CHAPTER 3. PROBABILITY THEORY IN A NUTSHELL

1. Sample Space

The starting point for probability theory is the concept of a *state of Nature*, which is a description of everything that has happened and will happen in the universe. In particular, this description includes the outcomes of all probability and sampling experiments. The set of all possible states of Nature is called the *sample space*. Let s denote a state of Nature, and $\mathbb{S}$ the sample space. These are abstract objects that play a conceptual rather than a practical role in the development of probability theory. Consequently, there can be considerable flexibility in thinking about what goes into the description of a state of Nature and into the specification of the sample space; the only critical restriction is that there be enough states of Nature so that distinct observations are always associated with distinct states of Nature. In elementary probability theory, it is often convenient to think of the states of Nature as corresponding to the outcomes of a particular experiment, such as flipping coins or tossing dice, and to suppress the description of everything else in the universe.

2. Event Fields and Information

2.1. An *event* is a set of states of Nature with the property that one can in principle determine whether the event occurs or not. If states of Nature describe all happenings, including the outcome of a particular coin toss, then one event might be the set of states of Nature in which this coin toss comes up heads. The family of potentially observable events is denoted by $\mathfrak{F}$. This family is assumed to have the

following properties:

(i) The "anything can happen" event $\mathbb{S}$ is in $\mathfrak{F}$.

(ii) If event $\mathbb{A}$ is in $\mathfrak{F}$, then "not $\mathbb{A}$" (denoted $\mathbb{A}^c$) is in $\mathfrak{F}$.

(iii) If $\mathbb{A}$ and $\mathbb{B}$ are events in $\mathfrak{F}$, then the event "both $\mathbb{A}$ and $\mathbb{B}$" (denoted $\mathbb{A}\cap\mathbb{B}$) is in $\mathfrak{F}$.

(iv) If $\mathbb{A}_1,\mathbb{A}_2,...$ is a finite or countable sequence of events in $\mathfrak{F}$, then the event "one or more of $\mathbb{A}_1$ or $\mathbb{A}_2$ or ..." (denoted $\bigcup_{i=1}^{\infty} \mathbb{A}_i$) is in $\mathfrak{F}$.

A family $\mathfrak{F}$ with these properties is called a σ-*field* (or *Boolean* σ-*algebra*) of subsets of $\mathbb{S}$. The pair $(\mathbb{S},\mathfrak{F})$ consisting of an abstract set $\mathbb{S}$ and a σ-field of subsets of $\mathbb{S}$ is called a *measurable space*, and the sets in $\mathfrak{F}$ are called the *measurable* subsets of $\mathbb{S}$. Two useful implications of the definition of a σ-field are

(1) If $\mathbb{A}_1,\mathbb{A}_2,...$ is a finite or countable sequence of events in $\mathfrak{F}$, then $\bigcap_{i=1}^{\infty} \mathbb{A}_i$ is also in $\mathfrak{F}$.

(2) If $\mathbb{A}_1,\mathbb{A}_2,...$ is a countable sequence of events in $\mathfrak{F}$ that is *monotone* (i.e., $\mathbb{A}_1 \subseteq \mathbb{A}_2 \subseteq ...$, then $\mathbb{A}_0 = \lim \mathbb{A}_i$ is also in $\mathfrak{F}$.

We will use a few concrete examples of sample spaces and σ-fields:

Example 1. [Two coin tosses] A coin is tossed twice, and for each toss a head or tail appears. Let HT denote the state of Nature in which the first toss yields a head and the second toss yields a tail. Then $\mathbb{S}$ = {HH,HT,TH,TT}. Let $\mathfrak{F}$ be the class of all possible subsets of $\mathbb{S}$; $\mathfrak{F}$ has 2^4 members.

Example 2. [Coin toss until a tail] A coin is tossed until a tail appears. The sample space is $\mathbb{S}$ = {T, HT, HHT, HHHT,...}. In this example, the sample space is infinite, but countable. Let $\mathfrak{F}$ be the σ-field generated by the finite subsets of $\mathbb{S}$. This σ-field contains events such as "Ten or more tosses without a tail", and "an even number of heads before a tail". A set that is not in $\mathfrak{F}$ will have the property that both the set and its complement are infinite. It is difficult to describe such a set, primarily because the language that we normally use to construct sets tends to correspond to elements in the σ-field. However, mathematical analysis shows that such sets exist.

Example 3. [Daily change in S&P stock index] The stock index change is a number in the real line $\mathbb{R}$, so $\mathbb{S} \equiv \mathbb{R}$. Take the σ-field of events to be the *Borel σ-field* $\mathfrak{B}$, which is defined as the smallest family of subsets of the real line that contains all the open intervals and satisfies the properties (i)-(iv) of a σ-field. The subset of $\mathbb{R}$ that are not in $\mathfrak{B}$ are said to be nonmeasurable.

Example 4. [Changes in S&P stock index on successive days] The set of states of Nature is the Cartesian product of the set of changes on day one and the set of changes on day 2, $\mathbb{S} = \mathbb{R}\times\mathbb{R}$ (also denoted $\mathbb{R}^2$). Take the σ-field of events to be the product of the one-dimensional σ-fields, $\mathfrak{F} = \mathfrak{B}_1 \otimes \mathfrak{B}_2$, where "⊗" denotes an operation that forms the smallest σ-field containing all sets of the form $\mathbb{A}\times\mathbb{C}$ with $\mathbb{A} \in \mathfrak{B}_1$ and $\mathbb{C} \in \mathfrak{B}_2$. In this example, $\mathfrak{B}_1$ and $\mathfrak{B}_2$ are identical copies of the Borel σ-field on the real line. Examples of events in $\mathfrak{F}$ are "an increase on day one", "increases on both days", and "a larger change the second day than the first day". (The operation "⊗" is different than the cartesian product "×", where $\mathfrak{B}_1\times\mathfrak{B}_2$ is the family of all

rectangles $\mathbb{A}\times\mathbb{C}$ formed from $\mathbb{A} \in \mathfrak{B}_1$ and $\mathbb{C} \in \mathfrak{B}_2$. This family is not itself a σ-field, and the σ-field that it generates is $\mathfrak{B}_1 \otimes \mathfrak{B}_2$.)

In the first example, the σ-field consisted of all possible subsets of the sample space. This was not the case in the last two examples, because the Borel σ-field does not contain all subsets of the real line. There are two reasons to introduce the complication of dealing with σ-fields that do not contain all the subsets of the sample space, one substantive and one technical. The substantive reason is that the σ-field can be interpreted as the potential information that is available by observation. If an observer is incapable of making observations that distinguish two states of Nature, then the σ-field cannot contain sets that include one of these states and excludes the other. Then, the specification of the σ-field will depend on what is observable in an application. The technical reason is that when the sample space contains an infinite number of states, it may be mathematically impossible to define probabilities on all subsets of the sample space that have the properties one would like probabilities to have. Restricting the definition of probabilities to appropriately chosen σ-fields solves this problem.

2.2. It is possible that more than one σ-field of subsets is defined for a particular sample space $\mathbb{S}$. If $\mathfrak{A}$ is an arbitrary collection of subsets of $\mathbb{S}$, then the smallest σ-field that contains $\mathfrak{A}$ is said to be the σ-field *generated* by $\mathfrak{A}$. If $\mathfrak{F}$ and $\mathfrak{G}$ are both σ-fields, and $\mathfrak{G} \subseteq \mathfrak{F}$, then $\mathfrak{G}$ is said to be a *sub-field* of $\mathfrak{F}$, and $\mathfrak{F}$ is said to *contain more information* or *refine* $\mathfrak{G}$. It is possible that neither $\mathfrak{F} \subseteq \mathfrak{G}$ nor $\mathfrak{G} \subseteq \mathfrak{F}$. However, there is always a smallest σ-field that refines both $\mathfrak{F}$ and $\mathfrak{G}$, which is simply the σ-field generated by the sets in the union of $\mathfrak{F}$ and $\mathfrak{G}$.

Example 1. (continued) Let $\mathfrak{F}$ denote the σ-field of all subsets of $\mathbb{S}$. Another σ-field is $\mathfrak{G} = \{\varnothing,\mathbb{S},\{HT,HH\},\{TT,TH\}\}$, containing all the events in which information is available only on the outcome of the first coin toss. Obviously, $\mathfrak{F}$ contains more information than $\mathfrak{G}$.

Example 3. (continued) Let $\mathfrak{F}$ denote the Borel σ-field. Let $\mathbb{P}$ denote the positive real line, and $\mathbb{N}$ the negative real line. Then $\mathfrak{G} = \{\varnothing,\mathbb{S},\mathbb{P},\mathbb{P}^c\}$ and $\mathfrak{D} = \{\varnothing,\mathbb{S},\mathbb{N},\mathbb{N}^c\}$ are both σ-fields, the first corresponding to the ability to observe whether price increases or not, the second corresponding to the ability to tell whether price decreases or not. Neither contains the other, both are contained in $\mathfrak{F}$, and the two have a smallest mutual refinement which is $\mathfrak{E} = \{\varnothing,\mathbb{S},\mathbb{P},\mathbb{N},\mathbb{P}^c,\mathbb{N}^c,\{0\}\}$; this corresponds to the ability to tell whether price is increasing, decreasing, or unchanged.

3. Probability

3.1. Given a sample space $\mathbb{S}$ and σ-field of subsets $\mathfrak{F}$, a *probability* (or *probability measure*) is defined as a function P from $\mathfrak{F}$ into the real line with the following properties:

(i) $P(\mathbb{A}) \geq 0$ for all $\mathbb{A} \in \mathfrak{F}$.

(ii) $P(\mathbb{S}) = 1$.

(iii) [Countable Additivity] If $\mathbb{A}_1, \mathbb{A}_2,\ldots$ is a finite or countable sequence of events in $\mathfrak{F}$ that are mutually exclusive (i.e., $\mathbb{A}_i \cap \mathbb{A}_j = \varnothing$ for all $i \neq j$), then $P(\bigcup_{i=1}^{\infty} \mathbb{A}_i) = \sum_{i=1}^{\infty} P(\mathbb{A}_i)$.

With conditions (i)-(iii), P has the following additional intuitive properties of

a probability when $\mathbb{A}$ and $\mathbb{B}$ are events in $\mathfrak{F}$:

(1) $P(\mathbb{A}) + P(\mathbb{A}^c) = 1$.

(2) $P(\mathbb{A}\cup\mathbb{B}) = P(\mathbb{A}) + P(\mathbb{B}) - P(\mathbb{A}\cap\mathbb{B})$.

(3) $P(\mathbb{A}) \geq P(\mathbb{B})$ when $\mathbb{B} \subseteq \mathbb{A}$.

(4) If a sequence $\mathbb{A}_i$ in $\mathfrak{F}$ approaches $\varnothing$ (in the sense that $\bigcap_{i=1}^{n} \mathbb{A}_i \to \varnothing$), then $P(\mathbb{A}_i) \to 0$.

(5) If $\mathbb{A}_i \in \mathfrak{F}$, not necessarily disjoint, then $P(\bigcup_{i=1}^{\infty} \mathbb{A}_i) \leq \sum_{i=1}^{\infty} P(\mathbb{A}_i)$.

The triplet $(\mathbb{S},\mathfrak{F},P)$ consisting of a measurable space $(\mathbb{S},\mathfrak{F})$ and a probability measure P is called a *probability space*.

3.2. If $\mathbb{A} \in \mathfrak{F}$ has $P(\mathbb{A}) = 1$, then $\mathbb{A}$ is said to occur *almost surely* (a.s.), or *with probability one* (w.p.1). If $\mathbb{A} \in \mathfrak{F}$ has $P(\mathbb{A}) = 0$, then $\mathbb{A}$ is said to occur with *probability zero* (w.p.0). Finite or countable intersections of events that occur almost surely again occur almost surely, and finite or countable unions of events that occur with probability zero again occur with probability zero.

Example 1. (continued) If the coin is fair so that heads and tails are equally likely, then each possible outcome HH,HT,TH,TT occurs with probability 1/4. The probability that the first coin is heads is the probability of the event {HH,HT}, which by countable additivity is P({HH,HT}) = P({HH}) + P({HT}) = 1/2.

Example 2. (continued) If the coin is fair, then the probability of k-1 heads followed by a tail is $1/2^k$. Then, the probability of "Ten or more heads" is $1/2^{10}$,

and the probability of "an even number of heads" is 2/3.

Example 3. (continued) Consider the function P defined on open sets (s,∞) by $P((s,\infty)) = 1/(1+e^{S})$. This function maps into the unit interval, and is nondecreasing as the length of the interval increases. It is then easy to show that P satisfies properties (i)-(iii) of a probability on the restricted family of open intervals, and a little work to show that when a probability is determined on this family of open intervals, then it is uniquely determined on the σ-field generated by these intervals. Each single point, such as {0}, is in $\mathfrak{F}$. Taking intervals that shrink to this point, each single point occurs with probability zero. Then, a countable set of points occurs w.p.0.

3.3. Often a measurable space $(\mathbb{S},\mathfrak{F})$ will have an associated *measure* ν that is a countably additive function from $\mathfrak{F}$ into the nonnegative real line; i.e., $\nu(\bigcup_{i=1}^{\infty} \mathbb{A}_i) = \sum_{i=1}^{\infty} \nu(\mathbb{A}_i)$ for any sequence of disjoint $\mathbb{A}_i \in \mathfrak{F}$. The measure is *positive* if $\nu(\mathbb{A}) \geq 0$ for all $\mathbb{A} \in \mathfrak{F}$; otherwise, it is *signed* and can be written $\nu(\mathbb{A}) = \nu^{+}(\mathbb{A}) - \nu^{-}(\mathbb{A})$, where ν^{+} and ν^{-} are both positive. The measure ν is *finite* if $|\nu(\mathbb{A})| \leq M$ for some constant M and all $\mathbb{A} \in \mathfrak{F}$, and σ-*finite* if there exists a countable sequence $\mathbb{A}_i \in \mathfrak{F}$ with $\nu^{+}(\mathbb{A}_i) + \nu^{-}(\mathbb{A}_i) < +\infty$ and $\bigcup_{i=1}^{\infty} \mathbb{A}_i = \mathbb{S}$. The measure ν may be a probability, but more commonly it is a measure of "length" or "volumn". For example, it is common when the sample space $\mathbb{S}$ is the countable set of positive integers to define ν to be *counting measure* with $\nu(\mathbb{A})$ equal to the number of points in $\mathbb{A}$. When the sample space $\mathbb{S}$ is the

real line, it is common to define ν to be *Lebesgue measure*, with $\nu((a,b)) = b - a$ for any open interval (a,b). Both of these examples are positive σ-finite measures.

A finite measure P on $(S,\mathfrak{F})$ is *absolutely continuous* with respect to a measure ν if $\mathbb{A} \in \mathfrak{F}$ with $\nu(\mathbb{A}) = 0$ implies $P(\mathbb{A}) = 0$. A fundamental result from analysis is the Radon-Nikodym theorem:

> If a finite measure P is absolutely continuous with respect to a positive σ-finite measure ν, then there exists a unique real-valued function p on $\mathbb{S}$ that is integrable, with $\int_{\mathbb{A}} p(s)\nu(ds) = P(\mathbb{A})$ for each $\mathbb{A} \in \mathfrak{F}$.

In the statement of this theorem, the symbol $\int_{\mathbb{A}} p(s)\nu(ds)$ denotes integration of p with respect to the measure ν over the set $\mathbb{A}$, and is defined to be the common limit as $n \to \infty$ of sums of the form $\sum_{k=-n^2}^{n^2} \frac{k+r}{n}\cdot\nu(\mathbb{C}_{kn})$, where $0 \le r \le 1$ and $\mathbb{C}_{kn}$ is the set of states of Nature in $\mathbb{A}$ that yield p(s) in the interval $(k/n^2,(k+1)/n^2)$. The requirement that p be integrable means is first that p is measurable, so that the sets $\mathbb{C}_{kn}$ are in $\mathfrak{F}$, and second that the common limit above exists, with $\lim \sum_{k=-n^2}^{n^2} |\frac{k+r}{n}|\cdot\nu(\mathbb{C}_{kn}) < +\infty$. In general, the measure ν can have point masses, or continuous measure, or both, so that the notation for integration with respect to ν includes sums and mixed cases. The integral $\int_{\mathbb{A}} p(s)\nu(ds)$ will sometimes be denoted $\int_{\mathbb{A}} p(s)d\nu$, and in the case of Lebesgue measure, $\int_{\mathbb{A}} p(s)ds$.

When P is a probability, the function p given by the theorem is nonnegative, and is called the *density*. The Radon-Nikodym result is often very useful in theoretical derivations, for example in the theory of choice under uncertainty. In basic econometrics, we will constantly characterize probabilities both in terms of the probability measure (or distribution) and the density. We will not need to refer explicitly to the Radon-Nikodym theorem. However, it may be helpful to remember that there is this fundamental mathematical result that makes the connection between probabilities and densities.

3.4. A probability that appears frequently in statistics is the *normal*, which is defined on $(\mathbb{R},\mathfrak{B})$, where $\mathbb{R}$ is the real line and $\mathfrak{B}$ the Borel σ-field, by the density $n(s\text{-}\mu,\sigma) \equiv \frac{1}{\sqrt{2\pi\sigma^2}}\cdot e^{-(s-\mu)^2/2\sigma^2}$, so that $P(\mathbb{A}) = \int_{\mathbb{A}} \frac{1}{\sqrt{2\pi\sigma^2}}\cdot e^{-(s-\mu)^2/2\sigma^2} ds$. In this probability, μ and σ are parameters that are interpreted as determining the *location* and *scale* of the probability, respectively. When $\mu = 0$ and $\sigma = 1$, this probability is called the *standard normal*.

Example 3. (continued) Given $P((s,\infty)) = 1/(1+e^s)$, one can use the differentiability of the function in s to argue that it is absolutely continuous with respect to Lebesgue measure on the line. Then, one can verify by integration that the density implied by the Radon-Nikodym theorem is $p(s) = e^s/(1+e^s)^2$.

3.5. Consider a probability space $(\mathbb{S},\mathfrak{F},P)$, and a σ-field $\mathfrak{G} \subseteq \mathfrak{F}$. If the event $\mathbb{B} \in \mathfrak{G}$ has $P(\mathbb{B}) > 0$, then the *conditional probability* of $\mathbb{A}$ given $\mathbb{B}$ is defined as $P(\mathbb{A}|\mathbb{B}) = P(\mathbb{A}\cap\mathbb{B})/P(\mathbb{B})$. Stated another way, $P(\mathbb{A}|\mathbb{B})$ is a real-valued function on $\mathfrak{F}\times\mathfrak{G}$

with the property that $P(\mathbb{A}\cap\mathbb{B}) = P(\mathbb{A}|\mathbb{B})P(\mathbb{B})$ for all $\mathbb{A} \in \mathfrak{F}$ and $\mathbb{B} \in \mathfrak{G}$. The concept of conditional probability can be extended to cases where $P(\mathbb{B}) = 0$ by defining $P(\mathbb{A}|\mathbb{B})$ as the limit of $P(\mathbb{A}|\mathbb{B}_i)$ for sequences $\mathbb{B}_i \in \mathfrak{G}$ that satisfy $P(\mathbb{B}_i) > 0$ and $\mathbb{B}_i \to \mathbb{B}$, provided the limit exists.

The idea behind conditional probabilities is that one has partial information of what the state of Nature may be, and one wants to calculate the probability of events using this partial information. One way to represent partial information is in terms of a subfield; e.g., $\mathfrak{F}$ is the field of events which distinguish outcomes in both the past and the future, and a subfield $\mathfrak{G}$ contains events which distinguish only past outcomes.

A conditional probability $P(\mathbb{A}|\mathbb{B})$ defined for $\mathbb{B} \subseteq \mathfrak{C}$ can be interpreted as a function from $\mathfrak{C}$ into [0,1]. To emphasize this, conditional probabilities are sometimes written $P(\mathbb{A}|\mathfrak{C})$, and $\mathfrak{C}$ is termed the *information set*, or a family of events with the property that you know whether or not they happened at the time you are forming the conditional probability.

Example 1. (continued) If $\mathfrak{G} = \{\varnothing,\{HH,HT\},\{TH,TT\},\mathbb{S}\}$, so that events in $\mathfrak{G}$ describe the outcome of the first coin toss, Then $P(HH|\{HH,HT\}) = \frac{P(HH)}{P(\{HH,HT\})} = 1/2$ is the probability of heads on the second toss, given heads on the first toss. In this example, the conditional probability of a head on the second toss equals the unconditional probability of this event. In this case, the outcome of the first coin toss provides no information on the probabilities of heads from the second coin, and the two tosses are said to be statistically independent. For a second case, take $\mathfrak{C} = \{\varnothing,\{HT,TH\},\{HH\},\{TT\},\{HH\}^c,\{TT\}^c,\mathbb{S}\}$, so that events in $\mathbb{C}$ describe the number of heads that occur in two tosses. Then, the conditional probability of heads on the

first toss, given at least one head, is $P(\{HT,HH\}\,|\,\{TT\}^c) = \frac{P(HT)+P(HH)}{P(HT)+P(HH)+P(TH)} = 2/3$. Then, the conditional probability of heads on the first toss given at least one head is not equal to the unconditional probability of heads on the first toss.

Example 2. (continued) If $\mathfrak{G} = \{\varnothing,\mathbb{P},\mathbb{P}^c,\mathbb{S}\}$ is the σ-field corresponding to the event that the price change is positive or not, then $P([-1,1]) = \frac{e-1}{e+1}$, $P((0,1]) = \frac{e-1}{2(e+1)}$, $P(\mathbb{P}) = 1/2$, and $P([-1,1])\,|\,\mathbb{P}) = \frac{e-1}{e+1}$. Here, the conditional and unconditional probability coincide, so that knowledge of the sign of the price change provides no information on the probability that the magnitude of the change does not exceed one.

4. Statistical Independence and Repeated Trials

4.1. Consider a probability space $(\mathbb{S},\mathfrak{F},P)$. Events $\mathbb{A}$ and $\mathbb{C}$ in $\mathfrak{F}$ are *statistically independent* if $P(\mathbb{A}\cap\mathbb{C}) = P(\mathbb{A})\cdot P(\mathbb{C})$. From the definition of conditional probability, if $\mathbb{A}$ and $\mathbb{C}$ are statistically independent and $P(\mathbb{A}) > 0$, then $P(\mathbb{C}\,|\,\mathbb{A}) = P(\mathbb{A}\cap\mathbb{C})/P(\mathbb{A}) = P(\mathbb{C})$. Thus, when $\mathbb{A}$ and $\mathbb{C}$ are statistically independent, knowing that $\mathbb{A}$ occurs is unhelpful in calculating the probability that $\mathbb{C}$ occurs. The idea of statistical independence of events has an exact analogue in a concept of statistical independence of subfields. Let $\mathfrak{A} = \{\varnothing,\mathbb{A},\mathbb{A}^c,\mathbb{S}\}$ and $\mathfrak{C} = \{\varnothing,\mathbb{C},\mathbb{C}^c,\mathbb{S}\}$. Verify as an exercise that if $\mathbb{A}$ and $\mathbb{C}$ are statistically independent, then so are any pair of events $\mathbb{A}' \in \mathfrak{A}$ and $\mathbb{C}' \in \mathfrak{C}$. Then, one can say that the subfields $\mathfrak{A}$ and $\mathfrak{C}$ are statistically independent. One can extend this idea and talk about statistical independence between pairs of subfields. The idea of statistical independence can also be extended to that of *mutual statistical independence* (MSI) among a family of events (or a family of

subfields). Let **N** denote an index set, which may be finite, countable, or non-countable. Let $\mathbf{N}_i$ denote the set **N**, excluding element i. Let $\mathfrak{F}_i$ denote a σ-subfield of $\mathfrak{F}$ ($\mathfrak{F}_i \subseteq \mathfrak{F}$) for each $i \in \mathbf{N}$. Then, MSI has the following definition:

$\{\mathfrak{F}_i | i \in \mathbf{N}\}$ are MSI if and only if, for all finite $\mathbf{K} \subseteq \mathbf{N}$ and $\mathbb{A}_j \in \mathfrak{F}_j$ for $j \in \mathbf{N}$,

one has $P(\bigcap_{j\in\mathbf{K}} \mathbb{A}_j) = \prod_{j\in\mathbf{K}} P(\mathbb{A}_j)$.

As in the case of statistical independence between two events (subfields), the concept of MSI can be stated in terms of conditional probabilities:

$\{\mathfrak{F}_i | i \in \mathbf{N}\}$ are mutually statistically independent (MSI) if, for all $i \in \mathbf{N}$, finite $\mathbf{K} \subseteq \mathbf{N}_i$ and $\mathbb{A}_j \in \mathfrak{F}_j$ for $j = i$ or $j \in \mathbf{K}$, one has $P(\mathbb{A}_i | \bigcap_{j\in\mathbf{K}} \mathbb{A}_j) = P(\mathbb{A}_i)$.

Example 1. (continued) Let $\mathbb{A}$ denote the event of a head for the first coin, $\mathbb{C}$ denote the event of a head for the second coin, $\mathbb{D}$ the event of a match, $\mathbb{G}$ the event of two heads. Verify that $\mathbb{A}$ and $\mathbb{C}$ are statistically independent and that $\mathbb{A}$ and $\mathbb{D}$ are statistically independent. Verify that $\mathbb{A}$, $\mathbb{C}$, and $\mathbb{D}$ are <u>not</u> MSI. Verify that $\mathbb{A}$ and $\mathbb{G}$ are not statistically independent.

Example 4. (continued) Recall that $\mathbb{S} = \mathbb{R}^2$ with $\mathfrak{F} = B \otimes B$, the *product* Borel σ-field. Define the subfields $\mathfrak{F}_1 = \{\mathbf{A}\times\mathbb{R} \,|\, \mathbf{A} \in B\}$, $\mathfrak{F}_2 = \{\mathbb{R}\times\mathbf{A} \,|\, \mathbf{A} \in B\}$ containing information on price changes on the first and second day, respectively. Define $\mathfrak{C} = \left\{\varnothing, \mathbb{S}, \mathbb{P}, \mathbb{N}, \mathbb{P}^c, \mathbb{N}^c, \{0\}\right\}$, the subfield of $\mathfrak{B}$ containing information on whether a price change is positive, negative, or zero. Define $\mathfrak{F}_3$ to be the σ-subfield of $\mathfrak{B}\otimes\mathfrak{B}$ generated by sets of the form $\mathbf{A}_1\times\mathbf{A}_2$ with $\mathbf{A}_1 \in \mathfrak{C}$ and $\mathbf{A}_2 \in B$. Suppose P is uniform on [-1,1]×[-1,1]. Then $\{\mathfrak{F}_1, \mathfrak{F}_2\}$ are MSI. However, $\{\mathfrak{F}_1, \mathfrak{F}_3\}$ are not independent.

Example 5. Consider $\mathbb{S}$ = {0, 1, 2, 3, 4, 5, 6, 7}, with $\mathfrak{F}$ equal to all subsets of $\mathbb{S}$. Define the subfield $\mathfrak{F}_1$ = {∅,0123,4567,$\mathbb{S}$}, where 0123 denotes {0,1,2,3}, etc. Also define the subfields $\mathfrak{F}_2$ = {∅,2345,0167,$\mathbb{S}$}, $\mathfrak{F}_3$ = {∅,0246,1357,$\mathbb{S}$},

$$\mathfrak{F}_4 = \{\varnothing,01,23,4567,0123,234567,014567,\mathbb{S}\},$$

$\mathfrak{F}_5$ = {∅,01,23,45,67,0123,0145,0167,2345,2367,4567,012345,012367,014567,234567,$\mathbb{S}$}, and

$\mathfrak{F}_6$ = {∅,06,17,24,35,0167,0246,0356,1247,1357,2345,123457,023456,013567,012467,$\mathbb{S}$}.

The field $\mathfrak{F}_4$ is a *refinement* of the field $\mathfrak{F}_1$ (i.e., $\mathfrak{F}_1 \subseteq \mathfrak{F}_4$), and can be said to contain more information than $\mathfrak{F}_1$. The field $\mathfrak{F}_5$ is a *mutual refinement* of $\mathfrak{F}_1$ and $\mathfrak{F}_2$ (i.e., $\mathfrak{F}_1 \cup \mathfrak{F}_2 \subseteq \mathfrak{F}_5$), and is in fact the smallest mutual refinement. It contains all the information available in either $\mathfrak{F}_1$ or $\mathfrak{F}_2$. Similarly, $\mathfrak{F}_6$ is a mutual refinement of $\mathfrak{F}_2$ and $\mathfrak{F}_3$. The intersection of $\mathfrak{F}_5$ and $\mathfrak{F}_6$ is the field $\mathfrak{F}_2$; it is the common information available in $\mathfrak{F}_5$ and $\mathfrak{F}_6$. If, for example, $\mathfrak{F}_5$ characterized the information available to one economic agent, and $\mathfrak{F}_6$ characterized the information available to a second agent, then $\mathfrak{F}_2$ would characterize the common information upon which they could make contingent contracts. Suppose P(i) = 1/8. Then $\{\mathfrak{F}_1, \mathfrak{F}_2, \mathfrak{F}_3\}$ are MSI. E.g., P(0123|2345) = P(0123|0246) = P(0123|2345∩0246) = P(0123) = 1/2. However, $\{\mathfrak{F}_1, \mathfrak{F}_4\}$ are not independent; e.g., 1 = P(0123|01) ≠ P(0123) = 1/2.

For $\mathbf{M} \subseteq \mathbf{N}$, let $\mathfrak{F}_M$ denote the smallest field containing $\mathfrak{F}_i$ for all $i \in \mathbf{M}$. Then MSI satisfies the following theorem:

If $\{\mathfrak{F}_i | i \in \mathbf{N}\}$ are MSI, and $i \notin \mathbf{M} \subseteq \mathbf{N}$, then $\{\mathfrak{F}_i, \mathfrak{F}_M\}$ are MSI. Further, $\{\mathfrak{F}_i | i \in \mathbf{N}\}$ are MSI if and only if, for all $i \in \mathbf{N}$, $\{\mathfrak{F}_i, \mathfrak{F}_{\mathbf{N}_i}\}$ are MSI.

Example 5. (continued) If **M** = {2,3}, then $\mathfrak{F}_{\mathbf{M}} \equiv \mathfrak{F}_6$, and P(0123|**A**) = 1/2 for each **A** $\in \mathfrak{F}_{23}$.

4.2. The idea of *repeated trials* is that an experiment, such as a coin toss, is replicated over and over. It is convenient to have common probability space in which to describe the outcomes of larger and larger experiments with more and more replications. The notation for repeated trials will be similar to that introduced in the definition of mutual statistical independence. Let **N** denote a finite or countable index set of trials, $\mathbb{S}_i$ a sample space for trial i, and $\mathfrak{F}_i$ a σ-field of subsets of $\mathbb{S}_i$. Note that $\mathbb{S}_i$ may be the same for all i. Assume that ($\mathbb{S}_i$, $\mathfrak{F}_i$) is the real line with the Borel field, or a countable set with the field of all subsets, or a pair with comparable mathematical properties (i.e., $\mathbb{S}_i$ is a complete separable metric space and $\mathfrak{F}_i$ is its Borel field). Let t = $(s_1,s_2,...)$ = $(s_i : i\in\mathbf{N})$ denote an ordered sequence of outcomes of trials, and $\mathbb{S}_N = \bigtimes_i \mathbb{S}_i$ denote the sample space of these sequences. Let $\mathfrak{F}_{\mathbf{N}} = \prod_{i\in\mathbf{N}} \mathfrak{F}_i$ denote the σ-field of subsets of $\mathbb{S}_{\mathbf{N}}$ generated by *finite rectangles*, where by a finite rectangle we mean a set of the form $(\bigtimes_{i\in\mathbf{K}} \mathbf{A}_i) \times (\bigtimes_{i\in\mathbf{K}^c} \mathbb{S}_i)$, where **K** is a finite subset of **N** and $\mathbf{A}_i \in \mathfrak{F}_i$ for i ∈ **K**.

Example 6. **N** = {1,2,3}, $\mathbb{S}_i$ = {0,1}, $\mathfrak{F}_i$ = {∅,0,1,01}, where {0} = 0, {0, 1} = 01, etc. Then $\mathbb{S}_{\mathbf{N}}$ = $\{s_1s_2s_3 | s_i \in \mathbb{S}_i\}$ = {000, 001, 010, 011, 100, 101, 110, 111} and $\mathfrak{F}_{\mathbf{N}}$ is the family of all subsets of $\mathbb{S}_{\mathbf{N}}$.

For any subset **K** of **N**, define $\mathbb{S}_{\mathbf{K}} = \bigtimes_{i\in\mathbf{K}} \mathbb{S}_i$ and $\mathfrak{F}_{\mathbf{K}} = \prod_{i\in\mathbf{K}} \mathfrak{F}_i$. Define $\mathfrak{T}_{\mathbf{K}}$ to be the family of subsets of $\mathbb{S}_{\mathbf{N}}$ of the form $\mathbf{A}\times\mathbb{S}_{\mathbf{M}}$ for $\mathbf{A} \in \mathfrak{F}_{\mathbf{K}}$ and **M** the set of indices in **N** but not in **K**. Then $\mathfrak{T}_{\mathbf{K}}$ is a subfield of $\mathfrak{F}_{\mathbf{N}}$. Suppose $P_{\mathbf{N}}$ is a probability on $(\mathbb{S}_{\mathbf{N}}, \mathfrak{F}_{\mathbf{N}})$. The *restriction* of $P_{\mathbf{N}}$ to $(\mathbb{S}_{\mathbf{K}}, \mathfrak{F}_{\mathbf{K}})$ is a probability $P_{\mathbf{K}}$ defined for for $\mathbf{A} \in \mathfrak{F}_{\mathbf{K}}$ by $P_{\mathbf{K}}(\mathbf{A}) = P_{\mathbf{N}}(\mathbf{A}\times\mathbb{S}_{\mathbf{M}})$, where **K** and **M** partition **N**. The following result establishes a link between different restrictions:

If $\mathbf{M} \subseteq \mathbf{K}$ and $P_{\mathbf{M}}$, $P_{\mathbf{K}}$ are restrictions of $P_{\mathbf{N}}$, then (letting **M** and **L** partition **K**) $P_{\mathbf{M}}$ and $P_{\mathbf{K}}$ satisfy the *compatability condition* that

$$P_{\mathbf{M}}(\mathbf{A}) = P_{\mathbf{K}}\,(\mathbf{A} \times \mathbb{S}_{\mathbf{L}}) \text{ for all } \mathbf{A} \in \mathfrak{F}_{\mathbf{M}}.$$

There is then a fundamental result that establishes that when probabilities are defined on all finite sequences of trials and are compatible, then there exists a probability defined on the infinite sequence of trials that yields each of the finite sequence probabilites as a restriction.

If $P_{\mathbf{K}}$ on $(\mathbb{S}_{\mathbf{K}},\mathfrak{F}_{\mathbf{K}})$ for all finite **K** satisfy the compatibility condition, then there exists a unique $P_{\mathbf{N}}$ on $(\mathbb{S}_{\mathbf{N}},\mathfrak{F}_{\mathbf{N}})$ such that each $P_{\mathbf{K}}$ is a restriction of $P_{\mathbf{N}}$.

This result guarantees that it makes sense to make probability statements about events such as "the probability of an infinite number of heads in repeated coin tosses is one" or "the frequency of heads approaches 1/2 with probability one as the number of repetitions increases to infinity".

One can apply the concept of mutual statistical independence in the case of a countable sequence of trials. Suppose the trials are independent (i.e., the σ-fields $\mathfrak{T}_i$ are mutually statistically independent). Consider **K** finite and $\mathbf{B}_i = \mathbf{A}_i \times \mathbb{S}_{\mathbf{N}_i} \in \mathfrak{T}_i$ for $\mathbf{A}_i \in \mathfrak{F}_i$, for $i \in \mathbf{K}$. Then $\bigcap_{i\in\mathbf{K}} B_i = (\bigtimes_{i\in\mathbf{K}} \mathbf{A}_i)\times\mathbb{S}_{\mathbf{L}}$, where **K** and **L** partition **N**. By

mutual statistical independence,

$$P_{\mathbf{N}}(\bigcap_{i\in\mathbf{K}} B_i) = \prod_{i\in\mathbf{K}} P_{\mathbf{N}}(B_i)$$

or

$$P_{\mathbf{K}}(\bigtimes_{i\in\mathbf{K}} \mathbf{A}_i) = \prod_{i\in\mathbf{K}} P_i(\mathbf{A}_i).$$

Then, the compatability condition is satisfied trivially, and the fundamental result establishes that the existence of a unique $P_{\mathbf{N}}$ on $\mathfrak{F}_N$ whose restrictions give the probabilities for the individual trials.

4.3. The assumption of statistically independent repeated trials is a natural one for many statistical and econometric applications where the data comes from random samples from the population, such as surveys of consumers or firms. This assumption has many powerful implications, and will be used to get most of the results of basic econometrics. However, it is also common in econometrics to work with aggregate time series data. In these data, each period of observation can be interpreted as a new trial. The assumption of statistical independence across these trials is unlikely in many cases, because in most cases real random effects do not conveniently limit themselves to single time periods. The question becomes whether there are weaker assumptions that time series data are likely to satisfy that are still strong enough to get some of the basic statistical theorems. It turns out that there are quite general conditions, called mixing conditions, that are enough to yield many of the key results. The idea behind these conditions is that usually events that are far apart in time are nearly independent, because intervening shocks overwhelm the older history in determining the later event. Mixing will be defined to formalize the concept of "nearly independent". Suppose t indexes time periods. Define $\mathbf{K}_1^t$ to be

the set of time indexes up through time t, and $\mathbf{K}^{\infty}_{t+n}$ to be the set of time indexes from time t+n on. Let $\mathfrak{I}^{t}_{1}$ be the product σ-field of events that use only information from time 1 to time t. Let $\mathfrak{I}^{\infty}_{t+n}$ denote the product σ-field of events that use only information from time t+n on. The idea of mixing is that when events are far enough removed in time, there is little information contained in the occurence of one that helps determine the probability of the second. The trials are *strong mixing* if there exists a scalar $\alpha(n)$ satisfying $\lim_{n\to\infty} \alpha(n) = 0$ such that $|P(\mathbb{A}\cap\mathbb{B}) - P(\mathbb{A})P(\mathbb{B})| \le \alpha(n)$ for all $\mathbb{A} \in \mathfrak{I}^{t}_{1}$ and $\mathbb{B} \in \mathfrak{I}^{\infty}_{t+n}$. They are *uniform mixing* if $|P(\mathbb{B}|\mathbb{A}) - P(\mathbb{B})| \le \phi(n)$ for all $\mathbb{A} \in \mathfrak{I}^{t}_{1}$ and $\mathbb{B} \in \mathfrak{I}^{\infty}_{t+n}$, and $\lim_{n\to\infty} \phi(n) = 0$; note that the inequality is equivalent to $|P(\mathbb{A}\cap\mathbb{B}) - P(\mathbb{A})P(\mathbb{B})| \le \phi(n)P(\mathbb{B})$. They are *strict mixing* if $|P(\mathbb{A}\cap\mathbb{B}) - P(\mathbb{A})P(\mathbb{B})| \le \psi(n)P(\mathbb{A})P(\mathbb{B})$ for all $\mathbb{A} \in \mathfrak{I}^{t}_{1}$ and $\mathbb{B} \in \mathfrak{I}^{\infty}_{t+n}$, and $\lim_{n\to\infty} \psi(n) = 0$. Note that the right-hand-side of each of these inequalities is zero for independent trials, so that the mixing assumptions correspond to near independence for remote events. Many economic time series model utilize assumptions that imply strong mixing, at least after suitable data transformation.

5. Random Variables, Distribution Functions, and Expectations

5.1. A *random variable* X is a measurable real-valued function on a probability space $(\mathbb{S},\mathfrak{F},P)$. The value of the function x = X(s) for an s that occurs is termed a *realization* of the random variable. One can have many random variables defined on the same probability space; another measurable function y = Y(s) defines a second random variable. It is very helpful in working with random variables to keep in mind that the random variable itself is a function of states of Nature, and that observations are of realizations of the random variable. Thus, when one talks about

convergence of a sequence of random variables, one is actually talking about convergence of a sequence of functions, and notions of distance and closeness need to be formulated as distance and closeness of functions.

5.2. The term *measurable* in the definition of a random variable means that for each set A in the Borel σ-field $\mathfrak{B}$ of subsets of the real line, the inverse image $X^{-1}(A) \equiv \{s \in \mathbb{S} \mid X(s) \in A\}$ is in the σ-field $\mathfrak{F}$ of subsets of the sample space $\mathbb{S}$. The assumption of measurability is a mathematical technicality that ensures that probability statements about the random variable are meaningful. We shall not make any explicit reference to measurability in basic econometrics, and shall always assume implicitly that the random variables we are dealing with are measurable.

5.3. The probability that a random variable X has a realization in a set $A \in \mathfrak{B}$ is given by

$$F(A) \equiv P(X^{-1}(A)) \equiv P(\{s \in \mathbb{S} \mid X(s) \in A\}).$$

The function F is a probability on $\mathfrak{B}$; it is defined in particular for half-open intervals of the form $A = (-\infty, x]$, in which case $F((-\infty, x))$ is abbreviated to $F(x)$ and is called the *distribution function* (or, *cumulative distribution function, CDF)* of X. From the properties of a probability, the distribution function has the properties

(i) $F(-\infty) = 0$ and $F(+\infty) = 1$.

(ii) F(x) is non-decreasing in x, and continuous from the right.

(iii) F(x) has at most a countable number of jumps, and is continuous except at these jumps. (Points without jumps are called *continuity* points.)

Conversely, any function F that satisfies (i) and (ii) determines uniquely a probability F on $\mathfrak{B}$. The *support* of the distribution F is the smallest closed set $A \in$

$\mathfrak{B}$ such that F(A) = 1.

5.4. If F is absolutely continuous with respect to a σ-finite measure ν on $\mathbb{R}$; i.e., F gives probability zero to any set that has ν-measure zero, then (by the Radon-Nikodym theorem) there exists a real-valued function f on $\mathbb{R}$, called the *density* (or *probability density function, pdf)* of X, such that

$$F(A) = \int_A f(x)\nu(dx)$$

for every $A \in \mathfrak{B}$. With the possible exception of a set of ν-measure zero, F is differentiable and the derivative of the distribution gives the density, f(x) = F′(x). When the measure ν is *Lebesgue measure*, so that the measure of an interval is its length, it is customary to simplify the notation and write

$$F(A) = \int_A f(x)dx.$$

If F is absolutely continuous with respect to counting measure on a countable subset $\mathbb{C}$ of $\mathbb{R}$, then it is called a *discrete* distribution, and there is a real-valued function f on $\mathbb{C}$ such that $F(A) = \sum_{x \in A} f(x)$.

5.5. If $(\mathbb{R},\mathfrak{B},F)$ is the probability space associated with a random variable X, and $g:\mathbb{R} \to \mathbb{R}$ is a measurable function, then Y = g(X) is another random variable. The random variable Y is *integrable* with respect to the probability F if

$$\int_{\mathbb{R}} |g(x)| F(dx) < +\infty;$$

if it is integrable, then the integral

$$\int_{\mathbb{R}} g(x)F(dx) \equiv \int_{\mathbb{R}} g \cdot dF$$

exists, is denoted **E** g(X), and is called *the expectation of* g(X). When necessary, this expectation will also be denoted $\mathbf{E}_X g(X)$ to identify the distribution used to form the expectation. When F is absolutely continuous with respect to Lebesgue measure, so that F has a density f, the notation for the expectation simplifies to

$$\mathbf{E}\ g(X) = \int_{\mathbb{R}} g(x)f(x)dx.$$

The expectation of X, if it exists, is called the *mean* of X. The expectation of $(X - \mathbf{E}X)^2$, if it exists, is called the *variance* of X. Define $\mathbf{1}(X \leq a)$ to be an indicator function that is one if $X(s) \leq a$, and zero otherwise. Then, $\mathbf{E}\ \mathbf{1}(X \leq a) = F(a)$, and the distribution function can be recovered from the expectations of the indicator functions.

Example 1. (continued) Define a random variable X by

$$X(s) = \begin{cases} 0 & \text{if } s = TT \\ 1 & \text{if } s = TH \text{ or } HT \\ 2 & \text{if } s = HH \end{cases}$$

Then, X is the number of heads in two coin tosses. For a fair coin, $\mathbf{E}\ X = 1$.

Example 2. (continued) Let X be a random variable defined to equal the number of heads that appear before a tail occurs. Then, possible values of X are the integers $\mathbb{C} = \{0,1,2,...\}$. For x real, define [x] to be the largest integer k satisfying $k \leq x$. A distribution function for X is the geometric, $F(x) = \begin{cases} 1 - 2^{-[x+1]} & \text{for } 0 \leq x \\ 0 & \text{for } 0 > x \end{cases}$; the associated density defined on $\mathbb{C}$ is

$f(k) = 2^{-k-1}$. The expectation of X is $\mathbf{E}\ X = \sum_{k=0}^{\infty} k \cdot 2^{-k-1} = 1$. (A geometric series $\sum_{k=0}^{\infty} r^k = \frac{1}{1-r}$ for $-1 < r < 1$. Differentiating with respect to r and then multiplying by r^2 yields $\sum_{k=0}^{\infty} k r^{k+1} = \frac{r^2}{(1-r)^2}$. Evaluate this at r = 1/2 to get the result.)

Example 3. (continued) Define a random variable X by $X(s) = |s|$. Then, X is the magnitude of the daily change in the price index. The inverse image of an interval (a,b) with $a < 0$ is $(-b,b) \in \mathfrak{F}$, and the inverse image of an interval (a,b) with $a \geq 0$ is $(-b,-a) \cup (a,b) \in \mathfrak{F}$. Then X is measurable. Another measurable random variable is Y defined by $Y(s) = \text{Max}\ \{0,s\}$, a third is Z defined by $Z(s) = s^3$.

5.6. Consider a random variable Y on $(\mathbb{R},\mathfrak{B})$. The expectation $\mathbf{E}_Y Y^k$ is the k-th (non-centered) moment of k, and $\mathbf{E}_Y (Y - \mathbf{E}_Y Y)^k$ is the k-th central moment. Sometimes moments fail to exist. However, if g(Y) is continuous and bounded, then $\mathbf{E}_Y g(Y)$ always exists. The expectation

$$m_Y(t) = \mathbf{E}_Y e^{tY}$$

is termed the *moment generating function* (mgf) of Y; it sometimes fails to exist. Call a mgf *proper* if it is finite for t in an interval around 0. When a proper mgf exists, the random variable has finite moments of all orders.

The expectation

$$\psi_Y(t) = \mathbf{E}_Y e^{\iota tY},$$

where $\iota = \sqrt{-1}$, is termed the *characteristic function* (cf) of Y. The characteristic function always exists.

5.7. A measurable function X from the probability space $(\mathbb{S},\mathfrak{F},P)$ into $(\mathbb{R}^n,\mathfrak{B}^n)$ is termed a *random vector*. (The notation $\mathfrak{B}^n$ means $\mathfrak{B}\otimes\mathfrak{B}\otimes...\otimes\mathfrak{B}$ n times, where $\mathfrak{B}$ is the Borel σ-field on the real line. This is also called the *product* σ-field, and is sometimes written $\mathfrak{B}^n = \prod_{i=1}^{n} \mathfrak{B}_i$, where the $\mathfrak{B}_i$ are identical copies of $\mathfrak{B}$.) The random vector can also be written $X = \begin{bmatrix} X_1 \\ X_2 \\ | \\ X_n \end{bmatrix}$, with each component X_i a random variable. The *distribution function* (CDF) of X is

$$F_X(x_1,...,x_n) = P(\{s\varepsilon\mathbb{S} \mid X_i(s) \leq x_i \text{ for } i = 1,...,n\}).$$

If $A \in \mathfrak{B}^n$, define $F_X(A) = P(\{s\varepsilon\mathbb{S} \mid X(s)\in A\})$. If $F_X(A) = 0$ for every set A of Lebesque measure zero, then there exists a *probability density function* (pdf) $f_X(x_1,...,x_n)$ such that

$$(1) \qquad F_X(x_1,...,x_n) = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2}...\int_{-\infty}^{x_n} f_X(y_1,...,y_n)\, dy_1...dy_n.$$

F_X and f_X are termed the *joint* or *multivariate* CDF and pdf, respectively, of X. The random variable X_1 has a distribution that satisfies

$$F_{X_1}(x_1) \equiv P(\{s\in S \mid X_1(s) \leq x_1\}) = F_X(x_1,+\infty,...,+\infty).$$

This random variable is measurable with respect to the σ-subfield $\mathfrak{G}_1$ containing the events whose occurance is determined by X_1 alone; i.e., $\mathfrak{G}_1$ is the family of sets of the form $A\times\mathbb{R}\times...\times\mathbb{R}$ with $A \in \mathfrak{B}$. If F_X is absolutely continuous with respect to Lebesque measure on $\mathfrak{B}^n$, then there is an associated density f_X and implied density

f_{X_1} satisfying

$$(2) \qquad F_{X_1}(x_1) = \int_{-\infty}^{x_1} f_{X_1}(y_1)\, dy_1$$

$$(3) \qquad f_{X_1}(x_1) = \int_{y_2=-\infty}^{+\infty} \cdots \int_{y_n=-\infty}^{+\infty} f_X(x_1,y_2,\ldots,y_n)\cdot dy_2\ldots dy_n.$$

F_{X_1} and f_{X_1} are termed the *marginal* CDF and pdf, respectively, of X_1.

5.8. Corresponding to the concept of a conditional probability, we can define a *conditional distribution*: Suppose C is an event in $\mathfrak{S}_1$ with P(C) > 0. Then, define $F_{X_2,\ldots,X_n}(x_2,\ldots,x_n \mid C) = F(\{y\in\mathbb{R}^n \mid y_1\in C, y_2\le x_2,\ldots,y_n\le x_n\})/F_{X_1}(C)$ to be the conditional distribution of $(X_2,\ldots,X_n)$ given $X_1 \in C$. When F_X is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^n$, the conditional distribution can be written in terms of the joint density,

$$F_{X_2,\ldots,X_n}(x_2,\ldots,x_n \mid C) = \frac{\int_{y_1\in C}\int_{y_2=-\infty}^{x_2}\cdots\int_{y_n=-\infty}^{x_n} f_X(y_1,y_2,\ldots,y_n)\cdot dy_1dy_2\ldots dy_n}{\int_{y_1\in C}\int_{y_2=-\infty}^{+\infty}\cdots\int_{y_n=-\infty}^{+\infty} f_X(y_1,y_2,\ldots,y_n)\cdot dy_1dy_2\ldots dy_n}.$$

Taking the limit as C shrinks to a point $X_1 = x_1$ where $f_{X_1}(x_1) > 0$, one obtains the conditional distribution of $(X_2,\ldots,X_n)$ given $X_1 = x_1$,

$$F_{X_2,\ldots,X_n}(x_2,\ldots,x_n \mid X_1=x_1) = \frac{\int_{y_2=-\infty}^{x_2}\cdots\int_{y_n=-\infty}^{x_n} f_X(x_1,y_2,\ldots,y_n)\cdot dy_1dy_2\ldots dy_n}{f_{X_1}(x_1)}.$$

Finally, associated with this conditional distribution is the conditional density

$$f_{X_2,...,X_n}(x_2,...,x_n \mid X_1=x_1) = \frac{f_X(x_1,x_2,...,x_n)}{f_{X_1}(x_1)} .$$

More generally, one could consider the marginal distributions of any subset, say $X_1,...X_k$, of the vector X, with $X_{k+1},...X_n$ integrated out; and the conditional distributions of one or more of the variables $X_{k+1},...X_n$ given one or more of the conditions $X_1 = x_1,...,X_k = x_k$.

5.9. Just as expectations are defined for a single random variable, it is possible to define expectations for a vector of random variables. For example, $\mathbf{E}(X_1 - \mathbf{E}X_1)(X_2-\mathbf{E}X_2)$ is called the *covariance* of X_1 and X_2, and $\mathbf{E}e^{t'X}$, where $t' = (t_1,...,t_n)$ is a vector of constants, is a (multivariate) moment generating function for the random vector X.

Some useful properties of expectations are

(a) If g(X) is a function of a random vector, then $\mathbf{E}g(X)$ is the integral of g with respect to the distribution of X. When g depends on a subvector of X, then $\mathbf{E}g(X)$ is the integral of g(y) with respect to the marginal distribution of this subvector.

(b) If X and Z are random vectors of length n, and *a* and *b* are scalars, then $\mathbf{E}(aX + bZ) = a\mathbf{E}X + b\mathbf{E}Z$.

(c) [Cauchy-Schwartz inequality] If X and Z are random vectors of length n, then $(\mathbf{E}X'Z)^2 \leq (\mathbf{E}X'X)(\mathbf{E}Z'Z)$.

When expectations exist, they can be used to bound the probability that a random variable takes on extreme values. We give three such bounds:

a. [Markov bound] If X is a random variable with $\mathbf{E}|X| < +\infty$ and ε is a positive

scalar, then $\Pr(|X| > \varepsilon) < \mathbf{E}|X|/\varepsilon$.

b. [Chebyshev bound] If X is a random vector with $\mathbf{E}X'X < +\infty$ and ε is a positive scalar, then $\Pr(\|X\|_2 > \varepsilon) < \mathbf{E}X'X/\varepsilon^2$.

c. [Chernoff bound] If X is a random vector with a proper moment generating function (i.e., $m(t) \equiv \mathbf{E}e^{t'X}$ exists for all vectors t in some neighborhood of zero), and ε is a positive scalar, then for some positive scalars α and M, $\Pr(\|X\|_2 > \varepsilon) < Me^{-\alpha\varepsilon}$.

All these inequalities are established by the same technique: If r(y) is a positive increasing function of $y > 0$, and $\mathbf{E}r(\|X\|) < +\infty$, then

$$\Pr(\|X\|_2 > \varepsilon) = \int_{\|x\|_2>\varepsilon} F(dx) \leq \int_{\|x\|_2>\varepsilon} [r(\|x\|_2)/r(\varepsilon)]F(dx) \leq \mathbf{E}r(\|X\|_2)/r(\varepsilon).$$

Taking r(y) = y, y^2 gives the results directly for the first two ineqalities. In the third case, first get a component-by-component inequality

$$\Pr(\|X\|_2 > \varepsilon) \leq \sum_{i=1}^{n} [\Pr(X_i > \varepsilon/\sqrt{n}) + \Pr(X_i < -\varepsilon/\sqrt{n})]$$

by showing that if the event on the left occurs, one of the events on the right must occur. Then apply the inequality $\Pr(|X_i| > \varepsilon) \leq \mathbf{E}r(|X_i|)/r(\varepsilon)$ with $r(y) = e^{y\alpha\sqrt{n}}$ to each term in the right-hand-side sum. The inequality for vectors is built up from a corresponding inequality for each component.

5.10. When the expectation of a random variable is taken with respect to a conditional distribution, it is called a *conditional expectation*. If $F(x|C)$ is the conditional distribution of a random vector X given the event C, then the conditional expectation of a function g(X) given C is defined as

$$\mathbf{E}_{X|C}g(X) = \int g(y)F(dy|C).$$

Another notation for this expectation is $\mathbf{E}_X(g(X)|C)$. The conditional expectation is actually a function on the σ-field $\mathfrak{C}$ of conditioning events, and is sometimes written $\mathbf{E}_{X|\mathfrak{C}}\, g(X)$ or $\mathbf{E}(g(X)|\mathfrak{C})$.

The concept of conditional expectations is very important in econometrics and in economic theory, so we will work out its properties in some detail for the case of two variables, using as an example random variables with a bivariate normal distribution.

Suppose random variables (U,X) have a joint density f(u,x). The marginal density of X is defined by

$$g(x) = \int f(u,x)du,$$

and the conditional density of U given X = x is defined by $f(u|x) = f(u,x)/g(x)$, provided $g(x) > 0$. The conditional expectation of a function h(U,X) satisfies $\mathbf{E}(h(U,X)|X=x) = \int h(u,x)f(u|x)du$, and is a function of x. The unconditional expectation of h(U,X) satisfies

$$\mathbf{E}h(U,X) = \int h(u,x)f(u,x)dudx = \int_x \left[\int_u h(u,x)f(u|x)du\right] g(x)dx$$

$$= \mathbf{E}_X\mathbf{E}_{U|X}h(U,X);$$

this is called the *law of iterated expectations*. The *conditional mean* of U given X=x is $\mathbf{M}_{U|X}(x) \equiv \mathbf{E}_{U|X=x}U$; by the law of iterated expectations, the conditional and unconditional mean are related by $\mathbf{E}_U U = \mathbf{E}_X\mathbf{E}_{U|X}U \equiv \mathbf{E}_X\mathbf{M}_{U|X}(X)$.

The *conditional variance* of U is $\mathbf{V}(U|x) = \mathbf{E}_{U|X}(U - \mathbf{M}_{U|X}(x))^2$. It is related to

the unconditional variance by the formula

$$\begin{aligned}\mathbf{E}_U(U - \mathbf{E}_U U)^2 &= \mathbf{E}_X\mathbf{E}_{U|X}(U - \mathbf{M}_{U|X}(X) + \mathbf{M}_{U|X}(X) - \mathbf{E}_U U)^2 \\ &= \mathbf{E}_X\mathbf{E}_{U|X}(U - \mathbf{M}_{U|X}(X))^2 + \mathbf{E}_X\mathbf{E}_{U|X}(\mathbf{M}_{U|X}(X) - \mathbf{E}_U U)^2 \\ &\quad + 2\mathbf{E}_X\mathbf{E}_{U|X}(U - \mathbf{M}_{U|X}(X))(\mathbf{M}_{U|X}(X) - \mathbf{E}_U U) \\ &= \mathbf{E}_X\mathbf{V}(U|X) + \mathbf{E}_X(\mathbf{M}_{U|X}(X) - \mathbf{E}_U U)^2 \\ &\quad + 2\mathbf{E}_X(\mathbf{M}_{U|X}(X) - \mathbf{E}_U U)\mathbf{E}_{U|X}(U - \mathbf{M}_{U|X}(X)) \\ &= \mathbf{E}_X\mathbf{V}(U|X) + \mathbf{E}_X(\mathbf{M}_{U|X}(X) - \mathbf{E}_U U)^2.\end{aligned}$$

Then, the unconditional variance equals the expectation of the conditional variance plus the variance of the conditional expectation.

Example: Suppose (U,X) are bivariate normal with $\mathbf{E}U = \mu_u$, $\mathbf{E}X = \mu_x$, $\mathbf{E}(U-\mu_u)^2 = \sigma_u^2$, $\mathbf{E}(X-\mu_x)^2 = \sigma_x^2$, and $\mathbf{E}(U-\mu_u)(X-\mu_x) = \sigma_{ux} \equiv \rho\sigma_u\sigma_x$. The bivariate normal density is $f(u,x) = \left(2\pi\sigma_u\sigma_x\sqrt{1-\rho^2}\right)^{-1}\exp(-Q/2)$, with

$$Q = \left(\frac{u-\mu_u}{\sigma_u}\right)^2 + \left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{u-\mu_u}{\sigma_u}\right)\left(\frac{x-\mu_x}{\sigma_x}\right).$$

The marginal density of X is normal with mean μ_x and variance σ_x^2:

$$n(x-\mu_x,\sigma_x) = (\sqrt{2\pi}\,\sigma_x)^{-1}e^{-(x-\mu_x)^2/2\sigma_x^2}.$$

This can be derived from the bivariate density by completing the square for u in Q and integrating over u. The conditional density then satisfies

$$f(u|x) = \left(\sqrt{2\pi}\cdot\sigma_u\sqrt{1-\rho^2}\right)^{-1}\exp\left(-Q/2(1-\rho^2) + (x-\mu_x)^2/2\sigma_x^2\right).$$

But

$$Q - (1-\rho)^2\left(\frac{x-\mu_x}{\sigma_x}\right)^2 = \left(\frac{u-\mu_u}{\sigma_u}\right)^2 - 2\rho\left(\frac{u-\mu_u}{\sigma_u}\right)\left(\frac{x-\mu_x}{\sigma_x}\right) + \rho^2\left(\frac{x-\mu_x}{\sigma_x}\right)^2$$

$$= \left[\left(\frac{u-\mu_u}{\sigma_u}\right) - \rho\left(\frac{x-\mu_x}{\sigma_x}\right)\right]^2,$$

so that

$$f(u|x) = \left(\sqrt{2\pi}\,\sigma_u\sqrt{1-\rho^2}\right)^{-1}\exp\left(-(u-\mu_u - \frac{\rho\sigma_u}{\sigma_x}(x-\mu_x))^2/2\sigma_u^2(1-\rho^2)\right).$$

Hence the conditional distribution of U given X = x is normal with conditional mean

$$\mathbf{E}(U|X=x) = \mu_u + \rho\frac{\rho\sigma_u}{\sigma_x}(x-\mu_x) \equiv \mu_u + \frac{\sigma_{ux}}{\sigma_x^2}(x-\mu_x)$$

and conditional variance $\mathbf{V}(U|X=x) \equiv \mathbf{E}\left((U-\mathbf{E}(U|X=x))^2|X=x\right) = \sigma_u^2(1-\rho^2) \equiv \sigma_u^2 - \sigma_{ux}^2/\sigma_x^2$. When U and X are joint normal random <u>vectors</u> with $\mathbf{E}U = \mu_u$, $\mathbf{E}X = \mu_x$, $\mathbf{E}(U-\mu_u)(U-\mu_u)' = \Omega_{uu}$, $\mathbf{E}(X-\mu_x)(X-\mu_x)' = \Omega_{xx}$, and $\mathbf{E}(U-\mu_u)(X-\mu_x)' = \Omega_{ux}$, then $(U|X=x)$ is normal with $\mathbf{E}(U|X=x) = \mu_u + \Omega_{ux}\Omega_{xx}^{-1}(x-\mu_x)$ and $V(U|X=x) = \Omega_{uu} - \Omega_{ux}\Omega_{xx}^{-1}\Omega_{xu}$.

5.11. Conditional densities satisfy $f(u,x) = f(u|x)g(x) = f(x|u)h(u)$, where $h(u)$ is the marginal density of U, and hence $f(u|x) = f(x|u)\,h(u)/g(x)$. This is called *Bayes Law*. When U and X are <u>independent</u>, $f(u,x) = h(u)\cdot g(x)$, or $f(u|x) = h(u)$ and $f(x|u) = g(x)$. For U and X independent, and $r(\cdot)$ and $s(\cdot)$ any functions, one has

$$\mathbf{E}(r(U)|X=x) = \int r(u)f(u|x)du \equiv \int r(u)h(u)du = \mathbf{E}r(U),$$

and

$\mathbf{E}(r(U)s(X)) = \int r(u)s(x)f(u,x)dudx = \int s(x)g(x)\int r(u)f(u|x)du\,dx$

$= \int s(x)g(x)\mathbf{E}r(U|x)dx = [\mathbf{E}s(X)][\mathbf{E}r(U)]$,

or $\text{cov}\left(r(U),s(X)\right) = 0$, provided $\mathbf{E}r(U)$ and $\mathbf{E}s(X)$ exist. In particular, if $r(u) = u - \mathbf{E}U$, then $\mathbf{E}(r(U)|X=x) = 0$ and $\text{cov}(U,X) = \mathbf{E}(U-\mathbf{E}U)X = 0$.

Conversely, suppose U and X are jointly distributed. If $\text{cov}(r(U),s(X)) = 0$ for all functions $r(\cdot)$, $s(\cdot)$ such that $\mathbf{E}r(U)$ and $\mathbf{E}s(X)$ exist, then X and U are independent. To see this, choose $r(u) = 1$ for $u \le u^*$, $r(u) = 0$ otherwise; choose $s(x) = 1$ for $x \le x^*$, $s(x) = 0$ otherwise. Then $\mathbf{E}r(U) = H(u^*)$ and $\mathbf{E}s(X) = G(x^*)$, where H and G are the marginal cumulative distribution functions, and $0 = \text{cov}\left(r(U), s(X)\right)$ $= F(u^*,x^*) - H(u^*)\cdot G(x^*)$, where F is the joint cumulative distribution function. Hence, $F(u,x) = H(u)\cdot G(x)$, and X, U are independent.

Note that $\text{cov}(U,X) = 0$ is not sufficient to imply U,X independent. For example, $g(x) = 1/2$ for $-1 \le x \le 1$ and $f(u|x) = 1/2$ for $-1 \le u-x^2 \le 1$ is nonindependent with $\mathbf{E}(U|X=x) = x^2$, but $\text{cov}(U,X) = \mathbf{E}X^3 = 0$. Furthermore, $\mathbf{E}(U|X=x) \equiv 0$ is not sufficient to imply U,X independent. For example, $g(x) = 1/2$ for $-1 \le x \le 1$ and $f(u|x) = 1/2(1 + x^2)$ for $-(1+ x^2) \le u \le (1 + x^2)$ is nonindependent with $\mathbf{E}(U^2|x) = (1 + x^2)^2 \neq \mathbf{E}U^2 = 28/15$, but $\mathbf{E}(U|X=x) \equiv 0$.

5.12. The discussion of expectations will be concluded with a list of detailed properties of characteristic functions and moment generating functions that will be useful later:

a. $\psi(t) = \mathbf{E}_Y e^{\imath tY} \equiv \mathbf{E}_Y \cos(tY) + \imath \mathbf{E}_Y \sin(tY)$,

b. $Z = a + bY$ has the cf $e^{\imath ta}\psi(bt)$,

c. If $\mathbf{E}_Y Y^k$ exists, then $\psi^{(k)}(t) \equiv d^k\psi(t)/dt^k$ exists and is uniformly continuous,

and $\mathbf{E}_Y Y^k = (-\iota)^k \psi^{(k)}(0)$. If $\psi^{(k)}(t)$ exists, then $\mathbf{E}_Y Y^k$ exists.

d. If Y has finite moments through order k, then $\psi(t)$ has a Taylor's expansion

$$\psi(t) = \sum_{j=0}^{k} \iota^j (\mathbf{E}_Y Y^j) t^j/j! + [\psi^{(k)}(\lambda t) - \psi^{(k)}(0)] t^k/k!$$

where λ is a scalar with $0 < \lambda < 1$; the Taylor's expansion satisfies the bounds

$$\left| \psi(t) - \sum_{j=0}^{k-1} \iota^j (\mathbf{E}_Y Y^j) t^j/j! \right| \leq |t|^k \mathbf{E}_Y |Y|^k/k!$$

and

$$\left| \psi(t) - \sum_{j=0}^{k} \iota^j (\mathbf{E}_Y Y^j) t^j/j! \right| \leq 2|t|^k \mathbf{E}_Y |Y|^k/k!$$

If $\mathbf{E}_Y Y^k$ exists, then the expression $\zeta(t) = \text{Ln } \psi(t)$, called the *second characteristic function* or *cumulant generating function*, has a Taylor's expansion

$$\zeta(t) = \sum_{j=0}^{k} \kappa_j \iota^j t^j/j! + [\zeta^{(k)}(\lambda t) - \zeta^{(k)}(t)],$$

where $\zeta^{(k)} \equiv d^k\zeta/dt^k$, and λ is a scalar with $0 < \lambda < 1$. The expressions κ_j are called the *cumulants* of the distribution, and satisfy $\kappa_1 = \mathbf{E}_Y Y$ and $\kappa_2 = \text{Var}(Y)$. The expression $\kappa_3/\kappa_2^{3/2}$ is called the *skewness*, and the expression $\kappa_4/\kappa_2^2 - 3$ is called the *kurtosis* (i.e., thickness of tails relative to center), of the distribution.

e. If Y is normally distributed with mean μ and variance σ^2, then its characteristic function is $e^{\iota\mu t - \sigma^2 t^2/2}$. The normal has cumulants $\kappa_1 = \mu$, $\kappa_2 = \sigma^2$, $\kappa_3 = \kappa_4 = 0$.

f. Random variables X and Y are distinct if and only if their characteristic functions are distinct.

g. If $Y_n \xrightarrow{d} Y$ (see Chap. 4), then the associated characteristic functions satisfy $\psi_n(t) \to \psi(t)$ for each t. Conversely, if Y_n has characteristic function $\psi_n(t)$ converging pointwise to a function $\psi(t)$ that is continuous at t = 0, then there exists Y such that $\psi(t)$ is the characteristic function of Y and $Y_n \xrightarrow{d} Y$.

h. The characteristic function of a sum of independent random variables equals the product of the characteristic functions of these random variables, and the second characteristic function of a sum of independent random variables is the sum of the second characteristic functions of these variables; the characteristic function of a mean of n independently identically distributed random variables, with characteristic function $\psi(t)$, is $\psi(t/n)^n$.

Similar properties hold for proper moment generating functions, with obvious modifications: Suppose a random variable Y has a proper mgf $m_Y(t)$, finite for $|t| < \tau$, where τ is a positive constant. Then, the following properties hold:

a. $m_Y(t) = \mathbf{E}_Y e^{tY}$ for $|t| < \tau$.

b. $Z = a + bY$ has the mgf $e^{ta} m_Y(bt)$.

c. $\mathbf{E}_Y Y^k$ exists for all $k > 0$, and $m_Y^{(k)} \equiv d^k m_Y(t)/dt^k$ exists and is uniformly continuous for $|t| < \tau$, with $\mathbf{E}_Y Y^k = m_Y^{(k)}(0)$.

d. $m_Y(t)$ has a Taylor's expansion (for any k)

$$m_Y(t) = \sum_{j=0}^{k} (\mathbf{E}Y^j)t^j/j! + [m_Y^{(k)}(\lambda t) - m_Y^{(k)}(0)]t^k/k!,$$

where λ is a scalar with $0 < \lambda < 1$.

e. If Y is normally distributed with mean μ and variance σ^2, then it has mgf $e^{\mu t+\sigma^2 t^2/2}$.

f. Random variables X and Y with proper mgf are distinct if and only if their mgf are distinct.

g. If $Y_n \xrightarrow{d} Y$ and the associated mgf satisfy $m_{Y_n}(t)$, $m_Y(t)$ finite for $|t| < \tau$, then $m_{Y_n}(t) \to m_Y(t)$. Conversely, if Y_n has proper mgf $m_{Y_n}(t)$ converging pointwise to a function $m_Y(t)$ finite for $|t| < \tau$, then there exists Y such that $m_Y(t)$ is the mgf of Y and $Y_n \xrightarrow{d} Y$.

h. The mgf of a sum of independent random variables equals the product of the mgf of these random variables; the mgf of the mean of n independently identically distributed random variables, each with proper mgf $m_Y(t)$, is $m_Y(t/n)^n$.

The properties of characteristic functions and moment generating functions are discussed and established in C. R. Rao Linear Statistical Inference, 2b.4, and W. Feller An Introduction to Probability Theory, II, Chap. 13 and 15.

6. Transformations of Random Variables

6.1. Suppose X is a measurable random variable on $(\mathbb{R},\mathfrak{B})$ with a distribution $F_X(x)$ that is absolutely continuous with respect to Lebesgue measure, so that X has a density $f_X(x)$. Consider an increasing transformation Y = H(X); then Y is another random variable. Let h denote the inverse function of H; i.e., y = H(x) implies x = h(y). The distribution function of Y is given by

$$F_Y(y) = \Pr(Y \le y) = \Pr(H(X) \le y) = \Pr(X \le h(y)) = F_X(h(y)).$$

When h(y) is differentiable, with a derivative $h'(y) = dh(y)/dy$, the density of Y is obtained by differentiating, and satisfies

$$f_Y(y) = f_X(h(y))h'(y).$$

Since $y \equiv H(h(y))$, one obtains by differentiation the formula $1 \equiv H'(h(y))h'(y)$, or $h'(y) = 1/H'(h(y))$. Substituting this formula gives

$$f_Y(y) = f_X(h(y))/H'(h(y)).$$

For example, suppose X has the distribution function $F_X(x) = 1\text{-}e^{-x}$ for $x > 0$, with $F_X(x) = 0$ for $x \leq 0$; then X is said to have an exponential distribution. Suppose $Y = H(X) \equiv \log X$, so that $X = h(Y) \equiv e^Y$. Then, $F_Y(y) = 1\text{-}\exp(-e^y)$ and $f_Y(y) = \exp(-e^y)e^y = \exp(y\text{-}e^y)$. This is called an extreme value distribution. A second example is X with some distribution function F_X and density f_X, and $Y = F_X(X)$, so that for any value of X, the corresponding value of Y is the proportion of all X that are below this value. Let x_p denote the solution to $F_X(x) = p$. The distribution function of Y is $F_Y(y) = F_X(x_y) = y$. Hence, Y has the uniform density on the unit interval.

The rule for an increasing transformation of a random variable X can be extended in several ways. If the transformation $Y = H(X)$ is <u>decreasing</u> rather than increasing, then

$$F_Y(y) = \Pr(Y \leq y) = \Pr(H(X) \leq y) = \Pr(X \geq h(y)) = 1\text{-}F_X(h(y)),$$

where h is the inverse function of H. Differentiating,

$$f_Y(y) = f_X(h(y))(-h'(y)).$$

Then, combining cases, one has the result that *for any one-to-one transformation* $Y = H(X)$ *with inverse* $X = h(Y)$*, the density of* Y *is*

$$f_Y(y) = f_X(h(y))\,|h'(y)| \equiv f_X(h(y))/|H'(h(y)|.$$

An example of a decreasing transformation is X with the exponential density e^{-x} for $x > 0$, and $Y = 1/X$. Show as an exercise that $F_Y(y) = e^{-1/y}$ and $f_Y(y) = e^{-1/y}/y^2$.

Consider a transformation $Y = H(X)$ that is <u>not</u> one-to-one. The interval $(-\infty,y)$

is the image of a set A_y of x values that may have a complicated structure. One can write

$$F_Y(y) = \Pr(Y \leq y) = \Pr(H(X) \leq y) = \Pr(X \in A_y) = F_X(A_y).$$

If this expression is differentiable, then its derivative gives the density. For example, if X has a distribution F_X and density f_X, and $Y = |X|$, then $A_y = [-y,y]$, implying $F_Y(y) = F_X(y) - F_X(-y)$ and $f_Y(y) = f_X(y) + f_X(-y)$. Another example is $Y = X^2$, implying $A_y = [-\sqrt{y},\sqrt{y}]$, $F_Y(y) = F_X(\sqrt{y}) - F_X(-\sqrt{y})$, and differentiating for $y \neq 0$, $f_Y(y) = (f_X(\sqrt{y}) + f_X(-\sqrt{y}))/2\sqrt{y}$.

6.2. Next consider transformations of random vectors. These transformations will permit us to analyze sums or other functions of random variables. Suppose X is a $n\times1$ random vector. Consider first the transformation $Y = AX$, where A is a nonsingular $n\times n$ matrix. The following result from multivariate calculus relates the densities of X and Y:

If X has density $f_X(x)$, and $Y = AX$, with A nonsingular, then the density of Y is

$$f_Y(y) = f_X(A^{-1}y)/|\det(A)| \ .$$

The following three examples prove this result in two dimensions:

Example 1. $\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} a_{11} & 0 \\ 0 & a_{22} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ with $a_{11} > 0$ and $a_{22} > 0$ has $F_Y(y_1,y_2) \equiv F_X(y_1/a_{11},y_2/a_{22})$. Differentiating with respect to y_1 and y_2, $f_Y(y_1,y_2) \equiv f_X(y_1/a_{11},y_2/a_{22})/a_{11}a_{22}$.

Example 2. $\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} a_{11} & 0 \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ with $a_{11} > 0$ and $a_{22} > 0$ has

$F_Y(y_1,y_2) \equiv \int_{x_1=-\infty}^{y_1/a_{11}} \int_{x_2=-\infty}^{(y_2-a_{21}x_1)/a_{22}} f_X(x_1,x_2)dx_2dx_1$. Differentiating with respect to y_1 and y_2 yields

$$\partial F_Y(y_1,y_2)/\partial y_2 = \int_{x_1=-\infty}^{y_1/a_{11}} a_{22}^{-1} f_X(x_1,(y_2-a_{21}x_1)/a_{22})dx_1.$$

$$\partial^2 F_Y(y_1,y_2)/\partial y_1 \partial y_2 \equiv f_Y(y_1,y_2) = (a_{11}a_{22})^{-1} f_X(y_1/a_{11},(y_2-y_1a_{21}/a_{11})/a_{22}).$$

Example 3. $\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ with $a_{11} > 0$ and $a_{11}a_{22}-a_{12}a_{21} > 0$ can be rewritten by making an intermediate transformation to Z as

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} a_{11} & 0 \\ a_{21} & a_{22}-a_{12}a_{21}/a_{22} \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \text{ and } \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = \begin{bmatrix} 1 & a_{12}/a_{11} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix},$$

since

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} a_{11} & 0 \\ a_{21} & a_{22}-a_{12}a_{21}/a_{22} \end{bmatrix} \begin{bmatrix} 1 & a_{12}/a_{11} \\ 0 & 1 \end{bmatrix}.$$

Then, using Example 2, $f_Z(z_1,z_2) = f_X(z_1-z_2a_{12}/a_{11},z_2)$ and $f_Y(y_1,y_2) = f_Z(y_1/a_{11},(y_2-y_1a_{21}/a_{11})/(a_{22}-a_{12}a_{21}/a_{11}))$. Substituting for f_Z in the last expression and simplifying (an exercise) gives

$$f_Y(y_1,y_2) = f_X((a_{22}y_1-a_{12}y_2)/\Delta,(a_{11}y_2-a_{21}y_1)/\Delta)/\Delta,$$

where $\Delta = a_{11}a_{22}-a_{12}a_{21}$ is the determinant of the transformation.

As an exercise, verify the formula for the density of Y = AX in the general case with A n×n and nonsingular. First, recall that A can be factored so that A = PLDUQ′,

where P and Q are permutation matrices, L and U are triangular with ones down the diagonal, and D is a nonsingular diagonal matrix. Write Y = PLDUQ′X. Then consider the series of intermediate transformations obtained by applying each matrix in turn, constructing the densities as was done in Example 3.

6.3. The extension from linear transformations to one-to-one nonlinear transformations of vectors is straightforward. Consider Y = H(X), with an inverse transformation X = h(Y). At a point y^o and $x^o = h(y^o)$, a first-order Taylor's expansion gives

$$y - y^o = A(x - x^o) + o(x - x^o),$$

where A is the *Jacobean* matrix

$$A = \begin{bmatrix} \partial H^1(x^o)/\partial x_1 & \dots & \partial H^1(x^o)/\partial x_n \\ | & & | \\ \partial H^n(x^o)/\partial x_1 & \dots & \partial H^n(x^o)/\partial x_n \end{bmatrix}$$

and the notation *o*(z) means some expression that is small relative to z. Then, the probability of Y in the little rectangle $[y^o, y^o+\Delta y]$ is approximately equal to the probability of X in the little rectangle $[x^o, x^o+A^{-1}\Delta y]$. This is the same situation as in the linear case, except there the equality was exact. Then, the formulas for the linear case carry over directly, with the Jacobean matrix of the transformation replacing the linear transformation matrix A.

In principle, it is possible to analyze n-dimensional nonlinear transformations that are <u>not</u> one-to-one in the same manner as the one-dimensional case, by working with the one-to-many inverse transformation. There are no general formulas, and each case needs to be treated separately.

Often in applications, one is interested in a transformation from a $n\times 1$ vector of random variables X to a lower dimension. For example, one may be interested in the scalar random variable $S = X_1 + ... + X_n$. If one "fills out" the transformation in a one-to-one way, so that the random variables of interest are components of the complete transformation, then the previous formulas can be applied. In the case of S, the transformation $Y_1 \equiv S$ filled out by $Y_i = X_i$ for i = 2,...,n is one-to-one, with

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ | \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & ... & 1 \\ 0 & 1 & 0 & ... & 0 \\ 0 & 0 & 1 & ... & 0 \\ | & | & | & & | \\ 0 & 0 & 0 & ... & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ | \\ X_n \end{bmatrix}.$$

7. Special Distributions

A number of special probability distributions appear frequently in statistics and econometrics, because they are convenient for applications or illustrations, because they are useful for approximations, or because they crop up in limiting arguments. The following tables list many of these distributions.

The first table lists discrete distributions. The binomial and geometric distributions are particularly simple, and are associated with statistical experiments such as coin tosses. The Poisson distribution is often used to model the occurence of rare events. The hypergeometric distribution is associated with various experiments of drawing red and white balls from urns, and is also used to approximate many other distributions.

The second table list a number of continuous distributions, including some basic

distributions such as the gamma and beta from which other distributions are constructed. The last table lists the normal distribution, and a number of other distributions that are related to it in statistical analysis. A series of notes follows the three tables, and provides more detail on some features.

SPECIAL DISCRETE DISTRIBUTIONS

NAME & DOMAIN	DENSITY	MOMENTS	CHAR. FN.
1. Binomial	$\binom{n}{k}p^k(1-p)^{n-k}$	$\mu = np$	$(1-p+pe^{\imath t})^n$
$k = 0,1,...,n$	$0 < p < 1$	$\sigma^2 = np(1-p)$	Note 1
2. Hypergeometric	$\binom{r}{k}\binom{w}{n-k} \div \binom{r+w}{n}$	$\mu = nr/(r+w)$	Note 2
k an integer	$r+w > n$	$\sigma^2 = \frac{nrw}{(r+w)^2}\cdot\frac{r+w-n}{r+w-1}$	
$\max\{0,n-w\} \le k$ & $k \le \min\{r,n\}$	r,w,n positive integers		
3. Geometric	$p(1-p)^k$	$\mu = (1-p)/p$	Note 3
$k = 0,1,2,...$	$0 < p < 1$	$\sigma^2 = (1-p)/p^2$	
4. Poisson	$e^{-\lambda}\lambda^k/k!$	$\mu = \lambda$	$\exp[\lambda(e^{\imath t}-1)]$
$k = 0,1,2,...$	$\lambda > 0$	$\sigma^2 = \lambda^2$	Note 4
5. Negative Binomial	$\binom{r+k-1}{k}p^r(1-p)^k$ r integer	$\mu = r(1-p)/p$	
$k = 0,1,2,...$	$r > 0$ & $0 < p < 1$	$\sigma^2 = r(1-p)/p^2$	Note 5

SPECIAL CONTINUOUS DISTRIBUTIONS

NAME & DOMAIN	DENSITY	MOMENTS	CHAR. FN.
1. Uniform $a \le x \le b$	$1/(b-a)$	$\mu = (a+b)/2$ $\sigma^2 = (b-a)^2/12$	$\frac{e^{\iota bt}-e^{\iota at}}{\iota t(b-a)}$ Note 6
2. Exponential $x \ge 0$	$e^{-x/\lambda}/\lambda$	$\mu = \lambda$ $\sigma^2 = \lambda^2$	$1/(1-\iota\lambda t)$ Note 7
3. Pareto $x \ge a$	$ba^b x^{-b-1}$	$\mu = ab/(b-1)$ if $b > 1$ $\sigma^2 = \frac{a^2 b}{(b-1)(b-2)}$ if $b > 2$	Note 8

NAME & DOMAIN	DENSITY	MOMENTS	CHAR. FN.
4. Gamma $x > 0$	$\frac{x^{a-1}e^{x/b}}{\Gamma(a)b^a}$	$\mu = ab$ $\sigma^2 = ab^2$	$(1-\iota bt)^{-a}$ Note 9
5. Beta $0 < x < 1$	$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}$	$\mu = a/(a+b)$ $\sigma^2 = \frac{ab}{(a+b)^2(a+b+1)}$	Note 10
6. Extreme Value $-\infty < x < +\infty$	$\frac{1}{b}\exp\left(-\frac{x-a}{b} - e^{-(x-a)/b}\right)$	$\mu = a + 0.57721 \cdot b$ $\sigma^2 = (\pi b)^2/12$	Note 11
7. Logistic $-\infty < x < +\infty$	$\frac{1}{b}\cdot\frac{\exp((a-x)/b)}{(1+\exp((a-x)/b))^2}$	$\mu = a$ $\sigma^2 = (\pi b)^2/6$	Note 12

NORMAL & RELATIVES	DENSITY	MOMENTS	CHAR. FN.
1. Normal	$n(x-\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$	μ = mean σ^2 = variance	$\exp(\iota t\mu-\sigma^2t^2/2)$ Note 13
$-\infty < x < +\infty$	$\sigma > 0$		
2. Standard Normal	$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$	$\mu = 0$	$e^{-t^2/2}$
$-\infty < x < +\infty$		$\sigma^2 = 1$	
3. Chi-Square	$\chi^2(x;k) = \frac{x^{(k/2)-1}.e^{x/2}}{\Gamma(k/2)2^{k/2}}$	$\mu = k$	$(1-\iota t/2)^{-k/2}$
$0 < x < +\infty$	$k = 1,2,...$	$\sigma^2 = 2k$	Note 14
4. F-distribution	F(x;k,n)	$\mu = \frac{n}{n-2}$ if $n > 2$	Note 15
$0 < x < +\infty$	k,n positive integers	$\sigma^2 = \frac{2n^2(k+n-2)}{k(n-2)^2(n-4)}$ if $n > 4$	
5. t-distribution	$\frac{\Gamma(\frac{k+1}{2})(1+x^2/k)^{-(k+1)/2}}{\sqrt{k}\Gamma(\frac{1}{2})\Gamma(\frac{1+2k}{2})}$	$\mu = 0$ if $k > 1$	Note 16
$-\infty < x < +\infty$		$\sigma^2 = \frac{k}{k-2}$ if $k > 2$	

NON-CENTRAL DISTRIBUTIONS ASSOCIATED WITH THE NORMAL

NAME & DOMAIN	DENSITY	MOMENTS	CHAR. FN.
1. Noncentral Chi-Squared $x > 0$	$\chi^2(x;k,\delta)$ k pos. integer $\delta \geq 0$	$\mu = k+\delta$ $\sigma^2 = 2(k+2\delta)$	Note 17
2. Noncentral F-distribution $x > 0$	$F(x;k,n,\delta)$ k,n pos. integers $\delta \geq 0$	$\mu = \frac{n(k+\delta)}{k(n-2)}$ if $n > 2$ $\sigma^2 = \frac{2(n/k)^2(k+\delta)^2+(k+2\delta)(n-2)}{(n-2)^2(n-4)}$ if $n > 4$	Note 18
3. Noncentral t-distribution	$t(x;k,\lambda)$ k pos. integer	$\mu = \sqrt{k/2}\,\frac{\Gamma((k-1)/2)\lambda}{\Gamma(k/2)}$ if $k > 1$ $\sigma^2 = (1+\lambda^2)k/(k-2) - \mu^2$ if $k > 2$	Note 19

NOTES TO THE TABLES

1. $\mu \equiv E_X X$ (the mean), and $\sigma^2 = E_X(X-\mu)^2$ (the variance). The density is often denoted b(k;n,p). The moment generating function is $(1-p+pe^t)^n$.

2. The characteristic and moment generating functions are complicated.

3. The characteristic function is $p/\left(1-(1-p)e^{\iota t}\right)$ and the moment generating function is $p/\left(1-(1-p)e^{t}\right)$, defined for $t < -\ln(1-p)$.

4. The moment generating function is $e^{\lambda(e^t-1)}$, defined for all t.

5. The characteristic function is $p^r/\left(1-(1-p)e^{\iota t}\right)^r$, and the moment generating function is $p^r/\left(1-(1-p)e^{t}\right)^r$, defined for $t < -\ln(1-p)$.

6. The moment generating function is $(e^{bt} - e^{at})/(b-a)t$, defined for all t.

7. The moment generating function is $1/(1 - \lambda t)$, defined for $t < 1/\lambda$.

8. The characteristic and moment generating functions are complicated.

9. For $a > 0$, $\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}dx$ is the gamma function. If a is an integer, $\Gamma(a) = (a-1)!$.

10. The moment generating function does not exist.

11. The moment generating function is $e^{at}\Gamma(1 - tb)$ for $t < 1/b$.

12. The moment generating function is $e^{at}\pi bt/\sin(\pi bt)$ for $|t| < 1/2b$.

13. The density is often denoted $n(x-\mu,\sigma^2)$, and the cumulative distribution referred to as $N(x-\mu,\sigma^2)$, or simply $N(\mu,\sigma^2)$. The moment generating function is $e^{at+b^2t^2/2}$, defined for all t. The standard normal density is often denoted $\phi(x)$, and the standard normal CDF is denoted $\Phi(x)$. The general normal and standard normal formulas are related by $n(x-\mu,\sigma^2) = \phi((x-\mu)/\sigma)/\sigma$ and $N(x-\mu,\sigma^2) = \Phi((x-\mu)/\sigma)$.

14. The moment generating function is $(1-tb)^{-a}$ for $t < 1/b$. The Chi-Square distribution with parameter k (≡ degrees of freedom) is the distribution of the sum of squares of k independent standard normal random variables. The Chi-Square density is the same as the gamma density with $b = 2$ and $a = k/2$.

15. The F-distribution is the distribution of the expression nU/kV, where U is a random variable with a Chi-square distribution with parameter k, and V is an independent random variable with a Chi-square distribution with parameter n. The density is $\frac{\Gamma(\frac{k+n}{2})}{\Gamma(\frac{k}{2})\Gamma(\frac{n}{2})}\cdot\frac{k^{k/2}n^{n/2}x^{k/2-1}}{(n+kx)^{(k+n)/2}}$. For $n \leq 2$, the mean does not exist, and for $n \leq 4$, the variance does not exist. The characteristic and moment generating functions are complicated.

16. If Y is standard normal and Z is independently Chi-squared distributed with parameter k, then $Y/\sqrt{Z/k}$ has a T-Distribution with parameter k (≡ degrees of freedom). The characteristic function is complicated; the moment generating function does not exist.

17. The Noncentral Chi-square is the distribution of the sum of squares of k independent normal random variables, each with variance one, and with means whose squares sum to δ. The Noncentral Chi-Square density is a Poisson mixture of (central) Chi-square densities, $\sum_{j=0}^{\infty} [e^{-\delta/2}(\delta/2)^j/j!]\chi^2(x;k+2j)$.

18. The Non-central F-distribution has a density that is a Poisson mixture of rescaled (central) F-distributed densities, $\sum_{j=0}^{\infty} [e^{-\delta/2}(\delta/2)^j/j!]\frac{k}{k+2j}F(\frac{kx}{k+2j};k+2j,n)$. It is the distribution of the expression nU′/kV, where U′ is a Noncentral Chi-Squared random variable with parameters k and δ, and V is an independent central Chi-Squared distribution with parameter n.

19. If Y is standard normal and Z is independently Chi-squared distributed with parameter k, then $(Y+\lambda)/\sqrt{(Z/k)}$ has a Noncentral T-Distribution with parameters k and λ. The density is a Poisson mixture of scaled Beta distributed densities, $\sum_{j=0}^{\infty} [e^{-\lambda^2/2}(\lambda^2/2)^j/j!]\frac{xk}{(k+x^2)^2}B(\frac{k}{k+x^2},\frac{k}{2},\frac{1+2j}{2})$. The square of a Noncentral T-Distributed random variable has a Noncentral F-Distribution with parameters 1, k, and $\delta = \lambda^2$.