# Chapter 1

# Probability and Expectations

## 1.1 Probability

**Definition 1.1.** *Statistics* is the science of extracting useful information from data.

This chapter reviews some of the tools from probability that are useful for statistics, and the following terms from set theory should be familiar. A *set* consists of distinct elements enclosed by *braces*, eg $\{1, 5, 7\}$. The *universal set* $S$ is the set of all elements under consideration while the *empty set* $\emptyset$ is the set that contains no elements. The set $A$ is a *subset* of $B$, written $A \subseteq B$, if every element in $A$ is in $B$. The *union* $A \cup B$ of $A$ with $B$ is the set of all elements in $A$ or $B$ or in both. The *intersection* $A \cap B$ of $A$ with $B$ is the set of all elements in $A$ and $B$. The *complement* of $A$, written $\overline{A}$ or $A^c$, is the set of all elements in $S$ but not in $A$.

**Theorem 1.1. DeMorgan's Laws:**
a) $\overline{A \cup B} = \overline{A} \cap \overline{B}$.
b) $\overline{A \cap B} = \overline{A} \cup \overline{B}$.

Sets are used in probability, but often different notation is used. For example, the universal set is called the sample space $S$. In the definition of an event below, the special field of subsets $\mathcal{B}$ of the sample space $S$ forming the class of events will not be formally given. However, $\mathcal{B}$ contains all "interesting" subsets of $S$ and every subset that is easy to imagine. The point is that not necessarily all subsets of $S$ are events, but every event $A$ is a subset of $S$.

**Definition 1.2.** The *sample space* $S$ is the set of all possible outcomes of an experiment.

**Definition 1.3.** Let $\mathcal{B}$ be a special field of subsets of the sample space $S$ forming the class of events. Then $A$ is an *event* if $A \in \mathcal{B}$.

**Definition 1.4.** If $A \cap B = \emptyset$, then $A$ and $B$ are *mutually exclusive* or *disjoint events.* Events $A_1, A_2, ...$ are *pairwise disjoint* or *mutually exclusive* if $A_i \cap A_j = \emptyset$ for $i \neq j$.

A *simple event* is a set that contains exactly one element $s_i$ of $S$, eg $A = \{s_3\}$. A *sample point* $s_i$ is a possible outcome.

**Definition 1.5.** A **discrete sample space** consists of a finite or countable number of outcomes.

**Notation.** Generally we will assume that all events under consideration belong to the same sample space $S$.

The *relative frequency interpretation of probability* says that the probability of an event $A$ is the proportion of times that event $A$ would occur if the experiment was repeated again and again infinitely often.

**Definition 1.6: Kolmogorov's Definition of a Probability Function.** Let $\mathcal{B}$ be the class of events of the sample space $S$. A **probability function** $P : \mathcal{B} \rightarrow [0, 1]$ is a set function satisfying the following three properties:
P1) $P(A) \geq 0$ for all events $A$,
P2) $P(S) = 1$, and
P3) if $A_1, A_2, ...$ are pairwise disjoint events, then $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

**Example 1.1.** Flip a coin and observe the outcome. Then the sample space $S = \{H, T\}$. If $P(\{H\}) = 1/3$, then $P(\{T\}) = 2/3$. Often the notation $P(H) = 1/3$ will be used.

**Theorem 1.2.** Let $A$ and $B$ be any two events of $S$. Then
i) $0 \leq P(A) \leq 1$.
ii) $P(\emptyset) = 0$ where $\emptyset$ is the empty set.
iii) **Complement Rule:** $P(A) = 1 - P(\overline{A})$.
iv) **General Addition Rule:** $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
v) If $A \subseteq B$, then $P(A) \leq P(B)$.

vi) **Boole's Inequality:** $P(\cup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i)$ for any events $A_1, A_2, ....$
vii) **Bonferroni's Inequality:** $P(\cap_{i=1}^{n} A_i) \geq \sum_{i=1}^{n} P(A_i) - (n-1)$ for any events $A_1, A_2, ..., A_n$.

The general addition rule for two events is very useful. Given three of the 4 probabilities in iv), the 4th can be found. $P(A \cup B)$ can be found given $P(A)$, $P(B)$ and that $A$ and $B$ are disjoint or independent. The addition rule can also be used to determine whether $A$ and $B$ are independent (see Section 1.3) or disjoint.

## 1.2 Counting

The *sample point method* for finding the probability for event $A$ says that if $S = \{s_1, ..., s_k\}$ then $0 \leq P(s_i) \leq 1$, $\sum_{i=1}^{k} P(s_i) = 1$, and $P(A) = \sum_{i:s_i \in A} P(s_i)$. That is, $P(A)$ is the sum of the probabilities of the sample points in $A$. If all of the outcomes $s_i$ are *equally likely,* then $P(s_i) = 1/k$ and $P(A) =$ (number of outcomes in $A)/k$ if $S$ contains $k$ outcomes.

Counting or combinatorics is useful for determining the number of elements in $S$. The *multiplication rule* says that if there are $n_1$ ways to do a first task, $n_2$ ways to do a 2nd task, ..., and $n_k$ ways to do a $k$th task, then the number of ways to perform the total act of performing the 1st task, then the 2nd task, ..., then the $k$th task is $\prod_{i=1}^{k} n_i = n_1 \cdot n_2 \cdot n_3 \cdots n_k$.

*Techniques for the multiplication principle:*
a) use a slot for each task and write $n_i$ above the $i$th task. There will be $k$ slots, one for each task.
b) Use a tree diagram.

**Definition 1.7.** A *permutation* is an ordered arrangements using $r$ of $n$ distinct objects and the *number of permutations* $= P_r^n$. A special case of permutation formula is

$$P_n^n = n! = n \cdot (n-1) \cdot (n-2) \cdot (n-3) \cdots 4 \cdot 3 \cdot 2 \cdot 1 =$$

$n \cdot (n-1)! = n \cdot (n-1) \cdot (n-2)! = n \cdot (n-1) \cdot (n-2) \cdot (n-3)! = \cdots$ . Generally $n$ is a positive integer, but define $0! = 1$. An application of the multiplication rule can be used to show that $P_r^n = n \cdot (n-1) \cdot (n-2) \cdots (n-r+1) = \dfrac{n!}{(n-r)!}$.

The quantity $n!$ is read "n factorial." Typical problems using $n!$ include the number of ways to arrange $n$ books, to arrange the letters in the word CLIPS (5!), et cetera.

*Recognizing when a story problem is asking for the permutation formula:* The story problem has $r$ slots and *order is important.* No object is allowed to be repeated in the arrangement. Typical questions include *how many ways* are there to "to choose $r$ people from $n$ and arrange in a line," "to make $r$ letter words with no letter repeated" or "to make 7 digit phone numbers with no digit repeated." Key words include *order, no repeated* and *different.*

**Notation.** The symbol $\equiv$ below means the first three symbols are equivalent and equal, but the fourth term is the formula used to compute the symbol. This notation will often be used when there are several equivalent symbols that mean the same thing. The notation will also be used for functions with subscripts if the subscript is usually omitted, eg $g_X(x) \equiv g(x)$. The symbol $\binom{n}{r}$ is read "$n$ choose $r$," and is called a binomial coefficient.

**Definition 1.8.** A *combination* is an unordered selection using $r$ of $n$ distinct objects. The *number of combinations* is

$$C(n,r) \equiv C^n_r \equiv \binom{n}{r} = \frac{n!}{r!(n-r)!}.$$

Combinations are used in story problems where *order is not important.* Key words include *committees, selecting* (eg 4 people from 10), *choose*, *random sample* and *unordered.*

## 1.3 Conditional Probability and Independence

**Definition 1.9.** The **conditional probability** of $A$ **given** $B$ is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

if $P(B) > 0$.

It is often useful to think of this probability as an experiment with sample space $B$ instead of $S$.

**Definition 1.10.** Two events $A$ and $B$ are **independent**, written $A \perp\!\!\!\perp B$, if

$$P(A \cap B) = P(A)P(B).$$

If $A$ and $B$ are not independent, then $A$ and $B$ are *dependent.*

**Definition 1.11.** A collection of events $A_1, ..., A_n$ are *mutually independent* if for *any* subcollection $A_{i_1}, ..., A_{i_k}$,

$$P(\cap_{j=1}^k A_{i_j}) = \prod_{j=1}^k P(A_{i_j}).$$

Otherwise the $n$ events are *dependent.*

**Theorem 1.3.** Assume that $P(A) > 0$ and $P(B) > 0$. Then the two events $A$ and $B$ are *independent* if any of the following three conditions hold:
i) $P(A \cap B) = P(A)P(B)$,
ii) $P(A|B) = P(A)$, or
iii) $P(B|A) = P(B)$.
If *any of these conditions fails to hold,* then $A$ and $B$ are dependent.

The above theorem is useful because only one of the conditions needs to be checked, and often one of the conditions is easier to verify than the other two conditions.

**Theorem 1.4.** a) *Multiplication rule:* If $A_1, ..., A_k$ are events and if the relevant conditional probabilities are defined, then $P(\cap_{i=1}^k A_i) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots P(A_k|A_1 \cap A_2 \cap \cdots \cap A_{k-1})$. In particular, $P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$.

b) *Multiplication rule for independent events:* If $A_1, A_2, ..., A_k$ are independent, then $P(A_1 \cap A_2 \cap \cdots \cap A_k) = P(A_1) \cdots P(A_k)$. If $A$ and $B$ are independent ($k = 2$), then $P(A \cap B) = P(A)P(B)$.

c) *Addition rule for disjoint events:* If $A$ and $B$ are disjoint, then $P(A \cup B) = P(A) + P(B)$. If $A_1, ..., A_k$ are pairwise disjoint, then $P(\cup_{i=1}^k A_i) = P(A_1 \cup A_2 \cup \cdots \cup A_k) = P(A_1) + \cdots + P(A_k) = \sum_{i=1}^k P(A_i)$.

**Example 1.2.** The above rules can be used to find the probabilities of more complicated events. The following probabilities are closed related to Binomial experiments. Suppose that there are $n$ independent identical trials, that $Y$ counts the number of successes and that $\rho$ = probability of success

for any given trial. Let $D_i$ denote a success in the $i$th trial. Then
i) P(none of the n trials were successes) $= (1-\rho)^n = P(Y=0) = P(\overline{D}_1 \cap \overline{D}_2 \cap \cdots \cap \overline{D}_n)$.
ii) P(at least one of the trials was a success) $= 1-(1-\rho)^n = P(Y \geq 1) = 1 - P(Y=0) = 1 - P(none) = P(\overline{\overline{D}_1 \cap \overline{D}_2 \cap \cdots \cap \overline{D}_n})$.
iii) P(all n trials were successes) $= \rho^n = P(Y=n) = P(D_1 \cap D_2 \cap \cdots \cap D_n)$.
iv) P(not all n trials were successes) $= 1-\rho^n = P(Y<n) = 1-P(Y=n) =$ 1 − P(all).
v) P(Y was at least k ) $= P(Y \geq k)$.
vi) P(Y was at most k) $= P(Y \leq k)$.

If $A_1, A_2, ...$ are pairwise disjoint and if $\cup_{i=1}^{\infty} A_i = S$, then the collection of sets $A_1, A_2, ...$ is a *partition* of $S$. By taking $A_j = \emptyset$ for $j > k$, the collection of pairwise disjoint sets $A_1, A_2, ..., A_k$ is a partition of $S$ if $\cup_{i=1}^{k} A_i = S$.

**Theorem 1.5: Law of Total Probability.** If $A_1, A_2, ..., A_k$ form a partition of $S$ such that $P(A_i) > 0$ for $i = 1, ..., k$, then

$$P(B) = \sum_{j=1}^{k} P(B \cap A_i) = \sum_{j=1}^{k} P(B|A_j)P(A_j).$$

**Theorem 1.6: Bayes' Theorem.** Let $A_1, A_2, ..., A_k$ be a partition of $S$ such that $P(A_i) > 0$ for $i = 1, ..., k$, and let $B$ be an event such that $P(B) > 0$. Then

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{k} P(B|A_j)P(A_j)}.$$

**Proof.** Notice that $P(A_i|B) = P(A_i \cap B)/P(B)$ and $P(A_i \cap B) = P(B|A_i)P(A_i)$. Since $B = (B \cap A_1) \cup \cdots \cup (B \cap A_k)$ and the $A_i$ are disjoint, $P(B) = \sum_{j=1}^{k} P(B \cap A_j) = \sum_{j=1}^{k} P(B|A_j)P(A_j)$. QED

**Example 1.3.** There are many medical tests for rare diseases and a positive result means that the test suggests (perhaps incorrectly) that the person has the disease. Suppose that a test for disease is such that if the person has the disease, then a positive result occurs 99% of the time. Suppose that a person without the disease tests positive 2% of the time. Also assume that 1 in 1000 people screened have the disease. If a randomly selected person tests positive, what is the probability that the person has the disease?

Solution: Let $A_1$ denote the event that the randomly selected person has the disease and $A_2$ denote the event that the randomly selected person does not have the disease. If $B$ is the event that the test gives a positive result, then we want $P(A_1|B)$. By Bayes' theorem,

$$P(A_1|B) = \frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2)} = \frac{0.99(0.001)}{0.99(0.001) + 0.02(0.999)}$$

$\approx 0.047$. Hence instead of telling the patient that she has the rare disease, the doctor should inform the patient that she is in a high risk group and needs further testing.

## 1.4 The Expected Value and Variance

**Definition 1.12.** A *random variable* (RV) $Y$ is a real valued function with a sample space as a domain: $Y : S \to \Re$ where the set of real numbers $\Re = (-\infty, \infty)$.

**Definition 1.13.** Let $S$ be the sample space and let $Y$ be a random variable. Then the *(induced) probability function* for $Y$ is $P_Y(Y = y_i) \equiv P(Y = y_i) = P_S(\{s \in S : Y(s) = y_i\})$. The sample space of $Y$ is
$S_Y = \{y_i \in \Re : \text{there exists an } s \in S \text{ with } Y(s) = y_i\}$.

**Definition 1.14.** The *population* is the entire group of objects from which we want information. The *sample* is the part of the population actually examined.

**Example 1.4.** Suppose that 5 year survival rates of 100 lung cancer patients are examined. Let a 1 denote the event that the $i$th patient died within 5 years of being diagnosed with lung cancer, and a 0 if the patient lived. Then outcomes in the sample space $S$ are 100-tuples (sequences of 100 digits) of the form $s = 1010111 \cdots 0111$. Let the random variable $X(s) =$ the number of 1's in the 100-tuple = the sum of the 0's and 1's = the number of the 100 lung cancer patients who died within 5 years of being diagnosed with lung cancer. Notice that $X(s) = 82$ is easier to understand than a 100-tuple with 82 ones and 18 zeroes.

For the following definition, $F$ is a right continuous function if for every real number $x$, $\lim_{y \downarrow x} F(y) = F(x)$. Also, $F(\infty) = \lim_{y\to\infty} F(y)$ and $F(-\infty) = \lim_{y\to-\infty} F(y)$.

**Definition 1.15.** The **cumulative distribution function** (cdf) of any RV $Y$ is $F(y) = P(Y \leq y)$ for all $y \in \Re$. If $F(y)$ is a cumulative distribution function, then $F(-\infty) = 0$, $F(\infty) = 1$, $F$ is a nondecreasing function and $F$ is right continuous.

**Definition 1.16.** A RV is **discrete** if it can assume only a finite or countable number of distinct values. The collection of these probabilities is the *probability distribution* of the discrete RV. The **probability mass function** (pmf) of a discrete RV $Y$ is $f(y) = P(Y = y)$ for all $y \in \Re$ where $0 \leq f(y) \leq 1$ and $\sum_{y:f(y)>0} f(y) = 1$.

**Remark 1.1.** The cdf $F$ of a discrete RV is a step function.

**Example 1.5: Common low level problem.** The sample space of $Y$ is $S_Y = \{y_1, y_2, ..., y_k\}$ and a table of $y_j$ and $f(y_j)$ is given with one $f(y_j)$ omitted. Find the omitted $f(y_j)$ by using the fact that $\sum_{i=1}^k f(y_i) = f(y_1) + f(y_2) + \cdots + f(y_k) = 1$.

**Definition 1.17.** A RV $Y$ is **continuous** if its distribution function $F(y)$ is continuous.

The notation $\forall y$ means "for all $y$."

**Definition 1.18.** If $Y$ is a continuous RV, then the **probability density function** (pdf) $f(y)$ of $Y$ is a function such that

$$F(y) = \int_{-\infty}^{y} f(t)dt \tag{1.1}$$

for all $y \in \Re$. If $f(y)$ is a pdf, then $f(y) \geq 0 \ \forall y$ and $\int_{-\infty}^{\infty} f(t)dt = 1$.

**Theorem 1.7.** If $Y$ has pdf $f(y)$, then $f(y) = \frac{d}{dy}F(y) \equiv F'(y)$ wherever the derivative exists (in this text the derivative will exist everywhere except possibly for a finite number of points).

**Theorem 1.8.** i) $P(a < Y \leq b) = F(b) - F(a)$.
ii) If $Y$ has pdf $f(y)$, then $P(a < Y < b) = P(a < Y \leq b) = P(a \leq Y < b) = P(a \leq Y \leq b) = \int_a^b f(y)dy = F(b) - F(a)$.
iii) If $Y$ has a probability mass function $f(y)$, then $Y$ is discrete and $P(a < Y \leq b) = F(b) - F(a)$, but $P(a \leq Y \leq b) \neq F(b) - F(a)$ if $f(a) > 0$.

**Definition 1.19.** Let $Y$ be a discrete RV with probability mass function

$f(y)$. Then the *mean* or **expected value** of $Y$ is

$$EY \equiv \mu \equiv E(Y) = \sum_{y: f(y)>0} y\ f(y) \tag{1.2}$$

if the sum exists when $y$ is replaced by $|y|$. If $g(Y)$ is a real valued function of $Y$, then $g(Y)$ is a random variable and

$$E[g(Y)] = \sum_{y: f(y)>0} g(y)\ f(y) \tag{1.3}$$

if the sum exists when $g(y)$ is replaced by $|g(y)|$. If the sums are not absolutely convergent, then $E(Y)$ and $E[g(Y)]$ do not exist.

**Definition 1.20.** If $Y$ has pdf $f(y)$, then the *mean* or **expected value** of $Y$ is

$$EY \equiv E(Y) = \int_{-\infty}^{\infty} y f(y) dy \tag{1.4}$$

and

$$E[g(Y)] = \int_{-\infty}^{\infty} g(y) f(y) dy \tag{1.5}$$

provided the integrals exist when $y$ and $g(y)$ are replaced by $|y|$ and $|g(y)|$. If the modified integrals do not exist, then $E(Y)$ and $E[g(Y)]$ do not exist.

**Definition 1.21.** If $E(Y^2)$ exists, then the *variance* of a RV $Y$ is

$$\text{VAR}(Y) \equiv \text{Var(Y)} \equiv \text{V Y} \equiv \text{V(Y)} = \text{E[(Y} - \text{E(Y))}^2]$$

and the *standard deviation* of $Y$ is $\text{SD}(Y) = \sqrt{V(Y)}$. If $E(Y^2)$ does not exist, then $V(Y)$ does not exist.

The following theorem is used over and over again, especially to find $E(Y^2) = V(Y) + (E(Y))^2$. The theorem is valid for all random variables that have a variance, including continuous and discrete RVs. If $Y$ is a Cauchy $(\mu, \sigma)$ RV (see Chapter 10), then neither $E(Y)$ nor $V(Y)$ exist.

**Theorem 1.9: Short cut formula for variance.**

$$V(Y) = E(Y^2) - (E(Y))^2. \tag{1.6}$$

If $Y$ is a discrete RV with sample space $S_Y = \{y_1, y_2, ..., y_k\}$ then

$$E(Y) = \sum_{i=1}^{k} y_i f(y_i) = y_1 f(y_1) + y_2 f(y_2) + \cdots + y_k f(y_k)$$

and $E[g(Y)] = \displaystyle\sum_{i=1}^{k} g(y_i) f(y_i) = g(y_1) f(y_1) + g(y_2) f(y_2) + \cdots + g(y_k) f(y_k)$. In particular,

$$E(Y^2) = y_1^2 f(y_1) + y_2^2 f(y_2) + \cdots + y_k^2 f(y_k).$$

Also

$$V(Y) = \sum_{i=1}^{k} (y_i - E(Y))^2 f(y_i) =$$

$$(y_1 - E(Y))^2 f(y_1) + (y_2 - E(Y))^2 f(y_2) + \cdots + (y_k - E(Y))^2 f(y_k).$$

For a continuous RV $Y$ with pdf $f(y)$, $V(Y) = \int_{-\infty}^{\infty} (y - E[Y])^2 f(y) dy$. Often using $V(Y) = E(Y^2) - (E(Y))^2$ is simpler.

**Example 1.6: Common low level problem.** i) Given a table of $y$ and $f(y)$, find $E[g(Y)]$ and the standard deviation $\sigma = SD(Y)$. ii) Find $f(y)$ from $F(y)$. iii) Find $F(y)$ from $f(y)$. iv) Given that $f(y) = c\ g(y)$, find $c$. v) Given the pdf $f(y)$, find $P(a < Y < b)$, et cetera. vi) Given the pmf or pdf $f(y)$ find $E[Y]$, $V(Y)$, $SD(Y)$, and $E[g(Y)]$. The functions $g(y) = y$, $g(y) = y^2$, and $g(y) = e^{ty}$ are especially common.

**Theorem 1.10.** *Let $a$ and $b$ be any constants and assume all relevant expectations exist.*
*i) $E(a) = a$.*
*ii) $E(aY + b) = aE(Y) + b$.*
*iii) $E(aX + bY) = aE(X) + bE(Y)$.*
*iv) $V(aY + b) = a^2 V(Y)$.*

**Definition 1.22.** The **moment generating function** (mgf) of a random variable $Y$ is

$$m(t) = E[e^{tY}] \tag{1.7}$$

if the expectation exists for $t$ in some neighborhood of 0. Otherwise, the mgf does not exist. If $Y$ is discrete, then $m(t) = \sum_y e^{ty} f(y)$, and if $Y$ is continuous, then $m(t) = \int_{-\infty}^{\infty} e^{ty} f(y) dy$.

**Definition 1.23.** The **characteristic function** (cf) of a random variable $Y$ is $c(t) = E[e^{itY}]$ where the complex number $i = \sqrt{-1}$.

This text does not require much knowledge of theory of complex variables, but know that $i^2 = -1$, $i^3 = -i$ and $i^4 = 1$. Hence $i^{4k-3} = i$, $i^{4k-2} = -1$, $i^{4k-1} = -i$ and $i^{4k} = 1$ for $k = 1, 2, 3, ....$ To compute the cf, the following result will be used. Moment generating functions do not necessarily exist in a neighborhood of zero, but a characteristic function always exists.

**Proposition 1.11.** Suppose that $Y$ is a RV with an mgf $m(t)$ that exists for $|t| < b$ for some constant $b > 0$. Then the cf of $Y$ is $c(t) = m(it)$.

**Definition 1.24.** Random variables $X$ and $Y$ are *identically distributed*, written $X \sim Y$ or $Y \sim F_X$, if $F_X(y) = F_Y(y)$ for all real $y$.

**Proposition 1.12.** Let $X$ and $Y$ be random variables. Then $X$ and $Y$ are identically distributed, $X \sim Y$, if any of the following conditions hold.
a) $F_X(y) = F_Y(y)$ for all $y$,
b) $f_X(y) = f_Y(y)$ for all $y$,
c) $c_X(t) = c_Y(t)$ for all $t$ or
d) $m_X(t) = m_Y(t)$ for all $t$ in a neighborhood of zero.

**Definition 1.25.** The *kth moment* of $Y$ is $E[Y^k]$ while the *kth central moment* is $E[(Y - E[Y])^k]$.

**Theorem 1.13.** Suppose that the mgf $m(t)$ exists for $|t| < b$ for some constant $b > 0$, and suppose that the $k$th derivative $m^{(k)}(t)$ exists for $|t| < b$. Then $E[Y^k] = m^{(k)}(0)$. In particular, $E[Y] = m'(0)$ and $E[Y^2] = m''(0)$.

**Notation.** The natural logarithm of $y$ is $\log(y) = \ln(y)$. If another base is wanted, it will be given, eg $\log_{10}(y)$.

**Example 1.7: Common problem.** Let $h(y)$, $g(y)$, $n(y)$ and $d(y)$ be functions. Review how to find the derivative $g'(y)$ of $g(y)$ and how to find $k$th derivative

$$g^{(k)}(y) = \frac{d^k}{dy^k} g(y)$$

for $k \geq 2$. Recall that the *product rule* is

$$(h(y)g(y))' = h'(y)g(y) + h(y)g'(y).$$

The **quotient rule** is

$$\left(\frac{n(y)}{d(y)}\right)' = \frac{d(y)n'(y) - n(y)d'(y)}{[d(y)]^2}.$$

The **chain rule** is

$$[h(g(y))]' = [h'(g(y))][g'(y)].$$

Know the derivative of $\log(y)$ and $e^y$ and know the chain rule with these functions. Know the derivative of $y^k$.

Then given the mgf $m(t)$, find $E[Y] = m'(0)$, $E[Y^2] = m^{''}(0)$ and $V(Y) = E[Y^2] - (E[Y])^2$.

**Definition 1.26.** Let $f(y) \equiv f_Y(y|\boldsymbol{\theta})$ be the pdf or pmf of a random variable $Y$. Then the set $\mathcal{Y}_{\boldsymbol{\theta}} = \{y | f_Y(y|\boldsymbol{\theta}) > 0\}$ is called the **support** of $Y$. Let the set $\Theta$ be the set of parameter values $\boldsymbol{\theta}$ of interest. Then $\Theta$ is the **parameter space** of $Y$. Use the notation $\mathcal{Y} = \{y | f(y|\boldsymbol{\theta}) > 0\}$ if the support does not depend on $\boldsymbol{\theta}$. So $\mathcal{Y}$ is the support of $Y$ if $\mathcal{Y}_{\boldsymbol{\theta}} \equiv \mathcal{Y}\ \forall \boldsymbol{\theta} \in \Theta$.

**Definition 1.27.** The **indicator function** $I_A(x) \equiv I(x \in A) = 1$ if $x \in A$ and 0, otherwise. Sometimes an indicator function such as $I_{(0,\infty)}(y)$ will be denoted by $I(y > 0)$.

**Example 1.8.** Often equations for functions such as the pmf, pdf or cdf are given only on the support (or on the support plus points on the boundary of the support). For example, suppose

$$f(y) = P(Y = y) = \binom{k}{y} \rho^y (1-\rho)^{k-y}$$

for $y = 0, 1, \ldots, k$ where $0 < \rho < 1$. Then the support of $Y$ is $\mathcal{Y} = \{0, 1, ..., k\}$, the parameter space is $\Theta = (0, 1)$ and $f(y) = 0$ for $y$ not $\in \mathcal{Y}$. Similarly, if $f(y) = 1$ and $F(y) = y$ for $0 \le y \le 1$, then the support $\mathcal{Y} = [0, 1]$, $f(y) = 0$ for $y < 0$ and $y > 1$, $F(y) = 0$ for $y < 0$ and $F(y) = 1$ for $y > 1$.

Since the pmf and cdf are defined for all $y \in \Re = (-\infty, \infty)$ and the pdf is defined for all but finitely many $y$, it may be better to use indicator functions when giving the formula for $f(y)$. For example,

$$f(y) = 1I(0 \le y \le 1)$$

is defined for all $y \in \Re$.

## 1.5 The Kernel Method

**Notation**. Notation such as $E(Y|\boldsymbol{\theta}) \equiv E_{\boldsymbol{\theta}}(Y)$ or $f_Y(y|\boldsymbol{\theta})$ is used to indicate that the formula for the expected value or pdf are for a family of distributions indexed by $\boldsymbol{\theta} \in \Theta$. A major goal of parametric inference is to collect data and estimate $\boldsymbol{\theta}$ from the data.

**Example 1.9.** If $Y \sim N(\mu, \sigma^2)$, then $Y$ is a member of the normal family of distributions with $\boldsymbol{\theta} = \{(\mu, \sigma) | -\infty < \mu < \infty \text{ and } \sigma > 0\}$. Then $E[Y|(\mu, \sigma)] = \mu$ and $V(Y|(\mu, \sigma)) = \sigma^2$. This family has uncountably many members.

The *kernel method* is a widely used technique for finding $E[g(Y)]$.

**Definition 1.28.** Let $f_Y(y)$ be the pdf or pmf of a random variable $Y$ and suppose that $f_Y(y|\boldsymbol{\theta}) = c(\boldsymbol{\theta})k(y|\boldsymbol{\theta})$. Then $k(y|\boldsymbol{\theta}) \geq 0$ is the **kernel** of $f_Y$ and $c(\boldsymbol{\theta}) > 0$ is the constant term that makes $f_Y$ sum or integrate to one. Thus $\int_{-\infty}^{\infty} k(y|\boldsymbol{\theta})dy = 1/c(\boldsymbol{\theta})$ or $\sum_{y \in \mathcal{Y}} k(y|\boldsymbol{\theta}) = 1/c(\boldsymbol{\theta})$.

Often $E[g(Y)]$ is found using "tricks" tailored for a specific distribution. The word "kernel" means "essential part." Notice that if $f_Y(y)$ is a pdf, then $E[g(Y)] = \int_{-\infty}^{\infty} g(y)f(y|\boldsymbol{\theta})dy = \int_{\mathcal{Y}} g(y)f(y|\boldsymbol{\theta})dy$. Suppose that after algebra, it is found that

$$E[g(Y)] = a\ c(\boldsymbol{\theta}) \int_{-\infty}^{\infty} k(y|\boldsymbol{\tau})dy$$

for some constant $a$ where $\boldsymbol{\tau} \in \Theta$ and $\Theta$ is the parameter space. Then the kernel method says that

$$E[g(Y)] = a\ c(\boldsymbol{\theta}) \int_{-\infty}^{\infty} \frac{c(\boldsymbol{\tau})}{c(\boldsymbol{\tau})} k(y|\boldsymbol{\tau})dy = \frac{a\ c(\boldsymbol{\theta})}{c(\boldsymbol{\tau})} \underbrace{\int_{-\infty}^{\infty} c(\boldsymbol{\tau})k(y|\boldsymbol{\tau})dy}_{1} = \frac{a\ c(\boldsymbol{\theta})}{c(\boldsymbol{\tau})}.$$

Similarly, if $f_Y(y)$ is a pmf, then

$$E[g(Y)] = \sum_{y: f(y) > 0} g(y)f(y|\boldsymbol{\theta}) = \sum_{y \in \mathcal{Y}} g(y)f(y|\boldsymbol{\theta})$$

where $\mathcal{Y} = \{y : f_Y(y) > 0\}$ is the support of $Y$. Suppose that after algebra, it is found that

$$E[g(Y)] = a\ c(\boldsymbol{\theta}) \sum_{y \in \mathcal{Y}} k(y|\boldsymbol{\tau})$$

for some constant $a$ where $\boldsymbol{\tau} \in \Theta$. Then the kernel method says that

$$E[g(Y)] = a\ c(\boldsymbol{\theta}) \sum_{y \in \mathcal{Y}} \frac{c(\boldsymbol{\tau})}{c(\boldsymbol{\tau})} k(y|\boldsymbol{\tau}) = \frac{a\ c(\boldsymbol{\theta})}{c(\boldsymbol{\tau})} \underbrace{\sum_{y \in \mathcal{Y}} c(\boldsymbol{\tau}) k(y|\boldsymbol{\tau})}_{1} = \frac{a\ c(\boldsymbol{\theta})}{c(\boldsymbol{\tau})}.$$

The kernel method is often useful for finding $E[g(Y)]$, especially if $g(y) = y$, $g(y) = y^2$ or $g(y) = e^{ty}$. The kernel method is often easier than memorizing a trick specific to a distribution because the kernel method uses the same trick for every distribution: $\sum_{y \in \mathcal{Y}} f(y) = 1$ and $\int_{y \in \mathcal{Y}} f(y)dy = 1$. Of course sometimes tricks are needed to get the kernel $f(y|\boldsymbol{\tau})$ from $g(y)f(y|\boldsymbol{\theta})$. For example, complete the square for the normal (Gaussian) kernel.

**Example 1.10.** To use the kernel method to find the mgf of a gamma $(\nu, \lambda)$ distribution, refer to Section 10.13 and note that

$$m(t) = E(e^{tY}) = \int_0^\infty e^{ty} \frac{y^{\nu-1} e^{-y/\lambda}}{\lambda^\nu \Gamma(\nu)} dy = \frac{1}{\lambda^\nu \Gamma(\nu)} \int_0^\infty y^{\nu-1} \exp[-y(\frac{1}{\lambda} - t)] dy.$$

The integrand is the kernel of a gamma $(\nu, \eta)$ distribution with

$$\frac{1}{\eta} = \frac{1}{\lambda} - t = \frac{1 - \lambda t}{\lambda} \quad \text{so} \quad \eta = \frac{\lambda}{1 - \lambda \mathrm{t}}.$$

Now

$$\int_0^\infty y^{\nu-1} e^{-y/\lambda} dy = \frac{1}{c(\nu, \lambda)} = \lambda^\nu \Gamma(\nu).$$

Hence

$$m(t) = \frac{1}{\lambda^\nu \Gamma(\nu)} \int_0^\infty y^{\nu-1} \exp[-y/\eta] dy = c(\nu, \lambda) \frac{1}{c(\nu, \eta)} =$$

$$\frac{1}{\lambda^\nu \Gamma(\nu)} \eta^\nu \Gamma(\nu) = \left(\frac{\eta}{\lambda}\right)^\nu = \left(\frac{1}{1 - \lambda t}\right)^\nu$$

for $t < 1/\lambda$.

**Example 1.11.** The zeta$(\nu)$ distribution has probability mass function

$$f(y) = P(Y = y) = \frac{1}{\zeta(\nu) y^\nu}$$

where $\nu > 1$ and $y = 1, 2, 3, ....$ Here the zeta function

$$\zeta(\nu) = \sum_{y=1}^{\infty} \frac{1}{y^\nu}$$

for $\nu > 1$. Hence

$$E(Y) = \sum_{y=1}^{\infty} y \frac{1}{\zeta(\nu)} \frac{1}{y^\nu}$$

$$= \frac{1}{\zeta(\nu)} \zeta(\nu - 1) \underbrace{\sum_{y=1}^{\infty} \frac{1}{\zeta(\nu-1)} \frac{1}{y^{\nu-1}}}_{1=sum\ \ of\ \ zeta(\nu-1)\ \ pmf} = \frac{\zeta(\nu-1)}{\zeta(\nu)}$$

if $\nu > 2$. Similarly

$$E(Y^k) = \sum_{y=1}^{\infty} y^k \frac{1}{\zeta(\nu)} \frac{1}{y^\nu}$$

$$= \frac{1}{\zeta(\nu)} \zeta(\nu - k) \underbrace{\sum_{y=1}^{\infty} \frac{1}{\zeta(\nu-k)} \frac{1}{y^{\nu-k}}}_{1=sum\ \ of\ \ zeta(\nu-k)\ \ pmf} = \frac{\zeta(\nu-k)}{\zeta(\nu)}$$

if $\nu - k > 1$ or $\nu > k + 1$. Thus if $\nu > 3$, then

$$V(Y) = E(Y^2) - [E(Y)]^2 = \frac{\zeta(\nu-2)}{\zeta(\nu)} - \left[\frac{\zeta(\nu-1)}{\zeta(\nu)}\right]^2.$$

**Example 1.12.** The generalized gamma distribution has pdf

$$f(y) = \frac{\phi y^{\phi\nu-1}}{\lambda^{\phi\nu}\Gamma(\nu)} \exp(-y^\phi/\lambda^\phi)$$

where $\nu, \lambda, \phi$ and $y$ are positive, and

$$E(Y^k) = \frac{\lambda^k \Gamma(\nu + \frac{k}{\phi})}{\Gamma(\nu)} \ \text{ if } \ k > -\phi\nu.$$

To prove this result using the kernel method, note that

$$E(Y^k) = \int_0^\infty y^k \frac{\phi y^{\phi\nu-1}}{\lambda^{\phi\nu}\Gamma(\nu)} \exp(-y^\phi/\lambda^\phi) dy = \int_0^\infty \frac{\phi y^{\phi\nu+k-1}}{\lambda^{\phi\nu}\Gamma(\nu)} \exp(-y^\phi/\lambda^\phi) dy.$$

This integrand looks much like a generalized gamma pdf with parameters $\nu_k$, $\lambda$ and $\phi$ where $\nu_k = \nu + (k/\phi)$ since

$$E(Y^k) = \int_0^\infty \frac{\phi y^{\phi(\nu+k/\phi)-1}}{\lambda^{\phi\nu}\Gamma(\nu)} \exp(-y^\phi/\lambda^\phi)dy.$$

Multiply the integrand by

$$1 = \frac{\lambda^k \Gamma(\nu + \frac{k}{\phi})}{\lambda^k \Gamma(\nu + \frac{k}{\phi})}$$

to get

$$E(Y^k) = \frac{\lambda^k \Gamma(\nu + \frac{k}{\phi})}{\Gamma(\nu)} \int_0^\infty \frac{\phi y^{\phi(\nu+k/\phi)-1}}{\lambda^{\phi(\nu+k/\phi)}\Gamma(\nu + \frac{k}{\phi})} \exp(-y^\phi/\lambda^\phi)dy.$$

Then the result follows since the integral of a generalized gamma pdf with parameters $\nu_k$, $\lambda$ and $\phi$ over its support is 1. Notice that $\nu_k > 0$ implies $k > -\phi\nu$.

## 1.6 Mixture Distributions

Mixture distributions are often used as outlier models. The following two definitions and proposition are useful for finding the mean and variance of a mixture distribution. Parts a) and b) of Proposition 1.14 below show that the definition of expectation given in Definition 1.30 is the same as the usual definition for expectation if $Y$ is a discrete or continuous random variable.

**Definition 1.29.** The distribution of a random variable $Y$ is a *mixture distribution* if the cdf of $Y$ has the form

$$F_Y(y) = \sum_{i=1}^k \alpha_i F_{W_i}(y) \tag{1.8}$$

where $0 < \alpha_i < 1$, $\sum_{i=1}^k \alpha_i = 1$, $k \geq 2$, and $F_{W_i}(y)$ is the cdf of a continuous or discrete random variable $W_i$, $i = 1, ..., k$.

**Definition 1.30.** Let $Y$ be a random variable with cdf $F(y)$. Let $h$ be a function such that the expected value $E[h(Y)]$ exists. Then

$$E[h(Y)] = \int_{-\infty}^\infty h(y)dF(y). \tag{1.9}$$

**Proposition 1.14.** a) If $Y$ is a discrete random variable that has a pmf $f(y)$ with support $\mathcal{Y}$, then

$$E[h(Y)] = \int_{-\infty}^{\infty} h(y)dF(y) = \sum_{y \in \mathcal{Y}} h(y)f(y).$$

b) If $Y$ is a continuous random variable that has a pdf $f(y)$, then

$$E[h(Y)] = \int_{-\infty}^{\infty} h(y)dF(y) = \int_{-\infty}^{\infty} h(y)f(y)dy.$$

c) If $Y$ is a random variable that has a mixture distribution with cdf $F_Y(y) = \sum_{i=1}^k \alpha_i F_{W_i}(y)$, then

$$E[h(Y)] = \int_{-\infty}^{\infty} h(y)dF(y) = \sum_{i=1}^k \alpha_i E_{W_i}[h(W_i)]$$

where $E_{W_i}[h(W_i)] = \int_{-\infty}^{\infty} h(y)dF_{W_i}(y)$.

**Example 1.13.** Proposition 1.14c implies that the pmf or pdf of $W_i$ is used to compute $E_{W_i}[h(W_i)]$. As an example, suppose the cdf of $Y$ is $F(y) = (1-\epsilon)\Phi(y) + \epsilon\Phi(y/k)$ where $0 < \epsilon < 1$ and $\Phi(y)$ is the cdf of $W_1 \sim N(0,1)$. Then $\Phi(x/k)$ is the cdf of $W_2 \sim N(0,k^2)$. To find $E[Y]$, use $h(y) = y$. Then

$$E[Y] = (1-\epsilon)E[W_1] + \epsilon E[W_2] = (1-\epsilon)0 + \epsilon 0 = 0.$$

To find $E[Y^2]$, use $h(y) = y^2$. Then

$$E[Y^2] = (1-\epsilon)E[W_1^2] + \epsilon E[W_2^2] = (1-\epsilon)1 + \epsilon k^2 = 1 - \epsilon + \epsilon k^2.$$

Thus $\text{VAR}(Y) = E[Y^2] - (E[Y])^2 = 1 - \epsilon + \epsilon k^2$. If $\epsilon = 0.1$ and $k = 10$, then $EY = 0$, and $\text{VAR}(Y) = 10.9$.

**Remark 1.2. Warning:** Mixture distributions and linear combinations of random variables are very different quantities. As an example, let

$$W = (1-\epsilon)W_1 + \epsilon W_2$$

where $\epsilon$, $W_1$ and $W_2$ are as in the previous example and suppose that $W_1$ and $W_2$ are independent. Then $W$, a linear combination of $W_1$ and $W_2$, has a normal distribution with mean

$$E[W] = (1-\epsilon)E[W_1] + \epsilon E[W_2] = 0$$

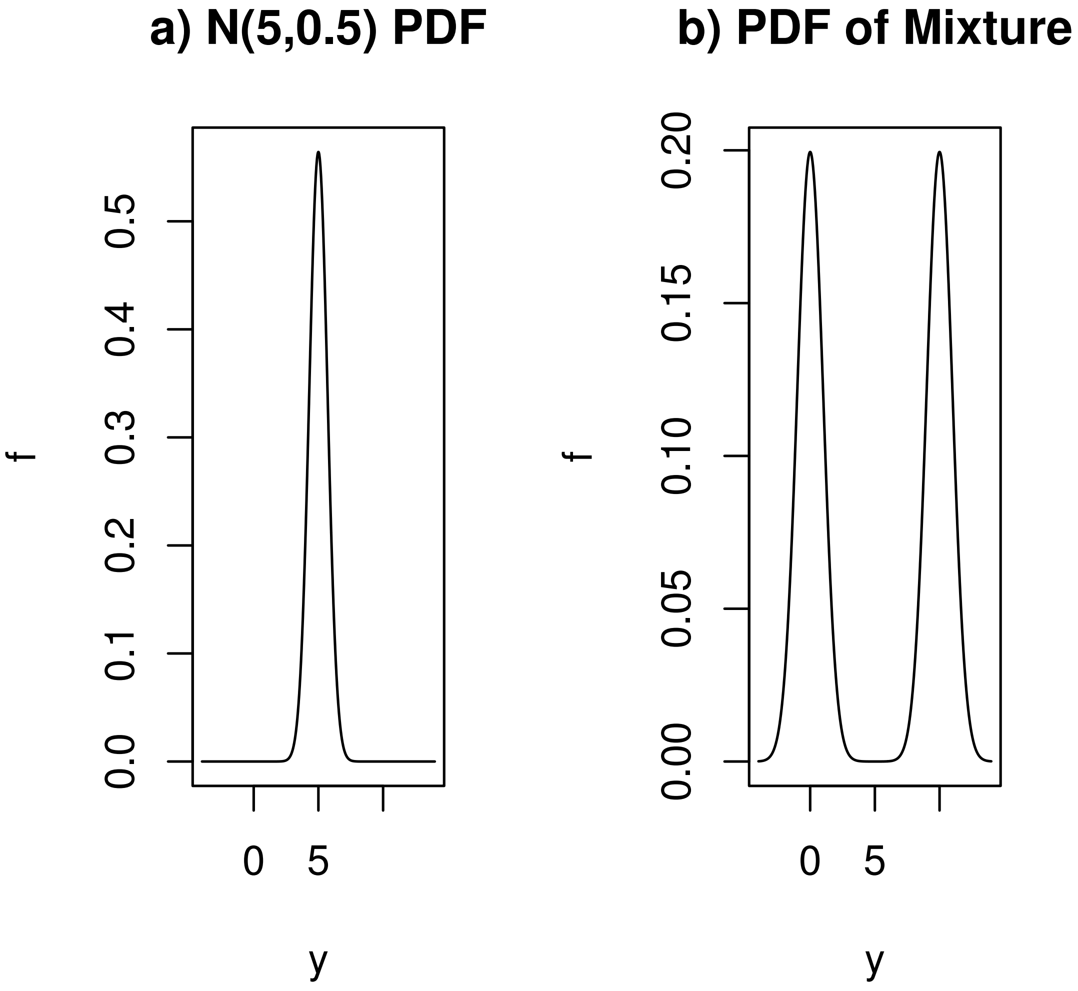

Figure 1.1: PDF $f$ of $(W_1 + W_2)/2$ and $f = 0.5f_1(y) + 0.5f_2(y)$

and variance

$$\text{VAR}(W) = (1-\epsilon)^2\text{VAR}(W_1) + \epsilon^2\text{VAR}(W_2) = (1-\epsilon)^2 + \epsilon^2 k^2 < \text{VAR}(Y)$$

where $Y$ is given in the example above. Moreover, $W$ has a unimodal normal distribution while $Y$ does not follow a normal distribution. In fact, if $W_1 \sim N(0,1)$, $W_2 \sim N(10,1)$, and $W_1$ and $W_2$ are independent, then $(W_1 + W_2)/2 \sim N(5, 0.5)$; however, if $Y$ has a mixture distribution with cdf

$$F_Y(y) = 0.5F_{W_1}(y) + 0.5F_{W_2}(y) = 0.5\Phi(y) + 0.5\Phi(y-10),$$

then the pdf of $Y$ is bimodal. See Figure 1.1.

## 1.7 Complements

Kolmogorov's definition of a probability function makes a probability function a normed measure. Hence many of the tools of measure theory can be used for probability theory. See, for example, Ash and Doleans-Dade (1999), Billingsley (1995), Dudley (2002), Durrett (1995), Feller (1971) and Resnick (1999). Feller (1957) and Tucker (1984) are good references for combinatorics.

Referring to Chapter 10, **memorize the pmf or pdf $f$, $E(Y)$ and $V(Y)$ for the following 10 RVs. You should recognize the mgf of the binomial, $\chi^2_p$, exponential, gamma, normal and Poisson distributions. You should recognize the cdf of the exponential and of the normal distribution.**

1) beta$(\delta, \nu)$

$$f(y) = \frac{\Gamma(\delta+\nu)}{\Gamma(\delta)\Gamma(\nu)} y^{\delta-1}(1-y)^{\nu-1}$$

where $\delta > 0$, $\nu > 0$ and $0 \le y \le 1$.

$$E(Y) = \frac{\delta}{\delta+\nu}.$$

$$\text{VAR}(Y) = \frac{\delta\nu}{(\delta+\nu)^2(\delta+\nu+1)}.$$

2) Bernoulli($\rho$) = binomial($k = 1, \rho$) $\quad f(y) = \rho(1-\rho)^{1-y}$ for $y = 0, 1$.
$E(Y) = \rho$.
$\text{VAR}(Y) = \rho(1-\rho)$.

$$m(t) = [(1-\rho) + \rho e^t].$$

3) binomial($k, \rho$)

$$f(y) = \binom{k}{y} \rho^y (1-\rho)^{k-y}$$

for $y = 0, 1, \ldots, k$ where $0 < \rho < 1$.
$E(Y) = k\rho$.
$\text{VAR}(Y) = k\rho(1-\rho)$.

$$m(t) = [(1-\rho) + \rho e^t]^k.$$

4) Cauchy($\mu, \sigma$)

$$f(y) = \frac{1}{\pi\sigma[1 + (\frac{y-\mu}{\sigma})^2]}$$

where $y$ and $\mu$ are real numbers and $\sigma > 0$.
$E(Y) = \infty = \text{VAR}(Y)$.

5) chi-square($p$) = gamma($\nu = p/2, \lambda = 2$)

$$f(y) = \frac{y^{\frac{p}{2}-1} e^{-\frac{y}{2}}}{2^{\frac{p}{2}} \Gamma(\frac{p}{2})}$$

$E(Y) = p$.
$\text{VAR}(Y) = 2p$.

$$m(t) = \left(\frac{1}{1-2t}\right)^{p/2} = (1-2t)^{-p/2}$$

for $t < 1/2$.

6) exponential($\lambda$)= gamma($\nu = 1, \lambda$)

$$f(y) = \frac{1}{\lambda} \exp{(-\frac{y}{\lambda})}\ I(y \geq 0)$$

where $\lambda > 0$.
$E(Y) = \lambda$,
$\text{VAR}(Y) = \lambda^2$.

$$m(t) = 1/(1 - \lambda t)$$

for $t < 1/\lambda$.

$$F(y) = 1 - \exp(-y/\lambda),\ \ y \geq 0.$$

7) gamma$(\nu, \lambda)$

$$f(y) = \frac{y^{\nu-1} e^{-y/\lambda}}{\lambda^{\nu}\Gamma(\nu)}$$

where $\nu, \lambda$, and $y$ are positive.
$E(Y) = \nu\lambda$.
VAR$(Y) = \nu\lambda^2$.

$$m(t) = \left(\frac{1}{1-\lambda t}\right)^{\nu}$$

for $t < 1/\lambda$.

8) $N(\mu, \sigma^2)$

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

where $\sigma > 0$ and $\mu$ and $y$ are real.
$E(Y) = \mu$. VAR$(Y) = \sigma^2$.

$$m(t) = \exp(t\mu + t^2\sigma^2/2).$$

$$F(y) = \Phi\left(\frac{y-\mu}{\sigma}\right).$$

9) Poisson$(\theta)$

$$f(y) = \frac{e^{-\theta}\theta^y}{y!}$$

for $y = 0, 1, \ldots,$ where $\theta > 0$.
$E(Y) = \theta = $ VAR$(Y)$.

$$m(t) = \exp(\theta(e^t - 1)).$$

10) uniform$(\theta_1, \theta_2)$

$$f(y) = \frac{1}{\theta_2 - \theta_1} I(\theta_1 \leq y \leq \theta_2).$$

$E(Y) = (\theta_1 + \theta_2)/2$.
VAR$(Y) = (\theta_2 - \theta_1)^2/12$.

The terms sample space S, events, disjoint, partition, probability function, sampling with and without replacement, conditional probability, Bayes' theorem, mutually independent events, random variable, cdf, continuous RV, discrete RV, identically distributed, pmf and pdf are important.

## 1.8 Problems

**PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL. Refer to Chapter 10 for the pdf or pmf of the distributions in the problems below.**

**1.1**$^*$**.** For the Binomial$(k, \rho)$ distribution,
a) find E $Y$.
b) Find Var $Y$.
c) Find the mgf $m(t)$.

**1.2**$^*$**.** For the Poisson$(\theta)$ distribution,
a) find E $Y$.
b) Find Var $Y$. (Hint: Use the kernel method to find E $Y(Y-1)$.)
c) Find the mgf $m(t)$.

**1.3**$^*$**.** For the Gamma$(\nu, \lambda)$ distribution,
a) find E $Y$.
b) Find Var $Y$.
c) Find the mgf $m(t)$.

**1.4**$^*$**.** For the Normal$(\mu, \sigma^2)$ (or Gaussian) distribution,
a) find the mgf $m(t)$. (Hint: complete the square to get a Gaussian kernel.)
b) Use the mgf to find E $Y$.
c) Use the mgf to find Var $Y$.

**1.5**$^*$**.** For the Uniform$(\theta_1, \theta_2)$ distribution
a) find E $Y$.
b) Find Var $Y$.
c) Find the mgf $m(t)$.

**1.6**$^*$**.** For the Beta$(\delta, \nu)$ distribution,
a) find E $Y$.
b) Find Var $Y$.

**1.7*.** See Mukhopadhyay (2000, p. 39). Recall integrals by u-substitution:

$$I = \int_a^b f(g(x))g'(x)dx = \int_{g(a)}^{g(b)} f(u)du = \int_c^d f(u)du =$$

$$F(u)|_c^d = F(d) - F(c) = F(u)|_{g(a)}^{g(b)} = F(g(x))|_a^b = F(g(b)) - F(g(a))$$

where $F'(x) = f(x)$, $u = g(x)$, $du = g'(x)dx$, $d = g(b)$, and $c = g(a)$.

This problem uses the Gamma function and u-substitution to show that the normal density integrates to 1 (usually shown with polar coordinates). When you perform the u-substitution, make sure you say what $u = g(x)$, $du = g'(x)dx$, $d = g(b)$, and $c = g(a)$ are.

a) Let $f(x)$ be the pdf of a $N(\mu, \sigma^2)$ random variable. Perform u-substitution on

$$I = \int_{-\infty}^{\infty} f(x)dx$$

with $u = (x - \mu)/\sigma$.

b) Break the result into two parts,

$$I = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} e^{-u^2/2}du + \frac{1}{\sqrt{2\pi}} \int_0^{\infty} e^{-u^2/2}du.$$

Then perform u-substitution on the first integral with $v = -u$.

c) Since the two integrals are now equal,

$$I = \frac{2}{\sqrt{2\pi}} \int_0^{\infty} e^{-v^2/2}dv = \frac{2}{\sqrt{2\pi}} \int_0^{\infty} e^{-v^2/2}\frac{1}{v}vdv.$$

Perform u-substitution with $w = v^2/2$.

d) Using the Gamma function, show that $I = \Gamma(1/2)/\sqrt{\pi} = 1$.

**1.8.** Let X be a $N(0,1)$ (standard normal) random variable. Use integration by parts to show that $EX^2 = 1$. Recall that integration by parts is used to evaluate $\int f(x)g'(x)dx = \int udv = uv - \int vdu$ where $u = f(x)$, $dv = g'(x)dx$, $du = f'(x)dx$ and $v = g(x)$. When you do the integration, clearly state what these 4 terms are (eg $u = x$).

**1.9.** Verify the formula for the cdf $F$ for the following distributions. That is, either show that $F'(y) = f(y)$ or show that $\int_{-\infty}^{y} f(t)dt = F(y)\ \forall y \in \Re$.
a) Cauchy $(\mu, \sigma)$.
b) Double exponential $(\theta, \lambda)$.
c) Exponential $(\lambda)$.
d) Logistic $(\mu, \sigma)$.
e) Pareto $(\sigma, \lambda)$.
f) Power $(\lambda)$.
g) Uniform $(\theta_1, \theta_2)$.
h) Weibull $W(\phi, \lambda)$.

**1.10.** Verify the formula for the expected value $E(Y)$ for the following distributions. a) Double exponential $(\theta, \lambda)$.
b) Exponential $(\lambda)$.
c) Logistic $(\mu, \sigma)$. (Hint from deCani and Stine (1986): Let $Y = [\mu + \sigma W]$ so $E(Y) = \mu + \sigma E(W)$ where $W \sim L(0, 1)$. Hence

$$E(W) = \int_{-\infty}^{\infty} y \frac{e^y}{[1 + e^y]^2} dy.$$

Use substitution with

$$u = \frac{e^y}{1 + e^y}.$$

Then

$$E(W^k) = \int_0^1 [\log(u) - \log(1 - u)]^k du.$$

Also use the fact that

$$\lim_{v \to 0} v \log(v) = 0$$

to show $E(W) = 0$.)
d) Lognormal $(\mu, \sigma^2)$.
e) Pareto $(\sigma, \lambda)$.
f) Weibull $(\phi, \lambda)$.

**1.11.** Verify the formula for the variance VAR($Y$) for the following distributions.
a) Double exponential $(\theta, \lambda)$.
b) Exponential $(\lambda)$.
c) Logistic $(\mu, \sigma)$. (Hint from deCani and Stine (1986): Let $Y = [\mu + \sigma X]$ so $V(Y) = \sigma^2 V(X) = \sigma^2 E(X^2)$ where $X \sim L(0,1)$. Hence

$$E(X^2) = \int_{-\infty}^{\infty} y^2 \frac{e^y}{[1+e^y]^2} dy.$$

Use substitution with

$$v = \frac{e^y}{1+e^y}.$$

Then

$$E(X^2) = \int_0^1 [\log(v) - \log(1-v)]^2 dv.$$

Let $w = \log(v) - \log(1-v)$ and $du = [\log(v) - \log(1-v)]dv$. Then

$$E(X^2) = \int_0^1 w du = uw|_0^1 - \int_0^1 u dw.$$

Now

$$uw|_0^1 = [v\log(v) + (1-v)\log(1-v)] \; w|_0^1 = 0$$

since

$$\lim_{v \to 0} v \log(v) = 0.$$

Now

$$-\int_0^1 u dw = -\int_0^1 \frac{\log(v)}{1-v} dv - \int_0^1 \frac{\log(1-v)}{v} dv = 2\pi^2/6 = \pi^2/3$$

using

$$\int_0^1 \frac{\log(v)}{1-v} dv = \int_0^1 \frac{\log(1-v)}{v} dv = -\pi^2/6.)$$

d) Lognormal $(\mu, \sigma^2)$.
e) Pareto $(\sigma, \lambda)$.
f) Weibull $(\phi, \lambda)$.

**Problems from old quizzes and exams.**

**1.12.** Suppose the random variable $X$ has cdf $F_X(x) = 0.9\ \Phi(x - 10) + 0.1\ F_W(x)$ where $\Phi(x - 10)$ is the cdf of a normal $N(10, 1)$ random variable with mean 10 and variance 1 and $F_W(x)$ is the cdf of the random variable $W$ that satisfies $P(W = 200) = 1$.
a) Find $E\ W$.
b) Find $E\ X$.

**1.13.** Suppose the random variable $X$ has cdf $F_X(x) = 0.9\ F_Z(x) + 0.1\ F_W(x)$ where $F_Z$ is the cdf of a gamma$(\alpha = 10, \beta = 1)$ random variable with mean 10 and variance 10 and $F_W(x)$ is the cdf of the random variable $W$ that satisfies $P(W = 400) = 1$.
a) Find $E\ W$.
b) Find $E\ X$.

**1.14.** Suppose the cdf $F_X(x) = (1 - \epsilon)F_Z(x) + \epsilon F_W(x)$ where $0 \le \epsilon \le 1$, $F_Z$ is the cdf of a random variable Z, and $F_W$ is the cdf of a random variable W. Then $E\ g(X) = (1 - \epsilon)E_Z\ g(X) + \epsilon E_W\ g(X)$ where $E_Z\ g(X)$ means that the expectation should be computed using the pmf or pdf of $Z$. Suppose the random variable $X$ has cdf $F_X(x) = 0.9\ F_Z(x) + 0.1\ F_W(x)$ where $F_Z$ is the cdf of a gamma$(\alpha = 10, \beta = 1)$ random variable with mean 10 and variance 10 and $F_W(x)$ is the cdf of the RV $W$ that satisfies $P(W = 400) = 1$.

a) Find $E\ W$.
b) Find $E\ X$.

**1.15.** Let $A$ and $B$ be positive integers. A hypergeometric random variable $X = W_1 + W_2 + \cdots + W_n$ where the random variables $W_i$ are identically distributed random variables with $P(W_i = 1) = A/(A + B)$ and $P(W_i = 0) = B/(A + B)$.

a) Find $E(W_1)$.
b) Find $E(X)$.

**1.16.** Suppose $P(X = x_o) = 1$ for some constant $x_o$.

a) Find $E\ g(X)$ in terms of $x_o$.

b) Find the moment generating function $m(t)$ of $X$.

c) Find $m^{(n)}(t) = \frac{d^n}{dt^n} m(t)$. (Hint: find $m^{(n)}(t)$ for $n = 1, 2,$ and 3. Then the pattern should be apparent.)

**1.17.** Suppose $P(X = 1) = 0.5$ and $P(X = -1) = 0.5$. Find the moment generating function of $X$.

**1.18.** Suppose that $X$ is a discrete random variable with pmf $f(x) = P(X = x)$ for x = 0, 1, ..., n so that the moment generating function of $X$ is $m(t) = \sum_{x=0}^{n} e^{tx} f(x)$.

a) Find $\frac{d}{dt} m(t) = m'(t)$.
b) Find $m'(0)$.
c) Find $m''(t) = \frac{d^2}{dt^2} m(t)$.
d) Find $m''(0)$.
e) Find $m^{(k)}(t) = \frac{d^k}{dt^k} m(t)$. (Hint: you found $m^{(k)}(t)$ for $k = 1, 2$, and the pattern should be apparent.)

**1.19.** Suppose that the random variable $W = e^X$ where $X \sim N(\mu, \sigma^2)$. Find $E(W^r) = E[(e^X)^r]$ by recognizing the relationship of $E[(e^X)^r]$ with the moment generating function of a normal$(\mu, \sigma^2)$ random variable.

**1.20.** Let $X \sim N(\mu, \sigma^2)$ so that $EX = \mu$ and Var $X = \sigma^2$.

a) Find $E(X^2)$.
b) If $k \geq 2$ is an integer, then $E(X^k) = (k-1)\sigma^2 E(X^{k-2}) + \mu E(X^{k-1})$. Use this recursion relationship to find $E(X^3)$.

**1.21*.** Let $X \sim$ gamma$(\nu, \lambda)$. Using the kernel method, find $EX^r$ where $r > -\nu$.

**1.22.** Find $\int_{-\infty}^{\infty} \exp(-\frac{1}{2} y^2) dy$.
(Hint: the integrand is a Gaussian kernel.)

**1.23.** Let $X$ have a Pareto $(\sigma, \lambda = 1/\theta)$ pdf

$$f(x) = \frac{\theta \sigma^\theta}{x^{\theta+1}}$$

where $x > \sigma$, $\sigma > 0$ and $\theta > 0$. Using the kernel method, find $EX^r$ where $\theta > r$.

**1.24.** Let $Y \sim$ beta $(\delta, \nu)$. Using the kernel method, find $EY^r$ where $r > -\delta$.

**1.25.** Use the kernel method to find the mgf of the logarithmic $(\theta)$ distribution.

**1.26.** Suppose that $X$ has pdf

$$f(x) = \frac{h(x)e^{\theta x}}{\lambda(\theta)}$$

for $x \in \mathcal{X}$ and for $-\infty < \theta < \infty$ where $\lambda(\theta)$ is some positive function of $\theta$ and $h(x)$ is some nonnegative function of $x$. Find the moment generating function of $X$ using the kernel method. Your final answer should be written in terms of $\lambda, \theta$ and $t$.

**1.27.** Use the kernel method to find $E(Y^r)$ for the chi $(p, \sigma)$ distribution. (See Section 10.6.)

**1.28.** Suppose the cdf $F_X(x) = (1-\epsilon)F_Z(x) + \epsilon F_W(x)$ where $0 \le \epsilon \le 1$, $F_Z$ is the cdf of a random variable Z, and $F_W$ is the cdf of a random variable W. Then $E\ g(X) = (1-\epsilon)E_Z\ g(X) + \epsilon E_W\ g(X)$ where $E_Z\ g(X)$ means that the expectation should be computed using the pmf or pdf of $Z$.

Suppose the random variable $X$ has cdf $F_X(x) = 0.9\ F_Z(x) + 0.1\ F_W(x)$ where $F_Z$ is the cdf of a gamma$(\nu = 3, \lambda = 4)$ random variable and $F_W(x)$ is the cdf of a Poisson(10) random variable.

a) Find $E\ X$.

b) Find $E\ X^2$.

**1.29.** If $Y$ has an exponential distribution truncated at 1, $Y \sim TEXP(\theta, 1)$, then the pdf of $Y$ is

$$f(y) = \frac{\theta}{1 - e^{-\theta}} e^{-\theta y}$$

for $0 < y < 1$ where $\theta > 0$. Find the mgf of $Y$ using the kernel method.