Mathematical Probability Theory

**Main Definition (6).** \((\Omega, \mathcal{E}, P)\) is a probability space if \(\Omega\) is a non-empty set (called the "sample space"), \(\mathcal{E}\) is a non-empty \(\sigma\)-algebra of subsets of \(\Omega\) (members of \(\mathcal{E}\) are called "events") and \(P\) is real-valued function on \(\mathcal{E}\) such that

1. For all \(E \in \mathcal{E}\), \(P(E) \geq 0\).
2. \(P(\Omega) = 1\).
3. For all \(E \in \mathcal{E}\) and for all \(F \in \mathcal{E}\), if \(E \cap F = \emptyset\) then \(P(E \cup F) = P(E) + P(F)\).
4. For all sequences \(\{E_k\}_{k=1}^{\infty}\) with \(E_k \in \mathcal{E}\) for all \(k\), if \(E_i \cap E_j = \emptyset\) for all \(i \neq j\), then
   \[
   P\left(\bigcup_{k=1}^{\infty} E_k\right) = \sum_{k=1}^{\infty} P(E_k).
   \]

\(P(E)\) is usually read as "the probability of the event \(E\".

**Definition 1.** \(\mathcal{F}\) is an algebra of subsets of \(S\) means

1. For all \(F \in \mathcal{F}\), \(F \subset S\).
2. For all \(E \in \mathcal{F}\) and all \(F \in \mathcal{F}\), \(E \cup F \in \mathcal{F}\).
3. For all \(E \in \mathcal{F}\), \(E' \in \mathcal{F}\) (where \(E'\) is the complement of \(E\) with respect to \(S\)).

To paraphrase Definition 1, \(\mathcal{F}\) consists of some of the subsets of \(S\) and is closed under complements and unions of two sets.

**Theorem 2.** If \(\mathcal{F}\) is a non-empty algebra of subsets of \(S\), then

1. \(S \in \mathcal{F}\).
2. \(\emptyset \in \mathcal{F}\).
3. For all \(E \in \mathcal{F}\) and all \(F \in \mathcal{F}\), \(E \cap F \in \mathcal{F}\).
4. Given \(\{E_i\}_{i=1}^{n}\) with \(E_i \in \mathcal{F}\), \(\bigcup_{i=1}^{n} E_i \in \mathcal{F}\) (algebras are closed under a finite number of unions).
5. For all \(E \in \mathcal{F}\) and all \(F \in \mathcal{F}\), \(E \cap F \in \mathcal{F}\).
6. Given \(\{E_i\}_{i=1}^{n}\) with \(E_i \in \mathcal{F}\), \(\bigcap_{i=1}^{n} E_i \in \mathcal{F}\) (algebras are closed under a finite number of intersections).

**Definition 3.** \(\mathcal{F}\) is a \(\sigma\)-algebra of subsets of \(S\) if it is an algebra of subsets of \(S\) and, for all sequences \(\{E_k\}_{k=1}^{\infty}\) with \(E_k \in \mathcal{F}\) for all \(k\), \(\bigcup_{k=1}^{\infty} E_k \in \mathcal{F}\).

**Theorem 4.** If \(\mathcal{F}\) is a \(\sigma\)-algebra of subsets of \(S\), and if a sequence \(\{E_k\}_{k=1}^{\infty}\) has \(E_k \in \mathcal{F}\) for all \(k\), then \(\bigcap_{k=1}^{\infty} E_k \in \mathcal{F}\).

**Theorem 5 (The Easiest \(\sigma\)-Algebra).** Given any set \(S\), if we let \(\mathcal{F}\) consist of ALL possible subsets of \(S\), then \(\mathcal{F}\) is a non-empty \(\sigma\)-algebra of subsets of \(S\).

Theorem 5 is the obvious thing to do, when defining \(P(E)\) presents no difficulties for any subset \(E\) of \(S\). Unfortunately, several of our favorite and most useful probabilities (lengths of the intervals, areas of figures in the plane, and those involving integration) DO PRESENT MAJOR DIFFICULTIES SO THAT WE CAN’T USE THEOREM 5 TO AVOID FIGURING OUT A GOOD \(\sigma\)-ALGEBRA TO USE FOR THEM. The good news is that Theorem 5 takes care of 90% of the Math 371 examples, and we ignore the difficulties in the remaining 10% (the difficulties are deep, and are explored in Math 631 and 632).
Theorem 6 (The Easiest Probability Space). Let $S$ be a non-empty set and $f : S \to \mathbb{R}$ be a non-negative function defined on $S$ so that
\begin{enumerate}
\item For all $x \in S$, $f(x) \geq 0$.
\item $\sum_{x \in S} f(x) = 1$.
\end{enumerate}
If we let $\mathcal{E}$ consist of all subsets of $S$ and define $P(E) = \sum_{x \in E} f(x)$ for all $E \subseteq S$, then $(S, \mathcal{E}, P)$ is a probability space. For examples such as this, $f$ is called the "density function".

Definition 7. Let $(\Omega, \mathcal{E}, P)$ be a probability space and let $f : \Omega \to \mathbb{R}$. If, for all $\alpha \in \mathbb{R}$,
\[ \{w \in \Omega : f(w) < \alpha\} \in \mathcal{E}, \]
then we say that $f$ is a (real-valued) random variable (for $(\Omega, \mathcal{E}, P)$) or $f$ is an rvrv on $\Omega$ for short.

An example is that $\Omega$ could be all possible poker hands when playing cards, and $f$ might be the win (or loss) associated with each hand (when playing poker on a slot machine, for example). The key idea is this: we want $P(f < \alpha)$ to make sense, namely as
\[ P(\{w \in \Omega : f(w) < \alpha\}). \]
It won’t make sense (won’t be defined) unless $\{w \in \Omega : f(w) < \alpha\} \in \mathcal{E}$. So when we use the words "random variable", we are signalling our reader that we’ve checked this condition and guarantee that it does make sense to talk about $P(f < \alpha)$.

Theorem 8 (The Easiest Random Variables). Let $(\Omega, \mathcal{E}, P)$ be a probability space with $\mathcal{E}$ consisting of all subsets of $\Omega$. Then EVERY $f$ such that $f : \Omega \to \mathbb{R}$ is an rvrv for $(\Omega, \mathcal{E}, P)$.

Theorem 9 below assures us that, if $f$ is an rvrv, the probability of any inequality such as $\beta < f \leq \alpha$ is defined.

Theorem 9. Let $f$ be a real-valued random variable for $(\Omega, \mathcal{E}, P)$. Then, for all real numbers $\alpha$ and $\beta$,
\begin{enumerate}
\item $\{w \in \Omega : f(w) \leq \alpha\} \in \mathcal{E}$.
\item $\{w \in \Omega : f(w) > \alpha\} \in \mathcal{E}$.
\item $\{w \in \Omega : f(w) \geq \alpha\} \in \mathcal{E}$.
\item $\{w \in \Omega : f(w) = \alpha\} \in \mathcal{E}$.
\item $\{w \in \Omega : \beta \leq f(w) \leq \alpha\} \in \mathcal{E}$.
\item $\{w \in \Omega : \beta < f(w) \leq \alpha\} \in \mathcal{E}$.
\item $\{w \in \Omega : \beta \leq f(w) < \alpha\} \in \mathcal{E}$.
\item $\{w \in \Omega : \beta < f(w) < \alpha\} \in \mathcal{E}$.
\end{enumerate}

Theorem 10 below only provides new information in the cases where Theorem 8 does not apply, but is important for the 10% of difficult probability spaces where defining $P(E)$ can’t be done for all subsets $E$ of the sample space.

Theorem 10. Let $\{f_i\}_{i=1}^n$ be random-variables for $(\Omega, \mathcal{E}, P)$, and suppose that $H(x_1, \ldots, x_n)$ is a continuous real-valued function at $H(f_1(w), f_2(w), \ldots, f_n(w))$ for all $w \in \Omega$. Then $H(f_1, \ldots, f_n)$ is a random-variable for $(\Omega, \mathcal{E}, P)$. 

Theorem 10 takes care of the following examples:

1. \( f_1 + f_2 \). Let \( H(x,y) = x + y \).
2. \( f_1 - f_2 \). Let \( H(x,y) = x - y \).
3. \( f_1 f_2 \). Let \( H(x,y) = xy \).
4. \( f_1 / f_2 \) provided \( f_2(w) \neq 0 \) for all \( w \in \Omega \). Let \( H(x,y) = x/y \) which is continuous where \( y \neq 0 \).
5. \( f_1^n \) for any positive integer \( n \). Let \( H(x,y,...) = x^n \). I could also let \( H(x_1,x_2,... x_n) = x_1 x_2 ... x_n \) and set \( f_j = f_1 \) for \( 1 < j \leq n \).
6. \( f_1 f_2^2 \), provided \( f_1(w) > 0 \) for all \( w \in \Omega \). Let \( H(x,y) = x^y \), which is continuous for \( x > 0 \).
7. \( \cos(f_1 + f_2) \). Let \( H(x,y) = \cos(x + y) \).
8. \( \log(f_1) \), provided \( f_1(w) > 0 \) for all \( w \in \Omega \). Let \( H(x,y) = \log(x) \), which is continuous for \( x > 0 \).
9. \( |f_1| \). Let \( H(x,y) = |x| \).
10. \( \sqrt{f_1^2 + f_2^2} \). Let \( H(x,y) = \sqrt{x^2 + y^2} \).

All of (1) through (10) are rvrv’s under the stated conditions.

The next Theorem describes a special case of Theorem 6; this special case applies to about 50% of the problems in Math 371. It is the case where each point of the sample space has equal density. Thus, for about 50% of the problems in Math 371, the probability of an event \( E \) is computed by counting the things in \( E \) (that number is \( \#E \)) and dividing by the number of things in the sample space.

**Theorem 11.** Let \( S \) be a non-empty, finite set and let \( \mathcal{E} \) consist of all subsets of \( S \). Let \( f : S \to \mathbb{R} \) satisfy conditions (1) and (2) of Theorem 6. Assume also that \( f(x) = f(y) \) for all \( x \) and \( y \) in \( S \). Then

\[
(3) \quad \text{For all } x \in S, \quad f(x) = \frac{1}{\#S}.
\]

If we define \( P(E) \) as in Theorem 6, then

\[
(4) \quad \text{For all } E \subset S, \quad P(E) = \frac{\#E}{\#S}.
\]

When one choose \( \mathcal{E} \) to consist of ALL subsets of a sample space \( S \), the number of subsets is huge, according to the next theorem.

**Theorem 12.** If \( S \) is a finite set and \( \mathcal{E} \) consists of all subsets of \( S \), then

\[
\#\mathcal{E} = 2^{\#S}.
\]

**Theorem 13.** For all probability spaces \((\Omega, \mathcal{E}, P)\),

1. \( P(\emptyset) = 0 \).
2. \( \text{If } A \in \mathcal{E}, \text{ then } P(A') = 1 - P(A) \).
3. \( \text{If } A \in \mathcal{E} \text{ and } B \in \mathcal{E}, \text{ then } P(A) = P(A - B) + P(A \cap B). \) So, \( P(A - B) = P(A) - P(A \cap B) \).
4. \( \text{If } B \subset A, \text{ then } P(A) \geq P(B) \).
Theorem 14. For all probability spaces \((\Omega, \mathcal{E}, P)\), if \(\{A_i\}_{i=1}^n\) is a finite sequence of disjoint events [meaning: \(A_i \cap A_j = \emptyset\) for \(i \neq j\)], then

\[ P(\bigcup_{i=1}^n A_i) = \sum_{i=1}^n P(A_i). \]

Theorem 15. For all probability spaces \((\Omega, \mathcal{E}, P)\), if \(A\) and \(B\) are events, then

\[ P(A \cup B) = P(A) + P(B) - P(A \cap B). \]

If \(A, B\) and \(C\) are events, then

\[ P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C). \]

In general, if \(\{A_i\}_{i=1}^n\) is a finite sequence of events, then

\[ P(\bigcup_{i=1}^n A_i) = \sum_{k=1}^n (-1)^{k-1} \sum_{1 \leq j_1 < j_2 < \ldots < j_k \leq n} P(\bigcap_{t=1}^k A_{j_t}). \]

Definition 16. If \(A\) and \(B\) are events for a probability space \((\Omega, \mathcal{E}, P)\) with \(P(B) > 0\), we define

\[ P(A/B) = \frac{P(A \cap B)}{P(B)}. \]

In many books this is written as \(P(A|B)\); it is read as "the conditional probability of \(A\) given \(B\)".

Theorem 17. Let \(B\) be an event for a probability \((\Omega, \mathcal{E}, P)\) with \(P(B) > 0\). If we define \(P^\#(A) = P(A/B)\) for all events \(A \in \mathcal{E}\), then \((\Omega, \mathcal{E}, P^\#)\) is a probability space. Note that \(P^\#(B) = 1\) and \(P^\#(B') = 0\).

Theorem 18. Let \(A\) and \(B\) be events for a probability space \((\Omega, \mathcal{E}, P)\) with \(P(B) > 0\). Then \(P(A \cap B) = P(A/B) \cdot P(B)\).

Theorem 19. Let \(\{A_i\}_{i=1}^n\) be a sequence of disjoint events for a probability space \((\Omega, \mathcal{E}, P)\) with these two properties: \(P(A_i) > 0\) for \(1 \leq i \leq n\) and

\[ \sum_{i=1}^n P(A_i) = 1. \]

[For example, this equation will hold for disjoint events \(A_i\) whose union is all of \(\Omega\).] Then, for any event \(B\) of this probability space,

\[ P(B) = \sum_{i=1}^n P(B/A_i)P(A_i). \]

The next theorem is Theorem 19 for an infinite sequence of events.
Theorem 20. Let \( \{A_i\}_{i=1}^{\infty} \) be a sequence of disjoint events for a probability space \((\Omega, \mathcal{E}, P)\) with these two properties: \( P(A_i) > 0 \) for all \( i \) and
\[
\sum_{i=1}^{\infty} P(A_i) = 1.
\]
[For example, this equation will hold for disjoint events \( A_i \) whose union is all of \( \Omega \).] Then, for any event \( B \) of this probability space,
\[
P(B) = \sum_{i=1}^{\infty} P(B|A_i)P(A_i).
\]

Theorem 21. Let \( \{A_i\}_{i=1}^{n} \) be a sequence of events for a probability space \((\Omega, \mathcal{E}, P)\).
If
\[
P(\bigcap_{i=1}^{n-1} A_i) > 0,
\]
then
\[
P(\bigcap_{i=1}^{n} A_i) = P(A_1) \cdot \prod_{j=2}^{n} P(A_j/\bigcap_{t=1}^{j-1} A_t).
\]

Definition 22. Let \( \{A_i\}_{i=1}^{n} \) consist of events for a probability space \((\Omega, \mathcal{E}, P)\). The events \( A_i, 1 \leq i \leq n \), are said to be mutually independent if and only if, for all \( 2 \leq k \leq n \) and \( 1 \leq j_1 < j_2 < \ldots < j_k \leq n \),
\[
(P) \quad P(\bigcap_{t=1}^{k} A_{j_t}) = \prod_{t=1}^{k} P(A_{j_t}).
\]
The events \( A_i, 1 \leq i \leq n \) are said to be pairwise independent if and only if, for all \( 1 \leq i < j \leq n \),
\[
P(A_i \cap A_j) = P(A_i) \cdot P(A_j).
\]

Theorem 23. Let \( \{A_i\}_{i=1}^{n} \) be a sequence of events for a probability space \((\Omega, \mathcal{E}, P)\).
The definition of mutual independence for events \( A_i, 1 \leq i \leq n \), specifies \( 2^n - n - 1 \) equations of the form \((P)\) above.

The Grand Theorem (24). Let \( \{A_i\}_{i=1}^{n} \) be a sequence of mutually independent events for a probability space \((\Omega, \mathcal{E}, P)\). Let \( S \subseteq \{1, 2, \ldots, n\} \) and \( T \subseteq \{1, 2, \ldots, n\} \) with \( S \cap T = \emptyset \). Let \( E \) be formed from the \( A_i \)'s with \( i \in S \), by repeated applications of \( \cap, \cup, -, ' \), etc. (any finite number of times). Let \( F \) be formed from the \( A_i \)'s with \( i \in T \) in the same way. Then
\[
P(E \cap F) = P(E) \cdot P(F).
\]
Theorem 25. If \( E \) and \( F \) are independent events for \((\Omega, \mathcal{E}, P)\) and \( P(F) > 0 \), then

\[
P(E/F) = P(E).
\]

This last theorem captures the essence of the idea of independence: knowing \( F \) does not affect the probability \( E \) at all. Knowledge of \( E \) is “independent” of knowledge of \( F \).

Definition 26. A function \( f \) on a probability space \((\Omega, \mathcal{E}, P)\) is a discrete real-valued random variable if and only if, \( f \) is a real-valued random variable for \((\Omega, \mathcal{E}, P)\) and there is a sequence (finite or infinite) of distinct numbers, \( \{x_i\}_{i=1}^{n} \) or \( \{x_i\}_{i=1}^{\infty} \), such that

\[
\sum_{i=1}^{n} P(\{w \in \Omega : f(w) = x_i\}) = 1
\]

or

\[
\sum_{i=1}^{\infty} P(\{w \in \Omega : f(w) = x_i\}) = 1
\]

Theorem 27. If \( f \) is a real-valued random variable on a probability space \((\Omega, \mathcal{E}, P)\) and if the range of \( f \) is finite or enumerable as a sequence, then \( f \) is a discrete real-valued random variable.

Definition 28. Let \( f \) be a discrete real-valued random variable on a probability space \((\Omega, \mathcal{E}, P)\). For each real \( x \), define \( g_f(x) \) by

\[
g_f(x) = P(\{w \in \Omega : f(w) = x\}).
\]

[We often write \( g_f(x) = P(f = x) \) as a short-hand for this definition.] Then \( g_f \) is called the probability distribution function of \( f \).

Theorem 29. Let \( f \) be a real-valued random variable on a probability space \((\Omega, \mathcal{E}, P)\). Suppose that there is a sequence \( \{x_i\}_{i=1}^{n} \) or \( \{x_i\}_{i=1}^{\infty} \) of real numbers such that

\[
\sum_{i=1}^{n} P(\{w \in \Omega : f(w) = x_i\}) = 1 \quad \text{or} \quad \sum_{i=1}^{\infty} P(\{w \in \Omega : f(w) = x_i\}) = 1, \quad \text{resp.}
\]

[By the previous definition, this would make \( f \) be discrete.] Then, if \( x \) is any real number which is not a value of the sequence \( \{x_i\}_{i=1}^{n} \) or \( \{x_i\}_{i=1}^{\infty} \) resp., then

\[
P(\{w \in \Omega : f(w) = x\}) = 0.
\]

More generally, if \( S \subset \mathbb{R} \),

\[
P(\{w \in \Omega : f(w) \in S\}) = \sum_{x \in S} P(\{w \in \Omega : f(w) = x\}).
\]
Theorem 30. Let $f$ be a discrete real-valued random variable on a probability space $(\Omega, \mathcal{E}, P)$, with probability distribution function $g_f$. Then

$$\sum_{x \in \mathbb{R}} g_f(x) = 1.$$ 

Definition 31. Let $\{f_j\}_{j=1}^n$ be a sequence of discrete real-valued random variables on a probability space $(\Omega, \mathcal{E}, P)$. Let

$$g(x_1, x_2, \ldots, x_n) = P(\{w \in \Omega : f_j(w) = x_j \text{ for } 1 \leq j \leq n\}).$$

The function $g$ is called the joint probability distribution function of $\{f_j\}_{j=1}^n$.

Theorem 32. Let $\{f_j\}_{j=1}^n$ be a sequence of discrete real-valued random variables on a probability space $(\Omega, \mathcal{E}, P)$. Let $g$ be the joint probability distribution function of $\{f_j\}_{j=1}^n$. Then

(1) For all real $x_i$’s, $g(x_1, \ldots, x_n) \geq 0$.

Also,

(2) For all real $x_i$’s

$$\sum_{x \in \mathbb{R}} g(x_1, \ldots, x_n) = 1.$$ 

Moreover, for any $j_0$ in $[1, n]$, and for any (fixed) $x_i$ with $1 \leq i \leq n$ but $i \neq j_0$,

(3) $\sum_{x \in \mathbb{R}} g(x_1, \ldots, x_{j_0-1}, x, x_{j_0+1}, \ldots x_n) = \tilde{g}(x_1, \ldots, x_{j_0-1}, x_{j_0+1}, \ldots x_n),$

where $\tilde{g}$ is the joint probability distribution function of the sequence of random variables which is obtained from $\{f_j\}_{j=1}^n$ by deleting the $j_0$-th random variable (and concatenating the remaining random variables into one list). Finally, by using equation 3 and induction, one has this formula: for any $1 \leq j_1 < j_2 < \ldots < j_t < n$, and for (fixed) $x_i$’s (fixed for $i$ distinct from every $j_s$),

(4) $\sum_{\text{all real } x_{j_s}, 1 \leq s \leq t} g(x_1, \ldots, x_n) = g^*(v),$

where $v$ is the vector $(x_1, \ldots, x_n)$ with entries $x_{j_s}$ deleted (and concatenating the remaining entries into one list) and $g^*$ is the joint probability distribution function of the sequence of random variables which is obtained from $\{f_i\}_{i=1}^n$ by deleting the $j_s$-th random variables (for $1 \leq s \leq t$) and concatenating the remaining random variables into one list.

Definition 33. Let $\{f_j\}_{j=1}^n$ be a sequence of discrete real-valued random variables on a probability space $(\Omega, \mathcal{E}, P)$. Let $g$ be the joint probability distribution function of $\{f_j\}_{j=1}^n$, and let $g_{f_j}$ be the individual probability distribution function for each $f_j$. We say that the $f_j$’s are mutually independent if and only if, for all real $x_i$’s,

$$g(x_1, \ldots, x_n) = \prod_{i=1}^n g_{f_i}(x_i).$$
Grand Theorem 34. Let \( \{f_j\}_{j=1}^n \) be a sequence of discrete, mutually independent, real-valued random variables on a probability space \((\Omega, \mathcal{E}, P)\). Let \( S \) and \( T \) be disjoint subsets of the index set \( \{1, 2, \ldots, n\} \). Let \( E \) be an event defined by the values of random variables \( f_i \) for \( i \in S \) and let \( F \) be any event defined by the values of random variables \( f_j \) for \( j \in T \). Then

\[
P(E \cap F) = P(E)P(F).
\]

An example of an \( E \) could be

\[
E = \{ w \in \Omega : f_1(w) < 2 \text{ and } f_3(w) - f_4(w) \geq 1.5 \}
\]

while an example of \( F \) could be

\[
F = \{ w \in \Omega : |f_2(w) \cdot f_6(w)| < 3 \}.
\]

Definition 35. A discrete real-valued random variable \( f \) (on a probability space \((\Omega, \mathcal{E}, P)\)) has an expectation, \( E(f) \), if and only if

\[
\sum_{x \in \mathbb{R}} |x|g_f(x) < \infty,
\]

where \( g_f \) is the probability distribution function for \( f \). If condition (5) holds, we define

\[
E(f) = \sum_{x \in \mathbb{R}} xg_f(x) < \infty.
\]

Theorem 36. Let \( f \) be a discrete real-valued random variable on a probability space \((\Omega, \mathcal{E}, P)\), with probability distribution function \( g_f \). If \( g_f(x) > 0 \) for at most finitely many \( x \), then \( E(f) \) exists (because the summation in condition (5) above has only finitely many non-zero terms, and hence must be finite). In particular, if the range of \( f \) is finite, then \( E(f) \) exists. Even more particularly, if \( \Omega \) is finite, then \( E(f) \) exists.

Below is a table of famous examples of expectations:

<table>
<thead>
<tr>
<th>rvrv, f</th>
<th>( E(f) )</th>
<th>( E(f^2) )</th>
<th>( Var(f) )</th>
<th>pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>constant, ( c )</td>
<td>( c )</td>
<td>( c^2 )</td>
<td>0</td>
<td>183</td>
</tr>
<tr>
<td>indicator, ( A )</td>
<td>( P(A) )</td>
<td>( P(A) )</td>
<td>( P(A)[1 - P(A)] )</td>
<td>182</td>
</tr>
<tr>
<td>( b(n, p) )</td>
<td>( np )</td>
<td>( np(1-p) + n^2p^2 )</td>
<td>( np(1-p) )</td>
<td>207-210</td>
</tr>
<tr>
<td>geometric, ( p )</td>
<td>( p^{-1} )</td>
<td>( 2p^{-2} - p^{-1} )</td>
<td>( p^{-2} - p^{-1} )</td>
<td>320-321</td>
</tr>
<tr>
<td>Poisson, ( \lambda )</td>
<td>( \lambda )</td>
<td>( \lambda + \lambda^2 )</td>
<td>( \lambda )</td>
<td>326-327</td>
</tr>
<tr>
<td>one fair die</td>
<td>( \frac{7}{2} )</td>
<td>( \frac{91}{6} )</td>
<td>( \frac{35}{12} )</td>
<td>199</td>
</tr>
<tr>
<td>hypergeometric, ( N, k, n )</td>
<td>( \frac{nK}{N} )</td>
<td>( \frac{nK}{N} + k(k-1)\frac{n(n-1)}{N(N-1)} )</td>
<td>( \frac{nk(N-k)(N-n)}{N^2(N-1)} )</td>
<td>211-213</td>
</tr>
</tbody>
</table>
Theorem 37. Let \( f \) be a discrete real-valued random variable on a probability space \((\Omega, \mathcal{E}, P)\), with probability distribution function \( g_f \). Suppose that \( H \) is a function such that \( H \circ f \) is defined and real-valued. Then

1. \( H \circ f \) is a discrete, real-valued random variable on \((\Omega, \mathcal{E}, P)\).
2. \( E(H \circ f) \) exists if and only if
   \[
   \sum_{x \in \mathbb{R}} |H(x)|g_f(x) < \infty.
   \]
3. If \( E(H \circ f) \) exists, then
   \[
   E(H \circ f) = \sum_{x \in \mathbb{R}} H(x)g_f(x).
   \]

Theorem 38. Let \( f \) be a discrete real-valued random variable on a probability space \((\Omega, \mathcal{E}, P)\) and let \( c \) be any real constant. Then \( f + c \) is a discrete real-valued random variable on \((\Omega, \mathcal{E}, P)\). Moreover, if \( E(f) \) exists, then \( E(f + c) \) exists and
   \[
   E(f + c) = E(f) + c.
   \]

Theorem 39. Let \( f \) be a discrete real-valued random variable on a probability space \((\Omega, \mathcal{E}, P)\). Then \( E(f) \) exists if and only if, for any real constant \( c \), \( E(f + c) \) exists. When either exists (and hence both), \( E(f + c) = E(f) + c \).

Theorem 40. Let \( f \) be a discrete real-valued random variable on a probability space \((\Omega, \mathcal{E}, P)\) and let \( c \) be any real constant. Then \( cf \) is a discrete real-valued random variable. Moreover, if \( E(f) \) exists, then \( E(cf) \) exists and
   \[
   E(cf) = cE(f).
   \]

Theorem 41. Let \( f \) be a discrete real-valued random variable on a probability space \((\Omega, \mathcal{E}, P)\). Then \( E(f) \) exists if and only if, for any real constant \( c \neq 0 \), \( E(cf) \) exists.

Theorem 42. Let \( f \) be a discrete real-valued random variable on a probability space \((\Omega, \mathcal{E}, P)\). If \( E(f^2) \) exists, then \( E(f) \) exists.

Jensen’s Inequality (43). Let \( f \) be a discrete real-valued random variable on a probability space \((\Omega, \mathcal{E}, P)\). Let \( H \) be a real-valued function defined on a subset of \( \mathbb{R} \) which is concave up [e.g., \( H'' \geq 0 \)] on an interval containing the range of \( f \) (which implies that \( H \circ f \) is defined and real-valued). If both \( E(f) \) and \( E(H \circ f) \) exist, then
   \[
   H(E(f)) \leq E(H \circ f).
   \]

Examples of \( H \) that are concave up on \( \mathbb{R} \) include \( H(x) = x^2 \), \( H(x) = e^x \), \( H(x) = |x| \), and \( H(x) = |x|^a \) for any \( a \geq 1 \). Thus, if both \( E(|f|^3) \) and \( E(f) \) exist, we have
   \[
   |E(f)|^3 \leq E(|f|^3).
   \]
**Theorem 44.** Let $f$ be a discrete real-valued random variable on a probability space $(\Omega, \mathcal{E}, P)$ and $c$ any real constant. Then $f^2$ and $(f-c)^2$ are discrete, real-valued random variables. Moreover, $E(f^2)$ exists if and only if $E((f-c)^2)$ exists. If $E(f^2)$ exists, then

\[
E((f-c)^2) = E(f^2) - 2cE(f) + c^2 = E(f^2) - [E(f)]^2 + (E(f) - c)^2.
\]

$E((f-c)^2)$ reaches a minimum when $c = E(f)$, and that minimum is $E(f^2) - (E(f))^2$.

**Definition 45.** Let $f$ be a discrete real-valued random variable on a probability space $(\Omega, \mathcal{E}, P)$. If $E(f^2)$ exists, we define $\text{Var}(f)$, read as the variance of $f$, as $E((f-E(f))^2)$. By equation 9,

\[
\text{Var}(f) = E((f-E(f))^2) = E(f^2) - [E(f)]^2.
\]

We define the standard deviation $\sigma(f)$ as $\sqrt{\text{Var}(f)}$.

**Theorem 46.** Let $f$ be a discrete real-valued random variable on a probability space $(\Omega, \mathcal{E}, P)$; assume that $E(f^2)$ exists. Then $E((f+c)^2)$ and $E((cf)^2)$ exist for any real constant $c$, and

\[
\text{Var}(f+c) = \text{Var}(f) \quad \text{and} \quad \text{Var}(cf) = c^2 \text{Var}(f).
\]

**Chebyshev’s Inequality (47).** Let $f$ be a discrete real-valued random variable on a probability space $(\Omega, \mathcal{E}, P)$; assume that $E(f^2)$ exists. For any $\epsilon > 0$,

\[
P(\{w \in \Omega : |f(w)| \geq \epsilon\}) \leq \frac{E(f^2)}{\epsilon^2}.
\]

Since $E((f-E(f))^2)$ also exists, the inequality above can be applied to $f - E(f)$:

\[
P(\{w \in \Omega : |f(w) - E(f)| \geq \epsilon\}) \leq \frac{\text{Var}(f)}{\epsilon^2}.
\]

**Theorem 48.** Let $\{f_i\}_{i=1}^n$ be a sequence of discrete, real-valued random variables on a probability space $(\Omega, \mathcal{E}, P)$. Suppose that $g(x_1, x_2, \ldots, x_n)$ be the joint probability distribution function for $\{f_i\}_{i=1}^n$, and that $H$ is a function such that $Y(w) = H(f_1(w), f_2(w), \ldots, f_n(w))$ is a real number for all $w \in \Omega$. Then

1. $Y$ is a discrete, real-valued random variable.
2. $E(Y)$ exists if and only if

\[
\sum_{\text{all real } x_i's} |H(x_1, \ldots, x_n)|g(x_1, \ldots, x_n) < \infty.
\]

3. If $E(Y)$ exists, then

\[
E(Y) = \sum_{\text{all real } x_i's} H(x_1, \ldots, x_n)g(x_1, \ldots, x_n) < \infty.
\]

Examples of $H$ include $H(x, y) = x + y$ (so that $Y$ is $f_1 + f_2$), $H(x, y) = x - y$ (so that $Y = f_1 - f_2$), and $H(x, y) = xy$ (so that $Y$ is $f_1 \cdot f_2$).
Theorem 49. Let $f_1$ and $f_2$ be discrete, real-valued random variables on a probability space $(\Omega, \mathcal{E}, P)$. For any real numbers $a$ and $b$, then $af_1 + bf_2$ is a discrete random variable. Moreover, if $E(f_1)$ and $E(f_2)$ exist, then $E(af_1 + bf_2)$ exists and

$$E(af_1 + bf_2) = aE(f_1) + bE(f_2).$$

Theorem 50. Let $f_1$ and $f_2$ be discrete, real-valued random variables on a probability space $(\Omega, \mathcal{E}, P)$. If $E(f_1^2)$ and $E(f_2^2)$ exist, then $E(f_1f_2)$ exists and

$$|E(f_1f_2)|^2 \leq E(f_1^2)E(f_2^2).$$

Definition 51. Let $f_1$ and $f_2$ be discrete, real-valued random variables on a probability space $(\Omega, \mathcal{E}, P)$. If $E(f_1^2)$ and $E(f_2^2)$ exist, we know that $E((f_1 - E(f_1))^2)$ and $E((f_2 - E(f_2))^2)$ both exist [see previous theorems] and hence $E((f_1 - E(f_1))(f_2 - E(f_2)))$ exists by Theorem 48. Thus, when $E(f_1^2)$ and $E(f_2^2)$ exist, we can make the following definition:

$$\text{cov}(f_1, f_2) = E(((f_1 - E(f_1))(f_2 - E(f_2))).$$

This is called the covariance of $f_1$ and $f_2$. If, in addition, $\text{Var}(f_1) \neq 0$ and $\text{Var}(f_2) \neq 0$, we define

$$r(f_1, f_2) = \frac{\text{cov}(f_1, f_2)}{\sqrt{\text{Var}(f_1)\text{Var}(f_2)}}.$$

This is called the correlation coefficient of $f_1$ and $f_2$.

Theorem 52. Let $f_1$ and $f_2$ be discrete, real-valued random variables on a probability space $(\Omega, \mathcal{E}, P)$, with both $E(f_1^2)$ and $E(f_2^2)$ existing. Then

$$\text{cov}(f_1, f_2) = E(f_1f_2) - E(f_1)E(f_2)$$

and

$$|\text{cov}(f_1, f_2)|^2 \leq \text{Var}(f_1)\text{Var}(f_2).$$

In addition, if $\text{Var}(f_1) \neq 0$ and $\text{Var}(f_2) \neq 0$, then

$$|r(f_1, f_2)| \leq 1.$$

Equality in statements 18 and 19 occurs if and only if there are real numbers $a$ and $b$ with $b \neq 0$ such that $f_2 = bf_1 + a$ almost everywhere (equality holds for all $w \in \Omega$ except possibly on a subset of probability 0).

If you think of $f_1$ and $f_2$ as vectors, you can interpret $r(f_1, f_2)$ as the cosine of the angle between them (notice that $-1 \leq r(f_1, f_2) \leq 1$ when it is defined). When $r(f_1, f_2) = 0$ (this implies $E(f_1f_2) = E(f_1)E(f_2)$), $f_1$ and $f_2$ are said to be uncorrelated.
Theorem 53. Let $f_1$ and $f_2$ be independent, discrete, real-valued random variables on a probability space $(\Omega, \mathcal{E}, P)$. If $E(f_1)$ and $E(f_2)$ exist, then $E(f_1f_2)$ exists and $E(f_1f_2) = E(f_1)E(f_2)$.

There are examples in the book to show that random variables can be uncorrelated and yet not independent; thus Theorem 51 cannot be made into an "if and only if" theorem.

Theorem 54. Let $\{f_i\}_{i=1}^n$ be a sequence of discrete, real-valued random variables on a probability space $(\Omega, \mathcal{E}, P)$. Assume $E(f_i^2)$ exists for $1 \leq i \leq n$. Then $Y = \sum_{i=1}^n f_i$ is a discrete random variable on $(\Omega, \mathcal{E}, P)$ and $E(Y^2)$ exists. Moreover,

\begin{equation}
\text{Var}(Y) = \sum_{i=1}^n \text{Var}(f_i) + 2 \sum_{1 \leq i < j \leq n} \text{cov}(f_i, f_j).
\end{equation}

If, in addition, the $f_i$’s are mutually independent (and hence $\text{cov}(f_i, f_j) = 0$ for $i \neq j$), then

\begin{equation}
\text{Var}(Y) = \sum_{i=1}^n \text{Var}(f_i).
\end{equation}

Even more particular, if the $f_i$’s all have $\text{Var}(f_i) = \sigma^2$, and $Z = Y/n$ (the average of the $f_i$’s), then

\begin{equation}
\text{Var}(Z) = \frac{n \sigma^2}{n^2} = \frac{\sigma^2}{n}.
\end{equation}