Review of Probability

1 Definition of Probability

The *Sample Space* is the set of all possible outcomes of an experiment. Note that this is sometimes subjective and will depend on context.

An *Event* is a subset of the sample space. We are generally interested in the probability of a particular event.

Recall the definition and notation used for union, intersections, and complements. For two events, E and F,

- The union of E and F is written as $E \bigcup F$ and consists of all the elements of E or F or both.
- The intersection of E and F is written as $E \bigcap F$ and consists of the elements that are in both E and F.
- The complement of E is denoted as E^c and includes all of the elements in the sample space which are not members of E.

Now, we would like to assign probabilities on all possible events we can make from the sample space. We assume that a function, $P(\cdot)$, from the possible events of the sample space to the real line satisfies the following:

1. $0 \leq P(E) \leq 1$
2. $P(S) = 1$
3. For any sequence of events $E_1, E_2, ...$ that are mutually exclusive, then

$$P\left(\bigcup_{n=1}^{\infty} E_n\right) = \sum_{n=1}^{\infty} P(E_n)$$

Note that these imply the following:

- $P(S) = P(E) + P(E^c) = 1$ which implies that $P(E^c) = 1 - P(E)$.
- $P(E \bigcup F) = P(E) + P(F) - P(E \bigcap F)$.

2 Conditional Probability

2.1 Definition

Now, let us define the conditional probabilities. Conditional probabilities allow us to formulate probabilities under partial information about the outcome of the experiment. The probability of E given F is defined as

$$P(E|F) = \frac{P(E \bigcap F)}{P(F)}$$

Some quick examples. Turning this definition around gives us a general definition for the probability of the intersection of two events.

$$P(E \bigcap F) = P(E|F)P(F)$$

This greatly simplifies certain calculations such as the probability of drawing two aces in sequence without replacement from a full deck of cards

2.2 Independence

Now, we may define a very important concept, that of independence. Events E and F are independent if

$$P(E \bigcap F) = P(E)P(F)$$

Note that this means that $P(E|F) = P(E)$ and $P(F|E) = P(F)$. Also, note that when this is expanded to more than two events that pairwise independent events are not necessarily independent (Example 1.10 of Ross).

2.3 Bayes' Rule

Another important formula that is useful for dealing with conditional probabilities is Bayes' rule. Bayes' rule allows us to essentially reverse the order of conditioning which can be convenient for some calculations.

$$P(F|E) = \frac{P(F \bigcap E)}{P(E)} = \frac{P(E|F)P(F)}{P(E \bigcap F) + P(E \bigcap F^c)} = \frac{P(E|F)P(F)}{P(E|F)P(F) + P(E|F^c)P(F^c)}$$

This can easily be generalized to any partition of the sample space, S. In other words, for events $F_1, ..., F_n$ which are mutually exclusive of one another and $\bigcup_{i=1}^{n} F_i$. Now, for any particular event, F_j,

$$P(F_j|E) = \frac{P(E|F_j)P(F_j)}{\sum_{i=1}^{n} P(E|F_i)P(F_i)}$$

Example Suppose you have three boxes filled with microchips. The first box has 5 defective and 20 good chips. The second box has 10 defective chips and 25 good chips. The third has 5 defective chips and 35 good chips. Suppose a chip is drawn at random from the three boxes. What is the probability of a defective chip?

$$P(E) = \left(\frac{1}{3}\right)\left(\frac{5}{25}\right) + \left(\frac{1}{3}\right)\left(\frac{10}{35}\right) + \left(\frac{1}{3}\right)\left(\frac{5}{40}\right) = \frac{57}{280}$$

Now, we have drawn a bad microchip. We want to blame one of the boxes. Let's calculate the probability that the bad chip came from the first box

$$P(F_1|E) = \frac{\left(\frac{1}{3}\right)\left(\frac{5}{25}\right)}{\left(\frac{1}{3}\right)\left(\frac{5}{25}\right) + \left(\frac{1}{3}\right)\left(\frac{10}{35}\right) + \left(\frac{1}{3}\right)\left(\frac{5}{40}\right)} = \frac{56}{171} = 0.327$$

3 Random Variables

Random variables are functions from the sample space to numbers–a mapping from the possible outcomes to numeric results. An easy way to see the relationship is through the mapping of a sample space generated by rolling two dice to the sum of those dice. The original sample space consists of all the ordered pairs of the integers from 1 to 6. These pairs are mapped to the single numbers 2 through 12. If we use X to denote the sum of the two dice, we can calculate an example of a probability associated with this random variable as follows:

$$P\{X = 3\} = P\{(1,2),(2,1)\}$$

Often, we will not even write down the sample space–this may be too complicated. Take as an example, the random variable counting the number of cars to arrive at a rural intersection in an hour. This might be modeled as a Poisson random variable. What would be the original sample space? It is hard to spell out all the contributing factors to the arrival of cars.

Random variables are generally categorized as *discrete* or *continuous* depending upon the possible values of the random variable. The former take on values that are a subset of the integers, and the latter take on values in an uncountable set–generally some interval of real numbers. Both types of random variables have an associated cumulative distribution function (CDF). This function is defined as

$$F(b) = P(X \leq b)$$

This function has a number of important properties.

1. $F(b)$ is a nondecreasing function of b
2. $\lim_{b\to\infty} F(b) = 1$
3. $\lim_{b\to\infty} F(b) = 0$

We may calculate intervals using the following rule:

$$P(a < X \leq b) = F(b) - F(a)$$

To calculate $P(X < b)$ we need only to take a limit

$$P(X < b) = lim_{h\to 0^+} P(X \leq b - h) = lim_{h\to 0^+} P(X \leq b - h)$$

We this last situation frequently arises when dealing with discrete random variables.

4 Discrete Random Variables

For a discrete random variable X, we define the *probability mass function* $p(a)$ by

$$p(a) = P(X = a)$$

We have for a discrete set of values, $x_1, x_2, ...$ that

$$p(x_i) > 0, \quad i = 1, 2, ...$$

and

$$p(x) = 0$$

for all other x.

Also,

$$\sum_{i=1}^{\infty} p(x_i) = 1$$

and

$$F(a) = \sum_{x_i \leq a} p(x_i)$$

We can also define expectation for discrete random variables. In general, expectation is defined as

$$Eg(X) = \sum_{i=1}^{\infty} g(x_i)p(x_i)$$

We can use this definition to define both the mean and variance of a random variable. The mean is

$$EX = \sum_{i=1}^{\infty} x_i p(x_i)$$

and the variance

$$Var(X) = E(X - EX)^2 = E(X^2) - (EX)^2$$

In addition, the following facts are easily shown (Corollary 2.2). If a and b are constants, then

$$E(aX + b) = aEX + b$$

and

$$Var(aX + b) = a^2 Var(X)$$

4.1 Bernoulli Random Variable

Perhaps the simplest random variable is the Bernoulli random variable. This is a common model for a coin toss. The random variable is one with a success and zero with a failure. The pmf is given as

$$p(0) = 1 - p \quad p(1) = p$$

with parameter p With this parameterization, the mean of a Bernoulli random variance is

$$EX = p$$

and the variance

$$Var(X) = p(1 - p)$$

4.2 Binomial Random Variable

A Binomial Random Variable is used to represent the total number of successes in n independent Bernoulli trials. (Equivalently, such random variables represent the sum of n iid Bernoulli random variables.) The pmf for a Binomial Random Variable is given as

$$p(i) = \binom{n}{i} p^i (1 - p)^{n-i} i = 0, 1, ..., n$$

The mean of a Binomial Random Variable is np, and the variance is $np(1-p)$. (These are obvious if you think about the Binomial Random Variable as a sum of independent Bernoulli random variables.)

4.3 Poisson Random Variable

The Poisson Random Variable is often used to model the number of rare events. (We'll talk about this later.) The pmf of a Poisson Random Variable is given as

$$p(i) = \frac{e^{-\lambda}\lambda^i}{i!}, i = 0, 1, 2, 3, ...$$

The mean and variance of a Poisson Random Variable is λ.

5 Continuous Random Variables

A continuous random variable may take on an uncountable number of values. X is a continuous random variable if there is a non-negative function, $f(x)$ such that for any (reasonable) set B of real numbers

$$P(X \in B) = \int_B f(x)dx$$

The function $f(x)$ is called the probability density function. Obviously, if we integrate $f(x)$ over the real line, we must obtain one to ensure proper probabilities.

Question: Is the value of a pdf at a particular value a probability?

We can also define expectation for cts random variables. In general, expectation is defined as

$$Eg(X) = \int_{-\infty}^{\infty} g(x)f(x)dx$$

We can use this definition to define both the mean and variance of a random variable. The mean is

$$EX = \int_{-\infty}^{\infty} xf(x)dx$$

and the variance

$$Var(X) = E(X - EX)^2 = E(X^2) - (EX)^2$$

5.1 Uniform Random Variable

The Uniform Random Variable places models equal probabilities over a fixed interval of real numbers. It is especially important as the distribution approximated by pseudo-random number generators. The density for the uniform random variable is given by

$$f(x) = \frac{1}{b-a} \; if \; a < x < b$$

and zero otherwise.

The mean using this parameterization is $\frac{a+b}{2}$ and the variance is $\frac{(b-a)^2}{12}$.

5.2 Exponential Random Variable and Gamma Random Variable

The Exponential Random Variable will be commonly used in this course due to its memoryless property (which we will see later). It is in some respect a continuous analog to the geometric distribution. The pdf is given as

$$f(x) = \lambda e^{-\lambda x} \ for \ x \geq 0$$

and zero otherwise. Note that with this parameterization the mean is $1/\lambda$ and the variance is $1/\lambda^2$. Sometimes λ is called the rate (on average λ events per time period).

The exponential distribution is a special case of the gamma distribution which has a pdf given by

$$f(x) = \frac{\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}}{\Gamma(\alpha)}$$

for non-negative x and zero otherwise. The gamma function is defined by

$$\Gamma(\alpha) = \int_0^\infty e^{-x} x^{\alpha-1}$$

The mean of such a random variable is $\frac{\alpha}{\lambda}$ and the variance is $\frac{\alpha}{\lambda^2}$. The sum of α independent exponential random variables each with mean $1/\lambda$ is gamma with the obvious parameters. Also, the gamma function has the following property

$$\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$$

for $\alpha > 1$. This combined with the fact that $\gamma(0) = 1$ yields

$$\Gamma(\alpha) = (\alpha - 1)!$$

for positive integer α.

5.3 Normal Random Variable

The Normal distribution is very important largely due to the generality of the Central Limit Theorem (see below).

The pdf of a Normal Random Variable is

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

for $-\infty < x < \infty$. With this parameterization, the mean is μ and the variance is σ^2.

6 Joint Distributions

We have so far we have discussed a single random variable. However, random variable frequently occur in groups, and random variable may depend on one another. We define for joint random variables X and Y the joint CDF,

$$F(a,b) = P(X \leq a, Y \leq b), \quad -\infty < a, b < \infty$$

We may return to a single distribution with

$$\begin{aligned} F_X(a) &= P(X \leq a) \\ &= P(X \leq a, Y < \infty \\ &= \lim_{b \to \infty} F(a,b) \end{aligned}$$

similarly we may determine $F_Y(b)$.

For discrete random variables X and Y, we may define the joint probability mass function

$$p(x,y) = P(X = x, Y = y)$$

From this we can calculate the (marginal) pmf for X with

$$p_X(x) = \sum_{y:p(x,y)>0} p(x,y)$$

Similarily, X,Y are jointly continuous if there is a joint density function, $f(x,y)$, such that for all (reasonable) sets

$$P(X \in A, Y \in B) = \int_B \int_A f(x,y) dx dy$$

Like the case of discrete random variables, we may recapture the (marginal) density of X alone by integrating the joint density over all possible values of Y.

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

A similar calculation may be done for Y.

Now, we will define expectation for pairs of random variables. For discrete random variables,

$$E(g(X,Y)) = \sum_x \sum_y g(x,y)p(x,y).$$

For continuous random variables,

$$E(g(X,Y)) = \int_x \int_y g(x,y)p(x,y)dydx.$$

A special case of this allows us to simply the calculation

$$E[aX + bY] = aE[X] + bE[Y]$$

Obviously, this can be generalized to n random variables.

6.1 Independent Random Variables

Just as events can be independent, so too can random variables. Random variables, X and Y are independent if for all a and b,

$$P(X \leq a, Y \leq b) = P(X \leq a)P(Y \leq b)$$

Obviously, this is equivalent to

$$F(a,b) = F_X(a)F_Y(b)$$

If X and Y are independent, then for discrete RV's

$$p(x,y) = p_X(x)p_Y(y)$$

or for continuous RVs

$$f(x,y) = f_X(x)f_Y(y)$$

From these facts it is easy to derive the following

$$E(g(X)h(Y)) = E(g(X))E(h(Y))$$

which will be useful in calculating the covariance of independent random variables.

6.2 Covariance

Covariance is a measure of dependency between random variables. The covariance of random variables X and Y is defined as

$$Cov(X,Y) = E[(X - EX)(Y - EY)] = E(XY) - EXEY$$

It is fairly easy to obtain the following properties of covariance:

1. Cov(X,X)=Var(X)
2. Cov(X,Y)=Cov(Y,X)
3. Cov(cX,Y)=c Cov(X,Y)
4. Cov(X,Y+Z)=Cov(X,Y)+Cov(X,Z)

This last property can be extendended to give

$$Cov(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j) = \sum_{i=1}^{n}\sum_{j=1}^{m} Cov(X_i, Y_j)$$

The concept of covariance allows us to calculate the variance of the sum of random variables.

$$\begin{aligned} Var(\sum_{i=1}^{n} X_i) &= Cov(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} Cov(X_i, X_j) \\ &= \sum_{i=1}^{n} Cov(X_i, X_i) + \sum_{i=1}^{n}\sum_{j\neq i} Cov(X_i, X_j) \\ &= \sum_{i=1}^{n} Cov(X_i, X_i) + 2\sum_{i=1}^{n}\sum_{j<i} Cov(X_i, X_j) \end{aligned}$$

Since independent random variables have zero covariance(Check this yourself.), the last property gives that the variance of the sum of random variables is equal to the sum of the variances of the random variables.

6.3 Limit Theorems

We have already discussed the central limit theorem in the context of the importance of the normal distribution. Perhaps an even more important result is the Strong Law of Large Numbers.

Let $X_1, ...$ be a sequenced of iid random variables where $EX_i = \mu$. Then, with probability 1,

$$\frac{X_1 + ... + X_N}{n} \to \mu$$

There are a set of inequalities that may help in proving limit theorems. One is Markov's inequality, for X which takes only non-negative values, then for any value $a > 0$,

$$P(X \geq a) \leq \frac{EX}{a}$$

From this we may derive Chebyshev's inequality,

$$P(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2}$$

This can be used, for instance to prove a simplified version of the law of large numbers (sometimes called the weak law of large numbers). The inequality implies that

$$P(|\bar{X} - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}$$

If we take the limit as n approaches infinity, then this quantity goes to zero for every ϵ. This is the definition of convergence in probability. (Notice we also assumed finite variances which means this is a further weakening of the theorem.)

A version of the central limit theorem is as follows: Let $X_1, X_2, ...$ be a sequence of iid random variables each with mean μ and variance σ^2. Then

$$P\left[\frac{X_1 + X_2 + ...X_n - n\mu}{\sigma\sqrt{n}} \leq a\right] \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2} dx$$

6.4 Moment Generating Functions

Moment generating functions can be used for a number of purposes including of course calculating the moments of a distribution. It is defined as

$$\phi(t) = Ee^{tX}$$

We may obtain the moments by differentiating $\phi(t)$ and evaluating those derivatieves at zero. The first derivative yields the first moment, the second yields the second moment, etc. A quick example would be the exponential distribution; it is much easier to calculate the moment generating function than the direct calculation of the mean which requires integration by parts.

Another very important use is that the moment generating function of the sum of independent random variables is the product of individual random variables. Thus,

$$\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$$

We may show for instance that the sum of normal random variables is also normal. The MGF for a normal is

$$\phi(t) = e^{\frac{\sigma^2 t^2}{2} + \mu t}$$

What is the distribution of the sum of two iid normals, X and Y?

$$\phi_{X+Y}(t) = e^{\frac{\sigma^2 t^2}{2} + \mu t} e^{\frac{\sigma^2 t^2}{2} + \mu t} = e^{\frac{2\sigma^2 t^2}{2} + 2\mu t}$$

Therefore, the sum is also normal but with mean 2μ and variance $2\sigma^2$. The other way to do this calculation is using convolutions which can be very difficult.