/CDF of one variable given the value of (or range of values for) the other.

First, we can consider

$$F_Y(y \mid x_1 < X \le x_2) \triangleq P(Y \le y \mid x_1 < X \le x_2)$$

Note that $F_Y(y \mid x_1 < X \le x_2) = \dfrac{P(Y \le y,\ x_1 < X \le x_2)}{P(x_1 < X \le x_2)} = \dfrac{F(x_2, y) - F(x_1, y)}{F_X(x_2) - F_X(x_1)}$

In fact, we are often concerned with the CDF of Y given a particular value of X. In particular, let us define such a CDF in a limiting sense:

$$F_{Y|X}(y \mid X = x_1) \triangleq \lim_{\delta \to 0} F_Y(y \mid x_1 < X \le x_1 + \delta).$$

As always let us define the PDF of the random variable Y given X as the derivative of the CDF, i.e.,

$$f_Y(y \mid x_1 < X \le x_2) \triangleq \frac{d}{dy} F_Y(y \mid x_1 < X \le x_2) \text{ and}$$

$$f_{Y|X}(y \mid X = x) \triangleq \frac{d}{dy} F_{Y|X}(y \mid X = x)$$

With a little bit of work, we find that

$$f_{Y|X}(y \mid X = x) = \frac{f_{X,Y}(x, y)}{f_X(x)}$$

Similarly, note that $f_{X|Y}(x \mid Y = y) = \dfrac{f_{X,Y}(x, y)}{f_Y(y)}$ 11

Of course, the rule of total probabilities and Bayes' rule generalize naturally to the two-random-variable case:

$$\text{Total Probability: } f_Y(y) = \int_{-\infty}^{\infty} f_{Y|X}(y|X=x) f_X(x)\,dx$$

$$\text{Bayes' rule: } f_{X|Y}(x|Y=y) = \frac{f_{Y|X}(y|X=x) f_X(x)}{f_Y(y)}$$

Two random variables are said to be independent, if $f_{X|Y}(x_1|Y=y_1) = f_X(x_1)$, for all x_1 and y_1.

We can easily deduce that independence implies
$f_{Y|X}(y_1|X=x_1) = f_Y(y_1)$, and

$$f_{X,Y}(x_1,y_1) = f_X(x_1) f_Y(y_1),$$

and further that these conditions imply independence. Often, the product expression $(f_{X,Y}(x,y) = f_X(x) f_Y(y))$ is used to check independence.

Example

Two jointly Gaussian random variables have joint distribution

$$f_{X,Y}(x,y) = \frac{1}{2\pi \frac{\sqrt{3}}{2}} e^{-\left(\frac{1}{2\cdot\frac{3}{4}}\left[x^2 - xy + y^2\right]\right)}$$

What is $f_{Y|X=1}(y|X=1)$? Are X and Y independent?

- We know that $X \sim N(0,1)$, so $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, and $f_X(1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}}$

Thus, $f_{Y|X=1}(y|X=1) = \dfrac{f_{X,Y}(1,y)}{f_X(1)} = \dfrac{\left(e^{-\frac{2}{3}(y^2-y+1)}\right)\left(\frac{1}{2\pi\cdot\frac{\sqrt{3}}{2}}\right)}{e^{-\frac{1}{2}}\cdot\frac{1}{\sqrt{2\pi}}}$

$$= \frac{1}{\sqrt{2\pi}}\cdot\frac{2}{\sqrt{3}}\,\frac{e^{-\frac{2}{3}(y-\frac{1}{2})^2}\,\cancel{e^{-\frac{1}{2}}}}{\cancel{e^{-\frac{1}{2}}}}$$

$$\boxed{f_{Y|X=1}(y|x=1) = \frac{1}{\sqrt{2\pi}\cdot\frac{\sqrt{3}}{2}}\,e^{\frac{-(y-\frac{1}{2})^2}{2\cdot(3/4)}}}$$

Thus, given $X=1$, $Y \sim N(m=\frac{1}{2}, \sigma^2=\frac{3}{4})$

X and Y are not independent, since $Y \sim N(0,1)$, so $f_Y(y) \neq f_{Y|X=1}(y)$.

Single Functions of Two Random Variables

SECTION 3

Let us consider functions that map a pair of random variables to a single value, i.e. $Z = g(X,Y)$. Notice that Z is a random variable defined on the same sample space. We are interested in computing the PDF/CDF of Z, as well as its expectation. Such computation is valuable both in that new quantities of interest $(g(X,Y))$ can be characterized, and in that statistical information about X and Y is obtained.

As with functions of single random variables, the P.D.F.s of functions of two random variables can most easily found by finding the CDFs first. That is, we can find $F_Z(z) = P(Z \le z) = P(g(X,Y) \le z)$ by finding the set of x, y such that $g(x,y) \le z$, and integrating the joint PDF of (X,Y) over this region. The PDF for Z can then be found by taking the derivative of the CDF.

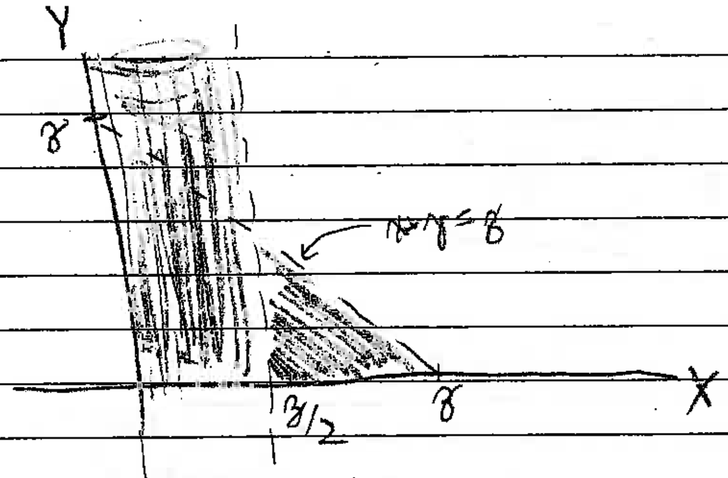
Example 4

Let X and Y be independent exponential random variables with means 1 and $\frac{1}{2}$, respectively. Please find the P.D.F. for $Z = \min(2X, X+Y)$.

$$f_X(x) = e^{-x},\ x \ge 0; \quad f_Y(y) = 2e^{-2y},\ y \ge 0.$$

$$f_{X,Y}(x,y) = 2e^{-x}e^{-2y},\ x \ge 0,\ y \ge 0.$$

$$\begin{aligned} F_Z(z) = P(Z \le z) &= P(\min(2X, X+Y) \le z) \\ &= P((2X \le z) \cup X+Y \le z) \\ &= P(X \le z/2 \cup X+Y \le z) \end{aligned}$$



We can find $P(Z \le z)$ by integrating the joint density over the shaded region. This integral is equal to

$$F_Z(z) = \int_0^{z/2} f_X(x)\,dx + \int_{z/2}^{z}\int_0^{z-x} f_{X,Y}(x,y)\,dy\,dx$$

Do you see why?

Substituting, we thus get

$$F_Z(z) = \int_0^{z/2} e^{-x}\,dx + \underbrace{\int_{z/2}^{z}\int_0^{z-x} 2e^{-x}e^{-2y}\,dy\,dx}$$

$$= -e^{-x}\Big]_0^{z/2}$$

$$= 1 - e^{-z/2}$$

$$= \int_{z/2}^{z} 2e^{-x}\int_0^{z-x} e^{-2y}\,dy\,dx$$

$$= \int_{z/2}^{z} 2e^{-x}\left(-\tfrac{1}{2}e^{-2y}\right)\Big]_0^{z-x}\,dx$$

$$= \int_{z/2}^{z} 2e^{-x}\left(\tfrac{1}{2} - \tfrac{1}{2}e^{-2(z-x)}\right)dx$$

$$= \int_{z/2}^{z}\left(e^{-x} - e^{-2z}e^{x}\right)dx$$

$$= -e^{-x}\Big]_{z/2}^{z} - e^{-2z}e^{x}\Big]_{z/2}^{z}$$

$$= e^{-z/2} - e^{-z} + e^{-3z/2} - e^{-z}$$

$$\boxed{\text{Thus, } F_Z(z) = 1 - 2e^{-z} + e^{-3z/2},\ z \ge 0}$$

Finally, taking the derivative with respect to z, we obtain that

$$\boxed{f_Z(z) = 2e^{-z} - \frac{3}{2}e^{-3z/2}}$$

Expectations of Functions of Two Random Variables

Now that we have figured out how to find PDFs for functions of two random variables, we might wonder if we can find the expectation of the function Z without actually having to find the PDF of Z (i.e., directly from the joint PDF of X and Y). In fact, we can, as follows:

$$E[Z] = E[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\,dy\,dx$$

The proof is similar to the proof for the expectation of a function of a single random variable, so we'll skip it. Let's practice using this expression, by finding the expectation of Z from the previous example:

$$E[Z] = E[\min(2X,2Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \min(2x, x+y) f_{X,Y}(x,y)\,dy\,dx$$

$$= \int_0^{\infty}\int_0^{\infty} \min(2x, x+y)\, 2e^{-x}e^{-2y}\,dy\,dx$$

$$= \int_0^{\infty}\int_0^{x} 2x \cdot 2e^{-x}e^{-2y}\,dy\,dx$$

$$+ \int_0^{\infty}\int_x^{\infty} (x+y)\, 2e^{-x}e^{-2y}\,dy\,dx.$$

I will let you take it from here!

Since we are thinking about expectations, let's also think about the expected value of one random variable X given the value of another, say $Y=y$. We use the notation $E[X|Y=y]$ for this expectation. Recalling that we have such a thing as a P.D.F. for X in the smaller universe that $Y=y$ (i.e., $f_{X|Y}(x|Y=y)$), it is natural for us to compute this expectation as $E[X|Y=y]=\int_{-\infty}^{\infty} x f_{X|Y}(x|Y=y)$. Notice that this is nothing more than the computation of the expectation of a single random variable, albeit given the value of the other.

You might be wondering if the expected value of X (not given anything) can be found from the expected value of X given Y, together with the P.D.F. of Y. In fact, it can:

$$\underline{E[X]=\int_{-\infty}^{\infty} E[X|Y=y]f_Y(y)\,dy}$$

Proof: $E(X)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X,Y}(x,y)\,dx\,dy$

$$=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X|Y}(x|Y=y)\,dx\, f_Y(y)\,dy$$

$$=\int_{-\infty}^{\infty} E[X|Y] f_Y(y)\,dy \quad \checkmark$$

There's actually another cute notation for this expectation. To come up with this cute notation, notice that $E[X|Y=y]$ is infact a function of y. There is nothing stopping us from plugging in the random variable Y in this function; this function of the random variable Y is itself a random variable — let's call it $E[X|Y]$. Notice that we can fine the expected value of this function of Y, or $E[E(X|Y)]$. However, this expectation (by definition) is $\int_{-\infty}^{\infty} E(X|Y=y) f_Y(y)\,dy$, which equals $E(X)$.

Thus, $\boxed{E(X)=E[E(X|Y)]}$. This rule is known as the law of iterated expectations. We note that similar approaches can be used to find $E[g(X,Y)]$ rather than simply $E(X)$.

There's one function of two random variables whose expectation is of particular interest: We call the expectation $E[(X-\mu_X)(Y-\mu_Y)]$, where $\mu_X=E[X]$ and $\mu_Y=E[Y]$, the covariance of X and Y. The covariance is an interesting quantity because it shows the interdependence of X and Y: if the covariance is large, then large X implies large Y, if the covariance is highly negative then larger X typically implies small Y, and if the covariance is near zero then on average X and Y are not interdependent.

We often find it convenient

to scale the covariance by the standard deviations of X and Y, in order to normalize this interdependency or correlation information. That is, we consider the <u>correlation coefficient</u>

$$\rho_{X,Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y},$$

Let's prove that the correlation coefficient is between -1 and 1:

As an example, let us find the correlation coefficient of jointly Gaussian random variables

Let's also explore the connection between independence and uncorrelation ($\rho_{X,Y}=0$).