CHAPTER 3. A REVIEW OF PROBABILITY THEORY

3.1. 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 **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. Sections 3.2-3.4 in this Chapter contain a few crucial definitions, for events, probabilities, conditional probabilities, and statistical independence. They also contain a treatment of measurability, the theory of integration, and probability on product spaces that is needed mostly for more advanced topics in econometrics. Therefore, readers who do not have a good background in mathematical analysis may find it useful to concentrate on the definitions and examples in these sections, and postpone study of the more mathematical material until it is needed.

3.2. EVENT FIELDS AND INFORMATION

3.2.1. 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 $\mathcal{F}$. This family is assumed to have the following properties:

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

(ii) If event **A** is in $\mathcal{F}$, then the event "not **A**", denoted $\mathbf{A}^c$ or $\mathbf{S}\backslash\mathbf{A}$, is in $\mathcal{F}$.

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

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

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

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

(vi) If $\mathbf{A}_1,\mathbf{A}_2,...$ is a countable sequence of events in $\boldsymbol{F}$ that is *monotone decreasing* (i.e., $\mathbf{A}_1 \supseteq \mathbf{A}_2 \supseteq ...$), then its limit, also denoted $A_i \searrow A_0$, is also in $\boldsymbol{F}$. Similarly, if a sequence in $\boldsymbol{F}$ is *monotone increasing* (i.e., $\mathbf{A}_1 \subseteq \mathbf{A}_2 \subseteq ...$), then its limit $\mathbf{A}_0 = \bigcup_{i=1}^{\infty} A_i$, is also in $\boldsymbol{F}$

(vii) The empty event $\boldsymbol{\varphi}$ is in $\boldsymbol{F}$.

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 $\mathbf{S}$ = {HH,HT,TH,TT}. Let $\boldsymbol{F}$ be the class of all possible subsets of $\mathbf{S}$; $\boldsymbol{F}$ has 2^4 members.

Example 2. [Coin toss until a tail] A coin is tossed until a tail appears. The sample space is $\mathbf{S}$ = {T, HT, HHT, HHHT,...}. In this example, the sample space is infinite, but countable. Let $\boldsymbol{F}$ be the σ-field generated by the finite subsets of $\mathbf{S}$. This σ-field contains events such as "At most ten heads", and also, using the monotone closure property (vi) above, 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 $\boldsymbol{F}$ 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 must exist, because the cardinality of the class of all possible subsets of $\mathbf{S}$ is greater than the cardinality of $\boldsymbol{F}$.

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

Example 4. [S&P stock index on successive days] The set of states of Nature is the Cartesian product of the set of calues on day one and the set of values on day 2, $\mathbf{S} = \mathbb{R}_+ \times \mathbb{R}_+$ (also denoted $\mathbb{R}_+^{\,2}$). Take the σ-field of events to be the product of the one-dimensional σ-fields, $\boldsymbol{F} = \boldsymbol{B}_1 \otimes \boldsymbol{B}_2$, where "⊗" denotes an operation that forms the smallest σ-field containing all sets of the form $\mathbf{A} \times \mathbf{C}$ with $\mathbf{A} \in \boldsymbol{B}_1$ and $\mathbf{C} \in \boldsymbol{B}_2$. In this example, $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ are identical copies of the Borel σ-field on $\mathbb{R}_+$. Assume that the index was normalized to be one at the beginning of the previous year. Examples of events in $\boldsymbol{F}$ are "below 1 on day 1", "at least 2 on both days", and "higher on the second day than the first day". The operation "⊗" is different than the cartesian product "×", where $\boldsymbol{B}_1 \times \boldsymbol{B}_2$ is the family of all rectangles $\mathbf{A}\times\mathbf{C}$ formed from $\mathbf{A} \in \boldsymbol{B}_1$ and $\mathbf{C} \in \boldsymbol{B}_2$. This family is not itself a σ-field, but the σ-field that it generates is $\boldsymbol{B}_1\otimes\boldsymbol{B}_2$. For example, the event "higher on the second day than the first day" is not a rectangle, but is obtained as a monotone limit of rectangles.

In the first example, the σ-field consisted of <u>all</u> 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 with sensible properties on all subsets of the sample space. Restricting the definition of probabilities to appropriately chosen σ-fields solves this problem.

3.2.2. It is possible that more than one σ-field of subsets is defined for a particular sample space **S**. If $\boldsymbol{A}$ is an arbitrary collection of subsets of **S**, then the smallest σ-field that contains $\boldsymbol{A}$ is said to be the σ-field *generated* by $\boldsymbol{A}$. It is sometimes denoted $\sigma(\boldsymbol{A})$. If $\boldsymbol{F}$ and $\boldsymbol{G}$ are both σ-fields, and $\boldsymbol{G} \subseteq \boldsymbol{F}$, then $\boldsymbol{G}$ is said to be a *sub-field* of $\boldsymbol{F}$, and $\boldsymbol{F}$ is said to *contain more information* or *refine* $\boldsymbol{G}$. It is possible that neither $\boldsymbol{F} \subseteq \boldsymbol{G}$ nor $\boldsymbol{G} \subseteq \boldsymbol{F}$. The intersection $\boldsymbol{F}\cap\boldsymbol{G}$ of two σ-fields is again a σ-field that contains the *common information* in $\boldsymbol{F}$ and $\boldsymbol{G}$. Further, the intersection of an arbitrary countable or uncountable collection of σ-fields is again a σ-field. The union $\boldsymbol{F}\cup\boldsymbol{G}$ of two σ-fields is not necessarily a σ-field, but there is always a smallest σ-field that refines both $\boldsymbol{F}$ and $\boldsymbol{G}$, which is simply the σ-field $\sigma(\boldsymbol{F}\cup\boldsymbol{G})$ generated by the sets in the union of $\boldsymbol{F}$ and $\boldsymbol{G}$, or put another way, the intersection of all σ-fields that contain both $\boldsymbol{F}$ and $\boldsymbol{G}$.

Example 1. (continued) Let $\boldsymbol{F}$ denote the σ-field of all subsets of **S**. Another σ-field is $\boldsymbol{G}$ = {**φ**,**S**,{HT,HH},{TT,TH}}, containing events with information only on the outcome of the first coin toss. Yet another σ-field contains the events with information only on the number of heads, but not their order, $\boldsymbol{H}$ = {**φ**,**S**,{HH},{TT},{HT,TH},{HH,TT},{HT,TH,TT},{HH,HT,TH}}. Then, $\boldsymbol{F}$ contains more information than $\boldsymbol{G}$ or $\boldsymbol{H}$. The intersection $\boldsymbol{G}\cap\boldsymbol{H}$ is the "no information" σ-field {**φ**,**S**}. The union $\boldsymbol{G}\cup\boldsymbol{H}$ is not a σ-field, and the σ-field $\sigma(\boldsymbol{G}\cup\boldsymbol{H})$ that it generates is $\boldsymbol{F}$. This can be verified constructively (in this finite S case) by building up $\sigma(\boldsymbol{G}\cup\boldsymbol{H})$ by forming intersections and unions of members of $\boldsymbol{G}\cup\boldsymbol{H}$, but is also obvious since knowing the outcome of the first toss and knowing the total number of heads reveals full information on both tosses.

Example 3. (continued) Let $\boldsymbol{F}$ denote the Borel σ-field. Then $\boldsymbol{G}$ = {**φ**,**S**,(1,∞),(-∞,1]} and $\boldsymbol{D}$ = {**φ**,**S**,{-∞,2),[2,∞)} are both σ-fields, the first corresponding to the ability to observe whether the index is above 1, the second corresponding to the ability to tell whether it is above 2. For shorthand, let a = $(-\infty,1]$, b = $(-\infty,2]$, c = $(1,+\infty)$, d = $(2,+\infty)$, and e = $(1,2]$. Neither $\boldsymbol{G}$ or $\boldsymbol{D}$ contains the other, both are contained in $\boldsymbol{F}$, and their intersection is the "no information" σ-field $\{\boldsymbol{\varphi},\mathbf{S}\}$. The σ-field generated by their union, corresponding to the ability to tell if the index is in a,e, or d, is $\sigma(\boldsymbol{G}\cup\boldsymbol{D}) = \{\boldsymbol{\varphi},\mathbf{S},a,b,c,d,e,a\cup d\}$.

An element $\mathbf{B}$ in a σ-field $\boldsymbol{G}$ of subsets of $\mathbf{S}$ is an *atom* if the only set in $\boldsymbol{G}$ that is a proper subset of $\mathbf{B}$ is the empty set $\boldsymbol{\varphi}$. In the last example, $\boldsymbol{D}$ has atoms b and d, and the atoms of $\sigma(\boldsymbol{G}\cup\boldsymbol{D})$ are a, d, and e, but not b = a∪e or c = e∪d. The atoms of the Borel σ-field are the individual real numbers. An economic interpretation of this concept that if the σ-field defining the common information of two economic agents contains an atom, then a contingent contract between them must have the same realization no matter what state of Nature within this atom occurs.

3.3. PROBABILITY

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

(i) $P(\mathbf{A}) \geq 0$ for all $\mathbf{A} \in \boldsymbol{F}$.
(ii) $P(\mathbf{S}) = 1$.
(iii) [Countable Additivity] If $\mathbf{A}_1, \mathbf{A}_2,...$ is a finite or countable sequence of events in $\boldsymbol{F}$ that are mutually exclusive (i.e., $\mathbf{A}_i\cap\mathbf{A}_j = \boldsymbol{\varphi}$ for all $i \neq j$), then $P(\bigcup_{i=1}^{\infty} \mathbf{A}_i) = \sum_{i=1}^{\infty} P(\mathbf{A}_i)$.

With conditions (i)-(iii), P has the following additional intuitive properties of a probability when $\mathbf{A}$ and $\mathbf{B}$ are events in $\boldsymbol{F}$:

(iv) $P(\mathbf{A}) + P(\mathbf{A}^c) = 1$.
(v) $P(\boldsymbol{\varphi}) = 0$.
(vi) $P(\mathbf{A}\cup\mathbf{B}) = P(\mathbf{A}) + P(\mathbf{B}) - P(\mathbf{A}\cap\mathbf{B})$.
(vii) $P(\mathbf{A}) \geq P(\mathbf{B})$ when $\mathbf{B} \subseteq \mathbf{A}$.
(viii) If $\mathbf{A}_i$ in $\boldsymbol{F}$ is monotone decreasing to $\boldsymbol{\varphi}$ (denoted $\mathbf{A}_i \searrow \boldsymbol{\varphi}$), then $P(\mathbf{A}_i) \to 0$.
(ix) If $\mathbf{A}_i \in \boldsymbol{F}$, not necessarily disjoint, then $P(\bigcup_{i=1}^{\infty} \mathbf{A}_i) \leq \sum_{i=1}^{\infty} P(\mathbf{A}_i)$.
(x) If $\{\mathbf{A}_i\}$ is a finite or countable *partition* of $\mathbf{S}$ (i.e., the events $\mathbf{A}_i \in \boldsymbol{F}$ are mutually exclusive and exhaustive, or $\mathbf{A}_i\cap\mathbf{A}_j = \boldsymbol{\varphi}$ for all $i \neq j$ and $\bigcup_{i=1}^{\infty} \mathbf{A}_i = \mathbf{S}$), then $P(\mathbf{B}) = \sum_{i=1}^{\infty} P(\mathbf{B}\cap\mathbf{A}_i)$.

The triplet (**S**,***F***,P) consisting of a measurable space (**S**,***F***) and a probability measure P is called a *probability space*.

Example 1 (continued). Consider the σ-field ***H*** containing information on the number of heads, but not their order. The table below gives three functions P_1, P_2, P_3 defined on ***H***. All satisfy properties (i) and (ii) for a probability. Functions P_2 and P_3 also satisfy (iii), and are probabilities, but P_1 violates (iii) since $P_1(\{HH\}\cup\{TT\}) \neq P_1(\{HH\}) + P_1(\{TT\})$. The probability P_2 is generated by fair coins, and the probability P_3 by one fair coin and one biased coin.

	φ	S	HH	TT	HT,TH	HH,TT	HT,TH,TT	HH,HT,TH
P_1	0	1	1/3	1/3	1/2	1/2	2/3	2/3
P_2	0	1	1/4	1/4	1/2	1/2	3/4	3/4
P_3	0	1	1/3	1/6	1/2	1/2	2/3	5/6

3.3.2. If $\mathbf{A} \in \boldsymbol{F}$ has P(**A**) = 1, then **A** is said to occur *almost surely* (a.s.), or *with probability one* (w.p.1). If $\mathbf{A} \in \boldsymbol{F}$ has P(**A**) = 0, then **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 2. (continued) If the coin is fair, then the probability of k-1 heads followed by a tail is $1/2^k$. Use the geometric series formulas in 2.1.10 to verify that the probability of "At most 3 heads" is 15/16, of "Ten or more heads" is $1/2^{10}$, and of "an even number of heads" is 2/3.

Example 3. (continued) Consider the function P defined on open sets $(s,\infty) \in \mathbb{R}_+$ by $P((s,\infty)) = e^{-s/2}$. This function maps into the unit interval. 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 {1}, is in ***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.

3.3.3. Often a measurable space (**S**,***F***) will have an associated *measure* ν that is a countably additive function from ***F*** into the nonnegative real line; i.e., $\nu(\cup_{i=1}^{\infty} \mathbf{A}_i) = \sum_{i=1}^{\infty} \nu(\mathbf{A}_i)$ for any sequence of disjoint $\mathbf{A}_i \in \boldsymbol{F}$. The measure is *positive* if $\nu(\mathbf{A}) \geq 0$ for all $\mathbf{A} \in \boldsymbol{F}$; we will consider only positive measures. The measure ν is *finite* if $|\nu(\mathbf{A})| \leq M$ for some constant M and all $\mathbf{A} \in \boldsymbol{F}$, and

σ-*finite* if $\boldsymbol{F}$ contains a countable partition $\{\mathbf{A}_i\}$ of **S** such that the measure of each partition set is finite; i.e., $\nu(\mathbf{A}_i) < +\infty$. The measure ν may be a probability, but more commonly it is a measure of "length" or "volume". For example, it is common when the sample space **S** is the countable set of positive integers to define ν to be *counting measure* with $\nu(\mathbf{A})$ equal to the number of points in **A**. When the sample space **S** is the real line, with the Borel σ-field $\boldsymbol{B}$, 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 set **A** is said to be of *ν-measure zero* if $\nu(\mathbf{A}) = 0$. A property that holds except on a set of measure zero is said to hold *almost everywhere* (a.e.). It will sometimes be useful to talk about a σ-finite measure space $(\mathbf{S},\boldsymbol{F},\mu)$ where μ is positive and σ-finite and may either be a probability measure or a more general counting or length measure such as Lebesgue measure.

3.3.4. Suppose f is a real-valued function on a σ-finite measure space $(\mathbf{S},\boldsymbol{F},\mu)$. This function is *measurable* if $f^{-1}(\mathbf{C}) \in \boldsymbol{F}$ for each open set **C** in the real line. A measurable function has the property that its contour sets of the form $\{s\in\mathbf{S}|a\le f(s)\le c\}$ are contained in $\boldsymbol{F}$. This implies that if $\mathbf{B} \in \boldsymbol{F}$ is an atom, then f(s) must be constant for all $s \in \mathbf{B}$.

The integral of measurable f on a set $\mathbf{A} \in \boldsymbol{F}$, denoted $\int_A f(s)\cdot\mu(ds)$, is defined for $\mu(\mathbf{A}) < +\infty$ as the limit as $n \to \infty$ of sums of the form $\sum_{k=-\infty}^{\infty} (k/n)\cdot\mu(\mathbf{C}_{kn})$, where $\mathbf{C}_{kn}$ is the set of states of Nature in **A** for which f(s) is contained in the interval (k/n,(k+1)/n]. A finite limit exists if $\sum_{k=\infty}^{\infty} |k/n|\cdot\mu(\mathbf{C}_{kn}) < +\infty$, in which case f is said to be *integrable* on **A**. Let $\{\mathbf{A}_i\} \in \boldsymbol{F}$ be a countable partition of **S** with $\mu(\mathbf{A}_i) < +\infty$, guaranteed by the σ-finite property of μ. The function f is integrable on a general set $\mathbf{A} \in \boldsymbol{F}$ if it is integrable on $\mathbf{A}\cap\mathbf{A}_i$ for each i and if $\int_A |f(s)|\cdot\mu(ds) = \lim_{n\to\infty} \sum_{i=1}^{n} \int_{A\cap A_i} |f(s)|\cdot\mu(ds)$ exists, and simply *integrable* if it is integrable for $\mathbf{A} = \mathbf{S}$. In general, the measure μ can have point masses (at atoms), or continuous measure, or both, so that the notation for integration with respect to μ includes sums and mixed cases. The integral $\int_A f(s)\mu(ds)$ will sometimes be denoted $\int_A f(s)d\mu$, or in the case of Lebesgue measure, $\int_A f(s)ds$.

3.3.5. For a σ-finite measure space $(\mathbf{S},\boldsymbol{F},\mu)$, define $\mathbf{L}_q(\mathbf{S},\boldsymbol{F},\mu)$ for $1 \le q < +\infty$ to be the set of measurable real-valued functions on **S** with the property that $|f|^q$ is integrable, and define $\|f\|_q =$

$[\int |f(s)|^q \mu(ds)]^{1/q}$ to be the *norm* of f. Then, $\mathbf{L}_q(\mathbf{S},\boldsymbol{F},\mu)$ is a linear space, since linear combinations of integrable functions are again integrable. This space has many, but not all, of familiar properties of finite-dimensional Euclidean space. The set of all linear functions on the space $\mathbf{L}_q(\mathbf{S},\boldsymbol{F},\mu)$ for $q > 1$ is the space $\mathbf{L}_r(\mathbf{S},\boldsymbol{F},\mu)$, where $1/r = 1 - 1/q$. This follows from an application of Holder's inequality, which generalizes from finite vector spaces to the condition

$$f \in \mathbf{L}_q(\mathbf{S},\boldsymbol{F},\mu) \text{ and } g \in \mathbf{L}_r(\mathbf{S},\boldsymbol{F},\mu) \text{ with } q^{-1} + r^{-1} = 1 \text{ imply } \int |f(s)\cdot g(s)|\, \mu(ds) \le \|f\|_q\cdot\|g\|_r.$$

The case $q = r = 2$ gives the Cauchy-Schwartz inequality in general form. This case arises often in statistics, with the functions f interpreted as random variables and the norm $\|f\|_2$ interpreted as a quadratic mean or variance.

3.3.6. There are three important concepts for the limit of a sequence of functions $f_n \in \mathbf{L}_q(\mathbf{S},\boldsymbol{F},\mu)$. First, there is *convergence in norm*, or strong convergence: f is a limit of f_n if $\|f_n - f\|_q \to 0$. Second, there is *convergence in μ-measure*: f is a limit of f_n if $\mu(\{s\in S|\ |f_n(s) - f(s)| > \varepsilon\}) \to 0$ for each $\varepsilon > 0$. Third, there is *weak convergence*: f is a limit of f_n if $\int (f_n(s) - f(s))\cdot g(s)\, \mu(ds) \to 0$ for each $g \in \mathbf{L}_r(\mathbf{S},\boldsymbol{F},\mu)$ with $1/r = 1 - 1/q$. The following relationship holds between these modes of convergence:

$$\text{Strong Convergence} \Longrightarrow \text{Weak Convergence} \Longrightarrow \text{Convergence in } \mu\text{-measure}$$

An example shows that convergence in μ-measure does not in general imply weak convergence: Consider $\mathbf{L}_2(\mathbf{[0,1]},\boldsymbol{B},\mu)$ where $\boldsymbol{B}$ is the Borel σ-field and μ is Lebesgue measure. Consider the sequence $f_n(s) = \cdot n\cdot\mathbf{1}(s \le 1/n)$. Then $\mu(\{s\in S|\ |f_n(s)| > \varepsilon\}) = 1/n$, so that f_n converges in μ-measure to zero, but for $g(s) = s^{-1/3}$, one has $\|g\|_2 = 3^{1/2}$ and $\int f_n(s)g(s)\, \mu(ds) = 3n^{1/3}/2$ divergent. Another example shows that weak convergence does not in general imply strong convergence: Consider $\mathbf{S} = \{1,2,...\}$ endowed with the σ-field generated by the family of finite sets and the measure μ that gives weight $k^{-1/2}$ to point k. Consider $f_n(k) = \cdot n^{1/4}\cdot\mathbf{1}(k = n)$. Then $\|f_n\|_2 \cdot= 1$. If g is a function for which $\sum_{k=1}^{\infty} f_n(k)g(k)\mu(\{k\}) = g(n)\cdot n^{1/4}$ does not converge to zero, then $g(k)^2\, \mu(\{k\})$ is bounded away from zero infinitely often, implying $\|g\|_2 = \sum_{k=1}^{\infty} g(k)^2\, \mu(\{k\}) = +\infty$. Then, f_n converges weakly, but not strongly, to zero. The following theorem, which is of great importance in advanced econometrics, gives a uniformity condition under which these modes of convergence coincide.

Theorem 3.1. (Lebesgue Dominated Convergence) If g and f_n for n = 1,2,... are in $\mathbf{L}_q(\mathbf{S},\boldsymbol{F},\mu)$ for $1 \le q < +\infty$ and a σ-finite measure space $(\mathbf{S},\boldsymbol{F},\mu)$, and if $|f_n(s)| \le g(s)$ almost everywhere, then f_n converges in μ-measure to a function f if and only if $f \in \mathbf{L}_q(\mathbf{S},\boldsymbol{F},\mu)$ and $\|f_n - f\|_q \to 0$.

One application of this theorem is a result for interchange of the order of integration and differentiation. Suppose $f(\cdot,t) \in \mathbf{L}_q(\mathbf{S},\boldsymbol{F},\mu)$ for t in an open set $\mathbf{T} \subseteq \mathbb{R}^n$. Suppose f is *differentiable*, meaning that there exists a function $\nabla_t f(\cdot,t) \in \mathbf{L}_q(\mathbf{S},\boldsymbol{F},\mu)$ for $t \in \mathbf{T}$ such that if $t+h \in \mathbf{T}$ and $h \ne 0$, then the remainder function $r(s,t,h) = [f(s,t+h) - f(s,t) - \nabla_t f(\cdot,t)\cdot h]/|h| \in \mathbf{L}_q(\mathbf{S},\boldsymbol{F},\mu)$ converges in μ-measure to zero as $h \to 0$. Define $F(t) = \int f(s,t)\mu(ds)$. If there exists $g \in \mathbf{L}_q(\mathbf{S},\boldsymbol{F},\mu)$ which dominates the remainder function (i.e., $|r(s,t,h)| \le g(s)$ a.e.), then Theorem 3.1 implies $\lim_{h\to 0}\|r(\cdot,t,h)\|_q = 0$, and F(t) is differentiable and satisfies $\nabla_t F(t) = \int \nabla_t f(s,t)\mu(ds)$.

A finite measure P on $(\mathbf{S},\boldsymbol{F})$ is *absolutely continuous* with respect to a measure ν if $\mathbf{A} \in \boldsymbol{F}$ and $\nu(\mathbf{A}) = 0$ imply $P(\mathbf{A}) = 0$. If P is a probability measure that is absolutely continuous with respect to the measure ν, then an event of measure zero occurs w.p.0, and an event that is true almost everywhere occurs almost surely. A fundamental result from analysis is the theorem:

Theorem 3.2. (Radon-Nikodym) If a finite measure P on a measurable space $(\mathbf{S},\boldsymbol{F})$ is absolutely continuous with respect to a positive σ-finite measure ν on $(\mathbf{S},\boldsymbol{F})$, then there exists an integrable real-valued function $p \in \mathbf{L}_1(\mathbf{S},\boldsymbol{F},\nu)$ such that

$$\int_{\boldsymbol{A}} p(s)\nu(ds) = P(\mathbf{A}) \text{ for each } \mathbf{A} \in \boldsymbol{F}.$$

When P is a probability, the function p given by the theorem is nonnegative, and is called the *probability density*. An implication of the Radon-Nikodym theorem is that if a measurable space $(\mathbf{S},\boldsymbol{F})$ has a positive σ-finite measure ν and a probability measure P that is absolutely continuous with respect to ν, then there exists a density p such that for every $f \in \mathbf{L}_q(\mathbf{S},\boldsymbol{F},P)$ for some $1 \le q < +\infty$, one has $\int_{\boldsymbol{S}} f(s)P(ds) = \int_{\boldsymbol{S}} f(s)\cdot p(s)\nu(ds)$.

3.3.7. In applications where the probability space is the real line with the Borel σ-field, with a probability P such that $P((-\infty,s]) = F(s)$ is continuously differentiable, the fundamental theorem of integral calculus states that $p(s) = F'(s)$ satisfies $F(\mathbf{A}) = \int_{\boldsymbol{A}} p(s)ds$. What the Radon-Nikodym theorem does is extend this result to σ-finite measure spaces and weaken the assumption from continuous differentiability to absolute continuity. In basic econometrics, we will often characterize probabilities both in terms of the probability measure (or distribution) and the density, and will usually need only the elementary calculus version of the Radon-Nikodym result. However, it is useful in theoretical discussions to remember that the Radon-Nikodym theorem makes the connection between probabilities and densities. We give two examples that illustrate practical use of the calculus version of the Radon-Nikodym theorem.

Example 3. (continued) Given $P((s,\infty)) = e^{-s/2}$, 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. Verify by integration that the density implied by the Radon-Nikodym theorem is $p(s) = e^{-s/2}/2$.

Example 5. A probability that appears frequently in statistics is the *normal*, which is defined on $(\mathbb{R},\boldsymbol{B})$, where $\mathbb{R}$ is the real line and $\boldsymbol{B}$ the Borel σ-field, by the density $n(s-\mu,\sigma) \equiv (2\pi\sigma^2)^{-1/2}\cdot e^{-(s-\mu)^2/2\sigma^2}$, so that $P(\mathbf{A}) = \int_A (2\pi\sigma^2)^{-1/2}\cdot e^{-(s-\mu)^2/2\sigma^2}ds$. 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*.

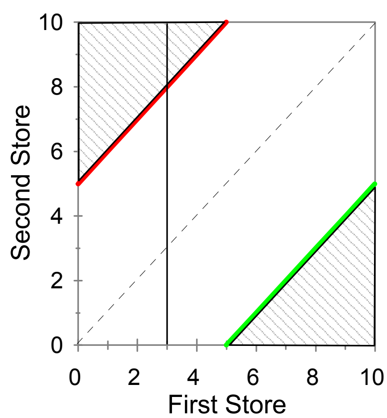
3.3.8. Consider a probability space $(\mathbf{S},\boldsymbol{F},P)$, and a σ-field $\boldsymbol{G} \subseteq \boldsymbol{F}$. If the event $\mathbf{B} \in \boldsymbol{G}$ has $P(\mathbf{B}) > 0$, then the *conditional probability* of $\mathbf{A}$ given $\mathbf{B}$ is defined as $P(\mathbf{A}|\mathbf{B}) = P(\mathbf{A}\cap\mathbf{B})/P(\mathbf{B})$. Stated another way, $P(\mathbf{A}|\mathbf{B})$ is a real-valued function on $\boldsymbol{F}\times\boldsymbol{G}$ with the property that $P(\mathbf{A}\cap\mathbf{B}) = P(\mathbf{A}|\mathbf{B})P(\mathbf{B})$ for all $\mathbf{A} \in \boldsymbol{F}$ and $\mathbf{B} \in \boldsymbol{G}$. When $\mathbf{B}$ is a finite set, the conditional probability of $\mathbf{A}$ given $\mathbf{B}$ is the ratio of sums

$$P(\mathbf{A}|\mathbf{B}) = \frac{\sum_{s\in A\cap B} P(\{s\})}{\sum_{s\in B} P(\{s\})} .$$

Example 6. On a quiz show, a contestant is shown three doors, one of which conceals a prize, and is asked to select one. Before it is opened, the host opens one of the remaining doors which he knows does not contain the prize, and asks the contestant whether she wants to keep her original selection or switch to the other remaining unopened door. Should the contestant switch? Designate the contestant's initial selection as door 1. The sample space consists of pairs of numbers ab, where a = 1,2,3 is the number of the door containing the prize and b = 2,3 is the number of the door opened by the host, with b ≠ a: S = {12,13,23,32}. The probability is 1/3 that the prize is behind each door. The conditional probability of b = 2, given a = 1, is 1/2, since in this case the host opens door 2 or door 3 at random. However, the conditional probability of b = 2, given a = 2 is zero and the conditional probability of b = 2 given a = 3 is one. Hence, P(12) = P(13) = (1/3)·(1/2), and P(23) = P(32) = 1/3. Let A = {12,13} be the event that door 1 contains the prize and B = {12,32} be the event that the host opens door 2. Then the conditional probability of A given B is P(12)/(P(12)+P(32)) = (1/6)/((1/6)+(1/3)) = 1/3. Hence, the probability of receiving the prize is 1/3 if the contestant stays with her original selection, 2/3 if she switches to the other unopened door.

Example 7. Two fast food stores are sited at random points along a street that is ten miles long. What is the probability that they are less than five miles apart? Given that the first store is located at the three mile marker, what is the probability that the second store is less than five miles away? The answers are obvious from the diagram below, in which the sample space is depicted as a rectangle of dimension 10 by 10, with the horizontal axis giving the location of the first store and the vertical axis giving the location of the second store. The shaded areas correspond to the event that the two are more than five miles apart, and the proportion of the rectangle in these areas is 1/4. Conditioned on the first store being at point 3 on the horizontal axis, the second store is located at random on a vertical line through this point, and the proportion of this line that lies in the shaded area is 1/5. Let x be the location of the first store, y the location of the second. The conditional probability of the event that $|x - y| > 5$, given x, is $|x-5|/10$. This could have been derived by forming the probability of the event $|x - y| > 5$ and $c < x < c+\delta$ for a small positive δ, taking the ratio of this probability to the probability of the event $c < x < c+\delta$ to obtain the conditional probability of the event $|x - y| > 5$ given $c < x < c+\delta$, and taking the limit $\delta \to 0$.

Location of Fast Food Stores



The idea behind conditional probabilities is that one has partial information on 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., ***F*** is the field of events which distinguish outcomes in both the past and the future, and a subfield ***G*** contains events which distinguish only past outcomes. A conditional probability P(**A**|**B**) defined for **B** ⊆ ***G*** can be interpreted for fixed **A** as a function from ***G*** into [0,1]. To emphasize this, conditional probabilities are sometimes written P(**A**|***G***), and ***G*** 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 ***G*** = {**φ**,**S**,{HT,HH},{TT,TH}}, so that events in ***G*** describe the outcome of the first coin toss, then P(HH|{HH,HT}) = P(HH)/(P(HH)+P(HT)) = ½ 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*. If ***G*** = {**φ**,**S**,{HT,TH},{HH},{TT},$\{HH\}^c$,$\{TT\}^c$}, the family of events that determine the number of heads that occur in two tosses without regard for order, then the conditional probability of heads on the first toss, given at least one head, is P({HT,HH}|$\{TT\}^c$) = (P(HT)+P(HH))/(1-P(TT))= 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 3. (continued) Suppose ***G*** = {**φ**,**S**,(1,∞),(-∞,1]} is the σ-field corresponding to the event that the index exceeds 1, and let ***B*** denote the Borel σ-field containing all the open intervals. The unconditional probability P((s,∞)) = $e^{-s/2}$ implies P((1,∞)) = $e^{-1/2}$ = 0.6065. The conditional probability of (2,∞) given (1,∞) satisfies P(((2,∞)|(1,∞)) = P((1,∞)∩(2,∞))/P((1,∞)) = $e^{-1}/e^{-1/2}$ = 0.6065 > P((2,∞)) = 0.3679. The conditional and unconditional probabilities are not the same, so that the conditioning event provides information on the probability of (2,∞).

For a probability space (**S**,***F***,P), suppose $\mathbf{A}_1,...,\mathbf{A}_k$ is a finite partition of **S**; i.e., $\mathbf{A}_i \cap \mathbf{A}_j = \boldsymbol{\varphi}$ and $\bigcup_{i=1}^{k} \mathbf{A}_i = \mathbf{S}$. The partition generates a finite field $\boldsymbol{G} \subseteq \boldsymbol{F}$. From the formula P(**A**∩**B**) = P(**A**|**B**)P(**B**) satisfied by conditional probabilities, one has for an event $\mathbf{C} \in \boldsymbol{F}$ the formula

$$P(\mathbf{C}) = \sum_{i=1}^{k} P(\mathbf{C}|\mathbf{A}_i)\cdot P(\mathbf{A}_i).$$

This is often useful in calculating probabilities in applications where the conditional probabilities are available.

3.3.9. In a probability space (**S**,***F***,P), the concept of a conditional probability P(**A**|**B**) of $\mathbf{A} \in \boldsymbol{F}$ given an event **B** in a σ-field $\boldsymbol{G} \subseteq \boldsymbol{F}$ can be extended to cases where P(**B**) = 0 by defining P(**A**|**B**) as the limit of P(**A**|$\mathbf{B}_i$) for sequences $\mathbf{B}_i \in \boldsymbol{G}$ that satisfy P($\mathbf{B}_i$) > 0 and $\mathbf{B}_i \to \mathbf{B}$, provided the limit exists. If we fix **A**, and consider P(**A**∩**B**) as a measure defined for $\mathbf{B} \in \boldsymbol{G}$, this measure obviously satisfies P(**A**∩**B**) ≤ P(**B**), so that it is absolutely continuous with respect to P(B). Then, Theorem 3.2

implies that there exists a function $P(\mathbf{A}|\cdot) \in \mathbf{L}_1(\mathbf{S},\boldsymbol{G},P)$ such that $P(\mathbf{A}\cap\mathbf{B}) = \int_{\boldsymbol{B}} P(\mathbf{A}|s)\cdot P(ds)$. We have written this function as if it were a conditional probability of A given the "event" {s}, and it can be given this interpretation. If $\mathbf{B} \in \boldsymbol{G}$ is an atom, then the measurability of $P(\mathbf{A}|\cdot)$ with respect to $\boldsymbol{G}$ requires that it be constant for $s \in \mathbf{B}$, so that $P(\mathbf{A}\cap\mathbf{B}) = P(\mathbf{A}|s)\cdot P(\mathbf{B})$ for any $s \in \mathbf{B}$, and we can instead write $P(\mathbf{A}\cap\mathbf{B}) = P(\mathbf{A}|\mathbf{B})\cdot P(\mathbf{B})$, satisfying the definition of conditional probability even if $P(\mathbf{B}) = 0$.

Example 4. (continued) Consider $\boldsymbol{F} = \boldsymbol{B}\otimes\boldsymbol{B}$, the product Borel σ-field on $\mathbb{R}_+$, and $\boldsymbol{G} = \boldsymbol{B}\otimes\{\boldsymbol{\varphi}.\mathbb{R}_+\}$, the σ-field corresponding to having complete information on the level of the index on the first day and no information on the second day. Suppose $P((s,\infty)\times(t,\infty)) = 2/(1+e^{s+t})$. This is a probability on these open intervals that extends to $\boldsymbol{F}$; verifying this takes some work. The conditional probability of $(s,\infty)\times(t,\infty)$ given the event $(r,\infty)\times(0,\infty) \in \boldsymbol{G}$ and $s \le r$ equals $P((r,\infty)\times(t,\infty))$ divided by $P((r,\infty)\times(0,\infty))$, or $(1+e^r)/(1+e^{r+t})$. The conditional probability of $(s,\infty)\times(t,\infty)$ given the event $(r,r+\delta)\times(0,\infty) \in \boldsymbol{G}$ and $s \le r$ is $[1/(1+e^{r+t}) - 1/(1+e^{r+\delta+t})]/[1/(1+e^r) - 1/(1+e^{r+\delta})]$. The limit of this expression as $\delta \to 0$ is $e^r\cdot(1+e^r)^2/(1+e^{r+t})^2 = P((s,\infty)\times(t,\infty)|\{r\}\times(0,\infty))$; this function of r is also the integrand that satisfies Theorem 3.2. Note that $P((s,\infty)\times(t,\infty)|\{r\}\times(0,\infty)) \neq P((s,\infty)\times(0,\infty)) = 1/(1+e^s)$, so that the conditioning event conveys information about the probability of $(s,\infty)\times(t,\infty)$.

3.4. STATISTICAL INDEPENDENCE AND REPEATED TRIALS

3.4.1. Consider a probability space $(\mathbf{S},\boldsymbol{F},P)$. Events **A** and **C** in $\boldsymbol{F}$ are *statistically independent* if $P(\mathbf{A}\cap\mathbf{C}) = P(\mathbf{A})\cdot P(\mathbf{C})$. From the definition of conditional probability, if **A** and **C** are statistically independent and $P(\mathbf{A}) > 0$, then $P(\mathbf{C}|\mathbf{A}) = P(\mathbf{A}\cap\mathbf{C})/P(\mathbf{A}) = P(\mathbf{C})$. Thus, when **A** and **C** are statistically independent, knowing that **A** occurs is unhelpful in calculating the probability that **C** occurs. The idea of statistical independence of events has an exact analogue in a concept of statistical independence of subfields. Let $\boldsymbol{A} = \{\boldsymbol{\varphi},\mathbf{A},\mathbf{A}^c,\mathbf{S}\}$ and $\boldsymbol{C} = \{\boldsymbol{\varphi},\mathbf{C},\mathbf{C}^c,\mathbf{S}\}$ be the subfields of $\boldsymbol{F}$ generated by **A** and **C**, respectively. Verify as an exercise that if **A** and **C** are statistically independent, then so are any pair of events $\mathbf{A}' \in \boldsymbol{A}$ and $\mathbf{C}' \in \boldsymbol{C}$. Then, one can say that the subfields $\boldsymbol{A}$ and $\boldsymbol{C}$ are statistically independent. One can extend this idea and talk about statistical independence in a collection of subfields. Let **N** denote an index set, which may be finite, countable, or non-countable. Let $\boldsymbol{F}_i$ denote a σ-subfield of $\boldsymbol{F}$ ($\boldsymbol{F}_i \subseteq \boldsymbol{F}$) for each $i \in \mathbf{N}$. The subfields $\boldsymbol{F}_i$ are *mutually statistically independence* (MSI) if and only if $P(\bigcap_{j\in\boldsymbol{K}} \mathbf{A}_j) = \prod_{j\in\boldsymbol{K}} P(\mathbf{A}_j)$ for all finite $\mathbf{K} \subseteq \mathbf{N}$ and $\mathbf{A}_j \in \boldsymbol{F}_j$ for $j \in \mathbf{K}$. As in the case of statistical independence between two events (subfields), the concept of MSI can be stated in terms of conditional probabilities: $\boldsymbol{F}_i$ for $i \in \mathbf{N}$ are mutually statistically independent (MSI) if, for all $i \in \mathbf{N}$, finite $\mathbf{K} \subseteq \mathbf{N}\backslash\{i\}$ and $\mathbf{A}_j \in \boldsymbol{F}_j$ for $j \in \{i\} \cup \mathbf{K}$, one has $P(\mathbf{A}_i | \bigcap_{j \in \boldsymbol{K}} \mathbf{A}_j) = P(\mathbf{A}_i)$, so the conditional and unconditional probabilities are the same.

Example 1. (continued) Let **A** = {HH,HT} denote the event of a head for the first coin, **C** = {HH,TH} denote the event of a head for the second coin, **D** = {HH,TT} denote the event of a match, **G** = {HH} the event of two heads. The table below gives the probabilities of various events.

Event	**A**	**C**	**D**	**G**	**A∩C**	**A∩D**	**C∩D**	**A∩C∩D**	**A∩G**
Prob.	½	½	½	1/4	1/4	1/4	1/4	1/4	1/4

The result $P(\mathbf{A} \cap \mathbf{C}) = P(\mathbf{A})P(\mathbf{C}) = 1/4$ establishes that **A** and **C** are statistically independent. Verify that **A** and **D** are statistically independent, and that **C** and **D** are statistically independent, but that $\mathbf{P}(\mathbf{A} \cap \mathbf{C} \cap \mathbf{D}) \neq P(\mathbf{A})P(\mathbf{C})P(\mathbf{D})$, so that **A, C**, and **D** are <u>not</u> MSI. Verify that **A** and **G** are not statistically independent.

Example 4. (continued) Recall that $\mathbf{S} = \mathbb{R}^2$ with $\boldsymbol{F} = \boldsymbol{B} \otimes \boldsymbol{B}$, the *product* Borel σ-field. Define $\boldsymbol{N} = \{\varphi, \mathbb{R}\}$ and the subfields $\boldsymbol{F}_1 = \boldsymbol{B} \times \boldsymbol{N}$ and $\boldsymbol{F}_2 = \boldsymbol{N} \times \boldsymbol{B}$, containing information on the index levels on the first and second day, respectively. Define $\boldsymbol{G}$ to be the σ-field generated by the rectangles $(0,1]\times(0,1]$, $(0,1]\times(1,\infty)$,$(1,\infty)\times(0,1]$, and$(1,\infty)\times(1,\infty)$. Then $\boldsymbol{G}$ is the subfield of $\boldsymbol{B}$ containing information on whether the indices on the two days are above one. Define $\boldsymbol{F}_3$ to be the σ-subfield of $\boldsymbol{B} \otimes \boldsymbol{B}$ generated by sets of the form $\mathbf{A}_1 \times \mathbf{A}_2$ with $\mathbf{A}_1 \in \boldsymbol{G}$ and $\mathbf{A}_2 \in \boldsymbol{B}$; then $\boldsymbol{F}_3$ contains full information on the second day index, but only the qualitative information on whether the first day index is above one. Suppose $P((s,\infty)\times(t,\infty)) = e^{-s-t}$. Then $\{\boldsymbol{F}_1, \boldsymbol{F}_2\}$ are MSI. However, $\{\boldsymbol{F}_1, \boldsymbol{F}_3\}$ are not independent.

Example 8. Consider **S** = {0, 1, 2, 3, 4, 5, 6, 7}, with $\boldsymbol{F}$ equal to all subsets of **S**. As a shorthand, let 0123 denote {0,1,2,3}, etc. Define the subfields

$\boldsymbol{F}_1$ = {φ,0123,4567,**S**}, $\boldsymbol{F}_2$ = {φ,2345,0167,**S**}, $\boldsymbol{F}_3$ = {φ,0246,1357,**S**},
$\boldsymbol{F}_4$ = {φ,01,23,4567,0123,234567,014567,**S**},
$\boldsymbol{F}_5$ = {φ,01,23,45,67,0123,0145,0167,2345,2367,4567,012345,012367,014567,234567,**S**},
$\boldsymbol{F}_6$ = {φ,06,17,24,35,0167,0246,0356,1247,1357,2345,123457,023456,013567,012467,**S**}.

The field $\boldsymbol{F}_4$ is a *refinement* of the field $\boldsymbol{F}_1$ (i.e., $\boldsymbol{F}_1 \subseteq \boldsymbol{F}_4$), and can be said to contain more information than $\boldsymbol{F}_1$. The field $\boldsymbol{F}_5$ is a *mutual refinement* of $\boldsymbol{F}_1$ and $\boldsymbol{F}_2$ (i.e., $\boldsymbol{F}_1 \cup \boldsymbol{F}_2 \subseteq \boldsymbol{F}_5$), and is in fact the smallest mutual refinement. It contains all the information available in either $\boldsymbol{F}_1$ or $\boldsymbol{F}_2$. Similarly, $\boldsymbol{F}_6$ is a mutual refinement of $\boldsymbol{F}_2$ and $\boldsymbol{F}_3$. The intersection of $\boldsymbol{F}_5$ and $\boldsymbol{F}_6$ is the field $\boldsymbol{F}_2$; it is the common information available in $\boldsymbol{F}_5$ and $\boldsymbol{F}_6$. If, for example, $\boldsymbol{F}_5$ characterized the information available to one economic agent, and $\boldsymbol{F}_6$ characterized the information available to a second agent, then $\boldsymbol{F}_2$ would characterize the common information upon which they could base contingent contracts. Suppose $P(i) = 1/8$. Then $\{\boldsymbol{F}_1, \boldsymbol{F}_2, \boldsymbol{F}_3\}$ are MSI. E.g., $P(0123|2345) = P(0123|0246) = P(0123|2345\cap0246) = P(0123) = 1/2$. However, $\{\boldsymbol{F}_1, \boldsymbol{F}_4\}$ are not independent; e.g., $1 = P(0123|01) \neq P(0123) = 1/2$.

For $\mathbf{M} \subseteq \mathbf{N}$, let $\boldsymbol{F}_{\mathbf{M}}$ denote the smallest σ-field containing $\boldsymbol{F}_i$ for all $i \in \mathbf{M}$. Then MSI satisfies the following theorem, which provides a useful criterion for determining whether a collection of subfields is MSI:

Theorem 3.3. If $\boldsymbol{F}_i$ are MSI for $i \in \mathbf{N}$, and $\mathbf{M} \subseteq \mathbf{N}\backslash\{i\}$, then $\{\boldsymbol{F}_i,\boldsymbol{F}_{\mathbf{M}}\}$ are MSI. Further, $\boldsymbol{F}_i$ for $i\in\mathbf{N}$ are MSI if and only if $\{\boldsymbol{F}_i,\boldsymbol{F}_{\mathbf{N}\backslash i}\}$ are MSI for all $i\in\mathbf{N}$.

Example 5. (continued) If $\mathbf{M} = \{2,3\}$, then $\boldsymbol{F}_M \equiv \boldsymbol{F}_6$, and $P(0123|\mathbf{A}) = ½$ for each $\mathbf{A} \in \boldsymbol{F}_{\mathbf{M}}$.

3.4.2. 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 $\mathbf{N}$ denote a finite or countable index set of trials, $\mathbf{S}_i$ a sample space for trial i, and $\boldsymbol{G}_i$ a σ-field of subsets of $\mathbf{S}_i$. Note that $(\mathbf{S}_i,\boldsymbol{G}_i)$ may be the same for all i. Assume that $(\mathbf{S}_i, \boldsymbol{G}_i)$ is the real line with the Borel σ-field, or a countable set with the field of all subsets, or a pair with comparable mathematical properties (i.e., $\mathbf{S}_i$ is a complete separable metric space and $\boldsymbol{G}_i$ 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 $\mathbf{S}_{\mathbf{N}} = \times_{i\in\mathbf{N}} \mathbf{S}_i$ denote the sample space of these sequences. Let $\boldsymbol{F}_{\mathbf{N}} = \otimes_{i\in\mathbf{N}}\boldsymbol{G}_i$ denote the σ-field of subsets of $\mathbf{S}_{\mathbf{N}}$ generated by the *finite rectangles* which are sets of the form $\left(\times_{i\in\mathbf{K}} \mathbf{A}_i\right)\times\left(\times_{i\in\mathbf{N}\backslash\mathbf{K}} \mathbf{S}_i\right)$ with $\mathbf{K}$ a finite subset of $\mathbf{N}$ and $\mathbf{A}_i \in \boldsymbol{G}_i$ for $i \in \mathbf{K}$. The collection $\boldsymbol{F}_{\mathbf{N}}$ is called the *product σ-field* of subsets of $\mathbf{S}_{\mathbf{N}}$.

Example 9. $\mathbf{N} = \{1,2,3\}$, $\mathbf{S}_i = \{0,1\}$, $\boldsymbol{G}_i = \{\varphi,\{0\},\{1\},\mathbf{S}\}$ is a sample space for a coin toss, coded "1" if heads and "0" if tails. Then $\mathbf{S}_{\mathbf{N}} = \{s_1s_2s_3|s_i \in \mathbf{S}_i\} = \{000, 001, 010, 011, 100, 101, 110, 111\}$, where 000 is shorthand for the event $\{0\}\times\{0\}\times\{0\}$, and so forth, is the sample space for three coin tosses. The field $\boldsymbol{F}_{\mathbf{N}}$ is the family of all subsets of $\mathbf{S}_{\mathbf{N}}$.

For any subset $\mathbf{K}$ of $\mathbf{N}$, define $\mathbf{S}_{\mathbf{K}} = \times_{i\in\mathbf{K}} \mathbf{S}_i$ and $\boldsymbol{G}_{\mathbf{K}} = \otimes_{i\in\mathbf{K}}\boldsymbol{G}_i$. Then, $\boldsymbol{G}_{\mathbf{K}}$ is the product σ-field on $\mathbf{S}_{\mathbf{K}}$. Define $\boldsymbol{F}_{\mathbf{K}}$ to be the σ-field on $\mathbf{S}_{\mathbf{N}}$ generated by sets of the form $\mathbf{A}\times\mathbf{S}_{\mathbf{N}\backslash\mathbf{K}}$ for $\mathbf{A} \in \boldsymbol{G}_{\mathbf{K}}$. Then $\boldsymbol{G}_{\mathbf{K}}$

and $\mathcal{F}_K$ contain essentially the same information, but $\mathcal{G}_K$ is a field of subsets of S_K and $\mathcal{F}_K$ is a corresponding field of subsets of S_N which contains no information on events outside of **K**. Suppose P_N is a probability on $(S_N, \mathcal{F}_N)$. The *restriction* of P_N to $(S_K,\mathcal{G}_K)$ is a probability P_K defined for $A \in \mathcal{G}_K$ by $P_K(A) = P_N(A\times S_{N\backslash K})$. The following result establishes a link between different restrictions:

Theorem 3.4. If $M \subseteq K$ and P_M, P_K are restrictions of P_N, then P_M and P_K satisfy the *compatibility condition* that $P_M(A) = P_K(A\times S_{K\backslash M})$ for all $A \in \mathcal{F}_M$.

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 probabilities for a finite sequence as a restriction.

Theorem 3.5. If P_K on $(S_K,\mathcal{G}_K)$ for all finite $K \subseteq N$ satisfy the compatibility condition, then there exists a unique P_N on $(S_N,\mathcal{F}_N)$ such that each P_K is a restriction of P_N.

This result guarantees that it is meaningful to make probability statements about events such as “an infinite number of heads in repeated coin tosses".

.

Suppose trials $(S_i,\mathcal{G}_i,P_i)$ indexed by i in a countable set **N** are mutually statistically independent. For finite $K \subseteq N$, let $\mathcal{G}_K$ denote the product σ-field on S_K. Then MSI implies that the probability of a set $\times_{i\in K} A_i \in \mathcal{G}_K$ satisfies $P_K(\times_{i\in K} A_i) = \prod_{j\in K} P_j(A_j)$. Then, the compatibility condition in Theorem 3.3 is satisfied, and that result implies the existence of a probability P_N on $(S_N,\mathcal{F}_N)$ whose restrictions to $(S_K,\mathcal{G}_K)$ for finite $K \subseteq N$ are the probabilities P_K.

3.4.3. 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. This idea is formalized in Chapter 4.

3.5. RANDOM VARIABLES, DISTRIBUTION FUNCTIONS, AND EXPECTATIONS

3.5.1. A *random variable* X is a measurable real-valued function on a probability space (**S**,***F***,P), or X:S → ℝ. Then each state of Nature s determines a value X(s) of the random variable, termed its *realization* in state s. When the functional nature of the random variable is to be emphasized, it is denoted X(·), or simply X. When its values or realizations are used, they are denoted X(s) or x. For each set **B** ∈ ***B***, the probability of the event that the realization of X is contained in **B** is well-defined and equals P'(**B**) ≡ P(X^{-1}(**B**)), where P' is termed the probability *induced* on ℝ by the random variable X. 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 important in working with random variables to keep in mind that the random variable itself is a <u>function</u> of states of Nature, and that observations are of <u>realizations</u> 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. Multiplying a random variable by a scalar, or adding random variables, results in another random variable. Then, the family of random variables forms a *linear vector space*. In addition, products of random variables are again random variables, so that the family of random variables forms an *Abelian group under multiplication*. The family of random variables is also closed under *majorization*, so that Z:S → ℝ defined by Z(s) = max(X(s),Y(s)) for random variables X and Y is again a random variable. Then, the family of random variables forms a *lattice* with respect to the partial order X ≤ Y(i.e., X(s) ≤ Y(s) almost surely).

3.5.2. The term *measurable* in the definition of a random variable means that for each set **A** in the Borel σ-field ***B*** of subsets of the real line, the inverse image X^{-1}(**A**) ≡ {s∈**S**|X(s)∈**A**} is in the σ-field ***F*** of subsets of the sample space **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.

3.5.3. The probability that a random variable X has a realization in a set **A** ∈ ***B*** is given by

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

The function F is a probability on ***B***; it is defined in particular for half-open intervals of the form **A** = (-∞,x], in which case F((-∞,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 ***B***. The *support* of the distribution F is the smallest closed set $\mathbf{A} \in \boldsymbol{B}$ such that F(**A**) = 1.

Example 5. (continued) The standard normal CDF is $\Phi(\mathbf{x}) = \int_{-\infty}^{x} (2\pi)^{-1/2} \cdot e^{-s^2/2} ds$, obtained by integrating the density $\varphi(s) = (2\pi)^{-1/2} \cdot e^{-s^2/2}$. Other examples are the CDF for the standard exponential distribution, $F(x) = 1 - e^{-x}$ for $x > 0$, and the CDF for the logistic distribution, $F(x) = 1/(1+e^{-x})$. An example of a CDF that has jumps is $F(x) = 1 - e^{-x}/2 - \sum_{k=1}^{\infty} \mathbf{1}(k \ge x)/2^{k+1}$ for $x > 0$.

3.5.4. If F is absolutely continuous with respect to a σ-finite measure ν on 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 ℝ, called the *density* (or *probability density function, pdf)* of X, such that

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

for every $A \in \boldsymbol{B}$. 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 $F(\mathbf{A}) = \int_A f(x)dx$.

If F is absolutely continuous with respect to counting measure on a countable subset **C** of ℝ, then it is called a *discrete* distribution, and there is a real-valued function f on **C** such that

$$F(\mathbf{A}) = \sum_{x \in A} f(x).$$

Recall that the probability is itself a measure. This suggests a notation $F(\mathbf{A}) = \int_A F(dx)$ that covers both continuous and counting cases. This is called a *Lebesgue-Stieltjes* integral.

3.5.5. If ($\mathbb{R}$,$\boldsymbol{B}$,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 expectation is written **E** g(X) = $\int_{\mathbb{R}} g(x)f(x)dx$.

Alternately, for counting measure on the integers with density f(k), **E** g(X) = $\sum_{k=-\infty}^{+\infty} g(k)f(k)$.

The expectation of X, if it exists, is called the *mean* of X. The expectation of (X - **E**X)^2, if it exists, is called the *variance* of X. Define $\mathbf{1}(X \le a)$ to be an indicator function that is one if X(s) $\le$ a, and zero otherwise. Then, **E** $\mathbf{1}(X \le a)$ = F(a), and the distribution function can be recovered from the expectations of the indicator functions. Most econometric applications deal with random variables that have finite variances. The space of these random variables is $\mathbf{L}_2(\mathbf{S},\boldsymbol{F},P)$, the space of random variables X for which **E** $X^2 = \int_S X(x)^2 P(ds) < +\infty$. The space $\mathbf{L}_2(\mathbf{S},\boldsymbol{F},P)$ is also termed the space of *square-integrable functions*. The norm in this space is root-mean-square, $\|X\|_2 = [\int_S X(s)^2 P(ds)]^{1/2}$. Implications of $X \in \mathbf{L}_2(\mathbf{S},\boldsymbol{F},P)$ are $\mathbf{E}|X| \le \int_S \max(X(s),1)P(ds) \le \int_S (X(s)^2+1)P(ds) = \|X\|_2^2 + 1 < +\infty$ and $\mathbf{E}(X - \mathbf{E}X)^2 = \|X\|_2^2 - (\mathbf{E}|X|)^2 \le \|X\|_2^2 < +\infty$, so that X has a well-defined, finite mean and variance.

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

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

Then, X is the number of heads in two coin tosses. For a fair coin, **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 **C** = {0,1,2,...}. Then **C** is the support of X. For x real, define [x] to be the largest integer k satisfying k $\le$ x. A distribution function for X, defined on the real line, is F(x) = $\begin{cases} 1 - 2^{-[x+1]} & \text{for } 0 \le x \\ 0 & \text{for } 0 > x \end{cases}$; the associated density defined on **C** is f(k) = 2^{-k-1}. The expectation of X, obtained using evaluation of a special series from 2.1.10, is **E** X = $\sum_{k=0}^{\infty} k \cdot 2^{-k-1} = 1$.

Example 3. (continued) Define a random variable X by X(s) = |s - 1|. Then, X is the magnitude of the deviation of the index from one. The inverse image of an interval (a,b) is (1-b,1-a)∪(1+a,1+b) ∈ ***F***, so that X is measurable. Other examples of measurable random variables are Y defined by Y(s) = Max {1,s} and Z defined by Z(s) = s^3.

3.5.6. Consider a random variable Y on (ℝ,***B***). The expectation $\mathbf{E}Y^k$ is the k-th *moment* of Y, and $\mathbf{E}(Y-\mathbf{E}Y)^k$ is the k-th *central moment*. Sometimes moments fail to exist. However, if g(Y) is continuous and bounded, then **E**g(Y) always exists. The expectation m(t) = $\mathbf{E}e^{tY}$ 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 ψ(t) = $\mathbf{E}e^{\iota tY}$, where ι is the square root of -1, is termed the *characteristic function* (cf) of Y. The characteristic function always exists.

Example 5. (continued) A density f(x) that is symmetric about zero, such as the standard normal, has $\mathbf{E}X^k = \int_{-\infty}^{+\infty} x^k f(x)dx = \int_{-\infty}^{0} x^k f(-x)dx + \int_{0}^{+\infty} x^k f(x)dx = \int_{0}^{+\infty} [1 + (-1)^k]x^k f(x)dx = 0$ for k odd. Integration by parts yields the formula $\mathbf{E}X^k = 2k \int_{0}^{+\infty} x^{k-1}[1-F(x)]dx$ for k even. For the standard normal, $\mathbf{E}X^{2k} = 2\cdot\int_0^{+\infty}(2\pi)^{-1/2}x^{2k-1}\cdot e^{-x^2/2}xdx = (2k-1)\cdot\mathbf{E}X^{2k-2}$ for k > 2 using integration by parts, and $\mathbf{E}X^2 = 2\cdot\int_0^{+\infty}(2\pi)^{-1/2}\cdot e^{-x^2/2}xdx = 2\cdot\Phi(0) = 1$. Then, $\mathbf{E}X^4 = 3$ and $\mathbf{E}X^6 = 15$. The moment generating function of the standard normal is m(t) = $\int_{-\infty}^{+\infty}(2\pi)^{-1/2}\cdot e^{tx}\cdot e^{-x^2/2}dx$. Completing the square in the exponent gives m(t) = $e^{t^2/2}\cdot\int_{-\infty}^{+\infty}(2\pi)^{-1/2}\cdot e^{-(x-t)^2/2}dx = e^{t^2/2}$.

3.5.7. If T random variables are formed into a vector, $X(\cdot) = (X(\cdot,1),...,X(\cdot,T))$, the result is termed a *random vector*. For each $s \in \mathbf{S}$, the realization of the random vector is a point $(X(s,1),...,X(s,T))$ in $\mathbb{R}^T$, and the random vector has an induced probability on $\mathbb{R}^T$ which is characterized by its multivariate CDF, $F_X(x_1,...,x_T) = P(\{s\in\mathbf{S}|X(s,1)\le x_1,...,X(s,T)\le x_T\})$. Note that all the components of a random vector are functions of the *same* state of Nature s, and the random vector can be written as a measurable function X from the probability space $(\mathbf{S},\boldsymbol{F},P)$ into $(\mathbb{R}^T,\boldsymbol{B}^{nT}$. (The notation $\boldsymbol{B}^T$ means $\boldsymbol{B}\otimes\boldsymbol{B}\otimes...\otimes\boldsymbol{B}$ T times, where $\boldsymbol{B}$ is the Borel σ-field on the real line. This is also called the *product* σ-field, and is sometimes written $\boldsymbol{B}^T = \bigotimes_{i=1,...,T} \boldsymbol{B}_i$, where the $\boldsymbol{B}_i$ are identical copies of $\boldsymbol{B}$.) The measurability of X requires $X^{-1}(\mathbf{C}) \in \mathbf{S}$ for each open rectangle $\mathbf{C}$ in $\mathbb{R}^T$. The independence or dependence of the components of X is determined by the fine structure of P on $\mathbf{S}$.

A useful insight comes from considering different representations of vectors in finite-dimensional spaces, and extending these ideas to infinite-dimensional situations. To be specific, consider $\mathbb{R}^2$. When we express a function X on $T = \{1,2\}$ as a point $(X(1),X(2))$ in this space, what we are really doing is defining two functions $Z_1 = (1,0)$ and $Z_2 = (0,1)$ with the property that Z_1 and Z_2 span the space, and then writing X as the linear combination $X = X(1)\cdot Z_1 + X(2)\cdot Z_2$. The pair of functions (points) Z_1 and Z_2 is called a *Hamel basis* for $\mathbb{R}^2$, and every point in the space has a unique representation in terms of this basis. However, there may be many different Hamel bases. For example, the unit function (1,1) and the function $\cos(\pi t)$ or (-1,1) also form a Hamel basis, and in terms of this basis X has the representation $X = ½(X(1)+X(2))\cdot(1,1) + ½(X(2)-X(1))\cdot(-1,1)$.

Another way to write a random vector X is to define an index set $\mathbf{T} = \{1,...,T\}$, and then define X as a real-valued function on $\mathbf{S}$ and $\mathbf{T}$, $X:\mathbf{S}\times\mathbf{T}\to\mathbb{R}$. Then, $X(\cdot,t)$ is a simple random variable for each $t \in \mathbf{T}$, and $X(s,\cdot)$ is a real vector that is a realization of X for each $s \in \mathbf{S}$. A function defined in this way is also called a *stochastic process*, particularly when $\mathbf{T}$ is not finite. The measurability requirement on X is the same as before, but can be written in a different form as requiring that the inverse image of each open interval in $\mathbb{R}$ be contained in $\boldsymbol{F}\otimes\boldsymbol{T}$, where $\boldsymbol{T}$ is a σ-field of subsets of $\mathbf{T}$ that can be taken to be the family of all subsets of $\mathbf{T}$ and "⊗" denotes the operation that forms the smallest σ-field containing all sets $\mathbf{A}\times\mathbf{B}$ with $\mathbf{A} \in \boldsymbol{F}$ and $\mathbf{B} \in \boldsymbol{T}$. There is then a complete duality between random vectors in a T-dimensional linear space and random functions on a T-dimensional index set. This duality between vectors and functions will generalize and provide useful insights into statistical applications in which $\mathbf{T}$ is a more general set indexing time. The *distribution function* (CDF) of X is

$$F(x_1,...,x_T) = P(\{s\varepsilon S|X_i(s) \le x_i \text{ for } i = 1,...,T\}).$$

If $\mathbf{A} \in \boldsymbol{B}^T$, define $F(\mathbf{A}) = P(\{s\varepsilon\mathbf{S}|X(s)\in\mathbf{A}\})$. If $F(\mathbf{A}) = 0$ for every set $\mathbf{A}$ of Lebesque measure zero, then there exists a *probability density function* (pdf) $f(x_1,...,x_T)$ such that

(1) $F(x_1,...,x_T) = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2}...\int_{-\infty}^{x_T} f(y_1,...,y_T)\, dy_1...dy_T.$

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

$$F_1(x_1) \equiv P(\{s\in\mathbf{S}|X_1(s) \le x_1\}) = F(x_1,+\infty,...,+\infty).$$

This random variable is measurable with respect to the σ-subfield $\boldsymbol{G}_1$ containing the events whose occurrence is determined by X_1 alone; i.e., $\boldsymbol{G}_1$ is the family generated by sets of the form $\mathbf{A}\times\mathbb{R}\times...\times\mathbb{R}$ with $\mathbf{A} \in \boldsymbol{B}$. If F is absolutely continuous with respect to Lebesque measure on $\boldsymbol{B}^T$, then there are associated densities f and f_1 satisfying

(2) $F_1(x_1) = \int_{y_1=-\infty}^{x_1} f_1(y_1)\, dy_1$

(3) $f_1(x_1) = \int_{y_2=-\infty}^{+\infty}\cdots\int_{y_n=-\infty}^{+\infty} f(x_1,y_2,...,y_n)\cdot dy_2...dy_n.$

F_1 and f_1 are termed the *marginal* CDF and pdf, respectively, of X_1.

3.5.8. Corresponding to the concept of a conditional probability, we can define a *conditional distribution*: Suppose $\mathbf{C}$ is an event in $\boldsymbol{G}_1$ with $P(\mathbf{C}) > 0$. Then, define $F_{(2)}(x_2,...,x_n|\mathbf{C}) = F(\{y\in\mathbb{R}^n|y_1\in\mathbf{C},y_2\le x_2,...,y_n\le x_n\})/F_1(\mathbf{C})$ to be the conditional distribution of $(X_2,...,X_n)$ given $X_1 \in \mathbf{C}$. When F is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^n$, the conditional distribution can be written in terms of the joint density,

$$F_{(2)}(x_2,...,x_n|\mathbf{C}) = \frac{\int_{y_1\in\boldsymbol{C}}\int_{y_2=-\infty}^{x_2}\int_{y_n=-\infty}^{x_n} f(y_1,y_2,...,y_n)\cdot dy_1dy_2...dy_n}{\int_{y_1\in\boldsymbol{C}}\int_{y_2=-\infty}^{+\infty}\int_{y_n=-\infty}^{+\infty} f(y_1,y_2,...,y_n)\cdot dy_1dy_2...dy_n}.$$

Taking the limit as $\mathbf{C}$ shrinks to a point $X_1 = x_1$, one obtains the conditional distribution of $(X_2,...,X_n)$ given $X_1 = x_1$,

$$F_{(2)}(x_2,...,x_n|X_1=x_1) = \frac{\int_{y_2=-\infty}^{x_2}\int_{y_n=-\infty}^{x_n} f(x_1,y_2,...,y_n)\cdot dy_1dy_2...dy_n}{f_1(x_1)},$$

provided $f_1(x_1) > 0$. Finally, associated with this conditional distribution is the conditional density $f_{(2)}(x_2,...,x_n|X_1=x_1) = f(x_1,x_2,...,x_n)/f_1(x_1)$. 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 $X_1 = x_1,...,X_k = x_k$.

3.5.9. 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. Here are some useful properties of expectations of vectors:

(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 $(\mathbf{E}X'Z)^2 \le (\mathbf{E}X'X)(\mathbf{E}Z'Z)$.
(d) [Minkowski Inequality] If X is a random vector of length n and $r \ge 1$ is a scalar, then $(\mathbf{E}|\sum_{i=1}^n X_i|^r)^{1/r} \le \sum_{i=1}^n (\mathbf{E}|X_i|^r)^{1/r}$.

(e) [Loeve Inequality] If X is a random vector of length n and $r > 0$, then $\mathbf{E}|\sum_{i=1}^n X_i|^r \le \max(1,n^{r-1}) \sum_{i=1}^n \mathbf{E}|X_i|^r$.

(f) [Jensen Inequality] If X is a random vector and g(x) is a convex function, then $\mathbf{E}\, g(X) \ge g(\mathbf{E}X)$. If g(x) is a concave function, the inequality is reversed.

When expectations exist, they can be used to bound the probability that a random variable takes on extreme values.

Theorem 3.6. Suppose X is a $n\times 1$ random vector and ε is a positive scalar.
a. [Markov bound] If $\max_i \mathbf{E}|X_i| < +\infty$, then $\max_i \Pr(|X_i| > \varepsilon) < \max_i \mathbf{E}|X_i|/\varepsilon$.
b. [Chebyshev bound] If $\mathbf{E}X'X < +\infty$, then $\Pr(\|X\|_2 > \varepsilon) < \mathbf{E}X'X/\varepsilon^2$.
c. [Chernoff bound] If $\mathbf{E}e^{t'X}$ exists for all vectors t in some neighborhood of zero, then for some positive scalars α and M, $\Pr(\|X\|_2 > \varepsilon) < Me^{-\alpha\varepsilon}$.

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

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

Taking $r(y) = y^2$ gives the result directly for the Chebyshev bound. In the remaining cases, first get a component-by-component inequality. For the Markov bound, $\Pr(|X_i| > \varepsilon) < E|X_i|/\varepsilon$ for each i gives the result. For the Chernoff bound,

$$\Pr(\|X\|_2 > \varepsilon) \le \sum_{i=1}^{n} [\Pr(X_i > \varepsilon \cdot n^{-1/2}) + \Pr(X_i < -\varepsilon \cdot n^{-1/2})]$$

since if the event on the left occurs, one of the events on the right must occur. Then apply the inequality $\Pr(|X_i| > \varepsilon) \le \mathbf{E}r(|X_i|)/r(\varepsilon)$ with $r(y) = n^{-1/2} \cdot e^{y\alpha}$ to each term in the right-hand-side sum. The inequality for vectors is built up from a corresponding inequality for each component. □

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

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

Another notation for this expectation is $\mathbf{E}(g(X)|\mathbf{C})$. When the distribution of the random variable X is absolutely continuous with respect to Lebesgue measure, so that it has a density f(x), the conditional density can be written as $f(x|\boldsymbol{C}) = f(x)\cdot\mathbf{1}(x\in\boldsymbol{C})/\int_{\boldsymbol{C}} f(s)ds$, and the conditional expectation can then be written

$$\mathbf{E}_{X|\mathbf{C}}g(X) = \int_{\boldsymbol{C}} g(x)\cdot f(x|\boldsymbol{C})dx = \frac{\int_{\boldsymbol{C}} g(x)\cdot f(x)dx}{\int_{\boldsymbol{C}} f(x)dx}.$$

When the distribution of X is discrete, this formula becomes

$$\mathbf{E}_{X|\mathbf{C}}g(X) = \frac{\sum_{k\in\boldsymbol{C}} g(k)\cdot f(k)}{\sum_{k\in\boldsymbol{C}} f(k)}.$$

The conditional expectation is actually a <u>function</u> on the σ-field $\boldsymbol{C}$ of conditioning events, and is sometimes written $\mathbf{E}_{X|\boldsymbol{C}}\, g(X)$ or $\mathbf{E}(g(X)|\boldsymbol{C})$ to emphasize this dependence.

Suppose $\mathbf{A}_1,...,\mathbf{A}_k$ *partition* the domain of X. Then the distribution satisfies

$$F(\mathbf{x}) = \sum_{i=1}^{k} F(x|\mathbf{A}_i)\cdot F(\mathbf{A}_i),$$

implying

$$\mathbf{E}g(X) = \int g(x)F(dx) = \sum_{i=1}^{k} \int g(x)F(dx\,|\,\mathbf{A}_i)\cdot F(\mathbf{A}_i) = \sum_{i=1}^{k} \mathbf{E}\{g(X)|\mathbf{A}_i\}\cdot F(\mathbf{A}_i).$$

This is called the *law of iterated expectations*, and is heavily used in econometrics.

Example 2. (continued) Recall that X is the number of heads that appear before a tail in a sequence of coin tosses, and that the probability of $X = k$ is 2^{-k-1} for $k = 0,1,...$. Let **C** be the event of an even number of heads. Then,

$$\mathbf{E}_{X|\mathbf{C}}X = \frac{\sum_{k=0,2,4,...} k\cdot 2^{-k-1}}{\sum_{k=0,2,4,...} 2^{-k-1}} = \frac{\sum_{j=0,1,2,...} j\cdot 4^{-j}}{\sum_{j=0,1,2,...} 4^{-j}/2} = 2/3,$$

where the second ratio is obtained by substituting $k = 2j$, and the value is obtained using the summation formulas for a geometric series from 2.1.10. A similar calculation for the event A of an odd number of heads yields $\mathbf{E}_{X|\mathbf{A}}X = 5/3$. The probability of an even number of heads is $\sum_{k=0,2,4,...} 2^{-k-1} = 2/3$. The law of iterated expectations then gives

$$\mathbf{E}\,X = \mathbf{E}\{X|\mathbf{C}\}\cdot P(\mathbf{C}) + \mathbf{E}\{X|\mathbf{A}\}\cdot P(\mathbf{A}) = (2/3)(2/3) + (5/3)(1/3) = 1,$$

which confirms the direct calculation of **E** X.

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

$$g(x) = \int_{u=-\infty}^{+\infty} f(u,x)du,$$

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

$$\mathbf{E}h(U,X) = \int\int h(u,x)f(u,x)dudx = \int_{x=-\infty}^{+\infty}\left(\int_{u=-\infty}^{\infty} h(u,x)f(u|x)du\right)g(x)dx = \mathbf{E}_X\mathbf{E}_{U|X}h(U,X);$$

another example of 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 defined by $\mathbf{V}(U|x) = \mathbf{E}_{U|X}(U - \mathbf{M}_{U|X}(x))^2$. It is related to the unconditional variance by the formula

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

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

Example 10: Suppose (U,X) are bivariate normal with means $\mathbf{E}U = \mu_u$ and $\mathbf{E}X = \mu_x$, and second moments $\mathbf{E}(U-\mu_u)^2 = \sigma_u^2$, $\mathbf{E}(X-\mu_x)^2 = \sigma_x^2$, and $\mathbf{E}(U-\mu_u)(X-\mu_x) = \sigma_{ux} \equiv \rho\sigma_u\sigma_x$. Define

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

and observe that

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

The bivariate normal density is $f(u,x) = [2\pi\sigma_u\sigma_x(1-\rho^2)^{1/2}]^{-1}\cdot\exp(-Q/2)$. The marginal density of X is normal with mean μ_x and variance σ_x^2: $n(x-\mu_x,\sigma_x) = (2\pi\sigma_x^2)^{-1}\cdot\exp(-(x-\mu_x)^2/2\sigma_x^2)$. This can be derived from the bivariate density by completing the square for u in Q and integrating over u. The conditional density of U given X then satisfies

$$f(u|x) = [2\pi\sigma_u\sigma_x(1-\rho^2)^{½}]^{-1}\cdot\exp(-Q/2)/(2\pi\sigma_x^2)^{-1}\cdot\exp(-(x-\mu_x)^2/2\sigma_x^2).$$

$$= [2\pi\sigma_u^2(1-\rho^2)]^{-½}\cdot \exp\left(\frac{-1}{2\cdot(1-\rho^2)}\cdot\left[\left(\frac{u-\mu_u}{\sigma_u}\right) - \rho\cdot\left(\frac{x-\mu_x}{\sigma_x}\right)\right]^2\right) .$$

Hence the conditional distribution of U, given X = x, is normal with conditional mean $\mathbf{E}(U|X=x) = \mu_u + \rho\sigma_u(x - \mu_x)/\sigma_x \equiv \mu_u + \sigma_{ux}(x-\mu_x)/\sigma_x^2$ and variance $\mathbf{V}(U|X=x) \equiv \mathbf{E}((U-\mathbf{E}(U|X=x))^2|X=x) = \sigma_u^2(1-\rho^2) \equiv \sigma_u^2 - \sigma_{ux}^2/\sigma_x^2$. 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 $\mathbf{V}(U|X=x) = \Omega_{uu} - \Omega_{ux}\Omega_{xx}^{-1}\Omega_{xu}$.

3.5.11. 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 independent, $f(u,x) = h(u)\cdot g(x)$, 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 $E(r(U)|X=x) = \int r(u)f(u|x)du \equiv \int r(u)h(u)du = \mathbf{E}r(U)$, 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 $cov(r(U),s(X)) = 0$, provided $\mathbf{E}r(U)$ and $\mathbf{E}s(X)$ exist. If $r(u) = u - \mathbf{E}U$, then $\mathbf{E}(r(U)|X=x) = 0$ and $cov(U,X) = \mathbf{E}(U\text{-}\mathbf{E}U)X = 0$. Conversely, suppose U and X are jointly distributed. If $cov(r(U),s(X)) = 0$ 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 $0 = cov = F(u^*,x^*) - H(u^*)\cdot G(x^*)$, where F is the joint cumulative distribution function. Hence, $F(u,x) = H(u)\cdot G(x)$, and X, U are independent.

Note that $cov(U,X) = 0$ is not sufficient to imply U,X independent. For example, $g(x) = ½$ for $-1 \le x \le 1$ and $f(u|x) = ½$ for $-1 \le u-x^2 \le 1$ is nonindependent with $\mathbf{E}(U|X=x) = x^2$, but $cov(U,X) = \mathbf{E}X^3 = 0$. Furthermore, $\mathbf{E}(U|X=x) \equiv 0$ is not sufficient to imply U,X independent. For example, $g(x) = ½$ 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 $\mathbf{E}(U^2|x) = (1 + x^2)^2 \ne \mathbf{E}\,U^2 = 28/15$, but $\mathbf{E}(U|X=x) \equiv 0$.

Example 11. Suppose monthly family income (in thousands of dollars) is a random variable Y with a CDF $F(y) = 1 - y^{-2}$ for $y > 1$. Suppose a random variable Z is one for home owners and zero otherwise, and that the conditional probability of the event Z = 1, given Y, is $(Y-1)/Y$. The unconditional expectation of Y is 2. The joint density of Y and Z is $f(y)\cdot g(z|y) = (2y^{-3})(1 - y^{-1})$ for $z = 1$. The unconditional probability of Z = 1 is then $\int_{y=1}^{+\infty} f(y)\cdot g(z|y)dy = 1/3$. Bayes Law gives the conditional density of Y given z = 1, $f(y|z) = f(y)\cdot g(z|y)/\int_{y=1}^{+\infty} f(y)\cdot g(z|y)dy = (6y^{-3})(1 - y^{-1})$, so that the conditional expectation of Y given z = 1 is $E(Y|Z=1) = \int_{y=1}^{+\infty} y\, f(y|z)dy = 3$.

Example 12. The problem of interpreting the results of medical tests illustrates Bayes Law. A blood test for prostate cancer is known to yield a "positive" with probability 0.9 if cancer is present, and a false "positive" with probability of 0.2 if cancer is not present. The prevalence of the cancer in the population of males is 0.05. Then, the conditional probability of cancer, given a "positive" test result, equals the joint probability of cancer and a positive test result, (0.05)(0.9), divided by the probability of a positive test result, (0.05)(0.9)+(0.95)(0.2), or 0.235. Thus, a "positive" test has a low probability of identifying a case of cancer, and if all "positive" tests were followed by surgery, about 75 percent of these surgeries would prove unnecessary.

3.5.12. The discussion of expectations will be concluded with a list of detailed properties of characteristic functions and moment generating functions:

a. $\psi(t) = \mathbf{E}e^{\iota tY} \equiv \mathbf{E}\cos(tY) + \iota\mathbf{E}\sin(tY)$.
b. $Z = a + bY$ has the cf $e^{\iota ta}\psi(bt)$ and $Z = f(Y)$ has the cf $\mathbf{E}e^{\iota t\cdot f(Y)}$.
c. If $\mathbf{E}Y^k$ exists, then $\psi^{(k)}(t) \equiv d^k\psi(t)/dt^k$ exists, satisfies the bound $|d^k\psi(t)/dt^k| \le \mathbf{E}|Y|^k$, and is uniformly continuous, and $\mathbf{E}Y^k = (\iota)^k\psi^{(k)}(0)$. If $\psi^{(k)}(t)$ exists, then $\mathbf{E}Y^k$ exists.
d. If Y has finite moments through order k, then $\psi(t)$ has a Taylor's expansion

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

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

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

and

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

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

$$\zeta(t) = \sum_{j=1}^{k} \kappa_j\iota^j t^j/j! + [\zeta^{(k)}(\lambda t) - \zeta^{(k)}(t)],$$

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 $\kappa_1 = \mathbf{E}Y$ 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 $\exp(\iota\mu t-\sigma^2t^2/2)$. The normal has cumulants $\kappa_1 = \mu$, $\kappa_2 = \sigma^2$, $\kappa_3 = \kappa_4 = 0$.
f. Random variables X and Y have identical distribution functions if and only if they have identical characteristic functions.
g. If $Y_n \rightarrow_p Y$ (see Chap. 4.1), then the associated characteristic functions satisfy $\psi_n(t) \rightarrow \psi(t)$ 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 $Y_n \rightarrow_p Y$.
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(t), finite for $|t| < \tau$, where τ is a positive constant. Then, the following properties hold:

a. $m(t) = \mathbf{E}e^{tY}$ for $|t| < \tau$.
b. $Z = a + bY$ has the mgf $e^{ta}m(bt)$.
c. $\mathbf{E}Y^k$ exists for all $k > 0$, and $m \equiv d^k m(t)/dt^k$ exists and is uniformly continuous for $|t| < \tau$, with $\mathbf{E}Y^k = m_Y(0)$.
d. m(t) has a Taylor's expansion (for any k) $m_Y(t) = (\mathbf{E}Y^j)t^j/j! + [m(\lambda t) - m(0)]t^k/k!$, 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 $\exp(\mu t+\sigma 2t2)$.
f. Random variables X and Y with proper mgf have identical distribution functions if and only if their mgf are identical.
g. If $Y_n \rightarrow_p Y$ and the associated mgf are finite for $|t| < \tau$, then the mgf of Y_n converges pointwise to the MGF of Y. Conversely, if Y_n have proper MGF which converges pointwise to a function m(t) that is finite for $|t| < \tau$, then there exists Y such that m(t) is the mgf of Y and $Y_n \rightarrow_p Y$.
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(t), is $m(t/n)^n$.

The definitions of characteristic and moment generating functions can be extended to vectors of random variables. Suppose Y is a n×1 random vector, and let $\mathbf{t}$ be a n×1 vector of constants. Then $\psi(\mathbf{t}) = \mathbf{E}e^{\iota t'Y}$ is the characteristic function and $m(\mathbf{t}) = \mathbf{E}e^{t'Y}$ is the moment generating function. The properties of cf and mgf listed above also hold in their multivariate versions, with obvious modifications. For characteristic functions, two of the important properties translate to

(b') $Z = \mathbf{a} + \mathbf{B}Y$, where $\mathbf{a}$ is a m×1 vector and $\mathbf{B}$ is a m×n matrix, has cf $e^{\iota t'\mathbf{a}}\psi(\mathbf{B}\mathbf{t})$.
(e') if Y is multivariate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, then its characteristic function is $\exp(\iota\boldsymbol{\mu}'\mathbf{t} - \mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}/2)$.

A useful implication of (b') and (e') is that a linear transformation of a multivariate normal vector is again multivariate normal. Conditions (c) and (d) relating Taylor's expansions and moments for univariate cf have multivariate versions where the expansions are in terms of partial derivatives of various orders. Conditions (f) through (h) are unchanged in the multivariate version.

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.

3.6. TRANSFORMATIONS OF RANDOM VARIABLES

6.1. Suppose X is a measurable random variable on ($\mathbb{R}$,***B***) with a distribution F(x) that is absolutely continuous with respect to Lebesgue measure, so that X has a density f(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

$$G(y) = \Pr(Y \le y) = \Pr(H(X) \le y) = \Pr(X \le h(y)) = F(h(y)).$$

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

Example 13. Suppose X has the distribution function $F(x) = 1\text{-}e^{-x}$ for $x > 0$, with $F(x) = 0$ for $x \le 0$; 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, $G(y) = 1\text{-}\exp(-e^y)$ and $G(y) = \exp(-e^y)e^y = \exp(y-e^y)$ for $-\infty < y < +\infty$. This is called an extreme value distribution. A third example is X with some distribution function F and density f, and Y = F(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) = p. The distribution function of Y is $G(y) = F(x_y) = y$. 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 decreasing rather than increasing, then

$$G(y) = \Pr(Y \le y) = \Pr(H(X) \le y) = \Pr(X \ge h(y)) = 1\text{-}F(h(y)),$$

where h is the inverse function of H. Differentiating,

$$g(y) = f(h(y))(-h'(y)).$$

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*

$$g(y) = f(h(y))|h'(y)| \equiv f(h(y))/|H'(h(y)|.$$

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 $G(y) = e^{-1/y}$ and $g(y) = e^{-1/y}/y^2$.

Consider a transformation $Y = H(X)$ that is not 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

$$G(y) = \Pr(Y \le y) = \Pr(H(X) \le y) = \Pr(X \in A_y) = F(A_y).$$

If this expression is differentiable, then its derivative gives the density.

Example 14. If X has a distribution F and density f, and $Y = |X|$, then $A_y = [-y,y]$, implying $G(y) = F(y) - F(-y)$ and $f(y) = f(y) + f(-y)$.

Example 15. If $Y = X^2$, then $A_y = [-y^{1/2},y^{1/2}]$, $G(y) = F(y^{1/2}) - F(-y^{1/2})$. Differentiating for $y \ne 0$, $g(y) = (f(y^{1/2}) + f(-y^{1/2}))/2y^{1/2}$. Applying this to the standard normal with $F(x) = \Phi(x)$, the density of Y is $g(y) = \varphi(y^{1/2})/y^{1/2} = (2\pi y)^{-1/2}\cdot e^{-y/2}$, called the chi-square with one degree of freedom.

3.6.2. Next consider transformations of random vectors. These transformations will permit us to analyze sums or other functions of random variables. Suppose X is a $n\times 1$ 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:

Theorem 3.8. If X has density f(x), and $Y = \mathbf{A}X$, with $\mathbf{A}$ nonsingular, then the density of Y is

$$g(y) = f(\mathbf{A}^{-1}y)/|\det(\mathbf{A})| .$$

Proof: We will prove the result in two dimensions, leaving the general case to the reader. First, consider the case $\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} a_{11} & 0 \\ 0 & a_{22} \end{bmatrix}\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ with $a_{11} > 0$ and $a_{22} > 0$. One has $G(y_1,y_2) \equiv F(y_1/a_{11},y_2/a_{22})$. Differentiating with respect to y_1 and y_2, $g(y_1,y_2) \equiv f(y_1/a_{11},y_2/a_{22})/a_{11}a_{22}$. This establishes the result for diagonal transformations. Second, consider $\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} a_{11} & 0 \\ a_{21} & a_{22} \end{bmatrix}\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ with $a_{11} > 0$ and $a_{22} > 0$. Then

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

$$\partial^2 G(y_1,y_2)/\partial y_1\partial y_2 \equiv g(y_1,y_2) = (a_{11}a_{22})^{-1}f(y_1/a_{11},(y_2-y_1a_{21}/a_{11})/a_{22}).$$

This establishes the result for triangular transformations. Finally, consider the general transformation $\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ with $a_{11} > 0$ and $a_{11}a_{22}-a_{12}a_{21} > 0$. Apply the result for triangular transformations first to $\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = \begin{bmatrix} 1 & a_{12}/a_{11} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$, and second to $\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} a_{11} & 0 \\ a_{21} & a_{22}-a_{12}a_{21}/a_{11} \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix}$. This gives the general transformation, as $\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} a_{11} & 0 \\ a_{21} & a_{22}-a_{12}a_{21}/a_{11} \end{bmatrix} \begin{bmatrix} 1 & a_{12}/a_{11} \\ 0 & 1 \end{bmatrix}$. The density of Z is $h(z_1,z_2) = f(z_1-z_2a_{12}/a_{11},z_2)$, and of Y is $g(y_1,y_2) = h(y_1/a_{11},(y_2-y_1a_{21}/a_{11})/(a_{22}-a_{12}a_{21}/a_{11}))$. Substituting for h in the last expression and simplifying gives

$$g(y_1,y_2) = f((a_{22}y_1-a_{12}y_2)/D,(a_{11}y_2-a_{21}y_1)/D)/D,$$

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

We leave as an exercise the proof of the theorem 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 $A = PLDU'Q'$, where P and Q are permutation matrices, L and U are lower 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 previously. □

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

$$y - y^o = \mathbf{A}(x - x^o) + o(x - x^o),$$

where **A** is the *Jacobean* matrix

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

and the notation $o(z)$ means an expression that is small relative to z. Alternately, one has

$$\mathbf{B} = \mathbf{A}^{-1} = \begin{bmatrix} \partial h^1(y^o)/\partial y_1 & \ldots & \partial h^1(y^o)/\partial y_n \\ | & & | \\ \partial h^n(x^o)/\partial y_1 & \ldots & \partial h^n(y^o)/\partial y_n \end{bmatrix}.$$

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. If f(x) is the density of X, then $g(y) = f(h(y))\cdot|\det(\mathbf{B})| = f(h(y))/|\det(\mathbf{A})|$ is the density of Y.

Example 16. Suppose a random vector (X,Z) has a density f(x,z) for x,z > 0, and consider the nonlinear transformation $W = X\cdot Z$ and $Y = X/Z$, which has the inverse transformation $X = (WY)^{1/2}$ and $Z = (W/Y)^{1/2}$. The Jacobean matrix is $B = \begin{bmatrix} W^{-1/2}Y^{1/2}/2 & W^{1/2}Y^{-1/2}/2 \\ W^{-1/2}Y^{-1/2}/2 & -W^{1/2}Y^{-3/2}/2 \end{bmatrix}$, and $\det(\mathbf{B}) = 1/2y$. Hence, the density of (w,y) is $f((wy)^{1/2},(w/y)^{1/2})/2y$.

In principle, it is possible to analyze n-dimensional nonlinear transformations that are not 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 $S = X_1 + \ldots + X_n$. 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 Theorem 3.6 can be applied. In the case of S, the transformation $Y_1 \equiv S$ filled out by $Y_i = X_i$ for $i = 2,\ldots,n$ is one-to-one, with

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

Example 17. Consider a random vector (X,Z) with a density f(x,z), and the transformation S = X + Z and T = Z, or $\begin{bmatrix} S \\ T \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Z \end{bmatrix}$. The Jacobean of this transformation is one, and its inverse is $\begin{bmatrix} X \\ Z \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} S \\ T \end{bmatrix}$, so the density of (S,T) is g(s,t) = f(s-t,t). The marginal density of S is then $g_1(s)$ = $\int_{t=-\infty}^{+\infty}$ f(s-t,t)dt. If X and Z are statistically independent, so that their density is $f(x,z) = f_1(x) \cdot f_2(z)$, then this becomes $g_1(s) = \int_{t=-\infty}^{+\infty} f_1(s-t) \cdot f_2(t)dt$. This is termed a *convolution* formula.

3.7. SPECIAL DISTRIBUTIONS

3.7.1. 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 tables at the end of this Chapter list many of these distributions.

3.7.2. Table 3.1 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 occurrence of rare events. The hypergeometric distribution is associated with classical probability experiments of drawing red and white balls from urns, and is also used to approximate many other distributions.

3.7.3. Table 3.2 list a number of continuous distributions, including some basic distributions such as the gamma and beta from which other distributions are constructed. The extreme value and logistic distributions are used in the economic theory of discrete choice, and are also of statistical interest because they have simple closed form CDF's.

3.7.4. The normal distribution and its related distributions play a central role in econometrics, both because they provide the foundation for finite-sample distribution results for regression models with normally distributed disturbances, and because they appear as limiting approximations in large samples even when the finite sample distributions are unknown or intractable. Table 3.3 lists the normal distribution, and a number of other distributions that are related to it. The t and F distributions appear in the theory of hypothesis testing, and the chi-square distribution appears in large-sample approximations. The non-central versions of these distributions appear in calculations of the power of hypothesis tests.

It is a standard exercise in mathematical statistics to establish the relationships between normal, chi-square, F, and t distributions. For completeness, we state the most important result:

Theorem 3.9. Normal and chi-square random variables have the following properties:
(I) If $S = Y_1^2 + ... + Y_k^2$, where the Y_k are independent normal random variables with means μ_k and unit variances, then S has a non-central chi-square distribution with degrees of freedom parameter k and non-centrality parameter $\delta = \mu_1^2 + ... + \mu_k^2$, denoted $\chi'^2(k,\delta)$. If $\delta = 0$, this is a (central) chi-square distribution with degrees of freedom parameter k, denoted $\chi^2(k)$.
(ii) If Y and S are independent, Y is normal with mean λ and unit variance, and S is chi-square with k degrees of freedom, then $T = Y/(S/k)^{1/2}$ is non-central t-distributed with degrees of freedom parameter k and non-centrality parameter λ, denoted $t'(k,\lambda)$. If $\lambda = 0$, this is a (central) t-distribution with degrees of freedom parameter k, denoted t(k).
(iii) If R and S are independent, R is non-central chi-square with degrees of freedom parameter k and non-centrality parameter δ, and S is central chi-square with degrees of freedom parameter n, then F = nR/kS is non-central F-distributed with degrees of freedom parameters (k,n) and non-centrality parameter δ, denoted $F'(k,n,\delta)$. If $\delta = 0$, this distribution is F-distributed with degrees of freedom parameters (k,n), and is denoted F(k,n).
(iv) T is non-central t-distributed with degrees of freedom parameter k and non-centrality parameter λ if and only if $F = T^2$ is non-central F-distributed with degrees of freedom parameters (1,k) and non-centrality parameter $\delta = \lambda^2$.

Proof: These results can be found in most classical texts in mathematical statistics; see particularly Rao (1973), pp. 166-167, 170-172, 181-182, Johnson & Kotz (1970), Chap. 26-31, and Graybill (1961), Chap. 4.. □

In applied statistics, it is important to be able to calculate values $x = G^{-1}(p)$, where G is the CDF of the central chi-square, F, or t, distribution, and values p = G(x) where G is the CDF of the non-central chi-square, F, or t distribution. Selected points of these distributions are tabled in many books of mathematical and statistical tables, but it is more convenient and accurate to calculate these values within a statistical or econometrics software package. Most current packages, including TSP, STATA, and SST, can provide these values.

3.7.5. One of the most heavily used distributions in econometrics is the multivariate normal. We describe this distribution and summarize some of its properties. A n×1 random vector **Y** is multivariate normal with a vector of means $\boldsymbol{\mu}$ and a positive definite covariance matrix $\boldsymbol{\Sigma}$ if it has the density

$$n(\mathbf{y} - \boldsymbol{\mu},\boldsymbol{\Sigma}) = (2\pi)^{-n/2} \det(\boldsymbol{\Sigma})^{-1/2} \exp(-((\mathbf{y} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})/2).$$

This density is also sometimes denoted $n(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma})$, and the CDF denoted $N(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma})$. Its characteristic function is $\exp(\iota\boldsymbol{\mu}'\mathbf{t} - \mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}/2)$, and it has the moments E Y = $\boldsymbol{\mu}$ and E $(Y-\boldsymbol{\mu})(Y-\boldsymbol{\mu})' = \boldsymbol{\Sigma}$. From the characteristic function and the rule for linear transformations, one has immediately the property that a linear transformations of a multivariate normal vector is again multivariate normal. Specifically, if Y is distributed $N(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma})$, then the linear transformation $Z = \mathbf{a} + \mathbf{B}Y$, which has mean $\mathbf{a} + \mathbf{B}\boldsymbol{\mu}$ and covariance matrix $\mathbf{B}'\boldsymbol{\Sigma}\mathbf{B}$, is distributed $N(\mathbf{z};\mathbf{a} + \mathbf{B}\boldsymbol{\mu},\mathbf{B}'\boldsymbol{\Sigma}\mathbf{B})$. The dimension of Z need not be the same as the dimension of Y, nor does B have to be of maximum rank; if $\mathbf{B}'\boldsymbol{\Sigma}\mathbf{B}$ is less than full rank, then the distribution of Z is concentrated on an affine linear subspace of dimension n through the point $\mathbf{a} + \mathbf{B}\boldsymbol{\mu}$. Let $\sigma_k = (\Sigma_{kk})^{1/2}$ denote the standard deviation of Y_k, and let $\rho_{kj} = \Sigma_{kj}/\sigma_k\,\sigma_j$ denote the correlation of Y_k and Y_j. Then the covariance matrix $\boldsymbol{\Sigma}$ can be written

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \sigma_n \end{bmatrix} \begin{bmatrix} 1 & \varrho_{12} & .. & \varrho_{1n} \\ \varrho_{21} & 1 & \dots & \varrho_{2n} \\ \vdots & \vdots & & \vdots \\ \varrho_{n1} & \varrho_{n2} & .. & 1 \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \sigma_n \end{bmatrix} = \mathbf{DRD},$$

where $\mathbf{D} = \mathrm{diag}(\sigma_1,...,\sigma_n)$ and $\mathbf{R}$ is the array of correlation coefficients.

Theorem 3.10. Suppose Y is partitioned $\mathbf{Y}' = (\mathbf{Y}_1{}'\ \mathbf{Y}_2{}')$, where $\mathbf{Y}_1$ is m×1, and let $\boldsymbol{\mu}' = (\boldsymbol{\mu}_1{}'\ \boldsymbol{\mu}_2{}')$ and $\begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$ be commensurate partitions of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Then the marginal density of $\mathbf{Y}_1$ is multivariate normal with mean $\boldsymbol{\mu}_1$ and covariance matrix $\boldsymbol{\Sigma}_{11}$. The conditional density of $\mathbf{Y}_2$, given $\mathbf{Y}_1 = \mathbf{y}_1$, is multivariate normal with mean $\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{22}{}^{-1}\boldsymbol{\Sigma}_{21}(\mathbf{y}_1 - \boldsymbol{\mu}_1)$ and covariance matrix $\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\,\boldsymbol{\Sigma}_{11}{}^{-1}\boldsymbol{\Sigma}_{12}$. Then, the conditional mean of a multivariate normal is linear in the conditioning variables.

Proof: The easiest way to demonstrate the theorem is to recall from Chapter 2 that the positive definite matrix $\boldsymbol{\Sigma}$ has a Cholesky factorization $\boldsymbol{\Sigma} = \mathbf{LL}'$, where $\mathbf{L}$ is lower triangular, and that $\mathbf{L}$ has an inverse $\mathbf{K}$ that is again lower triangular. If $\mathbf{Z}$ is a n×1 vector of independent standard normal random variables (e.g., each $\mathbf{Z}_i$ has mean zero and variance 1), then $\mathbf{Y} = \boldsymbol{\mu} + \mathbf{LZ}$ is normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Conversely, if Y has density $n(\mathbf{y} - \boldsymbol{\mu},\boldsymbol{\Sigma})$, then $\mathbf{Z} = \mathbf{K}(Y - \boldsymbol{\mu})$ is a vector of i.i.d. standard normal random variables. These statement use the important property of normal random vectors that a linear transformation is again normal. This can be shown directly by using the formulas in Section 3.6 for densities of linear transformations, or by observing that the (multivariate) characteristic function of Y with density $n(\mathbf{y} - \boldsymbol{\mu},\boldsymbol{\Sigma})$ is $\exp(\,\iota\mathbf{t}'\boldsymbol{\mu} - \mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}/2)$, and the form of this characteristic function is unchanged by linear transformations.

The Cholesky construction $\mathbf{Y} = \boldsymbol{\mu} + \mathbf{LZ}$ provides an easy demonstration for the densities of marginal or conditional subvectors of $\mathbf{Y}$. Partition $\mathbf{L}$ and $\mathbf{Z}$ commensurately with $(\mathbf{Y}_1{}'\ \mathbf{Y}_2{}')$, so that $\mathbf{L} = \begin{bmatrix} L_{11} & \mathbf{0} \\ L_{21} & L_{22} \end{bmatrix}$ and $Z' = (\mathbf{Z}_1{}'\ \mathbf{Z}_2{}')$. Then $\boldsymbol{\Sigma}_{11} = \mathbf{L}_{11}\mathbf{L}_{11}{}'$, $\boldsymbol{\Sigma}_{21} = \mathbf{L}_{21}\,L_{11}{}'$, $\boldsymbol{\Sigma}_{22} = \mathbf{L}_{22}\mathbf{L}_{22}{}' + \mathbf{L}_{21}\mathbf{L}_{21}{}'$, and hence $\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}{}^{-1} = \mathbf{L}_{21}\mathbf{L}_{11}{}^{-1}$, implying $\mathbf{L}_{22}\mathbf{L}_{22}{}' = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}{}^{-1}\boldsymbol{\Sigma}_{12}$. Then, $\mathbf{Y}_1 = \boldsymbol{\mu}_1 + \mathbf{L}_{11}\mathbf{Z}_1$ has a marginal multivariate normal density with mean $\boldsymbol{\mu}_1$ and covariance matrix $\mathbf{L}_{11}\mathbf{L}_{11}{}' = \boldsymbol{\Sigma}_{11}$. Also, $\mathbf{Y}_2 = \boldsymbol{\mu}_2 + \mathbf{L}_{21}\mathbf{Z}_1 + \mathbf{L}_{22}Z_2$, implying $\mathbf{Y}_2 = \boldsymbol{\mu}_2 + \mathbf{L}_{21}L_{11}{}^{-1}(\mathbf{Y}_1 - \boldsymbol{\mu}_1) + \mathbf{L}_{22}\mathbf{Z}_2$. Conditioned on $\mathbf{Y}_1 = \mathbf{y}_1$, this implies $\mathbf{Y}_2 = \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}{}^{-1}(\mathbf{y}_1 - \boldsymbol{\mu}_1) + \mathbf{L}_{22}Z_2$ is multivariate normal with mean $\boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}{}^{-1}\boldsymbol{\mu}_1$ and covariance matrix $\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}{}^{-1}\boldsymbol{\Sigma}_{12}$. □

The next theorem gives some additional useful properties of the multivariate normal and of quadratic forms in normal vectors.

Theorem 3.11. Let Y be a n×1 random vector. Then,

(i) If $\mathbf{Y}' = (\mathbf{Y}_1{}'\ \mathbf{Y}_2{}')$ is multivariate normal, then $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are independent if and only if they are uncorrelated. However, $\mathbf{Y}_1$ and $\mathbf{Y}_2$ can be uncorrelated and each have a marginal normal distribution without necessarily being independent.

(ii) If every linear combination $\mathbf{c}'Y$ is normal, then Y is multivariate normal.

(iii) If Y is i.i.d. standard normal and $\mathbf{A}$ is an idempotent n×n matrix of rank k, then $Y'\mathbf{A}Y$ is distributed $\chi^2(k)$.

(iv) If Y is distributed $N(\boldsymbol{\mu},\mathbf{I})$ and $\mathbf{A}$ is an idempotent n×n matrix of rank k, then $Y'\mathbf{A}Y$ is distributed $\chi'^2(k,\delta)$ with $\delta = \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}$.

(v) If Y is i.i.d. standard normal and $\mathbf{A}$ and $\mathbf{B}$ are positive semidefinite n×n matrices, then $Y'\mathbf{A}Y$ and $Y'\mathbf{B}Y$ are independent if and only if $\mathbf{AB} = \mathbf{0}$.

(vi) If Y is distributed $N(\boldsymbol{\mu},\mathbf{I})$, and $\mathbf{A}_i$ is an idempotent n×n matrix of rank k_i for I = 1,...K, then the $Y'\mathbf{A}_i\,Y$ are mutually independent and distributed $\chi'^2(k_i,\delta_i)$ with $\delta_i = \boldsymbol{\mu}'\mathbf{A}_i\boldsymbol{\mu}$ if and only if either (a) $\mathbf{A}_i\mathbf{A}_j = 0$ for I ≠ j or (b) $\mathbf{A}_1 + ... + \mathbf{A}_K$ is idempotent.

(vii) If Y is distributed $N(\boldsymbol{\mu},\mathbf{I})$, $\mathbf{A}$ is a positive semidefinite n×n matrix, $\mathbf{B}$ is a k×n matrix, and $\mathbf{BA} = \mathbf{0}$, then $\mathbf{B}Y$ and $Y'\mathbf{A}Y$ are independent.

(viii) If Y is distributed $N(\boldsymbol{\mu},\mathbf{I})$ and $\mathbf{A}$ is a positive semidefinite n×n matrix, then $E\ Y'\mathbf{A}Y = \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} + \mathrm{tr}(\mathbf{A})$.

Proof: Results (i) and (ii) are proved in Anderson (1958), Thm. 2.4.2 and 2.6.2. For (iii) and (iv), write $\mathbf{A} = \mathbf{UU}'$, where this is its singular value decomposition with $\mathbf{U}$ a n×k column orthogonal matrix. Then $\mathbf{U}'Y$ is distributed $N(\mathbf{U}'\boldsymbol{\mu},\mathbf{I}_k)$, and the result follows from Theorem 3.8. For (v), let k be the rank of $\mathbf{A}$ and m the rank of $\mathbf{B}$. There exists a n×k matrix $\mathbf{U}$ of rank k and a n×m matrix $\mathbf{V}$ of rank m such that $\mathbf{A} = \mathbf{UU}'$ and $\mathbf{B} = \mathbf{VV}'$. The vectors $\mathbf{U}'Y$ and $\mathbf{V}'Y$ are uncorrelated, hence independent, if and only if $\mathbf{U}'\mathbf{V} = \mathbf{0}$. But $\mathbf{AB} = \mathbf{U}(\mathbf{U}'\mathbf{V})\mathbf{V}'$ is zero if and only if $\mathbf{U}'\mathbf{V} = \mathbf{0}$ since $\mathbf{U}$ and $\mathbf{V}'$ are of maximum rank. For (vi), use the SVD decomposition as in (iv). For (vii), write $\mathbf{A} = \mathbf{U}\mathbf{U}'$ with $\mathbf{U}$ of maximum rank as in (v). Then $\mathbf{BA} = (\mathbf{BU})\mathbf{U}' = \mathbf{0}$ implies $\mathbf{BU} = \mathbf{0}$, so that $\mathbf{B}Y$ and $\mathbf{U}'Y$ are independent by (i). For (vii), $E\, Y'\mathbf{A}Y = \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} + E\,(Y-\boldsymbol{\mu})'\mathbf{A}(Y-\boldsymbol{\mu}) = \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} + \mathrm{tr}(E\,(Y-\boldsymbol{\mu})'\mathbf{A}(Y-\boldsymbol{\mu})) = \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} + \mathrm{tr}(\mathbf{A})$. □

3.8. NOTES AND COMMENTS

The purpose of this chapter has been to collect the key results from probability theory that are used in econometrics. While the chapter is reasonably self-contained, it is expected that the reader will already be familiar with most of the concepts, and can if necessary refer to one of the excellent texts in basic probability theory and mathematical statistics, such as P. Billingsley, *Probability and Measure*, Wiley, 1986; or Y. Chow and H. Teicher, *Probability Theory*, 1997. A classic that provides an accessible treatment of fields of subsets, measure, and statistical independence is J. Neveu, *Mathematical Foundations of the Calculus of Probability*, Holden-Day, 1965. Another classic that contains many results from mathematical statistics is C. R. Rao (1973) *Linear Statistical Inference and Its Applications*, Wiley. A comprehensive classical text with treatment of many topics, including characteristic functions, is W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 1&2, Wiley, 1957. For special distributions, properties of distributions, and computation, a four-volume compendium by N. Johnson and S. Kotz, Distributions in Statistics, Houghton-Mifflin, 1970, is a good source. For the multivariate normal distribution, T. Anderson (1958) *An Introduction to Multivariate Statistical Analysis*, Wiley, and F. Graybill (1961) *An Introduction to Linear Statistical Models*, McGraw-Hill, are good sources. Readers who find some sections of this chapter unfamiliar or too dense may find it useful to first review an introductory text at the undergraduate level, such as K. Chung, *A Course in Probability Theory*, Academic Press, New York, or R. Larsen and M. Marx, *Probability Theory*, Prentice-Hall.

TABLE 3.1. SPECIAL DISCRETE DISTRIBUTIONS

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

NOTES

1. $\mu \equiv EX$ (the mean), and $\sigma^2 = E(X-\mu)^2$ (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 $p/(1-(1-p)e^{it})$ and the moment generating function is $p/(1-(1-p)e^t)$, defined for $t < -\ln(1-p)$.
4. The moment generating function is $\exp(\lambda(e^t-1))$, defined for all t.
5. The characteristic function is $p^r/(1-(1-p)e^{it})^r$, and the moment generating function is $p^r/(1-(1-p)e^t)^r$, defined for $t < -\ln(1-p)$.

TABLE 3.2. SPECIAL CONTINUOUS DISTRIBUTIONS

NAME & DOMAIN	DENSITY	MOMENTS	CHAR. FN.
1. Uniform $a \le x \le b$	1/(b-a)	$\mu = (a+b)/2$ $\sigma^2 = (b-a)^2/12$	$\dfrac{e^{\iota bt} - e^{\iota at}}{\iota t(b-a)}$ Note 1
2. Triangular $\lvert x\rvert < a$	$(1-\lvert x\rvert/a)/a$	$\mu = 0$ $\sigma^2 = a^2/6$	$2\,\dfrac{1-\cos at}{a^2t^2}$
3. Cauchy $-\infty < x < +\infty$	$a/\pi(a^2 + (x-\mu)^2)$	none	$e^{\iota t\mu - \lvert t\delta\rvert}$
4. Exponential $x \ge 0$	$e^{-x/\lambda}/\lambda$	$\mu = \lambda$ $\sigma^2 = \lambda^2$	$1/(1-\iota\lambda t)$ Note 2
5. Pareto $x \ge a$	ba^bx^{--b-1}	$\mu = ab/(b-1)$ $\sigma^2 = ba^2/(b-1)^2(b-2)$	Note 3
6. Gamma $x > 0$	$\dfrac{x^{a-1}e^{x/b}}{\Gamma(a)b^a}$	$\mu = ab$ $\sigma^2 = ab^2$	$(1-\iota bt)^{-a}$ Note 4
7. Beta $0 < x < 1$	$\dfrac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\,x^{a-1}(1-x)^{b-1}$	$\mu = a/(a+b)$ $\sigma^2 = \dfrac{ab}{(a+b)^2(a+b+1)}$	Note 5
8. Extreme Value $-\infty < x < +\infty$	$\dfrac{1}{b}\,\exp\left(-\dfrac{x-a}{b} - e^{-(x-a)/b}\right)$	$\mu = a + 0.57721\cdot b$ $\sigma^2 = (\pi b)^2/12$	Note 6
9. Logistic $-\infty < x < +\infty$	$\dfrac{1}{b}\cdot\dfrac{\exp((a-x)/b)}{(1+\exp((a-x)/b))^2}$	$\mu = a$ $\sigma^2 = (\pi b)^2/6$	Note 7

NOTES
1. The moment generating function is $(e^{bt} - e^{at})/(b-a)t$, defined for all t.
2. The moment generating function is $1/(1 - \lambda t)$, defined for $t < 1/\lambda$.
3. The moment generating function does not exist. The mean exists for $b > 1$, the variance exists for $b > 2$.
4. For $a > 0$, $\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}dx$ is the gamma function. If a is an integer, $\Gamma(a) = (a-1)!$.
5. For the characteristic function, see C. R. Rao, Linear Statistical Inference, Wiley, 1973, p. 151.
6. The moment generating function is $e^{at}\Gamma(1 - tb)$ for $t < 1/b$.
7. The moment generating function is $e^{at}\pi bt/\sin(\pi bt)$ for $\lvert t\rvert < 1/2b$.

TABLE 3.3. THE NORMAL DISTRIBUTION AND ITS RELATIVES

NAME & DOMAIN	DENSITY	MOMENTS	CHAR. FN.
1. Normal $n(x-\mu,\sigma)$ $-\infty < x < +\infty$, $\sigma > 0$	$(2\pi\sigma^2)^{-1/2}\cdot\exp(\frac{-(x-\mu)^2}{2\sigma^2})$	μ = mean σ^2 = variance	$\exp(\iota t\mu-\sigma^2t^2/2)$ Note 1
2. Standard Normal $-\infty < x < +\infty$	$\varphi(x) = (2\pi)^{-1/2}\cdot\exp(-x^2/2)$	$\mu = 0$ $\sigma^2 = 1$	$\exp(-t^2/2)$
3. Chi-Square $0 < x < +\infty$	$\chi^2(x;k) = \frac{x^{(k/2)-1}\cdot e^{-x/2}}{\Gamma(k/2)2^{k/2}}$	$\mu = k$ $\sigma^2 = 2k$ $k = 1,2,...$	$(1-2\iota t)^{-k/2}$ Note 2
4. F-distribution $0 < x < +\infty$	$F(x;k,n)$ k,n positive integers	$\mu =$ if $n > 2$ $\sigma^2 = \frac{2n^2(k+n-2)}{k(n-2)^2(n-4)}$ if $n > 4$	Note 3
5. t-distribution $-\infty < x < +\infty$	$\frac{\Gamma(\frac{k+1}{2})(1+x^2/k)^{-(k+1)/2}}{\sqrt{k}\ \Gamma(\frac{1}{2})\Gamma(\frac{1+2k}{2})}$	$\mu = 0$ if $k > 1$ $\sigma^2 = k/(k-2)$ if $k > 2$	Note 4
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 5
2. Noncentral F-distribution $x > 0$	$F(x;k,n,\delta)$ k,n positive integers $\delta \geq 0$	if $n > 2$, $\mu = n(k+\delta)/k(n-2)$ if $n > 4$, $\sigma^2 = \frac{2(n/k)^2(k+\delta)^2+(k+2\delta)(n-2)}{(n-2)^2(n-4)}$	Note 6
3. Noncentral t-distribution	$t(x;k,\lambda)$ k pos. integer	$\mu = \frac{\Gamma((k-1)/2)\lambda}{\Gamma(k/2)}$ if $k > 1$ $\sigma^2 = (1+\lambda^2)k/(k-2) - \mu^2$ if $k > 2$	Note 7

NOTES TO TABLE 3.3

1. 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 $\exp(\mu t+\sigma^2 t^2/2)$, defined for all t. The standard normal density is often denoted $\varphi(x)$, and the standard normal CDF is denoted $\Phi(x)$. The general normal and standard normal formulas are related by $n(x-\mu,\sigma^2) = \varphi((x-\mu)/\sigma)/\sigma$ and $N(x-\mu,\sigma^2) = \Phi((x-\mu)/\sigma)$.

2. The moment generating function is $(1-t/2)^{-k/2}$ for $t < 2$. 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.

3. 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 $\dfrac{\Gamma(\frac{k+n}{2})}{\Gamma(\frac{k}{2})\Gamma(\frac{n}{2})}\cdot\dfrac{k^{k/2}n^{n/2}x^{k/2-1}}{(n+kx)^{(k+n)/2}}$. For $n \le 2$, the mean does not exist, and for $n \le 4$, the variance does not exist.

The characteristic and moment generating functions are complicated.

4. 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.

5. 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, $\sum_{j=0}^{\infty} [e^{-\delta/2}(\delta/2)^j/j!]\chi^2(x;k+2j)$.

6. The Non-central F-distribution has a density that is a Poisson mixture of rescaled (central) F-distributed densities, $\sum_{j=0}^{\infty} [e^{-\delta/2}(\delta/2)^j/j!]\ \dfrac{k}{k+2j}\ F(\dfrac{kx}{k+2j};k+2j,n)$. 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.

7. 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 $\delta = \lambda^2$.

3.9 EXERCISES

1. In Example 1, write out all the members of $\boldsymbol{F}$.

2. Prove that a σ-field of events contains *countable* intersections of its members.

3. Example 2 claims that the class of all subsets of countable **S** has greater cardinality than **S** itself. Mathematically, this means that it is not possible to associate a unique element of **S** with each member of the class. Use the following device to convince yourself this is true: Write each number in the unit interval as a fraction in binary notation, $0.b_1b_2$.... Associate with each number the class member that contains the sequence with j heads if and only if bj = 1. Then, the real numbers, which are uncountable, map into unique members of the class, so the class is also uncountable.

4. In Example 4, show that the event "the change is the same on successive days" is not in $\boldsymbol{B}_1\times\boldsymbol{B}_2$, but is a monotone limit of sets in $\boldsymbol{B}_1\times\boldsymbol{B}_2$.

5. Economic agents can make contingent trades only if it is common knowledge if the contingency is realized. In Example 1, Agent 1 knows $\boldsymbol{F}$, Agent 2 knows $\boldsymbol{G}$, Agent 3 knows $\boldsymbol{H}$ = {∅,{HH,TT},{HT,TH},S}. What is the common knowledge of Agents 1 and 2? Of Agents 1 and 3?

6. Suppose, in Example 1, that instead of H and T being equally likely, the probability measure satisfies

HH	HT	TH	TT
0.2	0.4	0.1	0.3

What is the probability that the first coin is heads? That the second coin is heads? That the two coins give the same result?

7. Consider the sequence of functions $f_n(x) = x^{1/n}$ for $0 < x < 1$. These are square integrable. Do they converge to a limit, and if so, what is the convergence strong, in measure, or weak?

8. Consider the probability measure P([0,x]) = x1/2 on $0 \le x \le 1$. Does it meet the Radon-Nikodym conditions for the existence of a probability density?

9. It is known that 0.2 percent of the population is HIV-positive. It is known that a screening test for HIV has a 10 percent chance of incorrectly showing positive when the subject is negative, and a 2 percent chance of incorrectly showing negative when the subject is positive. What proportion of the population that tests positive has HIV?

10. John and Kate are 80 years old. The probability that John will die in the next year is 0.08, and the probability that Kate will die in the next year is 0.05. The probability that John will die, given that Kate dies, is 0.2. What is the probability that both will die? That at least one will die? That Kate will die, given that John dies?

11. The probability that a driver will have an accident next year if she has a Ph.D. is 0.2. The probability she will have an accident if she does not have a Ph.D. is 0.25. The probability the driver has a Ph.D. <u>and</u> an accident is 0.01. What is the probability the driver has a Ph.D.? What is the probability of a Ph.D. given an accident?

12. A quiz show offers you the opportunity to become a millionaire if you answer nine questions correctly. Questions can be easy (E), moderate (M), or hard (H). The respective probabilities that you will answer an E, M, or H question correctly are 2/3, 1/2, and 1/3. If you get an E question, your next question will be E, M, or H with probabilities 1/4, 1/2, and 1/4 respectively. If you get a M question, your next question will be E, M, or H with probabilities 1/3, 1/3, and 1/3 respectively. If you get a H question, your next question will be E, M, or H with probabilities 1/2, 0, and 1/2 respectively. The first question is always an E question. What is the probability that you will become a millionaire? [Hint: Show that the probability of winning if you reach question 9 is independent of whether this question is E, M, or H. Then use backward recursion.]

13. Show that if $\mathbf{A} \subseteq \mathbf{B}$ and $P(\mathbf{A}) > 0$, then $P(\mathbf{C}|\mathbf{A})$ can be either larger or smaller than $P(\mathbf{C}|\mathbf{B})$.

14. An airplane has 100 seats. The probability that a ticketed passenger shows up for the flight is 0.9, and the events that any two different passengers show up is statistically independent. If the airline sells 105 seats, what is the probability that the plane will be overbooked? How many seats can the airline sell, and keep the probability of overbooking to 5 percent or less?

15. Prove that the expectation $\mathbf{E}(X - c)^2$ is minimized when $c = \mathbf{E}X$.

16. Prove that the expectation $\mathbf{E}|X - c|$ is minimized when $c = \text{median}(X)$.

17. What value of c minimizes $\mathbf{E}\{\alpha\cdot\max(X-c,0) + (1-\alpha)\cdot\max(c-X,0)\}$? (Hint: describe the solution in terms of the distribution F of X.]

18. A sealed bid auction has a tract of land for sale to the highest of n bidders. You are bidder 1. Your experience is that the bids of each other bidder is distributed with a Power distribution $F(X) = X^\alpha$ for $0 \le X \le 1$. Your profit if you are successful in buying tract at price y is 1 - y. What should you bid to maximize your expected profit? What is your probability of winning the auction?

19. A random variable X has a normal distribution if its density is $f(x) = (2\pi\sigma^2)^{-1/2}\cdot\exp(-(x-\mu)^2/2\sigma^2)$, where μ and σ^2 are parameters. Prove that X has mean μ and variance σ^2. Prove that $\mathbf{E}(X-\mu)^3 = 0$ and $\mathbf{E}(X-\mu)^4 = 3\sigma^4$. [Hint: First show that $\int x\cdot\exp(-x^2/2)dx = -\exp(-x^2/2)$ and for $k > 1$, the integration by parts formula $\int x^k\cdot\exp(-x^2/2)dx = -x^{k-1}\cdot\exp(-x^2/2) + (k-1)\int x^{k-2}\cdot\exp(-x^2/2)dx$.]

20. Suppose the stock market has two regimes, Up and Down. In an Up regime, the probability that the market index will rise on any given day is P. In a Down regime, the probability that the market index will rise on any given day is Q, with Q < P. Within a regime, the probability that the market rises on a given day is independent of its history. The probability of being in a Up regime is 1/2, so that if you do not know which regime you are in, then all you can say is that the probability that the market will rise on any given day is $R = (P+Q)/2$. Assume that regimes persist far longer than runs of rises, so that when analyzing runs the regime can be treated as persisting indefinitely. Show that when you are in the Up regime, the probability of a run of k or more successive days in which the market rises is P^{k-1}, and that the probability of a run of exactly k days in which the market rises is $P^{k-1}(1-P)$. A similar formula with Q instead of P holds when you are in a Down regime. Show that expected length in an Up regime of a run of rises is $1/(1-P)$. Show that $\frac{1}{2}/(1-P) + \frac{1}{2}/(1-Q) \ge 1/(1-R)$.

21. The random vector (X_1,X_2) has the distribution function $\exp(-(\exp(-2x_1)+\exp(-2x_2))^{1/2})$. What is the marginal distribution of X_1? What is the conditional distribution of X_1 given $X_2 \le c$? Given $X_2 = c$?

22. The expectation $E(X+aZ)^2 \geq 0$ for random variables X, Z and any scalar a.. Use this property to prove the Cauchy-Schwartz inequality.

23. Prove Jensen's inequality for a probability concentrated at two points.

24. In Example 2, use the law of iterated expectations to calculate the expectation of the number of heads, given that the number exceeds one.

25. If X and Z are bivariate normal with means 1 and 2, and variances 1 and 4, respectively, and covariance , what is the density of X given Z = z? Use Bayes law to deduce the conditional density of Z given X = x?

26. Prove the formula for the characteristic function of a standard normal random variable.

27. What is the domain of the moment generating function of an exponentially distributed random variable with density f(x) = exp(-3x) for x > 0?

28. If (X,Z) is a random vector with density f(x,z) and z > 0, and S = X/Z, T = Z, what is the Jacobean of the transformation?

29. If X and Y are multivariate normal with zero means, $EXX' = A$, $EYY' = B$, and $EXY' = C$, show that X and $Z = Y - XB^{-1}C$ are independent.

30. For the binomial distribution b(k;n,p), what is the variance of the frequency f = k/n?

31. The hypergeometric distribution describes the probability that k of n balls drawn from an urn will be red, where the urn contains r red and w white balls, and sampling is without replacement. Calculate the same probability if sampling is with replacement. Calculate the probabilities, with and without replacement, when r = 10, w = 90, n = 5, k = 1.

32. In a Poisson distribution, what is the expected count conditioned on the count being positive?

33. Under what conditions is the characteristic function of a uniform distribution of [-a,b] real?

34. Show that if X and Y are independent identically distributed extreme value, then X - Y is logistic distributed.

35. Suppose that the duration of a spell of unemployment (in days) can be described by a geometric distribution, Prob(k) $= p^k(1-p)$, where $0 < p < 1$ is a parameter and k is a non-negative integer. What is the expected duration of unemployment? What is the probability of a spell of unemployment lasting longer than K days? What is the conditional expectation of the duration of unemployment, given the event that K > m, where m is a positive integer? [Hint: Use formulas for geometric series, see 2.1.10.]

36. Use the moment generating function to find $\mathbf{E}X^3$ when X has density $e^{-x/\lambda}/\lambda$, x > 0.

37. A log normal random variable Y is one that has log(Y) normal. If log(Y) has mean μ and variance σ^2, find the mean and variance of Y. [Hint: It is useful to find the moment generating function of Z = log(Y).]

38. If X and Y are independent normal, then X+Y is again normal, so that one can say that *the normal family is closed under addition*. (Addition of random variables is also called convolution, from the formula for the density of the sum.) Now suppose X and Y are independent and have extreme value distributions, $\text{Prob}(X \le x) = \exp(-e^{a-x})$ and $\text{Prob}(Y \le y) = \exp(-e^{b-y})$, where a and b are location parameters. Show that max(X,Y) once again has an extreme value distribution (with location parameter $c = \log(e^a+e^b)$), so that *the extreme value family is closed under maximization*.

39. If X is standard normal, derive the density and characteristic function of $Y = X^2$, and confirm that this is the same as the tabled density of a chi-square random variable with one degree of freedom. If X is normal with variance one and a mean μ that is not zero, derive the density of Y, which is non-central chi-square distributed with one degree of freedom and noncentrality parameter μ^2.

40. Random Variables X and Y are bivariate normal, with **E**X = 1, **E**Y = 3, and Var(X) = 4, Var(Y) = 9, Covariance(X,Y) = 5.
(a) What is the mean of Z = 2X - Y?
(b) What is the variance of Z = 2X - Y?
© What is the conditional mean of Z given X = 5?
(d) What is the conditional variance of Z given X = 5?

41. What is the probability that the larger of two random observations from any continuous distribution will exceed the population median?

42. If random variables X and Y are independent, with EX = 1, EY = 2, $EX^2 = 4$, $EY^2 = 9$, what is the unconditional mean and variance of $3{\cdot}X{\cdot}Y$? What is the conditional mean and variance of $3{\cdot}X{\cdot}Y$ given Y = 5?

43. Jobs are characterized by a wage rate W and a duration of employment X, and (W,X) can be interpreted as a random vector. The duration of employment has an exponential density $\lambda{\cdot}e^{-\lambda x}$, and the wage rate W has a exponential density, conditioned on X = x, equal to $(\alpha+\beta x)e^{-(\alpha+\beta x)w}$, where λ, α, and β are positive parameters. What is the marginal density of W? The conditional density of X given W?

44. Random Variables X and Y are bivariate normal, with **E**X = 1, **E**Y = 3, and Var(X) = 4, Var(Y) = 9, Covariance(X,Y) = 5.
(a) What is the mean of Z = 2X - Y?
(b) What is the variance of Z = 2X - Y?
© What is the conditional mean of Z given X = 5?
(d) What is the conditional variance of Z given X = 5?

45. The data set nyse.txt in the class data area of the class home page contains daily observations on stock market returns from Jan. 2, 1968 through Dec. 31, 1998, a total of 7806 observations corresponding to days the market was open. There are four variables, in columns delimited by spaces. The first variable (DAT) is the date in *yymmdd* format, the second variable (RNYSE) is the daily return to the NYSE market index, defined as the log of the ratio of the closing value of the index today to the closing index on the previous day the market was open, with distributions (dividends) factored in. The third variable (SP500) is the S&P500 market index, an index of a majority of the high market value stocks in the New York stock exchange. The fourth variable (RTB90) is the rate of interest in the secondary market for 90-day Treasury Bills, converted to a daily rate commensurate with RNYSE..

a. Let $\mathbf{E}_n$ denote a sample average (empirical expectation). Find the sample mean $\mu = \mathbf{E}_n X$, variance $\sigma^2 = \mathbf{E}_n(X - \mu)$, skewness $\mathbf{E}_n(X - \mu)^3/\sigma^3$, and kurtosis $\mathbf{E}_n(X - \mu)^4/\sigma^4 - 3$, for the variables RNYSE and RTB90. Normally distributed random variables have zero skewness and kurtosis in the population. Making an "eyeball" comparison, do the sample moments appear to be consistent with the proposition that RNYSE and RTB90 are normally distributed?

b. For RNYSE, form the *standardized* variable $Z = (RNYSE - \mu)/\sigma$, by subtracting this variable's sample mean and then dividing by the square root of its variance (or standard deviation). Sort the values of Z from low to high, and then construct a new variable Y that equals i/7806 for $1 \leq i \leq 7806$. The values of Z are called the *order statistics* of the sample, and Y is the empirical CDF, a CDF that puts 1/7806 probability at each observed value of RNYSE. Plot Y against $\Phi(Z)$, where Φ is the standard normal CDF. If RNYSE is normal, then these curves will differ only because of sampling noise in Y. Does it appear by eyeball comparison that they are likely to be the same? A particular issue is the theoretical question of whether the distribution of returns has fat tails, so that the variance and higher moments are hard to estimate precisely or may fail to exist. In a normal sample, one would expect that on average 99 percent of standardized observations are less than 2.575 in magnitude. Do the standardized values Z appear to be consistent with this frequency?

c. A claim in the analysis of stock market returns is that the introduction of financial derivatives and index funds through the 1980's made it easier for arbitragers to close windows of profit opportunity. The argument is made that the resulting actions of arbitragers have made the market more volatile. Compare the subsamples of NYSE excess returns (EXCESS = RNYSE - NRTB90) for the periods 1968-1978 and 1988-1998. By eyeball comparison, were there differences in mean excess return in the two decades? In the variance (or standard deviation) of excess return? Now do a 2×2 table of sample means classified by the two decades above and by whether or not the previous day's excess return was above its decade average. Does it appear that the gap between mean excess returns on days following previous rises and falls has increased or shrunk in the decade of the 90's?