Introduction to Probability Theory and Mathematical Statistics

Because the basis of Econometric Theory and its application is reliant on Statistical Theory, we will review and examine some concepts here before we proceed.

1 Random Variables

Let us first define an **Experiment**. It is a procedure, or act that we can in theory repeat or replicate ad infinitum, or as defined by the experimenter, and it must have a well defined set of **outcomes**. Here are some examples of experiments:

- Coin flipping to see how many times you get heads and tails.
- Rolling of a fair dice to see how many times a particular number appears.
- A reward of $1000 presented to you to spend as you wish for scoring at least an A- in a series of economics tests, and examining how much money you win, or how many times you actually score at least an A-.
- A game of chess with Grand Master Kasporov to see how many times you win or lose. Actually, is this really an experiment?

A **Random Variable** is a variable that takes on numerical values and has an outcome that is determined by an experiment. We will denote random variables (inclusive of all outcomes, e.g. Number of heads and tails) in upper case letters, while particular realizations of outcomes (e.g. Number of heads in one experiment) in lower case letters, in all our subsequent exposition. The following could be a possible representation of the outcomes from the examples above;

- Let X denote the number of times heads are realized in 8 tosses of a fair coin. Then possible outcomes of experiment X are $\{0, 1, 2, 3, 4, 5, 6, 7, 8\}$. X then represent one experiment. We can express this idea more succinctly as $X \in \{0, 1, 2, 3, 4, 5, 6, 7, 8\}$. Sometimes we may wish to repeat an experiment a few times. Suppose we replicate this experiment (of 8 tosses of the coin) n times (In fact we could replicate it any amount of time we want.) say once in each season, and under different levels of humidity. Just for the fun of it! Let us than denote each of this separate replication as subscript to X, such as $\{X_1, X_2, ...X_n\}$, and each one of them would have the same possible outcome of $X_i \in \{0, 1, 2, ..., 8\}$, $i \in \{1, 2, ..., n\}$. Let us denote each realization in each replication as x, and x can take on any number from $\{0, 1, 2, ..., 7, 8\}$. That is if on one of the replications, we got heads six times, then we can write it as $x = 6$. In situations when we run the experiment several times, each realization can be differentiated from the other by a subscript as well, $\{x_1, x_2, ..., x_8\}$.

- Write down the possible outcomes for yourself on say 3 rolls of a dice, where the outcome is the sum of the dice. And suppose you perform this experiment 4 times, one in each season. Does you luck change with the season, if getting all sixes is your aim?

- Do the same as in above, and let's say you take the test four times, in each replication. And you do it four times. What are the outcomes? What is the maximum amount of money you could earn? Why would I be interested in whether monetary rewards affect your test scores?

- An outcome of a game with Kasporov should be quite easy! It'd be a massacre?

Note:

1. A random variable that takes on either the values of 0 or 1 is called a **Bernoulli** (Binary) random variable. Suppose we have you toss a coin and just want to note if you get a head or a tail, and suppose the school gives you a loonie for each head you get. We can then denote heads as 1, and 0 otherwise, and the random variable of you getting heads can then be classified as a Bernoulli random variable. We typically think of 1 as success and 0 as a failure, such as when all you're concerned with when taking a test as whether you get A+ or not. However, there is nothing wrong with you switching the convention.

2. So far all the outcomes we have thought of are discrete, and we can call the random variables as discrete random variables. However, if we examining how education affects

your income in the future. Such an outcome can be thought of as continuous since its realization is on the positive real line, and we refer to them as continuous random variables. We will examine each in turn.

1.1 Discrete Random Variables

A **Discrete Random Variable** is one that takes on a finite or countably infinite number of values, such as in all of our above examples, and the Bernoulli random variable is the simplest form. We are in general of course not just concerned with the actual outcomes per se, but the probability or the regularity with which an event occurs.

Consider our monetary incentives example; what we are considering is whether the prospect of a monetary reward alters the amount of effort you place in your work. For some of you, such an incentive would work, while for others it might, or might not. Let us denote 1 as the event where a subject in the experiment scores at least an A-and 0 otherwise. Suppose there were 20 of you in the class, and the test was applied once, what would be the probability of you succeeding given the incentive? (Of course, the experiment is fully controlled, such that you cannot go back home, and ask your folks for the $1000. This highlights an important aspect of a real experiment. Control of the subjects environment.) Well, suppose before the incentives was applied, no one scored above A-, and afterwards, I note an increase from 0 to 10. Which means the probability of the incentives being a success is 1/2. More generally, we do not know how our experiments would turn out. Let us denote the probability of a successful realization, γ. We can express the probability for the event's of success as

$$\Pr(X = 1) = \gamma$$

This reads as "the probability of success, or X taking the value of 1 is γ". The probability of failure is

$$\Pr(X = 0) = 1 - \Pr(X = 1) = 1 - \gamma$$

More generally, suppose we have n possible discrete outcomes, using the previously noted notation convention, let X be the experiment in question, then $X = \{x_1, x_2, ..., x_{n-1}, x_n\}$. Let the probability of each event occurring can be written as,

$$\Pr(X = x_i) = p_i$$

for $i = 1, 2, ..., n$, and we say that "the probability the random variable X has an outcome of x_i is p_i". Further, since the outcomes we have are exhaustive, the sum of the probability

of each event occurring must be 1. In other words,

$$\sum_{i=1}^{n} p_i = 1$$

We typically call the function, or mathematical expression that summarizes the probability an outcome occurs as a **Probability Density Function** (pdf), or density in short. In the general case, the pdf is

$$\Pr(X = x_j) = f(x_j) = p_j$$

$j = 1, 2, ..., n$. When there are several random variables, it is useful to include the subscript. For example, when we have two random variables X, and Y, then the p.d.f. for the two variables can be written as f_X and f_Y respectively.

Let examine a specific example. Suppose we run an experiment that has 4 outcomes of $X = \{100, 300, 400, 1000\}$. After several replications, we note that each outcome has the following probabilities of occurring (let's say we ran 1 million replications and are pretty sure we have fully controlled the environment.) the following probabilities, $\{\frac{1}{8}, \frac{1}{4}, \frac{3}{8}, \frac{1}{4}\}$. Let the p.d.f. of X be f_X, then the p.d.f. can be expressed as

$$f_X = \begin{cases} \frac{1}{8} & \text{if} X = 100 \\ \frac{1}{4} & \text{if} X = 300 \\ \frac{3}{8} & \text{if} X = 400 \\ \frac{1}{4} & \text{if} X = 1000 \end{cases}$$

How would you depict this if the probability of the outcome is on the y axis, and the outcome is on the x axis? Refer to your text on page 731. What is the probability X is less than 400?

1.2 Continuous Random Variables

A variable X is a continuous random variable if it takes on any real value with probability 0. The idea is that when a random variable is continuous, there is an infinite amount of possible realizations such that it is not possible to match or count them, technically it each realization has a probability of 0 on the limit. Most of the variables we deal with are typically continuous variables, such as labor income, or a country's GDP etc. However, when we examine their probability of occurrence, we group them into bounds, such as when we consider your income on graduation of being somewhere between CAD 40,000 to perhaps CAD 70,000 annually.

This then makes matching and count. More generally, suppose Y is a random variable with the following support, or upper and lower bounds for its outcomes, $Y \in [-\infty, \infty]$. Suppose you are concerned with the event occurring between a and b, where a and b are real numbers. Then what we are concerned with is $\Pr(a \leq Y \leq b)$.

When working with continuous random variables, we typically calculate the cumulative distribution, and the function derived is called the **Cumulative Distribution Function** (c.d.f.). Let y be the realization we are concerned with, then the c.d.f. is just,

$$F(y) = \Pr(Y \leq y)$$

This then allows us to calculate the probabilities of different sets of events. Again note that the sum of the probabilities of the events must integrate (since we are dealing with continuous variables) to 1. Consider $\Pr(a \leq Y \leq b)$, we can calculate this by finding the c.d.f. for a and b, since

$$\Pr(a \leq Y \leq b) = \Pr(Y \leq b) - \Pr(Y \leq a) = F(b) - F(a)$$

Also,

$$\Pr(Y > b) = 1 - \Pr(Y \leq b) = 1 - F(b)$$

Also note the relationship B.9 and B.10 on page 733 of your text. Essentially, when dealing with continuous variables, it is not necessary to be pedantic about the inequality sign since the probability of any precise event occurring is zero. But it matters a lot for discrete random variables.

1.3 Formal Definition of Random Experiments

To formalize what we have discussed in a general form, and rather unimportant to you at this juncture of your academic career, but nonetheless useful as a summary,

Definition 1 *A random experiment is a triplet* $(\Omega, \mathcal{A}, P)$ *where:*

1. Ω *is the* ***set of all possible outcomes of the experiment****. For example, the possible outcomes an experiment involving the roll of a single fair dice has the* ***sample space****,*

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

2. *$\mathcal{A}$ is a σ-algebra, or a set of subsets called events of Ω containing the sets $\emptyset$ and Ω, and is closed under countable union and complement operations. Consider the following σ-algebra formed by the following subsets of Ω*

$$\begin{aligned} A_1 &= \{\text{Outcomes of at least } 3\} = \{1,2,3\} \\ A_2 &= \{\text{Outcomes strictly greater than } 3\} = \{4,5,6\} \\ B_1 &= A \cap A^c = \{\} = \emptyset \\ B_2 &= A \cup A^c = \Omega \end{aligned}$$

3. $\Pr$ *maps a set $\mathcal{A} \to \mathbb{R}^+$ such that $\Pr(\Omega) = 1$ and for a sequence $\{A_n\}$ of mutually exclusive events where $\sum_{i=1}^n A_n$ is their union, then $\Pr(\sum_{i=1}^n A_n) = \sum_{i=1}^n \Pr(A_n)$. The latter is referred to as the σ-additivity property, and P is called a probability measure or a probability distribution. Using the same example, note that,*

$$\Pr(A_1 + A_2) = \Pr(A_1) + \Pr(A_2) = 1$$

2 Joint and Conditional Distributions, and Independence

2.1 Joint Distribution and Independence

Of course our discussion so far pertaining to one variable is a highly specialized scenario since in economics, we typically interested in the probability of several events occurring. Consider the test-incentive example, suppose each experiment is for each one of you. Then it would be interesting to me the probability of everyone heading the reward system created, i.e. the probability of getting A- jointly by each of you. The must likely reason I would be interested is whether one of you putting in effort because of the reward affects the other, and of course if all of you do while, it is quite a substantial gain to you, but a huge loss to me!

Another example, consider I know the distribution of the random variable which is your test score, and the distribution of the random variable which is your true ability/IQ. It would be curious to know the joint distribution of your ability and your test score. If I were to arrange these two random variables in a matrix, with say your ability indexed for the rows (first row as lowest ability, and last row as highest ability), and test scores for the columns (With first column being the lowest test score, and last column being the highest test score),

then the values it takes in relation to each other has information, such as if ability implies a greater probability of doing well in the tests, I should see greater probability in the south-east corner of the matrix. Think about other possible situations, a priori, what do you think the matrix might look like? How about the joint density look like? Think about it.

Formally, let X and Y be two discrete random variables. Then (X, Y) have a joint distribution which is fully described by their joint probability density function;

$$f_{X,Y}(x, y) = \Pr(X = x, Y = y)$$

If these two random variables are *independent* of each other, such as when your ability has nothing to do with your test scores, the following is obtained;

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

How about when X and Y are continuous variables? Well, in stead of a two dimensional diagram as above, we now have three, but it does not change the fact that the point wise density is still 0. Suppose the support of X and Y are the same, and suppose we are wondering about the probability X falls between a and b, and Y fall between c and d, then the joint density of such an occurrence is, $\Pr(a < X < b, c < Y < d)$, and if they are independent, we have $\Pr(a < X < b, c < Y < d) = \Pr(a < X < b) \Pr(c < Y < d)$.

When random variables are independent, what we are saying is that one random variable does not affect and has no relationship with the other. Such as when we say that ability does not affect your test scores. When that is true, and we wish to examine the probability of the joint realization of the two random variables, all we need to know are the individual pdf's (Marginal Densities. We will discuss this shortly.) for each set of experiments. The relationships for independence holds true regardless of whether we have two or more random variables. Note that if the random variables are not independent, necessarily they are dependent on each other.

So what is the import of this idea of independence? Some of the classic distributions in statistics in derived from this idea, and you will see it has great importance in econometrics as well. Think of a roll of the dice. If the dice is a fair dice, that is equally weighted, each roll is totally independent of the following roll.

Consider the test-incentive experiment. For me to examine the true importance of incentives to you, without contamination by outside forces and peer effects, I must isolate you

from everyone else when performing the experiment and teach in exactly the same manner with absolutely no difference in no content nor methodology, which means each set of tests I administer to each of you in segregation is necessarily independent. Let the probability of you scoring more than A- be γ. An let it be that I wish to run the test n times. I would be interested to know the probability that out of the n times, how many times would you be able to get that reward, i.e. the pdf of the random variable which is the probability you will do well and get the incentive reward of \$1000 is

$$\begin{pmatrix} n \\ x \end{pmatrix} \gamma^x (1-\gamma)^{n-x}$$

$\forall x = 0, 1, 2, ..., n$. In our example of $n = 4$, what is the probability you will get \$4000?

2.2 Conditional Distributions

As noted earlier, in econometrics, what we are usually concerned with is given certain conditions, be it personal or environmental in nature, what is the probability of an event occurring. In our example of the testing experiment, if I know the distribution of ability among you, I would like to know how much of the probability of success in any one test is due to your ability, suppose it is observable (In truth, do you know your true ability or potential?). Using the notation suggested earlier, and letting X be for test attainment, and Y be for ability, what we want to know is the following,

$$\Pr(X = x|Y = y) = f_{X|Y} = \frac{\Pr(X = x, Y = y)}{\Pr(Y = y)} = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

How would you write this expression if X and Y are both continuous random variables? What if X and Y are independent of each other, i.e. Y does not affect X, or ability has no bearing on your attainment?

3 Features of Probability Distributions

3.1 The Expected Value/Expectation

One of the key summary statistics is the expected value of a random variable, or what you typically call an average. Taking your test-incentive example again, suppose there are 12 grade outcomes, and we know the p.d.f., the mean grade is just the average of all your grades.

If the p.d.f. I know is for any and every student that takes my class, I can safely say the mean grade (barring any change in quality of students across each year) is just

$$\mathbf{E}(x) = \sum_{i=1}^{12} x_i f(x_i)$$

More generally for n realizations for discrete random variable Y;

$$\mathbf{E}(y) = \sum_{i=1}^{n} y_i f(y_i)$$

If Y is continuous with a support between $-\infty$ and ∞,

$$\mathbf{E}(y) = \int_{-\infty}^{\infty} y f(y) dy$$

Further, if what we are concerned with not the random variable of Y per se, but a function (i.e. as part of an equation) of it, and let that function be $h(Y)$, then;

$$\mathbf{E}(h(y)) = \sum_{i=1}^{n} h(y_i) f(y_i)$$

and

$$\mathbf{E}(h(y)) = \int_{-\infty}^{\infty} h(y) f(y) dy$$

when Y is discrete and continuous respectively. This idea of expectation also generalizes into cases when we are concerned with the expected value of a function of two or more random variables.

The Expectation Operator, $\mathbf{E}(.)$ has several properties that you must know,

1. For any constant c, $\mathbf{E}(c) = c$.
2. For any constant a and b, $\mathbf{E}(aX + b) = aE(X) + b$.
3. If $\{a_1, a_2, ..., a_n\}$ are constants, and $\{X_1, X_2, ..., X_n\}$ are random variables, then,

$$\mathbf{E}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i \, \mathbf{E}(X_i)$$

We typically denote the mean or the expected value of a random variable as μ.

3.2 Variance and Standard Deviation

Concepts like mean (expected value), or median which are measures of central tendency conveys the idea of where most of our observations are concentrated around. But we are just as concerned with how far the observations can deviate from the mean. In terms of our test-incentive example, I would be concerned with how far anyone of you might choose not to exert any effort at all (It should not be construed that you should be concerned with monetary reward in making your effort choice in school!). If we are concerned with deviations, then what kind of deviation should we be concerned with? Typically, we are concerned with deviations of outcomes from the mean. But why should it be a concern? Consider the following distributions. If it depicts your p.d.f. for the random variable which is your test score (if it is random!), which would you prefer?

The variance measure is expressed as,

$$\mathbf{V}(X) = \mathbf{E}(X - \mu)^2$$

where $\mu = \mathbf{E}(X)$. It can be expressed simply for calculation as,

$$\begin{aligned} \mathbf{V}(X) &= \sigma^2 \\ &= \mathbf{E}(X - \mu)^2 \\ &= \mathbf{E}(X^2) - \mu^2 \end{aligned}$$

Properties of Variance

1. For any constant c, $\mathbf{V}(c) = 0$.
2. For any constant a, $\mathbf{V}(aX) = a^2\,\mathbf{V}(X)$
3. For any constant a and b, $\mathbf{V}(aX + b) = a^2\,\mathbf{V}(X)$

Standard Deviation is just the positive value of the square root of the variance;

$$\sigma = +\sqrt{\mathbf{V}(X)}$$

3.3 Standardizing a Random Variable

We can standardize a random variable by subtracting the mean from it, and dividing it by its standard deviation. This is to easy inference made on the random variable. Its import will be apparent subsequently.

$$Z \equiv \frac{X - \mu}{\sigma}$$

The neat properties about it can be seen easily,

$$\begin{aligned} \mathbf{E}(Z) &= \mathbf{E}\left(\frac{X-\mu}{\sigma}\right) \\ &= \sigma^{-1}(\mathbf{E}(X) - \mu) = 0 \end{aligned}$$

and

$$\begin{aligned} \mathbf{V}(Z) &= \mathbf{V}\left(\frac{X-\mu}{\sigma}\right) \\ &= \sigma^{-2}\,\mathbf{V}(X) = 1 \end{aligned}$$

which means it will have a mean of 0 and a variance and standard deviation of 1, regardless of the original p.d.f. of X (or the p.d.f. of the random variable it was transformed from).

4 Features of Joint and Conditional Distributions

As mentioned above, in econometrics we are often concerned with the connection between two random variables. Hence it is we want to be able to measure the manner in which two or more random variables vary in relation to each other.

4.1 Covariance and Correlation

Covariance measures how much two random variables varies in relation to each other. Let there be two random variables, X and Y, with their respective means μ_X and μ_Y. Then their covariance is just,

$$\begin{aligned} \text{cov}(X,Y) &= \mathbf{E}(X-\mu_X)(Y-\mu_Y) \\ &= \mathbf{E}(XY) - \mu_X\mu_Y \end{aligned}$$

Properties of Covariance

1. If X and Y are independent, then $\text{cov}(X, Y) = 0$.
2. For any constant a_1, a_2, b_1 and b_2, $\text{cov}(a_1X + b_1, a_2X + b_2) = a_1a_2\,\text{cov}(X, Y)$.
3. $\text{cov}(X, Y) \leq \sigma_X\sigma_Y$

However, the covariance measure is dependent on the unit of measure of the random variables. We can eliminate this by using the correlation measure which normalizes the covariance measure by the individual standard deviations. That is,

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X\sigma_Y} = \frac{\sigma_{X,Y}}{\sigma_X\sigma_Y}$$

There are several properties of the correlation coefficient that makes inferences easier;

1. $-1 \leq \rho_{X,Y} \leq 1$
2. For constants a_1, a_2, b_1 and b_2, if $a_1, a_2 > 0$, $\text{corr}(a_1X + b_1, a_2X + b_2) = \rho_{X,Y}$ and if $a_1, a_2 < 0$, $\text{corr}(a_1X + b_1, a_2X + b_2) = -\rho_{X,Y}$.

Variance of Sums of Random Variables

1. For constants a and b,

$$\mathbf{V}(aX + bY) = a^2\,\mathbf{V}(X) + b^2\,\mathbf{V}(Y) + 2ab\,\text{cov}(X, Y)$$

What is the variance if X and Y are independent?

2. The above generalizes to the case of n uncorrelated random variables. For $\{X_1, X_2, ..., X_n\}$, pairwise uncorrelated, the variance of their sum is

$$\mathbf{V}\left(\sum_{i=1}^{n} a_iX_i\right) = \sum_{i=1}^{n} a_i^2\,\mathbf{V}(X_i)$$

4.2 Conditional Expectation

A problem with the measure of expected value is that it is unconditional, in the sense that we do not know what the mean conditional on other influential factors is. Consider the test-incentive example again. What I would like to know is conditional on ability, what would the expected value be. That is would be higher for higher ability individuals' vis-à-vis lower ability individuals.

This is achieved by the measure of conditional expectation which technically is

$$\mathbf{E}(X|Y=y) \;=\; \sum_{i=1}^{n} x_i f_{X|Y}(x_i|y)$$

Intuitively, what this statistic is measuring is that it is measuring the average of X for observations where Y's realization is y, or in terms of our example it is measuring the average grade among individuals with say and average IQ score. There would no reason to expect that this measure is the same for individuals of different ability, or for different realizations of Y.

Some of the properties of Conditional Expectation are

1. Let $h(.)$ be a function. $\mathbf{E}(h(X)|X) = h(X)$. What this intuitively says is that if we know X, we know the function of X.
2. For functions h and g, $\mathbf{E}(h(X)Y + g(X)|X) = h(X)\,\mathbf{E}(Y|X) + g(X)$.
3. If X and Y are independent, then $\mathbf{E}(X|Y) = \mathbf{E}(X)$. Also if $\mathbf{E}(X|Y) = \mathbf{E}(X)$, than $\text{cov}(X,Y) = 0$, and therefore $\rho_{X,Y} = 0$.
4. The Law of Iterated Expectations, $\mathbf{E}(\mathbf{E}(X|Y)) = \mathbf{E}(X)$.
5. A more general version of property 4, $\mathbf{E}(X|Y) = \mathbf{E}(\mathbf{E}(X|Y,Z)|Y)$, for another random variable Z.
6. Let $\mathbf{E}(X|Y) = \mu(Y)$ which just says that the conditional mean of X is a function of Y. Then for a function h, if $\mathbf{E}(X^2) \leq \infty$ and $\mathbf{E}(h(Y)^2) \leq \infty$, then $\mathbf{E}((X-\mu(Y))^2|Y) \leq \mathbf{E}((X-h(Y))^2|Y)$ and $\mathbf{E}(X-\mu(Y))^2 \leq \mathbf{E}(X-h(Y))^2$.

4.3 Conditional Variance

Just as we have conditional expectation, we would also need to define conditional variance,

$$\mathbf{V}(X|Y=y) = \mathbf{E}(X^2|Y=y) - (\mathbf{E}(X|Y=y))^2$$

further if X and Y are independent, then $\mathbf{V}(X|Y) = \mathbf{V}(X)$.

5 Basic Distributions

Although much of the distributional assumptions we make in this class pertain to continuous distributions particularly the Normal Distribution and its related distributions, there are instances that other discrete and continuous random variable distributions come into consideration as you progress in the study of Econometrics. It is consequently worth our while to spend some time on some of the more common ones here.

5.1 Uniform Distribution

A continuous random variable X which is uniformly distributed on the support or sample space $[a, b] \in \mathbb{R}^+$ has a density function,

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if} \quad x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$

Its c.d.f. is

$$F(x) = \begin{cases} 0 & \text{if} \quad x \leq a \\ \frac{1}{b-a} & \text{if} \quad x \in [a, b] \\ 1 & \text{if} \quad x \geq b \end{cases}$$

What is the mean and variance of X?

This distribution has a discrete counterpart. Let discrete random variable X have the following sample space $\{1, 2, ..., n\}$ where $\Pr(X = i) = \frac{1}{n}$ $i = \{1, 2, ..., n\}$. **Show that the mean and variance of X is $\frac{n+1}{2}$ and $\frac{n^2-1}{12}$ respectively.**

5.2 Binomial and Multinomial Distributions

We had discussed the **Bernoulli** random variable briefly, and will formally introduce it here. Let X be a random variable with one of two mutually exclusive outcomes, $\{0, 1\}$, for example 0 denotes tails and 1 denotes heads. Let the probability of $\Pr(X = 1) = p_1 = p$ so that $\Pr(X = 0) = p_0 = 1 - p$. In this instance, we call X a Bernoulli random variable, and denote the distribution as $B(1, p)$, $p \in [0, 1]$. This random variable has a mean of p and a variance of $p(1 - p)$.

Consider an experiment where we repeat the above Bernoulli trial n times. The probability of a head and tail still remains the same. In such an instance, we may be more concerned

with the number of times we get heads. Let X be the random variable with the outcomes $x \in \{1, 2, ..., n\}$. We say X has a **Binomial Distribution**, and it has a density function of the form,

$$f(x) = \begin{cases} \begin{pmatrix} n \\ x \end{pmatrix} p^x (1-p)^{n-x} & \text{if} \quad x \in \{1, 2, ..., n\} \\ 0 & \text{otherwise} \end{cases}$$

A Binomial random variable has a mean and variance of np and $np(1-p)$.

Of course we would be hard pressed to fit all experiments into the framework of a Binomial Distribution given that it focuses on experiments with only two outcomes. More generally, we can imagine an experiment with possibly m outcomes repeated n times, and where we are concerned with the random variable of the number of times an outcome is obtained for each of the m possible outcomes from a single experiment. Let the probability in any single experiment of observing an outcome i $i \in \{1, 2, ..., m\}$ as p_i. We call such a distribution a **Multinomial Distribution**. Let X_i where $i \in \{1, 2, ..., m\}$ denotes the number of times outcome i is observed, and $x_i \in \{1, 2, ..., n\}$. Then the probability density function for the event that $X_1 = x_1$, $X_2 = x_2$,..., and $X_m = x_m$ is of the form,

$$\frac{n!}{x_1! x_2! ... x_{m-1}! x_m!} p_1^{x_1} ... p_{m-1}^{x_{m-1}} p_m^{x^m}$$

It should be noted that $x_m = n - \sum_{i=1}^{m-1} x_i$.

5.3 Poisson Distribution

Another discrete random variable distribution is that of the **Poisson Distribution**. The reason for its value is that this distribution models certain random variables, whose occurrence can be thought of as an accident, very well, for example the likelihood of defects in manufactured products. Let X be a random variable of observing a number of defects in a single manufactured product, where there are m possible defects. In other words, $x = \{0, 1, 2, ...\}$. Let $\lambda > 0$ be just a parameter. If X is described by a Poisson Process, then the probability density function of X is,

$$f(x) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!} & \text{if} \quad x = 0, 1, 2, ... \\ 0 & \text{otherwise} \end{cases}$$

6 Normal and Related Distributions

We will now describe some of the most common distributions briefly, most of which you will see through out the entire course.

- Normal Distribution A Normal Distribution has a p.d.f. of the following expression,

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

  where $x \in (-\infty, \infty)$, $\mathbf{E}(X) = \mu$ and $\mathbf{V}(X) = \sigma^2$. We say that X has a normal distribution with mean of μ and variance of σ^2, and write it as $X \sim N(\mu, \sigma^2)$. The normal distribution is also call the Gaussian Distribution. What is the cumulative distribution function of X?

- Standard Normal Distribution The standard normal distribution has a mean of 0, and a variance of 1, and we write it as $Z \sim N(0, 1)$. The pdf is

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

  where $z \in (-\infty, \infty)$. We can obtain the standard normal distribution by standardizing a random variable with a normal distribution. Verify for yourself.

  Note:

  1. If $X \sim N(\mu, \sigma^2)$, then $aX + b \sim N(a\mu + b, a^2\sigma^2)$, for constants a and b.
  2. If X and Y are jointly normally distributed, then they are independent if and only if $\text{cov}(X, Y) = 0$.
  3. Any linear combination of normally distributed random variables are also normally distributed.

- Chi-Square (χ^2) Distribution The χ^2 distribution is obtained from independent standard normal random variables. Let $Z_1, Z_2, ...Z_n$ be n standard normal random variables, then $\chi = \sum_{i=1}^{n} Z_i^2$ has χ^2 distribution with n degrees of freedom, and we write it as $\chi \sim \chi_n^2$. The χ^2 is a unimodal distribution skewed to the right, with the skewness decreasing with the degrees of freedom.

- t Distribution The t distribution is obtained from a standard normal and an a χ^2 distribution. Let Z have standard normal distribution, and χ has a χ^2 distribution. Then $T = \frac{Z}{\sqrt{\chi/n}}$ has a t distribution. This distribution is most often used in inferences, and you will see plenty of it. The t distribution is also a unimodal symmetric distribution.

- F Distribution The F distribution is derived from two χ^2 distributions. Let $\chi_1 \sim \chi^2_a$ and $\chi_2 \sim \chi^2_b$ be two random variables with χ^2 distribution with a and b degrees of freedom respectively. The $F = \frac{\chi_1/a}{\chi_2/b}$ has a F distribution, with (a, b) degrees of freedom, and we write it as $F \sim F_{a,b}$

What are the p.d.f.'s for the other distributions besides the normal and standard normal distributions? How about the c.d.f.'s?

7 Moment Generating Functions

Sometimes, it is may be difficult to solve for certain moments (for example the mean is the first moment) of a random variable. However, if the distribution has a **Moment Generating Function**, we can use it to find them far more easily. You can think of a moment generating function as a special type of expectation or a transformation[1].

Let $h > 0$ such that $t \in (-h, h)$, and $\mathbf{E}(e^{tX})$ exists. Then for a continuous random variable X, the moment generating function is,

$$\mathbf{E}(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f(x) dx$$

where $f(x)$ is the density function. For a discrete random variable X, the moment generating function is,

$$\mathbf{E}(e^{tX}) = \sum_{x} e^{tx} f(x)$$

You may occasionally see $e(tX) = M(t)$. You should keep in mind that not all distributions have a moment generating function. Its importance stems not only from the fact that it aids us in finding the mathematical expectations easily, but also that *a moment generating function is unique and completely determines the distribution of a random variable* so that if two random variables have the same moment generating function, by necessity, they must have the same distribution[2]. This section principally introduces the concept, and the use of this technique.

The key properties of a moment generating function are as follows,

[1] Sometimes you will see the Moment Generating function being referred to as a Bilateral Laplace Transform.

[2] The uniqueness of the moment generating function will not be proved in our class, but if through the course, you develop greater interest in Econometric Theory, it is recommended that you take additional courses in Probability Theory and Mathematical Statistics

1. If $M(t)$ exists for $t \in (-h, h)$, it implies that it is continuously differentiable of all order at $t = 0$.

2. Given $M(t)$ exists for $t \in (-h, h)$,

$$\frac{dM(t)}{dt} = M'(t) = \int_{-\infty}^{\infty} xe^{tx} f(x)dx \quad \text{If } X \text{ is a continuous r.v.}$$
$$\frac{dM(t)}{dt} = M'(t) = \sum_{x} xe^{tx} f(x)dx \quad \text{If } X \text{ is a discrete r.v.}$$

   Then note that $M'(0) = \mathbf{E}(X) = \mu$.

3. Given $M(t)$ exists for $t \in (-h, h)$,

$$\frac{d^2M(t)}{dt^2} = M''(t) = \int_{-\infty}^{\infty} x^2 e^{tx} f(x)dx \quad \text{If } X \text{ is a continuous r.v.}$$
$$\frac{d^2M(t)}{dt^2} = M''(t) = \sum_{x} x^2 e^{tx} f(x)dx \quad \text{If } X \text{ is a discrete r.v.}$$

   Then note that $M''(0) = \mathbf{E}(X^2)$, which together with the first moment about zero, $M'(0)$ allows you to solve for the variance.

It should be noted that it is not always true that using the moment generating function to find the moments is easier. Sometimes, directly finding them by using the density function is far simply. There is no general rule to help you decide in the choice of solution.

8 Transformation of Variable

In general, our concern may be for some function of a random variable. Under certain situations, it is sufficient to utilize the knowledge of the original distribution of the random variable, while other times, it may be easier to use the distribution of the new distribution defined through the function.

8.1 Discrete Variable

Let X be a discrete random variable with density function $f(x)$. Let the set of outcomes be denoted by $\mathcal{A}$, where $x \in \mathcal{A}$ and $f(x) > 0$. Let $y = g(x)$ where $y \in \mathcal{B}$ and let the inverse of g^{-1} be the inverse of the function $g(.)$ such that $x = g^{-1}(y) = h(y)$. Then define the new random variable $Y = g(X)$ so that if $y \in \mathcal{B}$ we have $x \in \mathcal{A}$, in other words the outcome

$Y = y$ is equivalent to the outcome $X = h(Y = y)$. The p.d.f. of this new random variable Y is,

$$w(y) = \begin{cases} \Pr(Y = y) = \Pr(X = h(Y = y)) = f(h(y)) & y \in \mathcal{B} \\ 0 & \text{otherwise} \end{cases}$$

More generally, you can think of the random variable X as a vector of random variables, and the same for Y. In that case, you can thinks of $f(.)$ and $w(.)$ as joint density functions, and likewise from the previous discussion, in order for you to obtain the marginal density of each of the random variables in question, it involves summing over the other random variables. It should be noted that in general the technique of transformation of variable involves the introduction of as many new variables as the original.

Consider the following example, let X be have a Binomial Distribution with the set of outcomes $\mathcal{A} = \mathbb{N}^+ + \{0\}$, in other words, $\mathcal{A}$ is the set of nonnegative integers.

$$f(x) = \begin{cases} \frac{n!}{x!(n-x)!} p^x (1-p)^{nx} & x \in \mathcal{A} \\ 0 & \text{otherwise} \end{cases}$$

Let $Y = X^2$, and that we wish the find the p.d.f. $g(y)$. First note that since the support of X is on $\mathbb{N}^+ + \{0\}$, the function is one-to-one transformation of A on the the set $\mathcal{B}$, where $Y \in \mathcal{B}$. Then since $y = x^2$ which implies that $x = \sqrt{y}$. Therefore,

$$g(y) = f(\sqrt{y}) = \begin{cases} \frac{n!}{\sqrt{y}!(n-\sqrt{y})!} p^{\sqrt{y}} (1-p)^{n-\sqrt{y}} & y \in \mathcal{B} \\ 0 & \text{otherwise} \end{cases}$$

8.2 Continuous Variable

The method has an analogue for continuous random variable. Let X be a continuous random variable with p.d.f. $f(x)$, and let $\mathcal{A}$ be the one dimensional space within which $f(x) > 0$. Let the random variable $Y = h(X)$ so that $y = h(x)$ is a one-to-one transformation that maps $\mathcal{A}$ onto $\mathcal{B}$ where $y \in \mathcal{B}$. Let the inverse of $h(.)$ be $w(.)$, and let it be twice continuously differentiable, so that $\frac{dx}{dy} = \frac{dw(y)}{dy} = w'(y)$. Then the p.d.f. of Y is,

$$g(y) = \begin{cases} f(w(y))|w'(y)| & y \in \mathcal{B} \\ 0 & \text{otherwise} \end{cases}$$

Generally, $w'(y)$ is known as the Jacobian of the inverse of the transformation. As in the discrete variable case, you can think of the above as vectors of variables, in which case, the Jacobian is a matrix of partial derivatives, and the p.d.f. is the product of the original p.d.f. as a function of the new random variable and the absolute value of the determinant of the Jacobian.

For example, let X_1 and X_2 be two continuous random variable with a joint p.d.f. $f(x_1, x_2)$. Let the set of outcomes for $\{x_1, x_2\}$ be in $\mathcal{A}$. Let $Y_1 = h_1(X_1, X_2)$ and $Y_2 = h_2(X_1, X_2)$ where the set of outcomes for $\{y_1, y_2\}$ is in $\mathcal{B}$, and the inverse gives $X_1 = w_1(Y_1, Y_2)$ and $X_2 = w_2(Y_1, Y_2)$. The Jacobian is,

$$J = \begin{bmatrix} \frac{\partial w_1(y_1,y_2)}{\partial y_1} & \frac{\partial w_1(y_1,y_2)}{\partial y_2} \\ \frac{\partial w_2(y_1,y_2)}{\partial y_1} & \frac{\partial w_2(y_1,y_2)}{\partial y_2} \end{bmatrix}$$

and the joint density function of Y_1 and Y_2 is,

$$g(y_1, y_2) = \begin{cases} f(w_1(y_1, y_2), w_2(y_1, y_2))|J| & \{y_1, y_2\} \in \mathcal{B} \\ 0 & \text{otherwise} \end{cases}$$