Lectures on EE 3025 Chapters 2-3

John Kieffer
Department of Electrical and Computer Engineering
University of Minnesota Twin Cities
Minneapolis, MN 55455

Lecture 7

Chapters 2-3 Part 1

In Lecture 7, I give terminology and examples regarding *types of random variables*, *probability mass functions*, and *probability density functions*. I also point out which probability distributions from Appendix A that we will be covering this semester.

7.1 Random Variable Definition and Notation

Definition

If you look back in Lecture 3 about *derived probability models*, you will see the beginnings of the random variable concept. A random variable (RV) X is a real-valued function defined on the sample space S. For each outcome $\omega \in S$, the RV X assigns a real value $X(\omega)$.

Let us recall from Lecture 3 the definition of the derived probability model P^X on the real line induced by the random variable X and the original probability model P on S. Let E be a subset of the real line. (Usually, we take E to be an interval.) Then the probability $P^X(E)$ of E is defined by

$$P^X(E) \triangleq P(\{\omega \in S : X(\omega) \in E\}).$$

[1] In other words, to find $P^X(E)$, you compute the P probability of the event back in S consisting of all outcomes for which the X value lies in E.

Notation

If E is a subset of the real line, then the notation

$$\{X \in E\}$$

[1] The Δ over the equal sign means that we are making a definition.

denotes the event back in the sample space S consisting of all outcomes in S which are mapped by X into a value in E. That is,

$$\{X \in E\} \triangleq \{\omega \in E : X(\omega) \in E\}.$$

It follows that the notation

$$P(X \in E)$$

is the same thing as $P^X(E)$, because

$$P(X \in E) = P(\{\omega \in S : X(\omega) \in E\}) = P^X(E).$$

In words, $P(X \in E)$ is the "probability with which the value of X falls in E".

Usually, E is an interval of some sort. We use the following notations for intervals:

- $[a, b]$ denotes the interval with left endpoint a and right endpoint b, including the two endpoints.
- (a, b) denotes the interval with endpoints a, b, respectively, where the two endpoints are excluded.
- $[a, b)$ denotes the interval with endpoints a, b, where endpoint a is included and endpoint b is excluded.
- $(a, b]$ denotes the interval with endpoints a, b, where endpoint a is excluded and endpoint b is included.

Here are some examples of notations for probabilities involving a RV X:

$$\begin{aligned} P(2 \le X \le 3) &= P^X(\,[2,3]\,) \\ P(3 < X \le 4) &= P^X(\,(3,4]\,) \\ P(X \ge 5) &= P(\,[5,\infty)\,) \\ P(X = 5) &= P^X(\{5\}) = P^X(\,[5,5]\,) \end{aligned}$$

The last example was a probability of the type $P^X(x)$, where x is a single point on the real line. Your book writes $P_X(x)$ instead of $P^X(x)$ (X as a subscript rather than a superscript). You can write it either way. (Obviously, it makes no difference; some books use the superscript notation and some use the subscript notation.)

We have obtained the induced probability model P^X on the real line from the original probability model P on the sample space S. Once the model P^X has been obtained in this way, you would only need the model P^X and not the original model P if you are dealing with just a single random variable.

7.2 Types of Random Variables

Discrete Random Variables

Suppose the possible values of the RV X are just a discrete set of values, written sequentially as

$$x_1, x_2, x_3, \cdots.$$

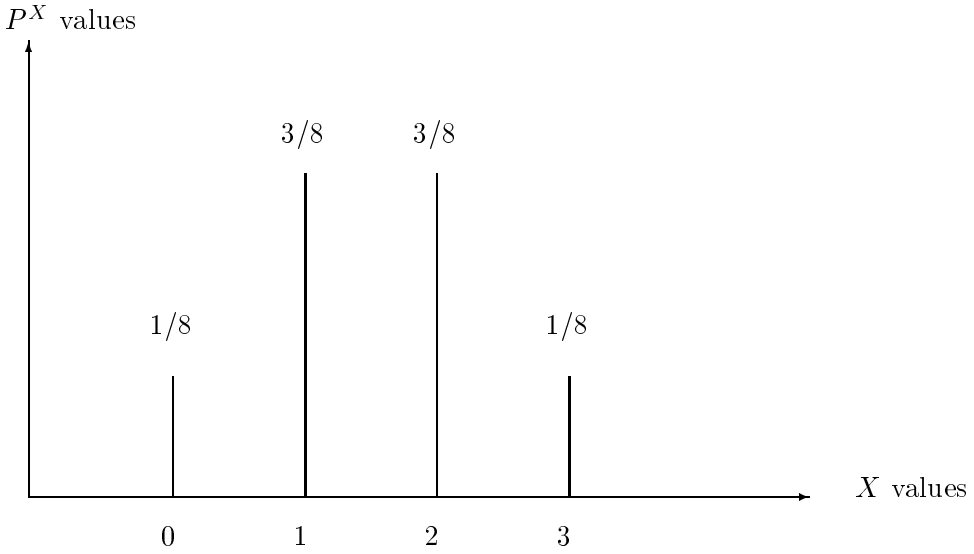
This sequence of values could be finite or infinite. Most of our discrete RV's will take just finitely many values, but there are some (like Poisson or geometric discrete RV's covered later on) that take infinitely many values. We must have

$$\sum_i P^X(x_i) = 1,$$

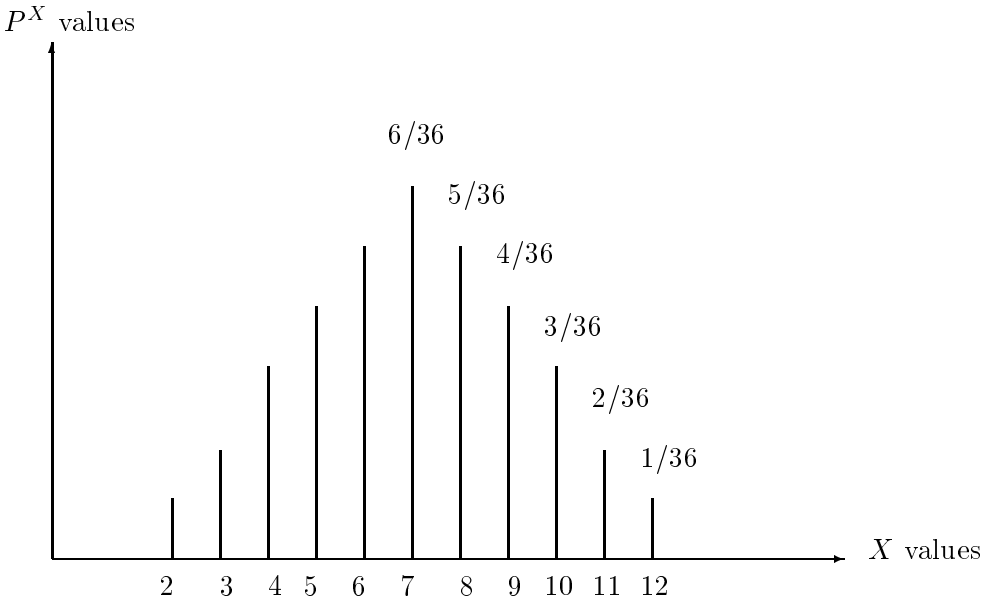
since the events of form $\{X = x_i\}$ form a partition of the sample space S and therefore their probabilities must add up to one.

Definition. The function which maps each value x_i into the probability $P^X(x_i)$ is called the *probability mass function* of the RV X. This is abbreviated as PMF.

Example 7.1. Let discrete RV X denote the total number of heads on 3 tosses of a fair coin. We plotted the PMF of X back in Lecture 3 notes:



Example 7.2. Let discrete RV X denote the sum of the numbers on the toss of two fair dies. We plotted the PMF of X back in Lecture 3 notes:



Example 7.3. Two equally matched sports teams play a best 3 of 5 championship series. Let X be the discrete RV giving the number of games that are played in the series. Then the PMF works out to be

$$\begin{aligned} P^X(3) &= 2/8 \\ P^X(4) &= 6/16 \\ P^X(5) &= 6/16 \end{aligned}$$

To see how I got this, look back at the tree for this experiment back in Example 3.1 of Lecture 3 notes. There are two paths in the tree corresponding to a 3 game series, with total probability $2/8$ (the probability of each such path is $1/8$); thus $P^X(3) = 2/8$. There are six paths in the tree corresponding to a 4 game series, with total probability $6/16$ (each such path has probability $1/16$); thus, $P^X(4) = 6/16$. The remaining PMF value can be found by the complementation rule:

$$P^X(5) = 1 - P^X(4) - P^X(3) = 6/16.$$

If you know the PMF of a discrete RV X, then you can compute the probability with which the value of X falls in any subset of the real line as follows:

$$P(X \in E) = \sum_{x_i \in E} P^X(x_i). \tag{7.1}$$

Example 7.4. For the PMF in Example 7.2, let us compute $P(5 \leq X \leq 8)$. We have

$$\begin{aligned} P(5 \leq X \leq 8) &= P^X(5) + P^X(6) + P^X(7) + P^X(8) \\ &= 4/36 + 5/36 + 6/36 + 5/36 = 20/36 \end{aligned}$$

Continuous Random Variables

A continuous RV takes values "continuously distributed" over some subset of the real line; "continuously distributed" refers to the fact that there can be no accumulation of probability at any particular real value of a continuous RV. Therefore, a continuous RV X will satisfy

$$P^X(x) = 0, \tag{7.2}$$

for every real number x! Because of the strange property (7.2), it will be impossible to compute probabilities of events associated with continuous RV's by summation as in formula (7.1). Instead, we will have to compute P^X probabilities for a continuous RV X via integration. Let us make this more precise. For a continuous RV X, you will have a so-called *probability density function* $f_X(x)$, defined for all real numbers x. (This function is abbreviated PDF, or sometimes simply called "density"). The PDF $f_X(x)$ satisfies the following properties:

(a): $0 \leq f_X(x) < \infty$ for all real x.

(b): $\int_{-\infty}^{\infty} f_X(x)dx = 1$.

(c): For any event E, a subset of the real line,

$$P^X(E) = \int_E f_X(x)dx. \tag{7.3}$$

Discussion. From property(a), if you plot the graph of a PDF $f_X(x)$, the graph will always lie on or above the x-axis as x varies. So, we can interpret property(b) as saying that the area lying under the entire density curve is equal to 1. In property(c), suppose we take E to be an interval $[a, b]$. Then equation (7.3) becomes

$$P(a \leq X \leq b) = P^X(\ [a, b]\ ) = \int_a^b f_X(x)dx. \tag{7.4}$$

In other words, $P(a \leq X \leq b)$ may be interpreted as the area under the density curve that is caught in between the vertical lines $x = a$ and $x = b$. Now in property(c), suppose we take E to consist of a single point a. Then (7.3) becomes

$$P(X = a) = P^X(a) = \int_a^a f_X(x)dx = 0,$$

and we now see why (7.2) is true. Since endpoints of intervals are assigned probability zero, we can ignore endpoints of intervals in probability calculations involving continuous RV's, that is, all four probabilities

$$P(a \leq X \leq b),\ \ P(a < X \leq b),\ \ P(a \leq X < b),\ \ P(a < X < b) \tag{7.5}$$

will be equal to the area given by the right side of (7.4). This will not be true for a discrete RV. In fact, for a discrete RV X, it can turn out that all four probabilities (7.5) are different! Finally, we point out that property(b) is a special case of property(c): in property(c), take E to be the entire real line, and then $P^X(E)$ must be 1 because $\{\omega \in S : X(\omega) \in E\}$ is all of S and S has probability 1.

Example 7.5. Look back at Example 3.7, where we gave an example of a PDF. Let us give a RV X with the PDF of Example 3.7 a name. We say that a continuous RV X having PDF

$$f_X(x) = 1, \;\; 0 \leq x \leq 1 \; (zero \; elsewhere),$$

has the *standard uniform distribution*, or we say that the RV is a *standard uniform RV*. In other words, the PDF of a RV having a standard uniform distribution is simply an amplitude 1 rectangular pulse over the interval $[0, 1]$. You simulate a value of a RV having the standard uniform distribution by executing the Matlab command

```
rand(1,1)
```

For any subinterval $[a, b]$ of the unit interval $[0, 1]$, the probability $P^X([a, b])$ for a standard uniform RV X works out according to formula (7.3) to be just $b - a$, the length of the interval $[a, b]$. There are also "nonstandard" uniform distributions, which refers to the scenario in which we have uniform RV's whose densities are rectangular pulses extending over finite intervals other than the unit interval $[0, 1]$. For finite interval $[c, d]$, one can obtain a uniform RV extending over $[c, d]$ by appropriately scaling and translating a standard uniform RV. Later, we will see how to do the scaling/translation.

Example 7.6. A continuous RV X is said to have the *standard Gaussian distribution* if its density is

$$f_X(x) = C \exp\left(-\frac{x^2}{2}\right), \quad -\infty < x < \infty, \tag{7.6}$$

where C is a unique positive real number whose value we will determine. The plot of the *standard Gaussian density* (7.6) is typically referred to as a "bell-shaped curve". (For a nice plot, see Figure 3.6 in your textbook.) We now compute C. Since the area under the density curve must be 1, it follows that

$$C = \frac{1}{\int_{-\infty}^{\infty} e^{-x^2/2} dx}.$$

Suppose we square the integral in the denominator. This gives us a double integral in rectangular coordinates that we can easily evaluate when we convert to polar coordinates r, θ:

$$\begin{aligned} \left(\int_{-\infty}^{\infty} e^{-x^2/2} dx\right)^2 &= \left(\int_{-\infty}^{\infty} e^{-x^2/2} dx\right)\left(\int_{-\infty}^{\infty} e^{-y^2/2} dy\right) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)/2} dx dy \end{aligned}$$

$$= \int_0^{2\pi} \int_0^{\infty} e^{-r^2/2} r dr d\theta = 2\pi.$$

It follows that $C = 1/\sqrt{2\pi}$, and our standard Gaussian density is therefore

$$f_X(x) = \left(\frac{1}{\sqrt{2\pi}}\right) \exp\left(-\frac{x^2}{2}\right)$$

Remark. In Matlab, it is easy to simulate an observation of a standard Gaussian random variable. You just execute the command

```
randn(1,1)
```

Statisticians call the Gaussian distribution the *normal distribution*. The `n` in `randn` stands for "normal".

Mixed Random Variables

A mixed random variable is neither purely discrete nor purely continuous. That is, a mixed RV takes a certain number of discrete values with positive probability, but the remaining possible values of the RV are taken on continuously. Here is a simple example of a continuous RV.

Example 7.7. Bill is a shotputter. With probability 0.1, when Bill throws the shot he will foul and throw the shot only 10 feet. Otherwise, with probability 0.9, Bill will not foul and will throw the shot a distance which is uniformly distributed between 60 and 70 feet. Let X be the distance that Bill throws the shot. This is clearly a mixed RV. (You have the discrete value $X = 10$, plus uniformly distributed values in the range $60 \leq X \leq 70$.)

We can describe a density function (PDF) for a mixed RV if we allow two additive components in the density function:

- one of the components will be a linear combination of delta functions concentrated at the discrete values; and
- the other component will be a finite density function scaled by the probability with which the RV takes its nondiscrete values.

With this idea in mind, let us see what the PDF $f_X(x)$ would be for the mixed RV X in Example 7.7. Here, we have one discrete value of X at $X = 10$ taken on with probability 0.1, and so we should have a delta function $(0.1)\delta(x - 10)$ as a component of the PDF. The rest of the values of X are uniformly distrbuted between 60 and 70, so we should make the other component of $f_X(x)$ be a suitably scaled rectangular pulse over the interval $[60, 70]$. In other words, the PDF $f_X(x)$ for Example 7.7 should take the form

$$f_X(x) = (0.1)\delta(x - 10) + (0.9)(1/10)[u(x - 60) - u(x - 70)].$$

The factor of 0.1 in front of the delta function $\delta(x-10)$ is correct because that will give us the following correct probability calculation for $P(X=10)$:

$$P(X=10) = \int_{10}^{10} f_X(x)dx = \int_{10}^{10} (0.1)\delta(x-10)dx = 0.1.$$

(Remember from EE 3015 that the integral of any delta function is 1!) The reader may be wondering why we have a factor of 1/10 in front of the rectangular pulse

$$u(x-60) - u(x-70).$$

This is the factor which makes the following integral equal to 1:

$$\int_{-\infty}^{\infty} (1/10)[u(x-60) - u(x-70)]dx = \int_{60}^{70} (1/10)dx = 1.$$

Then we have

$$\begin{aligned} \int_{-\infty}^{\infty} f_X(x)dx &= 0.1 + (0.9)\int_{-\infty}^{\infty} (1/10)[u(x-60) - u(x-70)]dx \\ &= 0.1 + (0.9)1 = 1, \end{aligned}$$

which makes $f_X(x)$ a bonafide density function.

Example 7.8. In the preceding example, we saw how to represent a discrete value of a mixed RV as a delta function component in the density function. It is interesting to note that we can use this same approach in handling a discrete RV. In other words, we can regard a discrete RV X as having a PDF $f_X(x)$ consisting entirely of delta function components concentrated at the different values of X. For example, we see that the discrete random variable X of Example 7.1 has a density $f_X(x)$ which can be written as

$$f_X(x) = (1/8)\delta(x) + (3/8)\delta(x-1) + (3/8)\delta(x-2) + (1/8)\delta(x-3). \tag{7.7}$$

Note the the constant appearing in front of each delta function term in (7.7) is equal to the probability with which RV X takes on the value represented by that delta function. The PDF $f_X(x)$ is correct because it will generate the correct PMF values upon integration:

$$\begin{aligned} P^X(0) &= \int_0^0 f_X(x)dx = (1/8)\int_0^0 \delta(x)dx + 0 + 0 + 0 = 1/8 \\ P^X(1) &= \int_1^1 f_X(x)dx = 0 + (3/8)\int_1^1 \delta(x-1)dx + 0 + 0 = 3/8 \\ P^X(2) &= \int_2^2 f_X(x)dx = 0 + 0 + (3/8)\int_2^2 \delta(x-2)dx + 0 = 3/8 \\ P^X(3) &= \int_3^3 f_X(x)dx = 0 + 0 + 0 + (1/8)\int_3^3 \delta(x-3)dx = 1/8 \end{aligned}$$

Remark. We now see that the probability behavior of *any* RV can be described in terms of a PDF, whether the RV is discrete, continuous, or mixed. If we look back at assumptions(a)-(c) placed on our PDF's for continuous RV's, we see that these are still valid for the PDF's of discrete or mixed RV's, except we have to allow $f_X(x)$ to take infinite values at discrete values of X (due to the delta functions concentrated at these values). The fact that we can use PDF's for all three types of RV's means later on that we can give a unified presentation of certain formulas involving RV concepts—this saves us from giving three separate derivations of such formulas!

7.3 Probability Distributions to be Covered

The textbook covers quite a number of commonly appearing discrete and continuous probability distributions for RV's. I will not be covering all of these. The discrete probability distributions I will be covering are:

- binomial
- Poisson
- geometric
- discrete uniform
- hypergeometric

All of these are in Appendix A of the textbook except for the hypergeometric distribution. The continuous probability distributions I will be covering are:

- Gaussian
- uniform
- exponential

Lecture 8

Chapters 2-3 Part 2

8.1 Binomial Distribution Via Binomial Theorem

The binomial theorem says

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}. \tag{8.1}$$

The notation $\binom{n}{k}$ denotes the number of combinations of n things taken k at a time, as every good college algebra student knows. (If you need to review combinations, see page 30 of your textbook.) The number $\binom{n}{k}$ is calculated as follows:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!}.$$

For fixed n, the numbers $\binom{n}{k}$ as k ranges from 0 to n are also called *binomial coefficients*. They are generated by Pascal's triangle:

$$\begin{array}{ccccccccccccccc}
 & & & & & & 1 & & & & & & & & \\
 & & & & & 1 & & 1 & & & & & & & \\
 & & & & 1 & & 2 & & 1 & & & & & & \\
 & & & 1 & & 3 & & 3 & & 1 & & & & & \\
 & & 1 & & 4 & & 6 & & 4 & & 1 & & & & \\
 & 1 & & 5 & & 10 & & 10 & & 5 & & 1 & & & \\
1 & & 6 & & 15 & & 20 & & 15 & & 6 & & 1 & &
\end{array}$$

Pascal's triangle extends infinitely down; we stopped with row 7. Each row of Pascal's triangle begins and ends with 1, and each entry in a row is the sum of the entries in the row above to the right and left. The entries in row $n+1$ of Pascal's triangle are the binomial coefficients $\binom{n}{k}$ as k

goes from 0 through n. For example, looking at row 5 of Pascal's triangle, we see the coefficients $1, 4, 6, 4, 1$. This tells us that

$$(a+b)^4 = a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4.$$

The coeeficient 6 in the middle is computed as

$$\binom{4}{2} = \frac{4*3}{2!} = 6.$$

In equation (8.1), substitute $a = p$ and $b = 1-p$, where p is a parameter between 0 and 1. Then

$$1 = (p+1-p)^n = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k}.$$

Since the terms on the right add up to 1, we can interpret these terms as forming the PMF of a random variable. Accordingly, suppose that n is any positive integer and p is any number between 0 and 1. We say that a RV X has the Binomial(n,p) distribution (or is a Binomial(n,p) RV) if

- X is a discrete RV taking the values $0, 1, 2, \cdots, n$.
- The PMF of X is

$$P^X(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, \cdots, n.$$

Useful Fact. Suppose you have any random experiment and any event E associated with that experiment that has probability $P(E) = p$. Suppose you perform n independent trials of the experiment and you define RV X to be the total number of trials in which E occurs. Then X is a Binomial(n,p) RV.

Proof. Each time you perform the n trials, form a binary n-tuple

$$(j_1, j_2, \cdots, j_n) \tag{8.2}$$

where, for $i = 1, 2, \cdots, n$, you choose the i-th entry j_i to be 1 if E occurs on trial i and $j_i = 0$ otherwise. You can regard the set of all 2^n possible binary n-tuples of the form (8.2) as forming a sample space, and you can regard RV X as being defined on this sample space as follows:

$$X(j_1, j_2, \cdots, j_n) = \textit{ number of ones in } n-\textit{tuple.}$$

Because we have independent trials, the probability model P on the set of n-tuples (8.2) is obtainable as an independent discrete probability model as covered on page 15 of Lecture 2 Notes. This means we have

$$P(j_1, j_2, \cdots, j_n) = P_1(j_1)P_2(j_2)\cdots P_n(j_n), \tag{8.3}$$

where

$$\begin{aligned} P_i(1) &= P(E \text{ occurs on } i^{th} \text{ trial}) = p \\ P_i(0) &= P(E^c \text{ occurs on } i^{th} \text{ trial}) = 1-p. \end{aligned}$$

The product (8.3) then simplifies to

$$P(j_1, j_2, \cdots, j_n) = p^k (1-p)^{n-k},$$

where k is the number of ones in the n-tuple (8.2), or equivalently, k is the number of times E occurs and therefore k is the value of X for this n-tuple. To obtain $P(X=k)$, we must add up all the $P(j_1, j_2, \cdots, j_n)$ terms in which the number of ones in $(j_1, j_2, \cdots, j_n)$ is equal to k. Each such $P(j_1, j_2, \cdots, j_n)$ term has probability $p^k(1-p)^{n-k}$ and there are $\binom{n}{k}$ such terms (you choose exactly k of the n positions of the n-tuple (8.1) in which to place the ones). We conclude that

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \cdots, n,$$

which is the Binomial(n,p) PMF.

Example 8.1. You flip a fair coin three times and let X be the number of heads. Then X is a Binomial(n,p) random variable in which $n=3$ and $p=1/2$. We have the following PMF:

$$\begin{aligned} P^X(0) &= \binom{3}{0}(1/2)^0(1/2)^3 = 1/8 \\ P^X(1) &= \binom{3}{1}(1/2)^1(1/2)^2 = 3/8 \\ P^X(2) &= \binom{3}{2}(1/2)^2(1/2)^1 = 3/8 \\ P^X(3) &= \binom{3}{3}(1/2)^3(1/2)^0 = 1/8 \end{aligned}$$

This confirms what we obtained earlier for this same RV. (See Lecture 7 Notes, Example 7.1.)

Example 8.2. The product items manufactured by a certain company are 5% defective. Ten items are selected at random from the product assembly line (with replacement) at the end of the day, and are tested. Let X be the number of defective items among the ten items selected. Since we did sampling with replacement, we have independent trials, and therefore X is Binomial(n,p) with parameters $n=10$ and $p=.05$. We compute the probabilities of some events associated with

X:

$$\begin{aligned} P(X=0) &= \binom{10}{0}(0.05)^0(0.95)^{10} = (0.95)^{10} = 0.5987 \\ P(X \leq 2) &= \sum_{k=0}^{2}\binom{10}{k}(0.05)^k(0.95)^{10-k} = 0.9885 \\ P(0<X<3) &= \binom{10}{1}(0.05)(0.95)^9 + \binom{10}{2}(0.05)^2(0.95)^8 = 0.3898 \\ P(X \geq 3) &= 1 - P(X \leq 2) = 1 - 0.9885 = 0.0115 \end{aligned}$$

Exercise. For the RV X in Example 8.2, the reader should

- verify that our figures at the end of Example 8.2 are correct by either using Matlab or a calculator.
- compute each of the eleven PMF values $P^X(k)$ for $k = 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$, and plot the PMF.

Remark 1. In Example 8.2, we selected the items to be tested with replacement. If we select them without replacement, then we do not have independent trials and the number of defective items in the sample is not binomially distributed. The number of defectives in this case will have a distribution called the *hypergeometric distribution*. I will talk about the hypergeometric distribution during a future lecture.

Remark 2. In Recitation 1, we saw how to estimate the probability of an event by taking independent trials. We can now get more insight into this procedure using the binomial distribution. Suppose we want to estimate the probability $P(E) = p$ of event E associated with some random experiment. We perform n trials of the experiment, n large, and count the number of trials on which E occurs. If we call this number X, then we now know that X is Binomial(n, p). The estimate of probability p is then the random variable X/n, and we can find the PMF of this random variable since we know the PMF of X. Later on in the course, we will examine how the values of X/n are distributed about p. We will be able to quantify how the distribution of these values becomes more and more closely concentrated about p as the number of trials n grows.

8.2 Poisson Distribution Via McClaurin Series Expansion of e^x

In calculus, everybody learns that the McClaurin Series expansion of the function e^x is:

$$e^x = \sum_{k=0}^{\infty}\frac{x^k}{k!} = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \cdots$$

Suppose we take $x = \alpha$, a positive parameter, and divide both sides by e^{α}. This gives us the summation formula

$$1 = \sum_{k=0}^{\infty} \frac{e^{-\alpha}\alpha^k}{k!} = e^{-\alpha} + \alpha e^{-\alpha} + (\alpha^2 e^{-\alpha}/2) + (\alpha^3 e^{-\alpha}/6) + (\alpha^4 e^{-\alpha}/24) + \cdots.$$

We can think of the terms on the right side as defining a discrete probability distribution. Accordingly, we say that a RV has the Poisson(α) distribution (or is a Poisson(α) RV) if:

- The values of X are the nonnegative integers $0, 1, 2, 3, \cdots$.
- The PMF of X is
$$P^X(x) = \frac{\alpha^x e^{-\alpha}}{x!}, \quad x = 0, 1, 2, \cdots$$

We explain how the Poisson distribution arises. Consider a random event which can occur anywhere on a time axis. (This random event could be a phone call, a hurricane, arrival of a customer at his/her bank, etc.). Let I be any finite time interval. Define X to be the random variable which is equal to the number of times that the random event of interest occurs in the interval I. It is fairly common to model such a random variable X as having a Poisson(α) distribution. What is the parameter α taken to be? We shall understand this better when we cover the concept of expected value. For the present, we say that α is typically taken to be the average of X over a large number of trials. For example, suppose that X is the number of phone calls coming into the ECE office (625-3300) between 1:00 and 2:00 PM on a randomly chosen working day. Suppose that we had observed phone calls over many previous days and had determined that the average number of phone calls in this time slot was 4.2. Then, we could model X as having a Poisson distribution with parameter $\alpha = 4.2$.

Here are some examples of random variables illustrating situations in which one could model the random variable as a Poisson variable:

- number of telephone calls arriving in a given time interval
- number of hurricanes hitting East coast in a given time period
- number of radioactive particles registered by a Geiger counter in a given time interval
- number of customers arriving at a bank's teller window in a given time period
- number of message packets arriving at an internet server in a given time interval
- number of imperfections in a certain length of magnetic tape (think of the "length of tape" in the same way you would think of a "time interval")

The reader can undoubtedly think of other examples.

Example 8.3. Let X be a Poisson random variable with parameter α. Then

$$\begin{aligned} P(X=0) &= e^{-\alpha} \\ P(X=1) &= \alpha e^{-\alpha} \\ P(X=2) &= \alpha^2 e^{-\alpha}/2 \\ P(X=3) &= \alpha^3 e^{-\alpha}/6 \\ P(X \le 2) &= (1+\alpha+\alpha^2/2)e^{-\alpha} \\ P(X \ge 4) &= 1-(1+\alpha+\alpha^2/2+\alpha^3/6)e^{-\alpha} \end{aligned}$$

Example 8.4. Let us model the number of phone calls X coming into 625-3300 between 1:00 and 2:00 PM on a randomly chosen working day as a Poisson random variable with parameter $\alpha = 4.2$. Then

$$P(X \ge 4) = 1 - P(X \le 3) = 1 - \sum_{k=0}^{3} \frac{(4.2)^k e^{-4.2}}{k!} = .6046.$$

(There is some Matlab code in Experiment 5 of Recitation 3 that I used to do this computation.) What does the probability $P(X \ge 4) = 0.6046$ mean physically? If the Poisson model is a good model of the physical situation here, then, if we observe the number of phone calls coming into 625-3300 over many working days, we will see that for approximately 60.5% of the days, the number of phone calls between 1:00 and 2:00 is at least four. Let us illustrate here another property of the Poisson model: suppose we now count the number of calls Y in the time interval from 1:00PM to 3:00PM. This new time interval is twice as big as before. As a consequence, we would typically model Y as a Poisson RV with parameter twice as much as before, namely, we would take $\alpha = 2 * 4.2 = 8.4$. This is because we would expect to have, on average, twice as many phone calls in a time interval twice as big. We would then have

$$P(Y \ge 8) = 1 - P(Y \le 7) = 1 - \sum_{k=0}^{7} \frac{(8.4)^k e^{-8.4}}{k!} = 0.6013.$$

Exercise. Let Z be the number of phone calls from 1:00PM to 3:30 PM. Compute $P[Z \ge 10]$.

8.3 Geometric Distribution Via Geometric Series Summation

By the time most students reach EE 3025, they have re-learned every year since the 9th grade the following formula for summing a geometric series:

$$\sum_{k=1}^{\infty} ar^{k-1} = a + ar + ar^2 + ar^3 + ar^4 + \cdots = \frac{a}{1-r}. \tag{8.4}$$

The parameter a is the first term of the geometric series, and the parameter r is the ratio between each term of the series and the preceding term. We assume that the ratio r is strictly between 0 and 1 in order for the summation formula to be valid. Let us now divide both sides of (8.4) by $\frac{a}{1-r}$ and then replace r by $1-p$, where p is a parameter strictly between 0 and 1. We obtain the formula

$$\sum_{k=1}^{\infty}(1-p)^{k-1}p = p + (1-p)p + (1-p)^2p + (1-p)^3p^2 + (1-p)^4p + \cdots = 1.$$

Since these terms sum to one, we can interpret these terms as the PMF of some discrete RV. Accordingly, we say that RV X has the Geometric(p) distribution (or that X is a Geometric(p) RV) if:

- The values of X are the positive integers $1, 2, 3, \cdots$
- The PMF is given by
$$P^X(x) = (1-p)^{x-1}p, \quad x = 1, 2, 3, \cdots.$$

Useful Fact. Here is the typical scenario in which a random variable with a geometric distribution arises. Suppose you have some random experiment and some event E associated with this experiment for which $P(E) = p$. Suppose you perform independent trials of this experiment, stopping after the first trial on which E occurs. Then the number of trials that are performed is a Geometric(p) RV.

Proof. Take the sample space of the repeated trials experiment as

$$S = \{E, E^c \cap E, E^c \cap E^c \cap E, E^c \cap E^c \cap E^c \cap E, \cdots\}.$$

By the product rule (applicable since we have independent trials), the respective probabilities of these outcomes are

$$p, (1-p)p, (1-p)^2p, (1-p)^3p, \cdots.$$

Letting X be the number of trials which are performed, we see that

$$P(X = k) = P(E^c \cap E^c \cap E^c \cap \cdots \ (k-1 \ times) \cap E) = (1-p)^{k-1}p, \quad k = 1, 2, 3, \cdots.$$

This is the geometric distribution.

Example 8.5. Suppose you have an unfair coin with $P(H) = 1/3$. Let X be the number of tosses it takes for us to obtain a head for the first time. We compute the following.

$$\begin{aligned} P(X \geq 3) &= 1 - P(X \leq 2) = 1 - (2/3)^0(1/3) - (2/3)^1(1/3) = 4/9 \\ P(X \geq 4) &= P(X \geq 3) - (2/3)^2(1/3) = 8/27 \\ P(X \geq 4 | X \geq 3) &= \frac{P(X \geq 4)}{P(X \geq 3)} = 2/3 \end{aligned}$$

Exercise. In the preceding example, notice that

$$P(X \geq 4 | X \geq 3) = P(X \geq 1) = 2/3.$$

Show that this is no coincidence by proving that

$$P(X \geq n+1 | X \geq n) = P(X \geq 1) = 2/3, \tag{8.5}$$

for every positive integer n. What is the intuitive meaning of the statement (8.5)?

8.4 Application to Coding

Suppose we want to represent every positive integer by a unique binary codeword. There is one simple-minded way to assign codewords that you may have seen earlier (especially in a computer science course): Just take the usual binary expansion of each integer as the codeword. For example, with this approach, the codeword for 9 would be 1001 and the codeword for 19 would be 10011. However, there is a drawback to assigning codewords in this way. Suppose you have coded a message consisting of a sequence of positive integers by simply replacing each integer in the message with its binary codeword, with no spaces between the codewords; that is, you just wind up with a seamless stream of bits. Suppose your coded message starts with

$$10011 \cdots .$$

Then you have no way of knowing whether the first integer in your message is 9 or 19. To get around this drawback, we require that *each binary codeword must not be a prefix of any other binary codeword.* This requirement on our coding method is called *the prefix condition.* We prove the following fact in EE 5585:

Useful Fact. If

$$p(i), \quad i = 1, 2, 3, \cdots$$

is a probability distribution, then there is a way to assign binary codewords satisfying the prefix condition so that the binary codeword assigned to i is of length

$$\lceil -\log_2 p(i) \rceil.$$

Example 8.6. Consider the probability distribution

$$p(i) = 2^{-i}, \quad i = 1, 2, \cdots .$$

(This is just the Geometric(p) distribution with $p = 1/2$.) The binary codeword assigned to i according to this distribution will then be of length

$$-\log_2(2^{-i}) = i.$$

Here is one possible coding method that satisfies this:

i	*codeword*
1	1
2	01
3	001
4	0001
5	00001

One continues in this way. The codeword for i is simply $i-1$ zeroes following by a one.

Example 8.7. Using EE 3015 tricks involving Fourier Series, one can prove that

$$\sum_{i=1}^{\infty} \frac{1}{i^2} = \frac{\pi^2}{6}.$$

Dividing both sides by $\frac{\pi^2}{6}$, we have

$$\sum_{i=1}^{\infty} \frac{6}{\pi^2 i^2} = 1.$$

This gives us a probability distribution

$$p(i) = \frac{6}{\pi^2 i^2}, \quad i = 1, 2, 3, \cdots.$$

There must therefore exist a coding method for the positive integers, satisfying the prefix condition, such that the codeword for i has length equal to

$$\left\lceil -\log_2 \left(\frac{6}{\pi^2 i^2} \right) \right\rceil.$$

I cannot give this method here, which is a type of code called an Elias code. (Take EE 5585 to see how to build Elias codes!) However, we can make an interesting observation about this particular code. For large i, it is easy to see that the length of the binary codeword assigned to i is approximately equal to

$$2 \log_2 i.$$

The simple-minded method we mentioned at the beginning of this section (coding each integer into its usual binary expansion) achieves codeword length for i approximately equal to

$$\log_2 i$$

for large i. So the Elias code is giving codewords of length roughly twice as long as the simple-minded code. Remember the Elias code satisfies the prefix condition and the simple-minded code

does not! Since we cannot use the simple-minded code, one's goal is to find a code satisfying the prefix condition in which the codeword lengths are not too much longer than the codeword lengths of the simple-minded code. The Elias code is one possible code to fulfill this goal.

Exercise. Find the codeword lengths for the first few positive integers using the simple-minded code. Find the codeword lengths for these same positive integers using the Elias code. Compare.

Lecture 9

Chapters 2-3 Part 3

In this lecture, I talk about

- concept of *cumulative distribution function* (CDF)
- introduction to the notions of *mean* and *variance* of a probability distribution
- define the *Nonstandard Gaussian*, *Nonstandard Uniform*, and *Exponential distributions*
- estimation of a PDF from data

9.1 Cumulative Distribution Function

The cumulative distribution function (CDF) of RV X is the function $F_X(x)$ defined for all real x by

$$F_X(x) \triangleq P(X \leq x).$$

In words, $F_X(x)$ is the probability that RV X will take a value less than or equal to x. Note that

$$F_X(x) = \int_{-\infty}^{x} f_X(u)du.$$

That is, if you plot the density function $f_X(x)$, locate x on the real line and draw a vertical line through x, then $F_X(x)$ will be the area under the density function that lies to the left of this vertical line. For example, look at Figure 3.6(a) on page 121 of your textbook; the shaded area is a value of the CDF.

It is sometimes useful to interpret the CDF from the following EE 3015 perspective, namely, the CDF is what you get when you pass the PDF through an integrator:

$$f_X(x) \rightarrow \boxed{\int_{-\infty}^{x}} \rightarrow F_X(x)$$

Recall the following properties of an integrator:

$$\delta(x-a) \rightarrow \boxed{\int_{-\infty}^{x}} \rightarrow u(x-a)$$

$$u(x-a) \rightarrow \boxed{\int_{-\infty}^{x}} \rightarrow r(x-a)$$

In case you've forgotten EE 3015 notation, $u(x-a)$ is the unit step function starting at $x=a$, and $r(x-a)$ is the ramp function starting at $x=a$:

$$u(x-a) = \begin{cases} 1, & x \geq a \\ 0, & elsewhere \end{cases}$$

$$r(x-a) = \begin{cases} x-a, & x \geq a \\ 0, & elsewhere \end{cases}$$

In other words, the response of an integrator to a delta function input is a unit step function, and the response of an integrator to a unit step function input is a ramp function.

CDF of a Discrete Random Variable

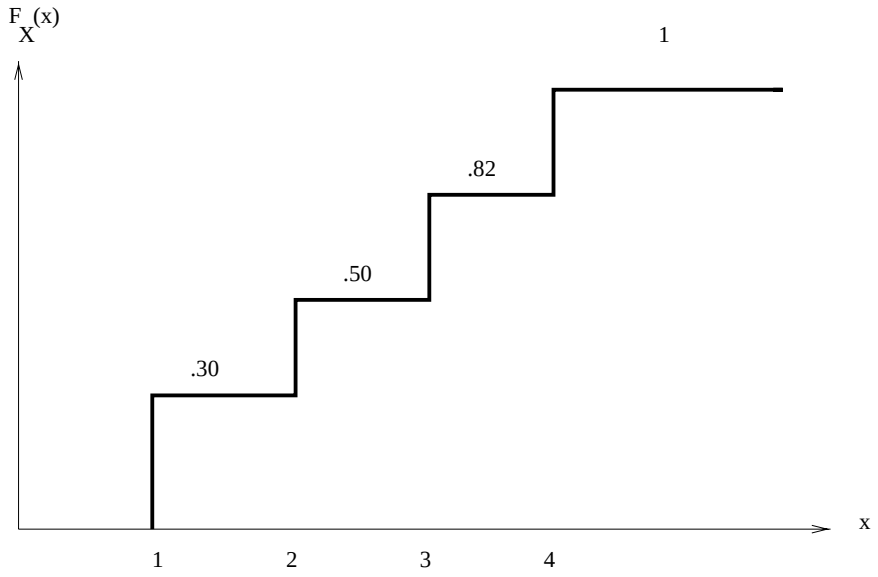
For a discrete RV X, the PDF $f_X(x)$ is a linear combination of delta functions, and therefore the CDF is a linear combination of unit step functions, which is a staircase function. We can say the following:

- *The CDF $F_X(x)$ of a discrete RV X is a staircase function. The jumps in the CDF plot occur at the values of X, and the magnitudes of the jumps are equal to the PMF values. As x goes from $-\infty$ to ∞, $F_X(x)$ increases from 0 to 1.*

Example 9.1. Suppose we have a discrete RV X taking the values $1, 2, 3, 4$, with the following PMF:

$$p^X(x) = \begin{cases} 0.30, & x=1 \\ 0.20, & x=2 \\ 0.32, & x=3 \\ 0.18, & x=4 \end{cases}$$

We have plotted the corresponding CDF $F_X(x)$ at the top of the next page:



Here is how we obtained the plot of $F_X(x)$ by inspection. The smallest value of X is $x = 1$ and so the plot of $F_X(x)$ is zero for $x < 1$. The first jump in the CDF is at $x = 1$ and the magnitude of the jump is $P^X(1) = 0.30$. The CDF takes the value 0.30 until you come to the next value of X at $x = 2$. At $x = 2$, the CDF jumps an amount equal to $P^X(2) = 0.20$, increasing up to

$$0.20 + 0.30 = 0.50,$$

the "cumulative probability" at $x = 2$. Similarly, you have a jump of $P^X(3) = 0.32$ at $x = 3$, taking the CDF up to the value

$$0.20 + 0.30 + 0.32 = 0.82,$$

the "cumulative probability" at $x = 3$. Finally, you have a jump of $P^X(4) = 0.18$ at $x = 4$, taking the CDF up to the value

$$0.20 + 0.30 + 0.32 + 0.18 = 1.$$

To the right of this largest value $x = 4$ of X, the CDF will take the constant value 1 (there are no more jumps, all the probability has been accumulated!).

Conversely, if I gave you the CDF plot above, you could go backwards and find the PMF on the preceding page: just see where the jumps occur and what the magnitudes of the jumps are in order to get the PMF values.

You can compute probabilities involving a discrete RV X directly from the CDF $F_X(x)$ as follows:

$$\begin{aligned} P(a < X \leq b) &= F_X(b^+) - F_X(a^+) && (9.1)\\ P(a < X < b) &= F_X(b^-) - F_X(a^+) && (9.2)\\ P(a \leq X < b) &= F_X(b^-) - F_X(a^-) && (9.3)\\ P(a \leq X \leq b) &= F_X(b^+) - F_X(a^-) && (9.4) \end{aligned}$$

In the preceding formulas, $F_X(x^+)$ denotes the right hand limit of $F_X(x)$ as you approach x from the right, and $F_X(x^-)$ denotes the left hand limit of $F_X(x)$ as you approach x from the left. If you find it hard to remember these four formulas, you can instead just remember the formula

$$P(X \in E) = \textit{ sum of jumps in } F_X(x) \textit{ occuring at values } x \in E. \quad (9.5)$$

Example 9.2. We take the same discrete RV X as in Example 9.1, with the CDF plotted on page 22. Using the CDF and the four formulas (9.1)-(9.4), we do the following calculations:

$$\begin{aligned} P(2 < X \leq 4) &= F_X(4) - F_X(2) = 1 - .50 = .50\\ P(2 < X < 4) &= F_X(4^-) - F_X(2) = .82 - .50 = .32\\ P(2 \leq X < 4) &= F_X(4^-) - F_X(2^-) = .82 - .30 = .52\\ P(2 \leq X \leq 4) &= F_X(4) - F_X(2^-) = 1 - .30 = .70\\ P(X \geq 3) &= F_X(\infty) - F_X(3^-) = 1 - .50 = .50 \end{aligned}$$

Or, using equation (9.5),

$$\begin{aligned} P(2 \leq X \leq 4) &= (jump\ at\ x = 2) + (jump\ at\ x = 3) + (jump\ at\ x = 4)\\ &= 0.20 + 0.32 + 0.18 = 0.70 \end{aligned}$$

CDF of a Continuous Random Variable

- *The CDF $F_X(x)$ of a continuous RV X is continuous at all values of x (i.e., the CDF has no jumps). As x goes from $-\infty$ to ∞, $F_X(x)$ increases from 0 to 1. Furthermore, the derivative of the CDF is the PDF:*

$$\frac{dF_X(x)}{dx} = f_X(x). \quad (9.6)$$

Example 9.3. Suppose X is a standard uniform RV. Then its density $f_X(x)$ is the unit rectangular pulse from $x = 0$ to $x = 1$. Consider this $f_X(x)$ as the input to an integrator. Then the

integrator response will be a ramp function from $x = 0$ to $x = 1$. It is pretty immediate, then, that the CDF in this case takes the form

$$F_X(x) = \begin{cases} 0, & x < 0 \\ x, & 1 \le x \le 1 \\ 1, & x > 1 \end{cases}$$

Example 9.4. This example illustrates how we can use equation (9.6) to obtain $f_X(x)$ from $F_X(x)$ for a continuous RV. Suppose our continuous RV X has the following CDF:

$$F_X(x) = \begin{cases} 0, & x < 0 \\ x^2, & 1 \le x \le 1 \\ 1, & x > 1 \end{cases}$$

Notice that the CDF comes in three pieces. Differentiating each piece, we get

$$f_X(x) = \begin{cases} 0, & x < 0 \\ 2x, & 1 \le x \le 1 \\ 0, & x > 1 \end{cases}$$

We can do direct probability calculation using the CDF much easier in the case of a continuous RV than in the case of a discrete RV. In the case of a discrete RV, we have to worry about the four separate equations (9.1)-(9.4). In the case of a continuous RV, since $F_X(x)$ has no discontinuities, the four equations (9.1)-(9.4) reduce to just a single calculation, as follows:

- *For a continuous RV X, you don't have to worry about the endpoints of intervals in probability computations. That is, all four of the probabilities*

$$P(a < X \le b), \;\; P(a < X < b), \;\; P(a \le X < b), \;\; P(a \le X \le b)$$

are equal to

$$F_X(b) - F_X(a).$$

Example 9.5. Suppose we have a continuous RV X with CDF

$$F_X(x) = \begin{cases} 0, & x < 0 \\ 1 - \exp(-x), & x \ge 0 \end{cases}$$

Then

$$P(1 < X < 2) = F_X(2) - F_X(1) = (1 - \exp(-2)) - (1 - \exp(-1)) = 0.2325.$$

CDF of a Mixed Random Variable

We have seen that for a discrete RV, the changes in the CDF values occur only as jumps occuring at the values of the RV. For a continuous RV, we have seen that the CDF values change continuously over the whole real line (no jumps). The CDF of a mixed RV combines these two features:

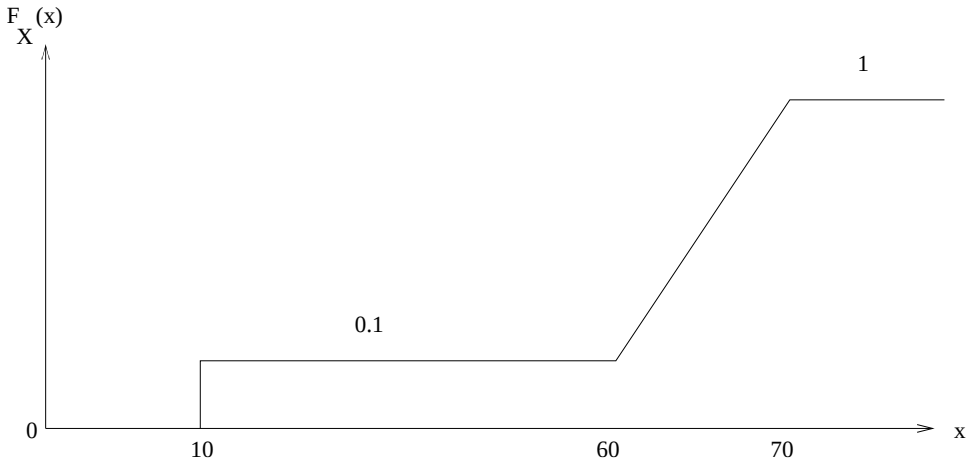
- *As x goes from $-\infty$ to ∞, the CDF $F_X(x)$ of a mixed RV X increases from 0 to 1. Some of this increase is due to jumps, where each jump occurs at a discrete values of X, with the magnitude of each jump being equal to the probability of occurence of the correponding discrete value of X. The rest of the increase in $F_X(x)$ is due to the continuous change in the values of $F_X(x)$ over the parts of the real line not containing the discrete values of X.*

Example 9.6. Let us go back to the mixed RV X of Example 7.7, whose density is

$$f_X(x) = (0.1)\delta(x-10) + (0.9)(1/10)[u(x-60) - u(x-70)].$$

Note that $f_X(x)$ consists of a delta function at $x = 10$ combined with a rectangular pulse from $x = 60$ to $x = 70$. The presence of the delta function will produce a unit step function component of $F_X(x)$ going from $x = 10$ to $x = 60$. The presence of the rectangular pulse will add on a ramp function component of $F_X(x)$ going from $x = 60$ to $x = 70$. At $x = 70$, $F_X(x)$ becomes 1 and remains at 1 as we move x further to the right. We therefore obtain

$$F_X(x) = \begin{cases} 0, & x < 10 \\ 0.1, & 10 \le x \le 60 \\ 0.1 + (0.9)(1/10)(x-60), & 60 < x \le 70 \\ 1, & x \ge 70 \end{cases}$$



9.2 Means and Variances

We denote the *mean* of a RV X using the notation μ_X or the notation $E[X]$ (also called the *expected value of* X). If it is clear from the context what the RV X is, then we can abbreviate μ_X as simply μ.

The mean μ_X is defined by

$$\mu_X \triangleq \int_{-\infty}^{\infty} x f_X(x) dx.$$

For a discrete RV X, this reduces to the following summation:

$$\mu_X = \sum_x x P^X(x).$$

We denote the *variance* of a RV X using the notation σ_X^2 or the notation $Var(X)$. The positive square root of the variance is denoted σ_X and is called the *standard deviation* of X. If it is clear from the context what the RV X is, then we can abbreviate σ_X^2 as simply σ^2 and can abbreviate σ_X as simply σ.

The variance σ_X^2 is defined by

$$\sigma_X^2 \triangleq \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x) dx.$$

9.2.1 Intuitive Meaning of μ_X, σ_X^2

Meaning of μ_X: If the probability model is a good one, then you expect the arithmetic average of a large number of observed values of X (from independent trials) to be close to μ_X a high percentage of the time. If `x` is the vector formed by your observations, then this arithmetic average is computed via Matlab as `mean(x)`. Thus, we expect that `mean(x)` (which varies randomly) will be close to μ_X most of the time.

Meaning of σ_X^2: Large variance means you have a significant chance of observing a value of X far from the mean μ_X. On the other hand, variance close to zero means most of the time the observed value of X will be close to μ_X. See Figure 3.5 on page 119 of the book to see the plot of a density for which the variance is large as well as the plot of a density for which the variance is close to zero.

I will do some μ_X computations in the following. I will defer variance computations to a later lecture, because we need some specialized tools to help us compute variance.

Example 9.7. A discrete RV X takes values $1, 2, 3, 4$ with PMF as follows:

$$P^X(x) = \begin{cases} 0.1, & x = 1 \\ 0.2, & x = 2 \\ 0.3, & x = 3 \\ 0.4, & x = 4 \end{cases}$$

We have

$$\mu_X = 1(0.1) + 2(0.2) + 3(0.3) + 4(0.4) = 3.$$

We expect that when we perform the experiment a large number of times, the average of the observed values of X on these trials will be close to 3.

Example 9.8. Let us compute the mean of a Poisson(α) RV X:

$$\begin{aligned}
\mu_X &= \sum_{x=0}^{\infty} x P^X(x) \\
&= \sum_{x=0}^{\infty} x \exp(-\alpha) \frac{\alpha^x}{x!} \\
&= \sum_{x=1}^{\infty} x \exp(-\alpha) \frac{\alpha^x}{x!} \\
&= \alpha \sum_{x=1}^{\infty} \exp(-\alpha) \frac{\alpha^{x-1}}{(x-1)!} \\
&= \alpha \sum_{x=0}^{\infty} \exp(-\alpha) \frac{\alpha^x}{x!} = \alpha
\end{aligned}$$

We conclude that the mean of a Poisson(α) RV is α. (See Appendix A, page 503.) Now go back to Example 8.4 of Lecture 8 Notes. There, we modeled the number of phone calls in a time interval as a Poisson(α) RV with $\alpha = 4.2$ by observing the average number of phone calls in this time interval over several days to be 4.2. The reason why we did this should now be clear to the reader.

Example 9.9. We compute the mean of an Exponential(a) random variable X. This means X has the density

$$f_X(x) = ae^{-ax}u(x).$$

We can use Laplace transforms in a clever way to compute μ_X. We have

$$\mu_X = \int_0^{\infty} xae^{-ax}dx. \tag{9.7}$$

Recall the following Laplace transform formula from EE 3015:

$$\int_0^{\infty} te^{-st}dt = \frac{1}{s^2}. \tag{9.8}$$

This formula says that the Laplace transform of the ramp function $tu(t)$ is $1/s^2$. In equation (9.8), the Laplace transform variable s can be any complex number in the ROC (region of convergence)

of the Lapace transform. In particular, s can be any positive number. Let $s = a$, the parameter of the Exponential(a) distribution. We conclude that

$$\int_0^\infty te^{-at}dt = \frac{1}{a^2}.$$

Multiplying both sides by a, we obtain

$$\int_0^\infty tae^{-at}dt = \frac{1}{a}.$$

Therefore, we have proved that for the Exponential(a) distribution,

$$\mu_X = \frac{1}{a}.$$

(See Appendix A of your textbook, page 504; notice that I am referring to the parameter of the exponential distribution as a instead of λ which the book uses.) Another way to do the integral on the right side of (9.7) is to integrate by parts.

9.2.2 Symmetry Rule for Computing μ_X

Useful Fact: If the density $f_X(x)$ of RV X is symmetric about some value $x = C$, then C must be the mean μ_X!

Example 9.10. Flip a fair die, and let X be the number that comes up. The PMF is equiprobable over the 6 values $1, 2, 3, 4, 5, 6$. This probability distribution is symmetric about $x = 3.5$. Therefore $\mu_X = 3.5$. If we flip this die hundreds of times and average up the numbers we get, we expect that this average will be close to 3.5. If you want to do this test yourself using Matlab, execute the Matlab command

```
mean(ceil(6*rand(1,50000)))
```

and see what you get.

Example 9.11. Consider the Binomial(n, p) distribution with $p = 1/2$. The PMF is symmetric about $n/2$. (You can sketch some of these PMF's for different n values to convince yourself of this fact.) Therefore the mean of the Binomial(n, p) distribution is $n/2$. For example, if you flip a fair coin three times and count the number of heads, the expected number of heads is $n/2 = 3/2 = 1.5$.

Example 9.12. We say that X has the Uniform(a, b) distribution if the density $f_X(x)$ is a rectangular pulse over the interval $[a, b]$. Since the area under this pulse must be 1, this forces the amplitude of the pulse to be $\frac{1}{b-a}$. That is,

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b \\ 0, & elsewhere \end{cases}$$

(If we take $[a, b]$ to be the unit interval $[0, 1]$, this is the *standard uniform distribution.* For all other cases, we have a *nonstandard uniform distribution.*) Let us compute the mean of X. The density of $f_X(x)$ is clearly symmetric about $x = \frac{a+b}{2}$, the midpoint of the interval $[a, b]$. We immediately conclude that

$$\mu_X = \frac{a+b}{2}.$$

(This coincides with the result in Appendix A, page 506.)

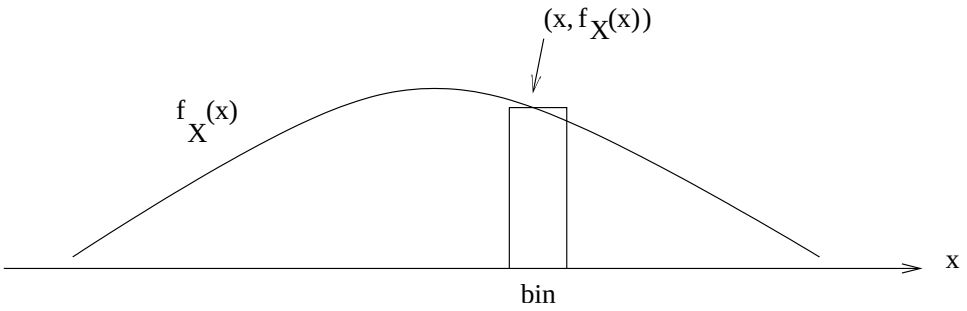
Example 9.13. We say X has the Gaussian(μ, σ) distribution if its density is

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{\sigma^2}\right), \quad -\infty < x < \infty.$$

By symmetry, the mean of X is μ. In a later lecture, we will show that the variance of X is σ^2. (This coincides with the result in Appendix A, page 505.) The special case in which $\mu = 0$, $\sigma = 1$ is called the *standard Gaussian distribution.* All the other cases are *nonstandard Gaussian.*

9.3 Estimating $f_X(x)$

Suppose we want to see what density $f_X(x)$ we should use as the probability model for RV X. Suppose we can take as many observations of the value of X (over independent trials) as we want. How do we use these observations to estimate $f_X(x)$? The figure below suggests a way to proceed.



Imagine n observations of X distributed along the x-axis. (We take n to be very large; in Recitation 3, we take $n = 100000$.) Suppose these observations go from a minimum value of a to a maximum value of b. Then we can partition up the interval $[a, b]$ into subintervals of $[a, b]$ of equal length, and we will call each such subinterval a "bin". (In Recitation 3, we take 100 bins.) Let the width of each bin be Δ. We have sketched one typical bin in the above figure. We chose point x to be the point on the x-axis in the center of the bin, and the point $(x, f_X(x))$ is the corresponding point on

the $f_X(x)$ curve. If the bin width Δ is sufficiently small, then calculus tells us that $\Delta f_X(x)$, the area of the rectangle erected above the bin, is approximately equal to

$$\int_{bin} f_X(x)dx = P(X \in \ bin). \tag{9.9}$$

Let k_{bin} be the number of the n observations that fall in the bin. Let us take the ratio k_{bin}/n as an estimate of the probability on the right side of (9.9). Then for each bin we have the approximate relationship

$$\Delta f_X(x) \approx \frac{k_{bin}}{n}.$$

Solving for $f_X(x)$, we obtain

$$f_X(x) \approx \frac{k_{bin}}{n\Delta}.$$

Here is our conclusion:

- *To obtain the estimated density curve $f_X(x)$, measure up from the center x of each bin a height equal to*

  $$\frac{k_{bin}}{n\Delta}$$

  to obtain the point

  $$(x, \frac{k_{bin}}{n\Delta}).$$

  You will obtain one of these points for each bin; connect up these points with straight lines. This is your estimated $f_X(x)$ curve.

We use this method in Experiment 1 of Recitation 3 to convince you that Matlab's pseudorandom number generators `rand` and `randn` are doing a good job. (We generate 100000 data points according to each of these, do the density estimate, and get an approximation of the standard uniform density and the standard Gaussian density, respectively.)

Lecture 10

Chapters 2-3 Part 4

In these Lecture 10 Notes, I talk a little bit more about the CDF, show you how to compute Gaussian probabilities, discuss the expectation operator and its properties, and do some variance computations.

10.1 More on CDF

In Lecture 9 Notes, I explained that for a continuous RV X, a probability can easily be computed from the CDF as

$$P(a \leq X \leq b) = F_X(b) - F_X(a).$$

For a discrete RV, this formula may not hold, and I gave four complicated formulas at the top of page 23 for handling the discrete case. However, there is one special case of discrete RV where the situation is a bit simpler, namely, the case in which the values of X are integers. In this case, we can say

$$P(a \leq X \leq b) = F_X(b) - F_X(a-1), \quad a, b \; integers. \tag{10.1}$$

If one or both of the inequalities on the left side of (10.1) is a strict inequality, you can just make the interval shorter in order to obtain an interval in which the endpoints are included. For example, you can write

$$P(3 < X < 7) = P(4 \leq X \leq 6) = F_X(6) - F_X(3).$$

Be sure to only use (10.1) for discrete RV's taking integer values!

Please refer to Problem 2.3 of the Chapter 2-3 Solved Problems for an example in which I use equation (10.1).

10.2 Apples and Oranges

At this point in the course, we typically find that some students have trouble keeping the six common distributions of Chapters 2-3 straight. The following table might help you:

Geometric	$1, 2, 3, \cdots$
Poisson	$0, 1, 2, 3, \cdots$
Binomial	$0, 1, 2, \cdots, n$
Uniform	$a \leq x \leq b$
Gaussian	$-\infty < x < \infty$
Exponential	$x \geq 0$

In the left column, I list the 6 types of probability distributions we are covering. In the right column, I give the values of the RV's having these distributions. You see that no two of these distributions are over the same values. Thus, if I tell you that a particular RV X will have one of these 6 distributions, all you have to do is examine what the possible values of X are in order to identify what the distribution is. In other words, these distributions are as unlike as "apples and oranges".

Example 10.1. Let RV X be the number of hurricanes that will hit Florida in 2010. The possible values of X are $0, 1, 2, \cdots$. If we model the distribution of X using one of our 6 common distributions, the Poisson distribution would be the only one that makes sense in this context.

Example 10.2. Suppose I take a large number of lightbulbs, let each of them burn until they burn out, and let RV X be the average lifetime (in hours) that I've observed for these bulbs. Suppose the expected lifetime of any of the bulbs is 1000 hours. Suppose I want to model the RV

$$Y = X - 1000$$

using one of the six common distributions. Which one would it make most sense to use? Notice that sometimes X will take a value less than 1000 and sometimes greater than 1000, and the value of Y therefore ranges over both positive and negative real numbers. Of the 6 distributions, only the Gaussian distribution does this. Later in the course, we will model Y as a Gaussian random variable.

10.2.1 Discrete Uniform Distribution

There is a 7th common distribution that is so trivial that I have not mentioned it yet, namely, the *discrete uniform distribution*. This is the distribution in which the values of the RV are finite in number, equally spaced on the real number axis, and equiprobable. For example, the number that comes up on one toss of a fair die has a discrete uniform distribution. Suppose you have a discrete uniform distribution in which the values are $0, 1, 2, 3, 4, 5$. I suppose there might exist a student somewhere who would get confused between that and the Binomial(n, p) distribution with

$n = 5$, since both of these distributions are over the integers $0, 1, 2, 3, 4, 5$. However, the PMF of the discrete uniform distribution is *flat* and the PMF of the Binomial(n, p) distribution is first increasing, then decreasing, reaching a peak somewhere around np (as we shall see later on). So, even though there are some discrete uniform distributions in which the values are the same as for a binomial distribution, one would not be likely to confuse the two.

10.3 Computing Gaussian Probabilities

In this section, I show you how to use the table on page 123 of your textbook to compute Gaussian probabilities. There are also Matlab ways to do this, using either the Matlab function "`erf`" or the Matlab symbolic integrator "`int`"; see Recitation 3 Instructions for these other methods.

Suppose X is a Gaussian(μ,σ) RV, and our goal is to compute a probability like $P(a \leq X \leq b)$. We can do this making use of the following useful fact that we shall prove in a later lecture:

Useful Fact: If $X \sim$ Gaussian(μ,σ), then $Z \sim$ Gaussian(0,1) (standard Gaussian), where Z is obtained from X by the following linear change of variable:

$$Z = \frac{X - \mu}{\sigma}.$$

Using the Useful Fact, we can do the following manipulations:

$$\begin{aligned} P(a \leq X \leq b) &= P\left(\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right) \\ &= P\left(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\right) \\ &= F_Z\left(\frac{b-\mu}{\sigma}\right) - F_Z\left(\frac{a-\mu}{\sigma}\right) \end{aligned}$$

In the preceding, $F_Z(z)$ denotes the CDF of the standard Gaussian distribution. This is such a common CDF that it has been given the notation $\Phi(z)$ in almost all statistics textbooks. That is, the function $\Phi(z)$ is defined by:

$$\Phi(z) \triangleq \int_{-\infty}^{z} \left(\frac{1}{\sqrt{2\pi}}\right) \exp(-u^2/2) du, \quad -\infty < z < \infty.$$

Here then is the formula you can use for evaluating Gaussian probabilities:

$$P(a \leq X \leq b) = \Phi\left(\frac{b-\mu_X}{\sigma_X}\right) - \Phi\left(\frac{a-\mu_X}{\sigma_X}\right). \tag{10.2}$$

On page 123, the function $\Phi(z)$ is tabulated between $z = 0$ to $z = 2.99$. For $z \geq 3$, $\Phi(z)$ is approximately equal to one. The reason that $\Phi(z)$ is not tabulated for $z < 0$ is because one can make use of symmetry:

$$\Phi(z) = 1 - \Phi(-z). \tag{10.3}$$

Here are a couple of worked examples in which I use the table on page 123.

Example 10.3. Let $X \sim \text{Gaussian}(\mu, \sigma)$. Let us compute the probability that X falls within one standard deviation of the mean. That is, let us compute the probability

$$P(\mu - \sigma \leq X \leq \mu + \sigma).$$

Using (10.2),

$$P(\mu - \sigma \leq X \leq \mu + \sigma) = \Phi(1) - \Phi(-1).$$

The value $\Phi(1)$ is looked up on page 123 to be 0.8413. Using equation (10.3),

$$\Phi(-1) = 1 - \Phi(1) = 0.1587.$$

Therefore

$$P(\mu - \sigma \leq X \leq \mu + \sigma) = (0.8413) - (0.1587) = 0.6826.$$

Example 10.4. Frequently, one has to use the table on page 123 in reverse, in order to find the particular z value for which $P[Z \leq z]$ is equal to some fixed probability. Here is an example of this type. Let $X \sim \text{Gaussian}(\mu, \sigma)$, where $\mu = 2$ and $\sigma^2 = 9$. Let us find the constant C such that

$$P[X \geq C] = 0.05.$$

We have $Z = (X - 2)/3$, and

$$P[X \geq C] = P\left[Z \geq \frac{C-2}{3}\right] = 1 - \Phi\left(\frac{C-2}{3}\right).$$

Using the table on page 123 in reverse, we see that

$$\begin{aligned} \Phi\left(\frac{C-2}{3}\right) &= 0.95 \\ \frac{C-2}{3} &= 1.645 \\ C = 3(1.645) + 2 &= 6.935 \end{aligned}$$

Exercise. For $X \sim \text{Gaussian}(\mu, \sigma)$, find the positive constant C such that

$$P(\mu - C\sigma \leq X \leq \mu + C\sigma) = 0.90.$$

(This constant C shall be very important to us later on in the course.)

For another example involving computation of Gaussian probabilities, see Problem 3.2 of the Chapter 2-3 Solved Problems.

10.4 Expectation Operator

We have used the expectation operator "E" in just one context so far, namely, in the expression $E[X]$ for expected value (mean μ_X) of a random variable X. However, we can take the expected value of any function of X according to the following definition:

$$E[\phi(X)] \triangleq \int_{-\infty}^{\infty} \phi(x) f_X(x) dx. \tag{10.4}$$

If we put $\phi(x) = x$ in the integrand on the right side of (10.4), then we get the formula for $E[X]$ as a special case. Another interesting thing to try is to put $\phi(x) = (x - \mu_X)^2$ in the right side of (10.4). Then you see that

$$E[(X - \mu_X)^2] = Var[X] = \sigma_X^2.$$

So, the expectation operator "E" gives us a convenient way to express variance as well as mean.

Properties of the Expectation Operator

Here are three useful properties of the expectation operator.

Property 1: $E[\phi_1(X) + \phi_2(X)] = E[\phi_1(X)] + E[\phi_2(X)]$.

Property 2: If C is a constant,

$$E[C\phi(X)] = CE[\phi(X)].$$

Property 3: If C is a constant,

$$E[C] = C.$$

These properties are trivial to prove. (For example, from calculus, the integral of a sum is the sum of the integrals, which yields Property 1.)

Using the Properties, one can prove many interesting things, and we will see some of these things in this course. For right now, I will prove for you the following Useful Fact.

Useful Fact. For any RV X and any constant C,

$$E[(X - C)^2] = \sigma_X^2 + (C - \mu_X)^2. \tag{10.5}$$

Equation (10.5) tells us that $E[(X-C)^2]$ is minimized when the constant C is chosen as $C = \mu_X$. Thus, the mean of a random variable is the unique real number which the values of the random variable are closest to on average.

Proof of Useful Fact. Let μ be the mean of X. The Properties are used repeatedly, in the following steps:

$$\begin{aligned} E[(X-C)^2] &= E[\{(X-\mu)+(\mu-C)\}^2] \\ &= E[(X-\mu)^2 + 2(X-\mu)(\mu-C) + (\mu-C)^2] \\ &= E[(X-\mu)^2] + 2(\mu-C)E[(X-\mu)] + E[(\mu-C)^2] \\ &= \sigma_X^2 + 2(\mu-C)(\mu-\mu) + (\mu-C)^2 \\ &= \sigma_X^2 + (C-\mu)^2 \end{aligned}$$

10.5 Variance Computations

I have not yet done much with variance computations. This section will remedy this situation.

Example 10.5. Let's verify that the variance of a Gaussian(μ, σ) RV X is σ^2. We have

$$\sigma_X^2 = \int_{-\infty}^{\infty} (x-\mu)^2 \left(\frac{1}{\sigma\sqrt{2\pi}}\right) \exp\left(-\frac{(x-\mu)^2}{\sigma^2}\right) dx.$$

Make the change of variable

$$\begin{aligned} y &= \frac{x-\mu}{\sigma} \\ dy &= dx/\sigma \end{aligned}$$

Then the preceding integral becomes

$$\sigma_X^2 = \sigma^2 \int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi}}\right) x^2 \exp(-x^2/2) dx.$$

The integral on the right is 1. (Use integration by parts or the method in Example 7.6.) Therefore,

$$\sigma_X^2 = \sigma^2.$$

Useful Result. Here is a nice way to evaluate variance:

$$\sigma_X^2 = E[X^2] - \mu_X^2. \tag{10.6}$$

$E[X^2]$ is called the *second moment* of RV X, and is calculated by

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x) dx.$$

You obtain formula (10.6) from (10.5) by plugging in $C = 0$.

Exercise. For an Exponential(a) RV X, prove that

$$\sigma_X^2 = \frac{1}{a^2},$$

as follows. First, look up the Laplace transform formula

$$\int_0^\infty t^2 \exp(-st)dt = \frac{2}{s^3}.$$

Then use this formula to conclude that

$$\int_0^\infty x^2 (a \exp(-ax))dx = \frac{2}{a^2}.$$

We know from Lecture 9 Notes that the mean is $1/a$. You are now ready to plug into formula (10.6).

Example 10.6. Let X be the discrete RV in Example 9.7. The mean was computed to be $\mu_X = 3$. The second moment is

$$E[X^2] = 1^2(0.1) + 2^2(0.2) + 3^2(0.3) + 4^2(0.4) = 10.$$

The variance is therefore one:

$$\sigma_X^2 = E[X^2] - \mu_X^2 = 10 - 9 = 1.$$

Example 10.7. Let X be the continuous RV with density

$$f_X(x) = \begin{cases} 2x, & 0 \le x \le 1 \\ 0, & elsewhere \end{cases}$$

We have

$$\mu_X = \int_0^1 x(2x)dx = 2/3.$$

The second moment is computed as:

$$E[X^2] = \int_0^1 x^2(2x)dx = 1/2.$$

We now have

$$\sigma_X^2 = E[X^2] - \mu_X^2 = 1/2 - 4/9 = 1/18.$$

Lecture 11

Chapters 2-3 Part 5

11.1 Gaussian Table Lookup Example

Let $X \sim \mathit{Gaussian}(\mu, \sigma)$. Using page 123 of your textbook, we find the constant C such that

$$P(\mu - C\sigma \leq X \leq \mu + C\sigma) = 0.95. \tag{11.1}$$

In other words, C is the constant such that X will fall within C standard deviations of the mean with probability 0.95.

Make the change of variable

$$Z = \frac{X - \mu}{\sigma},$$

obtaining standard Gaussian RV Z. Then (11.1) becomes

$$P(-C \leq Z \leq C) = 0.95.$$

At this point, it might help for you to sketch a plot of the standard Gaussian density $f_Z(z)$. The area caught under this bell-shaped density curve between $z = -C$ and $z = C$ is 0.95. The areas in each of the two tails (by symmetry) must be $0.05/2 = 0.025$. Therefore, the area to the left of the vertical line $z = C$ must be $0.95 + 0.025 = 0.975$. This area is a CDF value, namely, $F_Z(C)$, which is also written $\Phi(C)$. We have shown that

$$\Phi(C) = 0.975.$$

Using the table on page 123 in reverse, you see that $C = 1.96$. We conclude that

$$P(\mu - (1.96)\sigma \leq X \leq \mu + (1.96)\sigma) = 0.95.$$

This relationship will be useful to us later on in the course.

11.2 Linear Change of Variable

We have a given RV X. We make the following linear change of variable to obtain a new RV Y:

$$Y = aX + b. \tag{11.2}$$

In this equation, a, b are real constants with $a \neq 0$. The RV X has a density $f_X(x)$, a mean μ_X, and a variance σ_Y^2. The new RV Y has a density $f_Y(y)$, a mean μ_Y, and a variance σ_Y^2. We are going to find out how to obtain $f_Y(y)$, μ_Y, σ_Y^2 from $f_X(x)$, μ_X, σ_X^2, respectively. I will prove the following three formulas:

Formula 1: $f_Y(y) = \dfrac{1}{|a|} f_X\left(\dfrac{y-b}{a}\right)$.

Formula 2: $\mu_Y = a\mu_X + b$.

Formula 3: $\sigma_Y^2 = a^2\sigma_X^2$, or equivalently $\sigma_Y = |a|\sigma_X$.

Proof of Formula 1. Let us assume that $a > 0$, and then you can modify the argument to handle the case $a < 0$. Our approach is to obtain the CDF $F_Y(y)$ in terms of the CDF $F_X(x)$ and then to differentiate to obtain the density $f_Y(y)$ in terms of the density $f_X(x)$.

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(aX + b \leq y) \\ &= P\left(X \leq \frac{y-b}{a}\right) \\ &= F_X\left(\frac{y-b}{a}\right) \end{aligned}$$

We have shown that

$$F_Y(y) = F_X\left(\frac{y-b}{a}\right), \quad if\ a > 0. \tag{11.3}$$

Differentiate both sides of (11.3) with respect to y, where you differentiate the right side using the chain rule from calculus:

$$\frac{dF_X(x)}{dy} = \frac{dF_X(x)}{dx}\frac{dx}{dy} = f_X(x)\frac{dx}{dy}.$$

This gives us the following formula:

$$f_Y(y) = \frac{1}{a} f_X\left(\frac{y-b}{a}\right), \quad if\ a > 0. \tag{11.4}$$

If $a < 0$, then the preceding argument has to be modified because then we have

$$P(aX + b \leq y) = P\left(X \geq \frac{y-b}{a}\right) = 1 - F_X\left(\frac{y-b}{a}\right),$$

if $\frac{y-b}{a}$ is not a discrete value of X. The modified argument would yield

$$f_Y(y) = \left(\frac{1}{-a}\right) f_X\left(\frac{y-b}{a}\right), \quad if\ a < 0. \tag{11.5}$$

You can then combine equations (11.4)-(11.5) into the following single equation valid for all $a \neq 0$:

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right).$$

This completes the proof of Formula 1.

Proof of Formula 2. Apply the expectation operator "E" to both sides of equation (11.2) and then use the properties of the expectation operator given in the Lecture 10 Notes:

$$\begin{aligned} E[Y] &= E[aX+b] \\ &= E[aX] + E[b] \\ &= aE[X] + b \end{aligned}$$

We have proven that

$$E[Y] = aE[X] + b,$$

which is the same thing as Formula 2.

Proof of Formula 3. Again, we exploit properties of the expectation operator:

$$\begin{aligned} \sigma_Y^2 &= E[(Y-\mu_Y)^2] \\ &= E[(\{aX+b\} - \{a\mu_X + b\})^2] \\ &= E[(a\{X-\mu_X\})^2] \\ &= E[a^2(X-\mu_X)^2] \\ &= a^2 E[(X-\mu_X)^2] = a^2\sigma_X^2 \end{aligned}$$

Remarks. If $a > 0$, Formula 1 can be interpreted as saying that the plot of the new density is obtained from the plot of the old density by translating the old density to the right by b units (this is a translation to the left if $b < 0$!), and then scaling appropriately using the scaling factor a. (If $a = 1$, the new density is a pure translation of the old density, if $a > 1$, the scaling makes the new density more spread out than the old density, and if $a < 1$, the new density is more concentrated about its mean than the old density.) If $a < 0$, then the new density is obtained from the old density by a combination of translation, scaling, and reflection. (This is the sort of thing you did at the beginning of EE 3015.)

Example 11.1. Suppose X has the standard uniform distribution (i.e., the Uniform(0,1) distribution). Then its density function, mean, and variance are:

$$\begin{aligned} f_X(x) &= 1, \ \ 0 \le x \le 1 \ (zero \ elsewhere) \\ \mu_X &= 1/2 \\ \sigma_X^2 &= 1/12 \end{aligned}$$

Define new RV Y by the equation

$$Y = (b-a)X + a, \tag{11.6}$$

where a, b are any real constants for which $a < b$. Applying Formulas 1-3, it can be seen that the new density, new mean, and new variance are the following:

$$\begin{aligned} f_Y(y) &= \frac{1}{b-a}, \ \ a \le y \le b \ (zero \ elsewhere) \\ \mu_Y &= (b-a)\mu_X + a = (b-a)(1/2) + a = \frac{a+b}{2} \\ \sigma_Y^2 &= (b-a)^2\sigma_X^2 = \frac{(b-a)^2}{12} \end{aligned}$$

If you look in Appendix A, you will see that the new density, new mean, and new variance just obtained are what you get for a Uniform(a, b) distribution. Therefore, we have proved that the change of variable (11.6) converts a Uniform(0,1) random variable into a Uniform(a, b) random variable. This gives us a way to simulate n observations from a Uniform(a, b) distribution via the following Matlab one-liner:

```
y=(b-a)*rand(1,n)+a;
```

Example 11.2. Let Z be standard Gaussian. Then its density function, mean, and variance are:

$$\begin{aligned} f_Z(z) &= \frac{1}{\sqrt{2\pi}} \exp(-z^2/2) \\ \mu_Z &= 0 \\ \sigma_Z^2 &= 1 \end{aligned}$$

Define new RV X by the equation

$$X = \sigma Z + \mu, \tag{11.7}$$

where μ is any real parameter and σ is a positive real parameter. Applying Formulas 1-3, it can be seen that the new density, new mean, and new variance are the following:

$$\begin{aligned} f_X(x) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left(- \frac{(x-\mu)^2}{2\sigma^2} \right) \\ \mu_X &= \sigma\mu_Z + \mu = \sigma * 0 + \mu = \mu \\ \sigma_X^2 &= \sigma^2\sigma_Z^2 = \sigma^2 * 1 = \sigma^2 \end{aligned}$$

If you look in Appendix A, you will see that the new density, new mean, and new variance just obtained are what you get for a Gaussian(μ, σ) distribution. Therefore, we have proved that the change of variable (11.7) converts a Gaussian$(0, 1)$ random variable into a Gaussian(μ, σ) random variable. This gives us a way to simulate n observations from a Gaussian(μ, σ) distribution via the following Matlab one-liner:

```
x=sigma*randn(1,n)+mu;
```

Alternatively, suppose we solve equation (11.7) for Z in terms of X:

$$Z = \frac{X - \mu}{\sigma}. \tag{11.8}$$

Equation (11.8) gives us a way to convert a Gaussian(μ, σ) RV X into a standard Gaussian RV Z. This justifies our earlier technique for doing table lookups for nonstandard Gaussian distributions. (See, for example, Section 10.3 and Section 11.1, in which we exploited the change of variable (11.8).)

11.3 Moment Generating Functions

Here I skip ahead to cover Section 6.3, which gives us the useful notion of *moment generating function*. The moment generating function $\phi_X(s)$ of a RV X is defined by

$$\phi_X(s) \triangleq E[e^{sX}] = \int_{-\infty}^{\infty} e^{sx} f_X(x) dx. \tag{11.9}$$

In this definition, s plays the role of a parameter; s varies over all complex values for which the integral on the right side of (11.9) converges. Notice that if in this integral you replace s by $-s$, then what you have is simply the Laplace transform of $f_X(x)$. In other words,

$$\phi_X(s) = \mathcal{L}[f_X(x)]_{s \to -s}. \tag{11.10}$$

In formula (11.10), we simply mean that we can obtain the moment generating function of X by applying the Laplace transform operator $\mathcal{L}$ to $f_X(x)$, followed by a replacement of s by $-s$ in the Laplace transform. Formula (11.10) gives us an immediate way to find the moment generating function if the density $f_X(x)$ is a common type of signal considered in EE 3015 for which the Laplace transform is tabulated. The following example illustrates this.

Example 11.3. The PDF for the Exponential(a) probability distribution is

$$a \exp(-ax) u(x).$$

Recall from EE 3015 that the Laplace transform of the decaying exponential function $\exp(-ax)u(x)$ is $\frac{1}{s+a}$. Usinf formula (11.10), we then obtain the moment generating function of the Exponential(a) distribution by multiplying by a and replacing s by $-s$:

$$\phi_X(s) = \frac{a}{a-s}. \tag{11.11}$$

Discussion. The *moments* of a RV X are the quantities $E[X^k]$, as the power k ranges through the positive integers. For $k = 1$, we obtain the *first moment* $E[X]$, which is the same thing as the mean μ_X of RV X. For $k = 2$, we obtain the second moment $E[X^2]$. For $k = 3$, we obtain the third moment $E[X^3]$, etc. The reason $\phi_X(s)$ is called the moment generating function is that it gives us an easy way to compute the moments of X. This is what the following result shows us.

Useful Result: The moments of a RV X are computable from the moment generating function $\phi_X(s)$ as follows:

$$E[X] = \phi'_X(0) \tag{11.12}$$
$$E[X^2] = \phi''_X(0) \tag{11.13}$$
$$E[X^3] = \phi'''_X(0) \tag{11.14}$$

In these equations, the prime (') denotes differentiation with respect to s. Thus, you obtain the mean $\mu_X = E[X]$ by evaluating the first derivative of $\phi_X(s)$ at $s = 0$. You obtain the second moment by evaluating the second derivative of $\phi_X(s)$ at $s = 0$, etc.

Proof. We have

$$\begin{aligned}\frac{d\phi_X(s)}{ds} &= \frac{d}{ds}\left[\int_{-\infty}^{\infty} e^{sx} f_X(x)dx\right] \\ &= \int_{-\infty}^{\infty} \frac{d(e^{sx})}{ds} f_X(x)dx \\ &= \int_{-\infty}^{\infty} xe^{sx} f_X(x)dx\end{aligned}$$

Thus, we have proved that

$$\phi'_X(s) = \int_{-\infty}^{\infty} xe^{sx} f_X(x)dx. \tag{11.15}$$

Plugging $s = 0$ into both sides, we have the formula

$$\phi'_X(0) = \int_{-\infty}^{\infty} x f_X(x) = E[X].$$

This proves equation (11.12). By differentiating both sides of (11.15), the reader can easily obtain formula (11.13), and then formula (11.14) after another differentiation.

Example 11.4. Let X have the Exponential(a) probability distribution. In Example 11.3, we obtained formula (11.11) for the moment generating function. Differentiating this formula twice:

$$\begin{aligned} \phi'_X(s) &= \frac{a}{(a-s)^2} \\ \phi''_X(s) &= \frac{2a}{(a-s)^3} \end{aligned}$$

Plugging in $s = 0$,

$$\begin{aligned} \phi'_X(0) &= \frac{1}{a} \\ \phi''_X(0) &= \frac{2}{a^2} \end{aligned}$$

Applying the "Useful Result", we conclude that

$$\begin{aligned} \mu_X &= E[X] = \phi'_X(0) = \frac{1}{a} \\ E[X^2] &= \phi''_X(0) = \frac{2}{a^2} \\ \sigma_X^2 &= E[X^2] - \mu_X^2 = (2/a^2) - (1/a^2) = \frac{1}{a^2} \end{aligned}$$

The reader can look in Appendix A to verify that we have obtained the correct expressions for the mean and variance of the Exponential(a) distribution.

Example 11.5. Let X have the Binomial(n, p) distribution. If we try to directly evaluate the first and second moments, we wind up with the following sums:

$$\begin{aligned} E[X] &= \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} \\ E[X^2] &= \sum_{k=0}^{n} k^2 \binom{n}{k} p^k (1-p)^{n-k} \end{aligned}$$

It is not clear how to evaluate these complicated sums. (It can be done, but it takes a lot of "juggling", which we wish to avoid.) Instead, the moment generating function approach gives us an easy way to find these two moments. First, we evaluate the moment generating function:

$$\phi_X(s) = \sum_{k=0}^{n} \binom{n}{k} e^{sk} p^k (1-p)^{n-k}$$

$$\begin{aligned} &= \sum_{k=0}^{n} \binom{n}{k} (pe^s)^k (1-p)^{n-k} \\ &= (pe^s + 1 - p)^n \end{aligned}$$

(The last step used the well-known binomial theorem from college algebra, which says

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}.)$$

We have shown that

$$\phi_X(s) = (pe^s + 1 - p)^n.$$

If you check the table of MGF's on page 249 of your textbook, you will see that this is the correct expression for the MGF of the binomial distribution. Let us now compute the first two derivatives of the moment generating function:

$$\begin{aligned} \phi'_X(s) &= npe^s(pe^s + 1 - p)^{n-1} \\ \phi''_X(s) &= npe^s(pe^s + 1 - p)^{n-1} + n(n-1)(pe^s)^2(pe^s + 1 - p)^{n-2} \end{aligned}$$

Plugging in $s = 0$,

$$\begin{aligned} \phi'_X(0) &= np \\ \phi''_X(0) &= np + n(n-1)p^2 = np(1-p) + (np)^2 \end{aligned}$$

From the first equation, we see that $E[X] = np$. Subtracting the square of the mean $(np)^2$ from the second moment in the second equation, we obtain the variance to be $np(1-p)$. We have proved that for the Binomial(n, p) distribution, the mean and variance are given by the following formulas:

$$\begin{aligned} \mu_X &= np \\ \sigma_X^2 &= np(1-p) \end{aligned}$$

If you look in Appendix A, you will see that these expressions are correct.

Exercise. If you look in the table on page 249, you will see that the MGF of a Poisson(α) RV is given by the formula

$$\phi_X(s) = e^{-\alpha} \exp(\alpha e^s).$$

Try to use this formula to evaluate $\phi'_X(0)$ and $\phi''_X(0)$ and show that

$$\begin{aligned} \phi'_X(0) &= \alpha \\ \phi''_X(0) &= \alpha^2 + \alpha \end{aligned}$$

Show from these derivatives that the mean and variance of the Poisson(α) distribution are both α! (If you get stuck, look at Problem 6.4 in the Chapter 2-3 Solved Problems.)

Lecture 12

Chapters 2-3 Part 6

12.1 Some MGF Examples

Example 12.1. Let discrete RV X have the PMF

$$P^X(x) = \begin{cases} 1/2, & x = -1 \\ 1/2, & x = 1 \end{cases}$$

Let's find the MGF $\phi_X(s)$. The density is

$$f_X(x) = (1/2)\delta(x+1) + (1/2)\delta(x-1).$$

The Laplace transform of this is

$$(1/2)e^s + (1/2)e^{-s} = \cosh(s).$$

Replacing s by $-s$, you obtain $\phi_X(s)$, which is the same thing in this case. Thus,

$$\phi_X(s) = \cosh(s).$$

Taking the first two derivatives, we get

$$\begin{aligned} \phi'(s) &= \sinh(s) \\ \phi''(s) &= \cosh(s) \end{aligned}$$

Therefore,

$$\begin{aligned} \mu_X &= \phi'(0) = \sinh(0) = 0 \\ E[X^2] &= \phi''(0) = \cosh(0) = 1 \end{aligned}$$

The variance is

$$\sigma_X^2 = E[X^2] - \mu_X^2 = 1.$$

As an exercise, compute μ_X and σ_X^2 directly from the PMF and see if you get the same thing.

Example 12.2. Continuous RV X has density

$$f_X(x) = x \exp(-x) u(x).$$

Let's find the MGF. By the shift theorem for Laplace transforms from EE 3015, it is easy to see that the Laplace transform of $f_X(x)$ is

$$\frac{1}{(s+1)^2}.$$

Replacing s by $-s$, you get $\phi_X(s)$:

$$\phi_X(s) = \frac{1}{(1-s)^2}.$$

As an exercise, take two derivatives of $\phi_X(s)$ and use these derivatives to compute the mean and variance of X.

12.2 Estimation of μ, σ^2

Suppose you have a probability distribution with mean μ and variance σ^2, but you don't know the value of either of these parameters. To estimate these two parameters, you then gather a vector

$$\mathbf{x} = (x_i : i = 1, 2, 3, \cdots) \tag{12.1}$$

of actual or simulated observations x_i from this probability distribution, where the number of observations n is large. The *sample mean* of the observations (12.1) is defined by

$$sample\ mean \ \triangleq \frac{\sum_{i=1}^n x_i}{n},$$

the arithmetic average of the observations. In Matlab, the sample mean is computed as `mean(x)`. If you have a good probability model, then it would be highly likely that

$$sample\ mean \ \approx \mu,$$

that is, the sample mean would provide a good estimate of the theoretical mean μ of the probability distribution. (We prove this fact later in the course; it is called the *law of large numbers.*)

The *sample variance* of the observations (12.1) is defined by

$$sample\ variance \ \triangleq \frac{\sum_{i=1}^n (x_i - samplemean)^2}{n-1}, \tag{12.2}$$

which in Matlab is the same thing as `var(x)`. If you have a good probability model, then it would be highly likely that

$$\textit{sample variance} \approx \sigma^2,$$

that is, the sample variance would provide a good estimate of the theoretical variance σ^2 of the probability distribution.

At this point, we explain why people divide by $n-1$ instead of n in the formula (12.2) for the sample variance. Suppose you regard each observation x_i as the value of a random variable X_i (the observations are random, after all). Then the sample variance, which is actually a random variable, may be expressed as

$$\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1},$$

where $\bar{X}$ is the standard random variable notation for the sample mean. Later in the course, we prove that

$$E\left[\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}\right] = \sigma^2.$$

Therefore, if in (12.2) we divide by n instead of $n-1$, we would tend to *underestimate* the actual variance σ^2. However, if the number of samples n is large, it really does not matter very much whether we divide by n or $n-1$ in computing the sample variance, because in this case the two figures would be almost the same.

12.3 Hypergeometric Distribution

Given N items, of which N_1 are of "Type 1" and the remaining N_2 items are of "Type 2". (For example, in quality control, the Type 1 items could be the "defectives" and the Type 2 items could be the "nondefectives".) Choose n of the N items at random, without replacement. The n items you obtain will be called a "sample of size n". Define RV X to be the number of items of Type 1 in the sample of size n. The probability distribution of X is called *hypergeometric*, and is given by

$$P^X(x) = P(X = x) = \frac{\binom{N_1}{x}\binom{N_2}{n-x}}{\binom{N}{n}}, \quad x = 0, 1, \cdots, n, \tag{12.3}$$

where we adopt the convention that the number of combinations of r things taken s at a time is zero if $s > r$:

$$\binom{r}{s} = 0, \quad s > r.$$

Here is an easy way to derive the PMF probabilities in (12.3): Take the sample space for the experiment of selecting the sample of size n to be the set of all combinations of the N items taken n at a time, instead of the set of all permutations of the N items taken n at a time (in other

words, you do not care about the order in which the items in the sample are drawn). This is an equiprobable sample space containing $\binom{N}{n}$ outcomes. The probability of any event is therefore the number of outcomes in the event divided by $\binom{N}{n}$. There are $\binom{N_1}{x}$ ways to select x items of Type 1 from the total of N_1 items of Type 1, and there are $\binom{N_2}{n-x}$ ways to select $n-x$ items of Type 2 from the total of N_2 items of Type 2. Multiplying these two numbers together, you obtain the total number of ways of forming a sample of size n for which $X = x$; this is the numerator in the $P(X = x)$ expression in (12.3).

Example 12.3. You have 100 floppy discs, 5 of which have imperfections. 20 discs are chosen at random without replacement. Let X be the number of discs with imperfections in the sample of 20. Then

$$P(X = x) = \frac{\binom{5}{x}\binom{95}{20-x}}{\binom{100}{20}}, \quad x = 0, 1, 2, 3, 4, 5.$$

Example 12.4. Let us compute the probability of getting exactly two kings in a five card poker hand (dealt from a standard 52 card deck). This is a hypergeometric probability, given by

$$\frac{\binom{4}{2}\binom{48}{3}}{\binom{52}{5}} = 0.0399.$$

Now suppose we are dealt the five cards so that each card is dealt to us from a different 52 card deck. Then, we may view the five cards to be coming to us through independent trials. The probability of getting 2 kings in the 5 card hand is no longer a hypergeometric probability. Instead it is the binomial probability

$$\binom{5}{2}(4/52)^2(48/52)^3 = 0.0465.$$

Binomial Approximation to Hypergeometric

Again, suppose we have hypergeometric RV X with probability distribution according to (12.3). If the sample size n is small relative to the size N of the total pool of items, it can be shown that X has approximately a Binomial(n, p) distribution with $p = N_1/N$. Notice that this is exactly what the distribution of X would be if the sampling takes place with replacement (because of independent trials). Therefore, we are saying that if the total number of items N is large relative to the size of the sample n, sampling without replacement gives approximately the same probability distribution for X as sampling with replacement. This makes intuitive good sense, because even though the draws (trials) are dependent, the composition of the total pool of items being selected from changes very little in successive draws, and so the successive draws are almost like independent trials.

Example 12.5. Suppose a quality control engineer is faced with a pool of N items, 10% of which are defective. Suppose he/she chooses 50 items at random. The probability that there are 5

defectives in the sample of 50 is the exact hypergeometric probability

$$\frac{\binom{(0.1)N}{5}\binom{(0.9)N}{45}}{\binom{N}{50}} \tag{12.4}$$

The binomial approximation to the hypergeometric probability (12.4) is

$$\binom{50}{10}(0.1)^5 * (0.9)^{45} = 0.1849. \tag{12.5}$$

We can try different values of N in (12.4) to see for how big an N we might be able to say that (12.4) is close to the binomial probability (12.5). We compiled the following table using Matlab:

N	*hypergeometric prob* (12.4)
100	0.2593
200	0.2132
300	0.2025
400	0.1976
500	0.1949
1000	0.1897
2000	0.1873

We see that for $N = 2000$, the actual hypergeometric probability is 0.187 to three decimal places, as opposed to the binomial approximation of 0.185 (three decimal places). If we increase N still further, we'd see these two figures getting even closer.

12.4 Conditional Probability Distribution

Let X be a RV with density $f_X(x)$. Let B be a subset of the real line (usually, we will take B to be an interval). Suppose we are given that the value of X falls in B. In other words, we learn that the event $\{X \in B\}$ has occured. How should we modify our density $f_X(x)$ to reflect this partial information about the value of X? If we want to do probability, mean, or variance computations, it would no longer make sense to use the density $f_X(x)$, because in general $f_X(x)$ extends over the whole real line and we should instead do these computations using a density that extends only over B.

Given the event $\{X \in B\}$, we will use the so-called *conditional density of* X *given* $\{X \in B\}$ for our probability, mean, and variance computations. This conditional density is denoted $f_{X|B}(x)$ and is defined by

$$f_{X|B}(x) \triangleq \begin{cases} \frac{f_X(x)}{P(X \in B)}, & x \in B \\ 0, & x \notin B \end{cases}$$

Here is how we can justify this definition for $f_{X|B}(x)$. For simplicity of visualization, suppose B is an interval. Since we are given that the value of X falls in B, we should "chop off" that part of the plot of the density $f_X(x)$ that falls outside the interval B. The remaining part of the density $f_X(x)$ will take positive values only within B. If we have two subintervals E_1 and E_2 of B, then this remaining part of $f_X(x)$ could be used to judge the relative likelihoods with which X falls in E_1 or E_2, given that X falls in B, if we compare the areas under the remaining $f_X(x)$ curve that lie above E_1 and E_2, respectively. In other words, the following will be true:

$$\frac{P(X \in E_1|X \in B)}{P(X \in E_2|X \in B)} = \frac{\int_{E_1} f_X(x)dx}{\int_{E_2} f_X(x)dx}. \tag{12.6}$$

Because of (12.6), it follows that we can obtain our conditional density by properly scaling $f_X(x)$ for $x \in B$. This scaling factor should be $1/P(X \in B)$ because without that factor, that part of $f_X(x)$ extending over B will not yield area 1 over B but will yield area $P(X \in B)$ instead; applying the factor $1/P(X \in B)$ makes this area over B equal to 1.

Discussion. The conditional density function $f_{X|B}(x)$ is a bonfide density function in its own right, just as $f_X(x)$ is a density function. Therefore, whatever probability, mean, or variance computation we do with the density $f_X(x)$ can equally well be done with the density $f_{X|B}(x)$. Therefore, it stands to reason that we can compute *conditional probabilities*, *conditional expected value*, and *conditional variance* using the conditional density $f_{X|B}(x)$ as follows:

$$P(X \in E|X \in B) = \int_E f_{X|B}(x)dx \tag{12.7}$$

$$E(X|X \in B) = \int_{-\infty}^{\infty} x f_{X|B}(x)dx \tag{12.8}$$

$$Var(X|X \in B) = \int_{-\infty}^{\infty} (x - E(X|X \in B))^2 f_{X|B}(x)dx \tag{12.9}$$

- $P(X \in E|X \in B)$ is called the conditional probability that X falls in E given that X falls in B.
- $E(X|X \in B)$ is called the conditional expected value for X (conditional mean for X) given that X falls in B.
- $Var(X|X \in B)$ is called the conditional variance for X given that X falls in B.

In the formulas (12.7)-(12.8), we have explained how to compute conditional probability and conditional expected value using the conditional PDF $f_{X|B}(x)$. Alternatively, these quantitities can be computed using the original density $f_X(x)$ as follows:

$$P(X \in E|X \in B) = \frac{P(X \in E \cap B)}{P(X \in B)} = \frac{\int_{E \cap B} f_X(x)dx}{\int_B f_X(x)dx} \tag{12.10}$$

$$E(X|X \in B) = \frac{\int_B x f_X(x)dx}{P(X \in B)} = \frac{\int_B x f_X(x)dx}{\int_B f_X(x)dx} \tag{12.11}$$

Sometimes it can be easier to use the formulas (12.10)-(12.11) than the formulas (12.7)-(12.8).

Example 12.6. Let X be the discrete RV with PMF

$$P^X(x) = \begin{cases} 0.1, & x = 1 \\ 0.2, & x = 2 \\ 0.3, & x = 3 \\ 0.4, & x = 4 \end{cases}$$

Let us find the conditional density $f_{X|B}(x)$, where B is the event

$$B = \{2 \leq X \leq 3\}.$$

The density $f_X(x)$ is:

$$f_X(x) = (0.1)\delta(x-1) + (0.2)\delta(x-2) + (0.3)\delta(x-3) + (0.4)\delta(x-4).$$

Given $X \in B$, we know either $X = 2$ or $X = 3$. Therefore, the first and last terms of the density drop out and we have to re-normalize the remaining terms. Dropping out the 1st and last terms, we obtain

$$(0.2)\delta(x-2) + (0.3)\delta(x-3).$$

This expression integrates to 0.5. Therefore, we have to divide by 0.5 in order to obtain the conditional density:

$$f_{X|B}(x) = (0.4)\delta(x-2) + (0.6)\delta(x-3).$$

From this, we can read off the following two conditional probabilities:

$$\begin{aligned} P(X = 2|B) &= P(X = 2|2 \leq X \leq 3) = 0.4 \\ P(X = 3|B) &= P(X = 3|2 \leq X \leq 3) = 0.6 \end{aligned}$$

These two probabilities give us the so-called *conditional PMF for X given X falls in B*. Notice that these two probabilities add up to 1. (A conditional PMF is a PMF.) Using the conditional PMF, we can compute the conditional mean and the conditional variance:

$$\begin{aligned} E(X|X \in B) &= (0.4)(2) + (0.6)3 = 2.6 \\ E(X^2|X \in B) &= (0.4)(2^2) + (0.6)(3^2) = 7 \\ Var(X|X \in B) &= 7 - (2.6)^2 = 0.24 \end{aligned}$$

Remark. In the calculation just completed, we used the fact that

$$Var(X|X \in B) = E(X^2|X \in B) - E(X|X \in B)^2.$$

We already have this identity for unconditional probability distributions. It is true for conditional probability distributions, too. (Because a conditional probability distirbution is a probability distribution.)

Lecture 13

Chapters 2-3 Part 7

13.1 Conditional Mean,Variance Notations

- $\mu_{X|B}$ and $E(X|B)$ are alternate notations for the conditional mean (conditional expected value) $E(X|X \in B)$.
- $\sigma^2_{X|B}$ and $Var(X|B)$ are alternate notations for the conditional variance $Var(X|X \in B)$.
- $\sigma_{X|B}$ is our notation for the conditional standard deviation, given by

$$\sigma_{X|B} = \sqrt{\sigma^2_{X|B}}.$$

13.2 More Conditional Distribution Examples

Example 13.1. Suppose X has the Exponential(a) distribution. This means the density is

$$f_X(x) = a \exp(-ax)u(x).$$

Let C be a fixed positive real number, and let us condition on the event

$$B = \{X \geq C\}$$

that X takes a value $\geq C$. Notice that

$$P(X \geq C) = \int_C^\infty a \exp(-ax)dx = \exp(-aC).$$

To obtain the conditional density $f_{X|B}(x)$, we scale that portion of the curve $f_X(x)$ to the right of $x = C$ by the scaling factor $1/P(X \geq C)$. This gives us

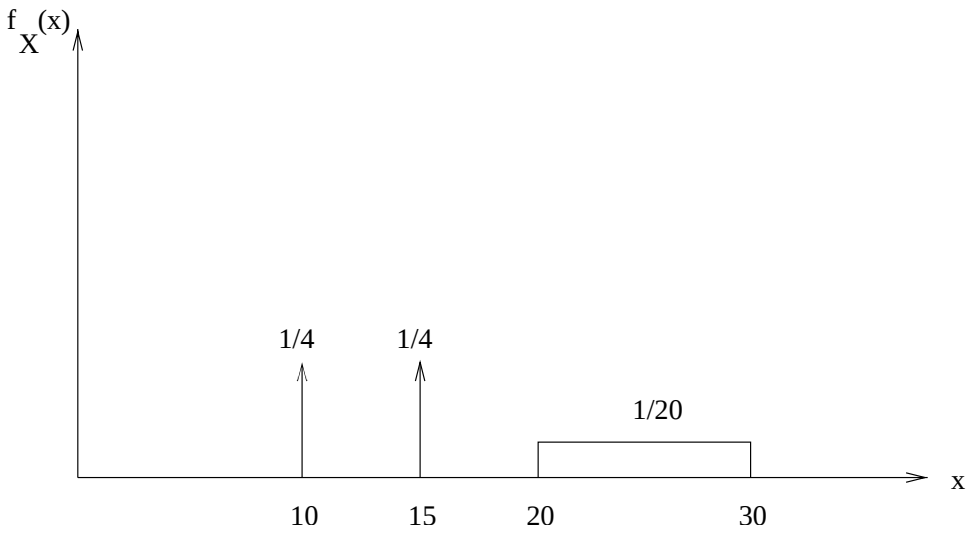
$$f_{X|B}(x) = \frac{a \exp(-ax)}{\exp(-aC)} = a \exp(-a(x - C)), \quad for\ x \geq C,$$

and zero elsewhere. In other words, we can write the conditional density $f_{X|B}(x)$ compactly as the one line expression

$$f_{X|B}(x) = f_X(x - C).$$

This is just the translation of the original density C units to the right!

Example 13.2. Let X be a mixed RV with the density $f_X(x)$ plotted below:



From this plot, we see that X takes the discrete values $10, 15$ and also takes continuous values between 20 and 30. If we condition appropriately, we can get conditional densities that encapsulate either the discrete behavior of X or the continuous behavior of X. To do this, consider the two events

$$B_1 = \{X < 17.5\}, \quad B_2 = \{X \geq 17.5\}.$$

If we condition on B_1, the conditional density will just involve the two discrete values and by inspection it therefore must be:

$$f_{X|B_1}(x) = (1/2)\delta(x - 10) + (1/2)\delta(x - 15).$$

(The heights of the two delta functions must be the same because they have the same height in $f_X(x)$; because a conditional density is a density, these two heights must each be $1/2$.) If we condition on B_2, the conditional density must be uniform from 20 to 30 and we already know what the uniform density from 20 to 30 is. So by inspection we get:

$$f_{X|B_2}(x) = (1/10)[u(x - 20) - u(x - 30)].$$

13.3 Renewal Property of Exponential Distribution

Let RV X have the Exponential(a) distribution. Let us compute a conditional probability of the form

$$P(X \geq u+v|X \geq u),$$

where $u \geq 0$ and $v \geq 0$. We can compute this conditional probability using the conditional density of X given $X \geq u$, which from Example 13.1 is the same thing as $f_X(x-u)$:

$$\begin{aligned} P(X \geq u+v|X \geq u) &= \int_{u+v}^{\infty} f_{X|\{X\geq u\}}(x)dx \\ &= \int_{u+v}^{\infty} f_X(x-u)dx \\ &= \int_{v}^{\infty} f_X(x)dx = P(X \geq v). \end{aligned}$$

We have proved that

$$P(X \geq u+v|X \geq u) = P(X \geq v), \quad for\ any\ u \geq 0, \quad v \geq 0. \tag{13.1}$$

Equation (13.1) is called the *renewal property* for the Exponential distribution.

Example 13.3. Suppose we model the lifetime X (in years) of a randomly chosen male US citizen as having the Exponential(a) distribution with $a = 1/70$. (This means we are assuming the mean lifetime is 70 years.) Then, by the renewal property

$$P(X \geq 75|X \geq 70) = P(X \geq 5) = \exp(-5/70) = 0.9311. \tag{13.2}$$

This tells us the following: Given that a male citizen survives to his 70th birthday, the probability he will survive at least another 5 years is the same as the probability that a newborn male baby will make it to his 5th birthday. This shows you some limitations in using the exponential distribution to model lifetimes. Instead, one might want to model the lifetime using a modification of the exponential distribution so that the probability on the left side of (13.2) is less than the probability on the right, which would fit physical reality better.

13.4 Density Decomposition Theorem

Let X be any random variable. Let $\{B_i\}$ be finitely many subsets of the real line which partition up the real line. (Typically, each B_i would be a finite or infinite interval.) The *density decomposition theorem* allows us to express the overall density $f_X(x)$ as a weighted average of the individual conditional densities $f_{X|B_i}(x)$. Specifically, the decomposition theorem says that

$$f_X(x) = \sum_i P(X \in B_i) f_{X|B_i}(x). \tag{13.3}$$

Discussion. The decomposition theorem is almost obvious, if you look at it from a geometric point of view. Suppose you just have two B_i's, namely B_1 and B_2. To easier visualize things, suppose B_1 consists of all real numbers to the left of some fixed point C on the real line, and B_2 consists of all real numbers to the right of C. Then the two functions

$$P(X \in B_1) f_{X|B_1}(x), \quad P(X \in B_2) f_{X|B_2}(x)$$

are respectively the part of the $f_X(x)$ curve to the left of C and to the right of C. Clearly, if you add up these two functions, you must get $f_X(x)$.

Consequences of the Density Decomposition Theorem

We make the same assumptions that we made above for the decomposition theorem. Then you easily obtain the following facts from the density decomposition formula (13.3):

(a) For any subset E of the real line,

$$P(X \in E) = \sum_i P(X \in B_i) P(X \in E | X \in B_i). \tag{13.4}$$

(b) The mean μ_X can be broken down as a combination of conditional means as follows:

$$\mu_X = E[X] = \sum_i P(X \in B_i) E[X|B_i]. \tag{13.5}$$

(c) More generally, any expected value $E[\phi(X)]$ can be broken down as

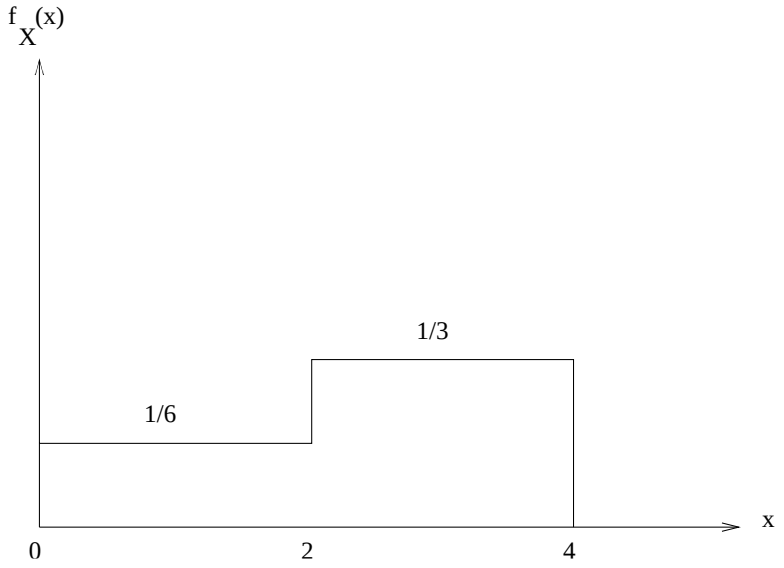
$$E[\phi(X)] = \sum_i P(X \in B_i) E[\phi(X)|B_i]. \tag{13.6}$$

Remarks.

- You obtain formula (13.4) by integrating both sides of (13.3) over the set E. You obtain formula (13.6) by multiplying both sides of (13.3) by $\phi(x)$, followed by an integration of both sides. Formula (13.5) is a special case of formula (13.6).
- Formula (13.4) is a special case of the *law on total probability* proved in Chapter 1.
- Formulas (13.5)-(13.6) are usually collectively referred to as the *law on total expectation.* We will have occasion to use this law many times in this course. (I will use it later in this lecture in the design of the one-bit quantizer.)
- In the above formulas, I found it more convenient to write expected values with bracket notation such as $E[\phi(X)]$ or $E[\phi(X)|B_i]$ instead of parenthesis notation $E(\phi(X))$ or $E(\phi(X)|B_i)$. It makes no difference whether you use parentheses or brackets.

- Warning: There is no decomposition formula for variances! That is, except in special (trivial) cases, you will have
$$Var(X) \neq \sum_i P(X \in B_i) Var(X|B_i). \tag{13.7}$$
The easiest way to see this is to take $f_X(x)$ to consist of two delta function spikes, and then to choose B_1, B_2 so that the two conditional densities are delta functions. Then each $Var(X|B_i) = 0$ and $Var(X) > 0$, whereupon the two sides of (13.7) will definitely not be equal to one another.

Example 13.4. Let X be a continuous RV with the following density function:



Let's compute μ_X and σ_X^2. One way to find these is a brute force approach which uses the above density to do a first and second moment calculation. Instead of doing that, I will exploit the decomposition theorem. Choose B_1, B_2 to be the events

$$B_1 = \{X < 2\}, \quad B_2 = \{X \geq 2\}.$$

Then from the above plot, it is obvious that the conditional density $f_{X|B_1}(x)$ will be the Uniform$(0, 2)$ density and that the conditional density $f_{X|B_2}(x)$ will be the Uniform$(2, 4)$ density. You can now go to Appendix A and look up the mean and variance figures for uniform densities. You immediately

obtain

$$\begin{aligned} E(X|B_1) &= \textit{midpoint of } [0,2] = 1 \\ Var(X|B_1) &= (2-0)^2/12 = 1/3 \\ E(X|B_2) &= \textit{midpoint of } [2,4] = 3 \\ Var(X|B_2) &= 1/3 \end{aligned}$$

The conditional second moments are now easily computable:

$$\begin{aligned} E(X^2|B_1) &= Var(X|B_1) + E(X|B_1)^2 = 4/3 \\ E(X^2|B_2) &= Var(X|B_2) + E(X|B_2)^2 = 28/3 \end{aligned}$$

Using formulas (13.5)-(13.6), we can now compute μ_X and $E(X^2)$:

$$\begin{aligned} \mu_X &= P(X \in B_1)E(X|B_1) + P(X \in B_2)E(X|B_2) \\ &= (1/3)(1) + (2/3)(3) = 2.3333 \\ E(X^2) &= (1/3)E(X^2|B_1) + (2/3)E(X^2|B_2) \\ &= (1/3)(4/3) + (2/3)(28/3) = 6.6667 \end{aligned}$$

We can now compute σ_X^2 because we know the first and second moments of X:

$$\sigma_X^2 = E(X^2) - \mu_X^2 = 1.2222.$$

Example 13.5. In Example 13.2, by inspection we have

$$\begin{aligned} E(X|X < 17.5) &= 12.5 \\ E(X|X \geq 17.5) &= 25 \end{aligned}$$

Therefore

$$\begin{aligned} E(X) &= P(X < 17.5)E(X|X < 17.5) + P(X \geq 17.5)E(X|X \geq 17.5) \\ &= (1/2)(12.5) + (1/2)(25) = 18.75. \end{aligned}$$

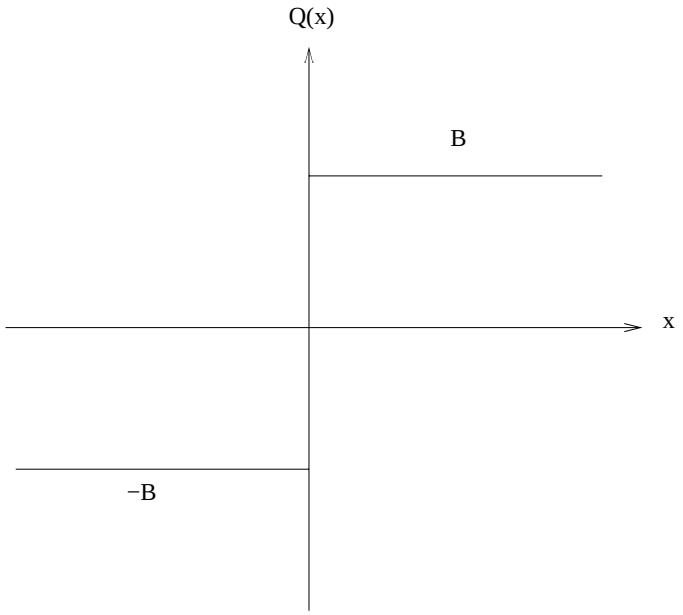
As an exercise, see if you can also compute $E(X^2)$ by the decomposition approach.

13.5 Design of the One-bit Quantizer

Suppose we have a RV X, and for simplicity we assume that its density function $f_X(x)$ is an even function. A so-called "one-bit quantizer" $Q(x)$ for X would take the form

$$Q(x) = \begin{cases} B, & x > 0 \\ -B, & x \leq 0 \end{cases}$$

where B is a positive constant that is to be chosen appropriately. The plot of $Q(x)$ looks like



The quantizer $Q(x)$ is called a one-bit quantizer because depending upon whether $Q(x)$ is equal to B or $-B$, you can send a bit (a zero or a one) to the user to indicate which of these two cases occurs. In other words, we have a kind of primitive communication system:

$$\boxed{\begin{array}{c}\text{data}\\ \text{source}\end{array}} \rightarrow X \rightarrow \boxed{\text{quantizer}} \rightarrow Q(X) = \pm B \rightarrow \boxed{\begin{array}{c}\text{bit}\\ \text{converter}\end{array}} \rightarrow \text{send zero or one} \rightarrow \boxed{\text{user}}$$

Depending upon whether the user's received bit is 0 or 1, the user will estimate X as either B or $-B$. (If you are allowed to send more bits to the user, you can make a finer quantization for greater accuracy. Using just one bit, I am considering the simplest such communication system.)

Our design goal is to choose B so that the so-called mean square quantization error

$$E[(X - Q(X))^2]$$

is minimized. (This is the choice of B which will give us the best fit of $Q(X)$ to X under the restriction that we are only allowed to use one bit to represent the value of X.) Using the decomposition theorem, we can write

$$\begin{aligned} E[(X - Q(X))^2] &= P(X \leq 0)E[(X - Q(X))^2 | X \leq 0] + P(X > 0)E[(X - Q(X))^2 | X > 0] \\ &= (1/2)E[(X + B)^2 | X \leq 0] + (1/2)E[(X - B)^2 | X > 0] \end{aligned}$$

The two conditional expected values on the right side of the preceding equation are equal, because of the fact that $f_X(x)$ is even. Therefore,

$$E[(X - Q(X))^2] = E[(X - B)^2 | X > 0].$$

In an earlier lecture, we developed a "moment of inertia" formula:

$$E[(X - B)^2] = E[(X - \mu_X)^2] + (B - \mu_X)^2.$$

The same formula is true for conditional distributions; that is,

$$E[(X - B)^2 | X > 0] = E[(X - \mu^*)^2 | X > 0] + (B - \mu^*)^2, \tag{13.8}$$

where μ^* must be the conditional mean:

$$\mu^* = E[X | X > 0].$$

From equation (13.8), it is clear that the left side is minimized when $B = \mu^*$. Here is what we have proved.

One-Bit Quantizer Design Rule: *The best one-bit quantizer $Q(x)$ for the given RV X is the one for which B is taken to be the conditional mean $E[X|X > 0]$. This conditional mean can be computed as*:

$$B = E[X | X > 0] = \int_0^\infty x f_{X|\{X>0\}}(x) dx = 2 \int_0^\infty x f_X(x) dx. \tag{13.9}$$

As an exercise, see if you understand where the factor of two comes from on the right side of (13.9).

Example 13.6. Let us take X to be Gaussian$(0, 1)$. The the best choice of B for the one-bit quantizer is

$$B = 2 \int_0^\infty x \left(\frac{1}{\sqrt{2\pi}} \right) \exp(-x^2/2) dx = \frac{2}{\sqrt{2\pi}} = 0.7979.$$

Lecture 14

Chapters 2-3 Part 8

14.1 Types of Change of Variable Problems

Suppose you are given a RV X with given density $f_X(x)$. You then define a new RV Y via a change of variable

$$Y = \phi(X).$$

We need to understand how to find the density $f_Y(y)$ of the new RV Y via some kind of procedure that uses $f_X(x)$ and the transformation function ϕ. The procedure that we would use would depend upon the various possible cases that could arise. Here are the possible cases to consider:

X discrete	Y discrete
X mixed	Y mixed or discrete
X continuous	Y continuous, mixed, or discrete

Notice that there are 6 different cases that can occur. It is not hard to find useful examples of each of these 6 types. (By "useful" examples, I mean an example that is not a "toy" example, but rather an example illustrative of a type of transformation people would use in engineering systems.)

Example 14.1. In this example, I illustrate the case in which X is continuous and Y is mixed. Let X be a Gaussian(0,1) RV. Define RV Y as follows:

$$Y = \begin{cases} -1, & X \leq -1 \\ 1, & X \geq 1 \\ X, & -1 < X < 1 \end{cases}$$

In a communication system, we'd say that Y is obtained from X by passing X through a *hard limiter*. (The purpose of the hard limiter is to produce an output that is limited to a finite range; in this case, the hard limiter produces output values ranging through the interval $[-1, 1]$.) Notice that Y is a mixed RV: it takes the discrete values ± 1 but it also takes continuously distributed

values between -1 and 1. You can use intuition to guess what the shape of the density $f_Y(y)$ would be: it would consist of two delta functions at $y = \pm 1$ of equal height, together with a continuous part from $y = -1$ to $y = 1$ that should look like a bell-shaped Gaussian density curve centered at $y = 0$ but "chopped off" to the right of $y = 1$ and to the left of $y = -1$. To find the precise mathematical representation of $f_Y(y)$, you'd use a method I am going to cover in this set of notes called the "CDF method"; if you go to Example 3.25 of the textbook, you will see this particular example worked out according to the CDF method.

Exercise. Think of examples of the other 5 types. (If you get stuck, by the end of this set of notes you will see many of these types of examples.)

14.2 Case when Y is discrete

Three of the six cases of change of variable are cases in which the RV Y that you obtain from the change of variable is *discrete*. These cases are very easy to handle. In fact, the section on "Derived Models" of Chapter 1 notes tells us what to do. It has been a while since we've talked about Derived Models; in the next paragraph, I review how you can use the Derived Model concept to handle the case when Y is discrete.

For each discrete value y of the discrete RV Y, you find the subset $S(y)$ of the real line defined by

$$S(y) \stackrel{\Delta}{=} \{real\ x : \phi(x) = y\}.$$

In other words, $S(y)$ is precisely that set of real numbers which are transformed into y by the transformation function ϕ. Then the following two events are identical:

$$\{Y = y\} = \{X \in S(y)\}. \tag{14.1}$$

(That is, if we have an outcome for which Y takes the value y, then X must take a value in $S(y)$ and vice-versa.) Since the two events in (14.1) are identical, they must have the same probability. Therefore,

$$P(Y = y) = P(X \in S(y)) = \int_{S(y)} f_X(x)dx. \tag{14.2}$$

Notice that we have computed the probability $P(X \in S(y))$ as an integral involving the density $f_X(x)$; it is possible to do this calculation because we are assuming that we know what $f_X(x)$ is.

Example 14.2. Let X be the discrete RV which is equiprobable over the following set of 7 values:

$$-3, \ -2, \ -1, \ 0, \ 1, \ 2, \ 3.$$

Our change of variable is $Y = X^2$. Let us find the PMF and PDF of Y. First, note that the values of Y are $0, 1, 4, 9$. The value 0 has to be handled separately because it has only one square root:

$$P(Y = 0) = P(X^2 = 0) = P(X = 0) = 1/7.$$

The values $y = 1, 4, 9$ can be handled in the same way because each such y value has two distinct square roots:

$$P(Y = y) = P(X^2 = y) = P(X = \pm\sqrt{y}) = P^X(\sqrt{y}) + P^X(-\sqrt{y}) = 2/7.$$

We have shown that the PMF of Y is:

$$P^Y(y) = \begin{cases} 1/7, & y = 0 \\ 2/7, & y = 1, 4, 9 \end{cases}$$

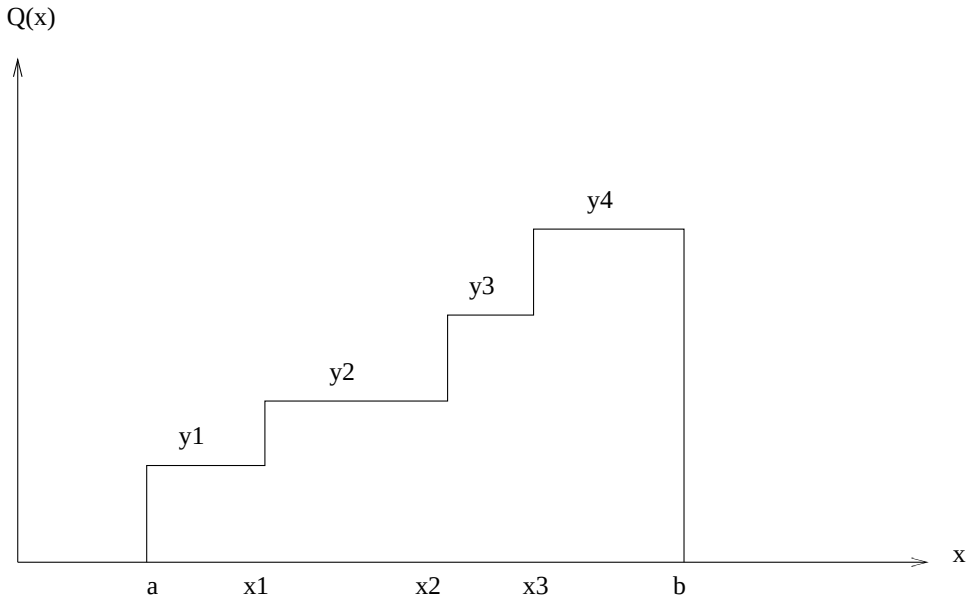
Notice that these four probabilities add up to one, so we do have a bonafide PMF. (This doesn't prove that our answer is correct, but if the four probabilities didn't add up to one, then we'd have known that there must have been a mistake.) The density of Y is therefore

$$f_Y(y) = (1/7)\delta(y) + (2/7)\delta(y-1) + (2/7)\delta(y-4) + (2/7)\delta(y-9).$$

Example 14.3. Let RV X be Uniform(a, b). Let's make the change of variable

$$Y = Q(X)$$

where $Q(x)$ is the "four-level quantizer" whose plot is the following:



Y is a discrete RV taking the four values y_1, y_2, y_3, y_4. Suppose you want to compute $P^Y(y_1)$. Looking at the plot, you'd compute the probability that X falls in the interval between a and x_1. In this way, we can determine the entire PMF of Y to be the following:

$$\begin{aligned} P^Y(y_1) &= \frac{x_1 - a}{b - a} \\ P^Y(y_2) &= \frac{x_2 - x_1}{b - a} \\ P^Y(y_3) &= \frac{x_3 - x_2}{b - a} \\ P^Y(y_4) &= \frac{b - x_3}{b - a} \end{aligned}$$

14.3 CDF Method

Suppose we make the change of variable $Y = \phi(X)$ to get new RV Y from old RV X. The "CDF Method" is a general method for finding the density $f_Y(y)$. It consists of two steps:

Step 1: For each value y of Y, compute the CDF value

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

as an integral

$$\int_{B(y)} f_X(x)dx,$$

where $B(y)$ is the subset of the real line for which the following two events are the same:

$$\{Y \leq y\} = \{X \in B(y)\}.$$

Step 2: Compute the density $f_Y(y)$ as the derivative of the CDF of Y:

$$f_Y(y) = \frac{d}{dy} F_Y(y).$$

Example 14.4. Let X be Uniform$(0, 1)$. Let Y be the RV

$$Y = -\log X, \tag{14.3}$$

where the logarithm is natural. Let us find $f_Y(y)$ by the CDF method. The values of Y are the positive real numbers $y > 0$. By Step 1, for each fixed $y > 0$, we must compute $P(Y \leq y)$. This gives us

$$P(Y \leq y) = P(-\log X \leq y) = P(X \geq e^{-y}).$$

The value e^{-y} falls in the interval $[0, 1]$, and so the probability that X takes a value $\geq e^{-y}$ must be $1 - e^{-y}$. As the result of Step 1, we have shown that

$$F_Y(y) = P(Y \leq y) = (1 - e^{-y})u(y).$$

Step 2 is to differentiate this expression in order to obtain the density $f_Y(y)$. This gives us

$$f_Y(y) = e^{-y}u(y).$$

That is, Y has the Exponential(1) distribution!

In Example 14.4, it is not hard to see that if we modify the transformation (14.3) by multiplying the right side by any positive constant, then Y will still have an Exponential distribution. This gives us the following result.

Useful Result. If X is Uniform$(0, 1)$ and $a > 0$ is a constant, then the RV

$$Y = -(\log_e X)/a$$

is Exponential(a).

Example 14.5. Suppose we model a randomly selected US male person as having an exponentially distributed lifetime with mean lifetime 70 years. To simulate the lifetimes of 50000 such people, we could execute the Matlab command

```
-70*log(rand(1,50000))
```

The parameter of the exponential distribution we should be using is $a = 1/70$. Flipping this over (see Useful Result), that is how I got the 70 in front of my Matlab line. We did this sort of thing in Recitation 2. Now we know why this works.

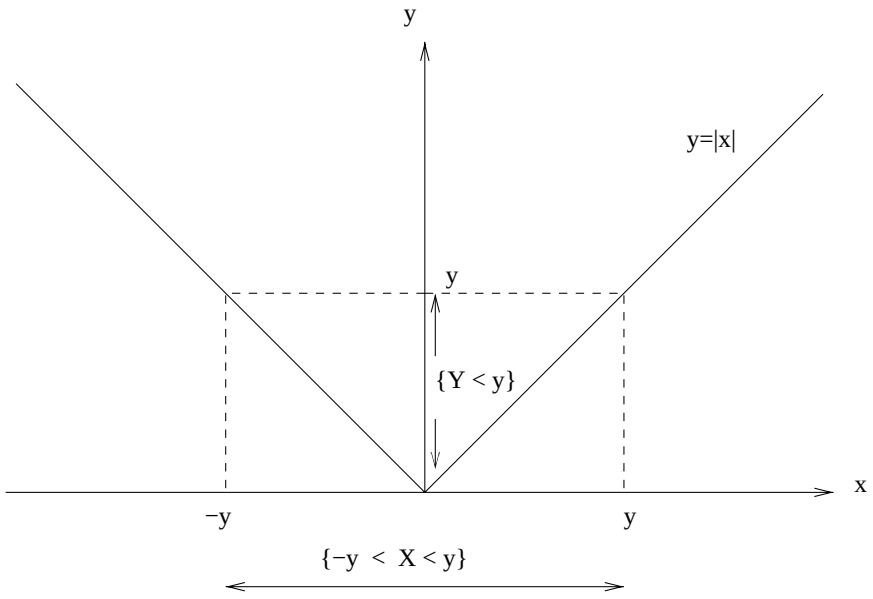
Example 14.6. Let X be Gaussian$(0, 1)$, and let Y be the RV

$$Y = |X|.$$

The values of Y are all real numbers $y \geq 0$. For each fixed $y \geq 0$, notice that

$$\{Y < y\} = \{-y < X < y\}.$$

To help you see that this is true, examine the following plot of the function $y = |x|$:



I've denoted the events $\{Y < y\}$ and $\{-y < X < y\}$ by arrows along the y-axis and x-axis, respectively. This gives us:

$$P(Y \leq y) = P(Y < y) = P(-y < X < y) = \int_{-y}^{y} f_X(x)dx. \tag{14.4}$$

We now have to differentiate this equation with respect to y. In calculus, you presumably learned that

$$\frac{d}{dy}\left[\int_{a(y)}^{b(y)} f(x)dx\right] = b'(y)f(b(y)) - a'(y)f(a(y)),$$

where $a'(y), b'(y)$ are the derivatives of $a(y), b(y)$ with respect to y. Applying this formula to (14.4), we see that

$$f_Y(y) = \frac{d}{dy}F_Y(y) = f_X(y) - (-1)f_X(-y) = 2f_X(y). \tag{14.5}$$

(In the last part of the preceding equation, I used the fact that the standard Gaussian density $f_X(x)$ is an even function, which allows us to say that $f_Y(-y) = f_Y(y)$.) Plugging in for $f_X(y)$ in the right side of (14.5), we have shown that

$$f_Y(y) = \frac{2}{\sqrt{2\pi}}\exp(-y^2/2)u(y). \tag{14.6}$$

If you plot this, you will see that your plot looks like the right half of a Gaussian bell-shaped density curve. What happened is that the original Gaussian density curve, under the transformation

$Y = |X|$, got folded in half along the vertical axis, with the left half coming over on top of the right half, giving the "2" in front of the density in (14.6).

14.4 Simulating A Continuous Random Variable

The following useful (and amazing!) result will allow us to use Matlab to simulate the values of any continuous RV whatsoever.

Useful Result. Let X be a continuous RV. Then the random variable

$$U = F_X(X)$$

is uniformly distributed between 0 and 1.

I will prove the "Useful Result" at the end of this section. But first, let me give you a couple of examples showing how to use this result for simulation.

Example 14.7. Suppose I want to use Matlab to simulate values of a continuous RV X having the density

$$f_X(x) = \begin{cases} x^2, & 0 \le x \le 1 \\ 0, & elsewhere \end{cases}$$

The CDF $F_X(x)$ is easily seen to satisfy

$$F_X(x) = x^2, \;\; 0 \le x \le 1.$$

Plug X into this CDF. We obtain the equation

$$U = X^2,$$

where, by the Useful Result, U is Uniform$(0, 1)$. Solving for X in terms of U, we obtain

$$X = \sqrt{U}.$$

Suppose now that we want to simulate 100000 values of RV X. Here is a Matlab one-liner that will do it:

```
x=sqrt(rand(1,100000));
```

Example 14.8. Let X be the continuous RV with density

$$f_X(x) = \begin{cases} \frac{1}{\pi\sqrt{1-x^2}}, & -1 < x < 1 \\ 0, & elsewhere \end{cases}$$

The CDF satisfies

$$F_X(x) = (1/2) + \left(\frac{1}{\pi}\right) \text{Sin}^{-1}(x), \quad -1 \leq x \leq 1.$$

Setting

$$U = (1/2) + \left(\frac{1}{\pi}\right) \text{Sin}^{-1}(X),$$

the RV U is Uniform$(0, 1)$. Solving for X in terms of U, we obtain

$$X = \sin(\pi[U - 0.5]).$$

The following Matlab one-liner could then be used to simulate 100000 values of X:

```
x=sin(pi*(-0.5+rand(1,100000)));
```

Proof of Useful Result

Letting

$$U = F_X(X),$$

we see that the values of U range from 0 to 1 because we know that the CDF of a continuous RV takes on all real values strictly between 0 and 1 (due to the fact that there can be no jumps in the CDF). Pick any fixed u satisfying $0 < u < 1$. Let x be the unique real number satisfying

$$u = F_X(x). \tag{14.7}$$

We have

$$\begin{aligned} F_U(u) &= P(U \leq u) \\ &= P(F_X(X) \leq F_X(x)) \\ &= P(X \leq x) = F_X(x) \end{aligned}$$

By equation (14.7), this last term $F_X(x)$ is equal to u. We have proved that

$$F_U(u) = u, \quad 0 < u < 1.$$

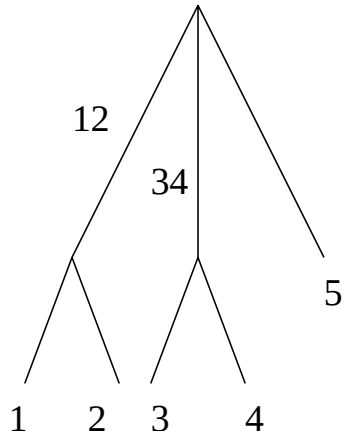
It follows from this that U must be uniformly distirbuted between 0 and 1.

14.5 Fun Examples

I work through a couple of examples showing situations in which you can use Chapter 2-3 tools as a means to figure out how to do something. These may be things you didn't know how to do before taking this course. I hope you find these examples fun; anyway, I had fun making them up!

14.5.1 Example: Which Coin is the Heavy Coin?

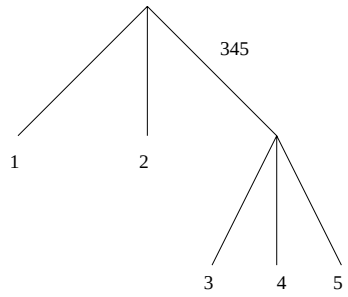
Suppose I have 5 coins which look identical. 4 of the coins weigh exactly the same but the 5th coin is heavier than each of the others. The coins are numbered 1,2,3,4,5 and any of these can be the heavy coin. We define a random variable X to be the number of the heavy coin; we assume that X is equiprobable. We are going to determine the heavy coin with a finite number of weighings. On each weighing, a balance beam scale is used: You can put a subset of the coins in the left pan of the scale, and an equal number of the remaining coins in the right pan. If the two sides balance the heavy coin belongs to the set of coins not in the pans; if one of the two pans is heavier, then the set of coins in that pan contains the heavy coin. The following tree denotes one possible weighing strategy:



The top level of the tree denotes the first weighing: Left pan of scale containing coins 1,2 is compared to right pan of scale containing coins 3,4 (coin 5 is held aside). If the total of coins 1,2 proves heavier than the total of coins 3,4, the left branch of the tree is followed; if the total of coins 3,4 is heavier, the middle branch is followed; if the two pans balance, the right branch is followed and we wind up at the leaf labeled 5, meaning that the heavy coin is automatically coin 5. If the left or middle branches were followed as a result of the first weighing, then a second weighing is necessary. If the left branch was followed as the result of the first weighing, coin 1 is weighed against coin 2 in the second weighing, and you follow the branch of the heavier of these two coins, winding up at a leaf labeled by the heavy coin. If the middle branch was followed as the result of the first weighing, coin 3 is weighed against coin 4, and then you follow the branch of the heavier of these two coins, winding up at a leaf for the heavy coin. The expected number of weighings is

$$1 * P(X = 5) + 2 * (1 - P(X = 5)) = 1.8.$$

The weighing strategy given by the following tree is better:

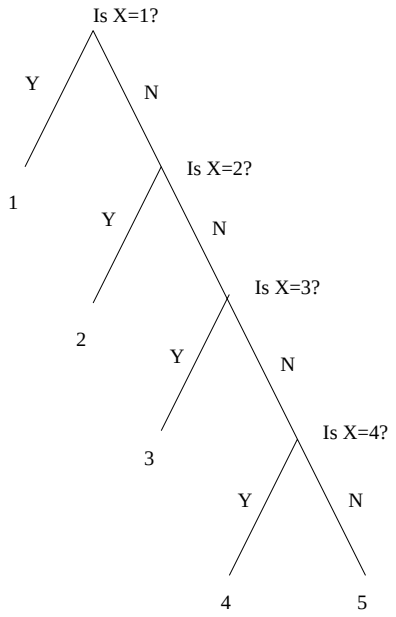


This is because the expected number of weighings is only 1.6 (show this).

14.5.2 Example: Questionaires

Ever play the game "20 questions"? What I show you in this example can be extended to give one a strategy for playing the 20 questions game in the best way possible.

Suppose I'm thinking of a number X belonging to the set $\{1, 2, 3, 4, 5\}$. We can take X as an equiprobable RV over this set. The following tree gives one possible questioning strategy ("questionaire") via which one can learn what X is through the asking of finitely many Yes-No questions:



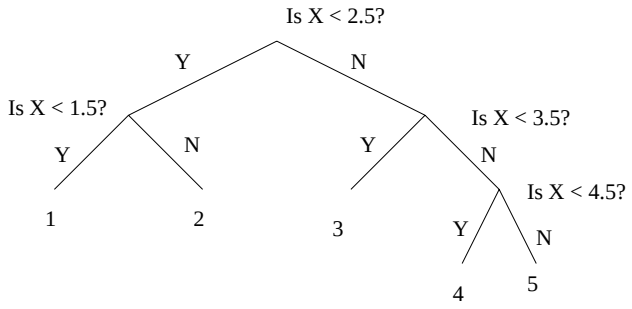
Let the RV Y denote the number of questions that have to be asked. Then Y takes the values 1,2,3,4. Its PMF is given by

$$\begin{aligned} P^Y(1) &= P(X=1) = 1/5 \\ P^Y(2) &= P(X=2) = 1/5 \\ P^Y(3) &= P(X=3) = 1/5 \\ P^Y(4) &= 1 - 3/5 = 2/5 \end{aligned}$$

The expected number of questions asked is

$$E(Y) = 1 * (1/5) + 2 * (1/5) + 3 * (1/5) + 4 * (2/5) = 2.8.$$

Here is another questionaire:



The number of questions that have to be asked is either 2 or 3. The expected number of questions is

$$2 * (P(X=1) + P(X=2) + P(X=3)) + 3 * (P(X=4) + P(X=5)) = 2.4.$$

The second questionaire is better than the first one because fewer questions are required, on average, to determine the value of X.