Chapter 1
Introduction

This chapter follows Olive (2014, ch. 1-3) closely. Much of the material can be skimmed, and then the reader can refer back to this chapter as needed.

Often large sample theory is taught after a course in probability and measure, and a probability space $(S, \mathcal{B}, P)$ is used where $\mathcal{B}$ is a σ-field. This text will usually ignore measure theoretic probability. Unless told otherwise, the notation $P(A)$ means that A is an event.

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

1.1 Probability, Expected Value, CDF

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 $A \in \mathcal{B}$.

In the definition of an event above, 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.

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. If $A = \emptyset$, then A and B are disjoint. In the following definition, disjoint events are often called pairwise disjoint events.

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

Definition 1.5. Let $\mathcal{B}$ be the class of events of the sample space S. A **probability function** $P : \mathcal{B} \to [0, 1]$ 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 disjoint events, then $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

If $A_1, ..., A_n$ are disjoint, then $P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$. This result follows from Definition 1.5 using $A_i = \emptyset$ for $i > n$.

Theorem 1.1. DeMorgan's Laws:
a) $\overline{A \cup B} = \overline{A} \cap \overline{B}$.
b) $\overline{A \cap B} = \overline{A} \cup \overline{B}$.
c) $(\bigcup_{i=1}^{\infty} A_i)^c = \bigcap_{i=1}^{\infty} A_i^c$.
d) $(\bigcap_{i=1}^{\infty} A_i)^c = \bigcup_{i=1}^{\infty} A_i^c$.

Proof. The proofs of a) and b) are similar to those of c) and d), and "iff" means "if and only if."

c) $(\bigcup_{i=1}^{\infty} A_i)^c$ occurred iff $\bigcup_{i=1}^{\infty} A_i$ did not occur, iff none of the A_i occurred, iff all of the A_i^c occurred, iff $\bigcap_{i=1}^{\infty} A_i^c$ occurred.

d) $(\bigcap_{i=1}^{\infty} A_i)^c$ occurred iff not all of the A_i occurred, iff at least one of the A_i^c occurred, iff $\bigcup_{i=1}^{\infty} A_i^c$ occurred. □

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

Note that A and $\overline{A}$ are disjoint and $A \cup \overline{A} = S$. Hence $1 = P(S) = P(A \cup \overline{A}) = P(A) + P(\overline{A})$, proving the complement rule. Note that S and $\emptyset$ are disjoint, so $1 = P(S) = P(S \cup \emptyset) = P(S) + P(\emptyset)$. Hence $P(\emptyset) = 0$. If $A \subseteq B$, let $C = \overline{A} \cap B$. Then A and C are disjoint with $A \cup C = B$. Hence $P(A) + P(C) = P(B)$, and $P(A) \leq P(B)$ by i).

Following Casella and Berger (2002, p. 13), $P(\bigcup_{i=1}^{n} A_i^c) = P[(\bigcap_{i=1}^{n} A_i)^c] = 1 - P(\bigcap_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} P(A_i^c) = \sum_{i=1}^{n} [1 - P(A_i)] = n - \sum_{i=1}^{n} P(A_i)$, where the first equality follows from DeMorgan's Laws, the second equality holds by the complement rule, and the inequality holds by Boole's inequality

$P(\bigcup_{i=1}^{n} A_i^c) \leq \sum_{i=1}^{n} P(A_i^c)$. Hence $P(\bigcap_{i=1}^{n} A_i) \geq \sum_{i=1}^{n} P(A_i) - (n-1)$, and Bonferonni's inequality holds.

If $A_1, A_2, ...$ are disjoint and if $\bigcup_{i=1}^{\infty} A_i = S$, 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 disjoint sets $A_1, A_2, ..., A_k$ is a partition of S if $\bigcup_{i=1}^{k} A_i = S$. The **conditional probability** of A **given** B is

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

if $P(B) > 0$.

Theorem 1.3: 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

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

Definition 1.6. A *random variable* Y is a real valued function with a sample space as a domain: $Y : S \to \mathbb{R}$ where the set of real numbers $\mathbb{R} = (-\infty, \infty)$.

Definition 1.7. The *population* is the entire group of objects from which we want information. The *sample* is the part of the population actually examined.

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 1.8. The **cumulative distribution function** (cdf) of any random variable Y is $F(y) = P(Y \leq y)$ for all $y \in \mathbb{R}$.

Cumulative distribution functions are very important for convergence in distribution. See Chapter 2. If $F(y)$ is a cumulative distribution function, then i) $F(-\infty) = \lim_{y \to -\infty} F(y) = 0$, ii) $F(\infty) = \lim_{y \to \infty} F(y) = 1$, iii) F is a nondecreasing function: if $y_1 < y_2$, then $F(y_1) \leq F(y_2)$, iv) F is right continuous: $\lim_{h \downarrow 0} F(y + h) = F(y)$ for all real y. v) Since a cdf is a probability for fixed y, $0 \leq F(y) \leq 1$ for all real y. vi) A cdf $F(y)$ can have at most countably many points of discontinuity, vii) $P(a < Y \leq b) = F(b) - F(a)$.

Definition 1.9. A random variable 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 random variable.

The **probability mass function** (pmf) of a discrete random variable Y is $f(y) = P(Y = y)$ for all $y \in \mathbb{R}$ 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 random variable is a step function with a jump of height $f(y)$ at values of y for which $f(y) > 0$.

Definition 1.10. A random variable Y is **continuous** if its distribution function $F(y)$ is absolutely continuous.

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

Definition 1.11. If Y is a continuous random variable, then a **probability density function** (pdf) $f(y)$ of Y is an integrable function such that

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

for all $y \in \mathbb{R}$. If $f(y)$ is a pdf, then $f(y)$ is continuous except at most a countable number of points, $f(y) \geq 0 \ \forall y$, and $\int_{-\infty}^{\infty} f(t)dt = 1$.

Theorem 1.4. 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 and be continuous except for at most a finite number of points in any finite interval).

Theorem 1.5. 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 1.12. Let Y be a discrete random variable with probability mass function $f(y)$. Then the *mean* or **expected value** of Y is

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

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

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

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 1.14. If $E(Y^2)$ exists, then the *variance* of a random variable Y is

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

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 notation $E(Y) = \infty$ or $V(Y) = \infty$ when the corresponding integral or sum diverges to ∞ can be useful. The following theorem is also used 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 random variables. If Y is a Cauchy (μ, σ) random variable, then neither $E(Y)$ nor $V(Y)$ exist.

Theorem 1.6: Short cut formula for variance.

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

If Y is a discrete random variable 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 $E[g(Y)] = \sum_{i=1}^{k} g(y_i)f(y_i) = g(y_1)f(y_1) + g(y_2)f(y_2) + \cdots + g(y_k)f(y_k)$.

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 random variable 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.

Theorem 1.7. 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) $V(aY + b) = a^2V(Y)$.

Definition 1.15. Random variables X and Y are *identically distributed*, written $X \sim Y$, $X \stackrel{D}{=} Y$, or $Y \sim F_X$, if $F_X(y) = F_Y(y)$ for all real y.

Theorem 1.8. 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 1.16. For positive integers k, the *kth moment* of Y is $E[Y^k]$ while the *kth central moment* is $E[(Y - E[Y])^k]$.

Definition 1.17. 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 *sample space* or **support** of Y. Let the set Θ be the set of parameter values $\boldsymbol{\theta}$ of interest. Then Θ 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 1.18. The **indicator function** $I_A(x) \equiv I(x \in A) = 1$ if $x \in A$ and $I_A(x) = 0$, otherwise. Sometimes an indicator function such as $I_{(0,\infty)}(y)$ will be denoted by $I(y > 0)$.

1.2 Multivariate Distributions

Often there are n random variables $Y_1, ..., Y_n$ that are of interest. For example, *age, blood pressure, weight, gender* and *cholesterol level* might be some of the random variables of interest for patients suffering from heart disease.

Notation. Let $\mathbb{R}^n$ be the n–dimensional Euclidean space. Then the vector $\boldsymbol{y} = (y_1, ..., y_n)^T \in \mathbb{R}^n$ if y_i is an arbitrary real number for $i = 1, ..., n$. Typically $\boldsymbol{y}$ is a column vector, but when $\boldsymbol{y}$ is the argument of a pdf, pmf, or cdf, then $\boldsymbol{y}$ is often a row vector, e.g., $f(\boldsymbol{y}) = f(y_1, ..., y_n)$. We may say $\boldsymbol{y} \in \mathbb{R}^n$ or $(y_1, ..., y_n) \in \mathbb{R}^n$.

Definition 1.19. If $Y_1, ..., Y_n$ are discrete random variables, then the **joint pmf** (probability mass function) of $Y_1, ..., Y_n$ is

$$f(y_1, ..., y_n) = P(Y_1 = y_1, ..., Y_n = y_n) \tag{1.7}$$

for any $(y_1, ..., y_n) \in \mathbb{R}^n$. A joint pmf f satisfies $f(\boldsymbol{y}) \equiv f(y_1, ..., y_n) \geq 0$ $\forall \boldsymbol{y} \in \mathbb{R}^n$ and

$$\sum_{\boldsymbol{y} \,:\, f(\boldsymbol{y}) > 0} \cdots \sum f(y_1, ..., y_n) = 1.$$

For any event $A \in \mathbb{R}^n$,

$$P[(Y_1, ..., Y_n) \in A] = \sum_{\boldsymbol{y} \,:\, \boldsymbol{y} \in A \text{ and } f(\boldsymbol{y}) > 0} \cdots \sum f(y_1, ..., y_n).$$

Definition 1.20. The **joint cdf** (cumulative distribution function) of $Y_1, ..., Y_n$ is $F(y_1, ..., y_n) = P(Y_1 \leq y_1, ..., Y_n \leq y_n)$ for any $(y_1, ..., y_n) \in \mathbb{R}^n$.

Definition 1.21. If $Y_1, ..., Y_n$ are continuous random variables, then the **joint pdf** (probability density function) of $Y_1, ..., Y_n$ is a function $f(y_1, ..., y_n)$ that satisfies $F(y_1, ..., y_n) = \int_{-\infty}^{y_n} \cdots \int_{-\infty}^{y_1} f(t_1, ..., t_n) dt_1 \cdots dt_n$ where the y_i are any real numbers. A joint pdf f satisfies $f(\boldsymbol{y}) \equiv f(y_1, ..., y_n) \geq 0$ $\forall \boldsymbol{y} \in \mathbb{R}^n$ and $\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(t_1, ..., t_n) dt_1 \cdots dt_n$
$= 1$. For any event $A \in \mathbb{R}^n$,

$$P[(Y_1, ..., Y_n) \in A] = \int_A \cdots \int f(t_1, ..., t_n) dt_1 \cdots dt_n.$$

Definition 1.22. If $Y_1, ..., Y_n$ has a joint pdf or pmf f, then the *sample space* or **support** of $Y_1, ..., Y_n$ is

$$\mathcal{Y} = \{(y_1, ..., y_n) \in \mathbb{R}^n : f(y_1, ..., y_n) > 0\}.$$

If $\boldsymbol{Y}$ comes from a family of distributions $f(\boldsymbol{y}|\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \Theta$, then the support $\mathcal{Y}_{\boldsymbol{\theta}} = \{\boldsymbol{y} : f(\boldsymbol{y}|\boldsymbol{\theta}) > 0\}$ may depend on $\boldsymbol{\theta}$.

Theorem 1.9. Let $Y_1, ..., Y_n$ have joint cdf $F(y_1, ..., y_n)$ and joint pdf $f(y_1, ..., y_n)$. Then

$$f(y_1, ..., y_n) = \frac{\partial^n}{\partial y_1 \cdots \partial y_n} F(y_1, ..., y_n)$$

wherever the partial derivative exists.

Definition 1.23. The **marginal pmf** of any subset $Y_{i1}, ..., Y_{ik}$ of the coordinates $(Y_1, ..., Y_n)$ is found by summing the joint pmf over all possible values of the other coordinates where the values $y_{i1}, ..., y_{ik}$ are held fixed. For example,

$$f_{Y_1,...,Y_k}(y_1,...,y_k) = \sum_{y_{k+1}} \cdots \sum_{y_n} f(y_1,...,y_n)$$

where $y_1,...,y_k$ are held fixed. In particular, if Y_1 and Y_2 are discrete random variables with joint pmf $f(y_1,y_2)$, then the marginal pmf for Y_1 is

$$f_{Y_1}(y_1) = \sum_{y_2} f(y_1,y_2) \tag{1.8}$$

where y_1 is held fixed. The marginal pmf for Y_2 is

$$f_{Y_2}(y_2) = \sum_{y_1} f(y_1,y_2) \tag{1.9}$$

where y_2 is held fixed.

Remark 1.2. For $n = 2$, double integrals are used to find marginal pdfs (defined below) and to show that the joint pdf integrates to 1. If the region of integration Ω is bounded on top by the function $y_2 = \phi_T(y_1)$, on the bottom by the function $y_2 = \phi_B(y_1)$ and to the left and right by the lines $y_1 = a$ and $y_1 = b$ then $\int\int_\Omega f(y_1,y_2)dy_1dy_2 = \int\int_\Omega f(y_1,y_2)dy_2dy_1 =$

$$\int_a^b \left[\int_{\phi_B(y_1)}^{\phi_T(y_1)} f(y_1,y_2)dy_2 \right] dy_1.$$

Within the inner integral, treat y_2 as the variable, anything else, including y_1, is treated as a constant.

If the region of integration Ω is bounded on the left by the function $y_1 = \psi_L(y_2)$, on the right by the function $y_1 = \psi_R(y_2)$ and to the top and bottom by the lines $y_2 = c$ and $y_2 = d$ then $\int\int_\Omega f(y_1,y_2)dy_1dy_2 = \int\int_\Omega f(y_1,y_2)dy_2dy_1 =$

$$\int_c^d \left[\int_{\psi_L(y_2)}^{\psi_R(y_2)} f(y_1,y_2)dy_1 \right] dy_2.$$

Within the inner integral, treat y_1 as the variable, anything else, including y_2, is treated as a constant.

Definition 1.24. The **marginal pdf** of any subset $Y_{i1},...,Y_{ik}$ of the coordinates $(Y_1,...,Y_n)$ is found by integrating the joint pdf over all possible values of the other coordinates where the values $y_{i1},...,y_{ik}$ are held fixed. For example, $f(y_1,...,y_k) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(t_1,...,t_n)dt_{k+1} \cdots dt_n$ where $y_1,...,y_k$ are held fixed. In particular, if Y_1 and Y_2 are continuous random variables with joint pdf $f(y_1,y_2)$, then the marginal pdf for Y_1 is

$$f_{Y_1}(y_1) = \int_{-\infty}^{\infty} f(y_1, y_2)dy_2 = \int_{\phi_B(y_1)}^{\phi_T(y_1)} f(y_1, y_2)dy_2 \tag{1.10}$$

where y_1 is held fixed (to get the region of integration, draw a line parallel to the y_2 axis and use the functions $y_2 = \phi_B(y_1)$ and $y_2 = \phi_T(y_1)$ as the lower and upper limits of integration). The marginal pdf for Y_2 is

$$f_{Y_2}(y_2) = \int_{-\infty}^{\infty} f(y_1, y_2)dy_1 = \int_{\psi_L(y_2)}^{\psi_R(y_2)} f(y_1, y_2)dy_1 \tag{1.11}$$

where y_2 is held fixed (to get the region of integration, draw a line parallel to the y_1 axis and use the functions $y_1 = \psi_L(y_2)$ and $y_1 = \psi_R(y_2)$ as the lower and upper limits of integration).

For independent random variables, the joint cdf is the product of the marginal cdfs, the joint pmf is the product of the marginal pmfs, and the joint pdf is the product of the marginal pdfs. Recall that $\forall$ is read "for all."

Definition 1.25. i) The random variables $Y_1, Y_2, ..., Y_n$ are **independent** if $F(y_1, y_2, ..., y_n) = F_{Y_1}(y_1)F_{Y_2}(y_2)\cdots F_{Y_n}(y_n)$ $\forall y_1, y_2, ..., y_n$.
ii) If the random variables have a joint pdf or pmf f then the random variables $Y_1, Y_2, ..., Y_n$ are independent if $f(y_1, y_2, ..., y_n) = f_{Y_1}(y_1)f_{Y_2}(y_2)\cdots f_{Y_n}(y_n)$ $\forall y_1, y_2, ..., y_n$.
If the random variables are not independent, then they are **dependent**.
In particular random variables Y_1 and Y_2 are **independent**, written $Y_1 \perp\!\!\!\perp Y_2$, if either of the following conditions holds.
i) $F(y_1, y_2) = F_{Y_1}(y_1)F_{Y_2}(y_2)$ $\forall y_1, y_2$.
ii) $f(y_1, y_2) = f_{Y_1}(y_1)f_{Y_2}(y_2)$ $\forall y_1, y_2$.
Otherwise, Y_1 and Y_2 are *dependent*.

The following theorem shows that finding the marginal pdfs or pmfs is simple if $Y_1, ..., Y_n$ are independent. Also **subsets of independent random variables are independent**: if $Y_1, ..., Y_n$ are independent and if $\{i_1, ..., i_k\} \subseteq \{1, ..., n\}$ for $k \geq 2$, then $Y_{i_1}, ..., Y_{i_k}$ are independent.

Theorem 1.10. Suppose that $Y_1, ..., Y_n$ are independent random variables with joint pdf or pmf $f(y_1, ..., y_n)$. Then the marginal pdf or pmf of any subset $Y_{i_1}, ..., Y_{i_k}$ is $f(y_{i_1}, ..., y_{i_k}) = \prod_{j=1}^{k} f_{Y_{i_j}}(y_{i_j})$. Hence $Y_{i_1}, ..., Y_{i_k}$ are independent random variables for $k \geq 2$.

Proof. The proof for a joint pdf is given below. For a joint pmf, replace the integrals by appropriate sums. The marginal

$$f(y_{i_1}, ..., y_{i_k}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left[\prod_{j=1}^{n} f_{Y_{i_j}}(y_{i_j}) \right] dy_{i_{k+1}} \cdots dy_{i_n}$$

$$= \left[\prod_{j=1}^{k} f_{Y_{i_j}}(y_{i_j}) \right] \left[\prod_{j=k+1}^{n} \int_{-\infty}^{\infty} f_{Y_{i_j}}(y_{i_j}) \, dy_{i_j} \right]$$

$$= \left[\prod_{j=1}^{k} f_{Y_{i_j}}(y_{i_j}) \right] (1)^{n-k} = \prod_{j=1}^{k} f_{Y_{i_j}}(y_{i_j}). \quad \square$$

Definition 1.26. Suppose that random variables $\boldsymbol{Y} = (Y_1, ..., Y_n)$ have support $\mathcal{Y}$ and joint pdf or pmf f. Then the **expected value** of the real valued function $h(\boldsymbol{Y}) = h(Y_1, ..., Y_n)$ is

$$E[h(\boldsymbol{Y})] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(\boldsymbol{y}) f(\boldsymbol{y}) \, d\boldsymbol{y} = \int \cdots \int_{\mathcal{Y}} h(\boldsymbol{y}) f(\boldsymbol{y}) \, d\boldsymbol{y} \qquad (1.12)$$

if f is a joint pdf and if

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} |h(\boldsymbol{y})| f(\boldsymbol{y}) \, d\boldsymbol{y}$$

exists. Otherwise the expectation does not exist. The expected value is

$$E[h(\boldsymbol{Y})] = \sum_{y_1} \cdots \sum_{y_n} h(\boldsymbol{y}) f(\boldsymbol{y}) = \sum_{\boldsymbol{y} \in \mathbb{R}^n} h(\boldsymbol{y}) f(\boldsymbol{y}) = \sum_{\boldsymbol{y} \in \mathcal{Y}} h(\boldsymbol{y}) f(\boldsymbol{y}) \qquad (1.13)$$

if f is a joint pmf and if $\sum_{\boldsymbol{y} \in \mathbb{R}^n} |h(\boldsymbol{y})| f(\boldsymbol{y})$ exists. Otherwise the expectation does not exist.

The notation $E[h(\boldsymbol{Y})] = \infty$ can be useful when the corresponding integral or sum diverges to ∞. The following theorem is useful since multiple integrals with smaller dimension are easier to compute than those with higher dimension.

Theorem 1.11. Suppose that $Y_1, ..., Y_n$ are random variables with joint pdf or pmf $f(y_1, ..., y_n)$. Let $\{i_1, ..., i_k\} \subset \{1, ..., n\}$, and let $f(y_{i_1}, ..., y_{i_k})$ be the marginal pdf or pmf of $Y_{i_1}, ..., Y_{i_k}$ with support $\mathcal{Y}_{Y_{i_1}, ..., Y_{i_k}}$. Assume that $E[h(Y_{i_1}, ..., Y_{i_k})]$ exists. Then

$$E[h(Y_{i_1}, ..., Y_{i_k})] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(y_{i_1}, ..., y_{i_k}) \; f(y_{i_1}, ..., y_{i_k}) \, dy_{i_1} \cdots dy_{i_k} =$$

$$\int \cdots \int_{\mathcal{Y}_{Y_{i_1}, ..., Y_{i_k}}} h(y_{i_1}, ..., y_{i_k}) \; f(y_{i_1}, ..., y_{i_k}) \, dy_{i_1} \cdots dy_{i_k}$$

if f is a pdf, and

$$E[h(Y_{i_1},...,Y_{i_k})] = \sum_{y_{i_1}} \cdots \sum_{y_{i_k}} h(y_{i_1},...,y_{i_k})\ f(y_{i_1},...,y_{i_k})$$

$$= \sum_{(y_{i_1},...,y_{i_k}) \in \mathcal{Y}_{Y_{i_1},...,Y_{i_k}}} h(y_{i_1},...,y_{i_k})\ f(y_{i_1},...,y_{i_k})$$

if f is a pmf.

Proof. The proof for a joint pdf is given below. For a joint pmf, replace the integrals by appropriate sums. Let $g(Y_1,...,Y_n) = h(Y_{i_1},...,Y_{i_k})$. Then $E[g(\boldsymbol{Y})] =$

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(y_{i_1},...,y_{i_k}) f(y_1,...,y_n)\ dy_1 \cdots dy_n =$$

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(y_{i_1},...,y_{i_k}) \left[\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(y_1,...,y_n)\ dy_{i_{k+1}} \cdots dy_{i_n} \right] dy_{i_1} \cdots dy_{i_k}$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(y_{i_1},...,y_{i_k}) f(y_{i_1},...,y_{i_k})\ dy_{i_1} \cdots dy_{i_k}$$

since the term in the brackets gives the marginal. □

Example 1.1. Typically $E(Y_i), E(Y_i^2)$ and $E(Y_iY_j)$ for $i \neq j$ are of primary interest. Suppose that (Y_1, Y_2) has joint pdf $f(y_1, y_2)$. Then $E[h(Y_1, Y_2)]$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(y_1, y_2) f(y_1, y_2) dy_2 dy_1 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(y_1, y_2) f(y_1, y_2) dy_1 dy_2$$

where $-\infty$ to ∞ could be replaced by the limits of integration for dy_i. **In particular**,

$$E(Y_1Y_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y_1 y_2 f(y_1, y_2) dy_2 dy_1 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y_1 y_2 f(y_1, y_2) dy_1 dy_2.$$

Since finding the marginal pdf is usually easier than doing the double integral, if h is a function of Y_i but not of Y_j, find the marginal for Y_i : $E[h(Y_1)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(y_1) f(y_1, y_2) dy_2 dy_1 = \int_{-\infty}^{\infty} h(y_1) f_{Y_1}(y_1) dy_1$. Similarly, $E[h(Y_2)] = \int_{-\infty}^{\infty} h(y_2) f_{Y_2}(y_2) dy_2$.

In particular, $E(Y_1) = \int_{-\infty}^{\infty} y_1 f_{Y_1}(y_1) dy_1$, and $E(Y_2) = \int_{-\infty}^{\infty} y_2 f_{Y_2}(y_2) dy_2$.

Suppose that (Y_1, Y_2) have a joint pmf $f(y_1, y_2)$. Then the **expectation** $E[h(Y_1, Y_2)]$
$= \sum_{y_2} \sum_{y_1} h(y_1, y_2) f(y_1, y_2) = \sum_{y_1} \sum_{y_2} h(y_1, y_2) f(y_1, y_2)$. **In particular**,

$$E[Y_1Y_2] = \sum_{y_1} \sum_{y_2} y_1 y_2 f(y_1, y_2).$$

Since finding the marginal pmf is usually easier than doing the double summation, if h is a function of Y_i but not of Y_j, find the marginal for pmf for Y_i: $E[h(Y_1)] = \sum_{y_2}\sum_{y_1} h(y_1)f(y_1,y_2) = \sum_{y_1} h(y_1)f_{Y_1}(y_1)$. Similarly, $E[h(Y_2)] = \sum_{y_2} h(y_2)f_{Y_2}(y_2)$. **In particular**, $E(Y_1) = \sum_{y_1} y_1 f_{Y_1}(y_1)$ and $E(Y_2) = \sum_{y_2} y_2 f_{Y_2}(y_2)$.

For pdfs it is sometimes possible to find $E[h(Y_i)]$, but for $k \geq 2$ these expected values tend to be very difficult to compute unless $f(y_1, ..., y_k) = c\ y_1^{i_1} \cdots y_k^{i_k}$ for small integers i_j on rectangular or triangular support. Independence makes finding some expected values simple.

Theorem 1.12. Let $Y_1, ..., Y_n$ be independent random variables. If $h_i(Y_i)$ is a function of Y_i alone and if the relevant expected values exist, then

$$E[h_1(Y_1)h_2(Y_2)\cdots h_n(Y_n)] = E[h_1(Y_1)]\cdots E[h_n(Y_n)].$$

In particular, $E[Y_iY_j] = E[Y_i]E[Y_j]$ for $i \neq j$.

Proof. The result will be shown for the case where $\boldsymbol{Y} = (Y_1, ..., Y_n)$ has a joint pdf f. For a joint pmf, replace the integrals by appropriate sums. By independence, the support of $\boldsymbol{Y}$ is a cross product: $\mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_n$. Since $f(\boldsymbol{y}) = \prod_{i=1}^n f_{Y_i}(y_i)$, the expectation $E[h_1(Y_1)h_2(Y_2)\cdots h_n(Y_n)] =$

$$\int \cdots \int_{\mathcal{Y}} h_1(y_1)h_2(y_2)\cdots h_n(y_n)\ f(y_1, ..., y_n)\ dy_1 \cdots dy_n$$

$$= \int_{\mathcal{Y}_n} \cdots \int_{\mathcal{Y}_1} \left[\prod_{i=1}^n h_i(y_i) f_{Y_i}(y_i)\right] dy_1 \cdots dy_n$$

$$= \prod_{i=1}^n \left[\int_{\mathcal{Y}_i} h_i(y_i) f_{Y_i}(y_i)\ dy_i\right] = \prod_{i=1}^n E[h_i(Y_i)]. \quad \square$$

Theorem 1.13. Let $Y_1, ..., Y_n$ be independent random variables. If $h_j(Y_{i_j})$ is a function of Y_{i_j} alone and if the relevant expected values exist, then

$$E[h_1(Y_{i_1})\cdots h_k(Y_{i_k})] = E[h_1(Y_{i_1})]\cdots E[h_k(Y_{i_k})].$$

Proof. Method 1: Take $X_j = Y_{i_j}$ for $j = 1, ..., k$. Then $X_1, ..., X_k$ are independent and Theorem 1.12 applies.

Method 2: Take $h_j(Y_{i_j}) \equiv 1$ for $j = k+1, ..., n$ and apply Theorem 1.12. $\square$

Theorem 1.14. Let $Y_1, ..., Y_n$ be independent random variables. If $h_i(Y_i)$ is a function of Y_i alone and $X_i = h_i(Y_i)$, then the random variables $X_1, ..., X_n$ are independent.

Definition 1.27. The **covariance** of Y_1 and Y_2 is

$$\text{Cov}(Y_1, Y_2) = E[(Y_1 - E(Y_1))(Y_2 - E(Y_2))]$$

provided the expectation exists. Otherwise the covariance does not exist.

Theorem 1.15: Short cut formula. If $\text{Cov}(Y_1, Y_2)$ exists then $\text{Cov}(Y_1, Y_2) = E(Y_1Y_2) - E(Y_1)E(Y_2)$.

Theorem 1.16. a) Let Y_1 and Y_2 be independent random variables. If $\text{Cov}(Y_1, Y_2)$ exists, then $\text{Cov}(Y_1, Y_2) = 0$.

b) **The converse is false**: $\text{Cov}(Y_1, Y_2) = 0$ does not imply $Y_1 \perp\!\!\!\perp Y_2$.

Definition 1.28. $\boldsymbol{Y} = (Y_1, ..., Y_p)^T$ is a $p \times 1$ **random vector** if Y_i is a random variable for $i = 1, ..., p$. $\boldsymbol{Y}$ is a discrete random vector if each Y_i is discrete, and $\boldsymbol{Y}$ is a continuous random vector if each Y_i is continuous. A random variable Y_1 is the special case of a random vector with $p = 1$.

In this section we will consider n random vectors $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n$. Often double subscripts will be used: $\boldsymbol{Y}_i = (Y_{i,1}, ..., Y_{i,p_i})^T$ for $i = 1, ..., n$.

Notation. In this text, $\boldsymbol{Y}$ is usually a column vector, and if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are both vectors, a phrase with $\boldsymbol{Y}$ and $\boldsymbol{X}^T$ means that $\boldsymbol{Y}$ is a column vector and $\boldsymbol{X}^T$ is a row vector where T stands for transpose. Arguments of pdfs, pmfs, and cdfs, are usually taken to be row vectors in this text.

Definition 1.29. The *population mean* or **expected value** of a random $p \times 1$ random vector $(Y_1, ..., Y_p)^T$ is

$$E(\boldsymbol{Y}) = (E(Y_1), ..., E(Y_p))^T$$

provided that $E(Y_i)$ exists for $i = 1, ..., p$. Otherwise the expected value does not exist. Now let $\boldsymbol{Y}$ be a $p \times 1$ column vector. The $p \times p$ *population covariance matrix*

$$\text{Cov}(\boldsymbol{Y}) = E(\boldsymbol{Y} - E(\boldsymbol{Y}))(\boldsymbol{Y} - E(\boldsymbol{Y}))^T = (\sigma_{i,j})$$

where the ij entry of $\text{Cov}(\boldsymbol{Y})$ is $\text{Cov}(Y_i, Y_j) = \sigma_{i,j}$ provided that each $\sigma_{i,j}$ exists. Otherwise $\text{Cov}(\boldsymbol{Y})$ does not exist.

The covariance matrix is also called the variance–covariance matrix and variance matrix. Sometimes the notation $\text{Var}(\boldsymbol{Y})$ is used. Note that $\text{Cov}(\boldsymbol{Y})$ is a symmetric positive semidefinite matrix. If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are $p \times 1$ random vectors, $\boldsymbol{a}$ a conformable constant vector and $\boldsymbol{A}$ and $\boldsymbol{B}$ are conformable constant matrices, then

$$E(\boldsymbol{a} + \boldsymbol{X}) = \boldsymbol{a} + E(\boldsymbol{X}) \quad \text{and} \quad E(\boldsymbol{X} + \boldsymbol{Y}) = E(\boldsymbol{X}) + E(\boldsymbol{Y}) \tag{1.14}$$

and

$$E(\boldsymbol{AX}) = \boldsymbol{A}E(\boldsymbol{X}) \text{ and } E(\boldsymbol{AXB}) = \boldsymbol{A}E(\boldsymbol{X})\boldsymbol{B}. \tag{1.15}$$

Thus

$$\text{Cov}(\boldsymbol{a} + \boldsymbol{AX}) = \text{Cov}(\boldsymbol{AX}) = \boldsymbol{A}\text{Cov}(\boldsymbol{X})\boldsymbol{A}^T. \tag{1.16}$$

Definition 1.30. Let $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n$ be random vectors with joint pdf or pmf $f(\boldsymbol{y}_1, ..., \boldsymbol{y}_n)$. Let $f_{\boldsymbol{Y}_i}(\boldsymbol{y}_i)$ be the marginal pdf or pmf of $\boldsymbol{Y}_i$. Then $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n$ are **independent random vectors** if

$$f(\boldsymbol{y}_1, ..., \boldsymbol{y}_n) = f_{\boldsymbol{Y}_1}(\boldsymbol{y}_1) \cdots f_{\boldsymbol{Y}_n}(\boldsymbol{y}_n) = \prod_{i=1}^{n} f_{\boldsymbol{Y}_i}(\boldsymbol{y}_i).$$

The following theorem is a useful generalization of Theorem 1.14.

Theorem 1.17. Let $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n$ be independent random vectors where $\boldsymbol{Y}_i$ is a $p_i \times 1$ vector for $i = 1, ..., n$. and let $\boldsymbol{h}_i : \mathbb{R}^{p_i} \to \mathbb{R}^{p_{j_i}}$ be vector valued functions and suppose that $\boldsymbol{h}_i(\boldsymbol{y}_i)$ is a function of $\boldsymbol{y}_i$ alone for $i = 1, ..., n$. Then the random vectors $\boldsymbol{X}_i = \boldsymbol{h}_i(\boldsymbol{Y}_i)$ are independent. There are three important special cases.
i) If $p_{j_i} = 1$ so that each h_i is a real valued function, then the random variables $X_i = h_i(\boldsymbol{Y}_i)$ are independent.
ii) If $p_i = p_{j_i} = 1$ so that each Y_i and each $X_i = h(Y_i)$ are random variables, then $X_1, ..., X_n$ are independent.
iii) Let $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$ and $\boldsymbol{X} = (X_1, .., X_m)^T$ and assume that $\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{X}$. If $\boldsymbol{h}(\boldsymbol{Y})$ is a vector valued function of $\boldsymbol{Y}$ alone and if $\boldsymbol{g}(\boldsymbol{X})$ is a vector valued function of $\boldsymbol{X}$ alone, then $\boldsymbol{h}(\boldsymbol{Y})$ and $\boldsymbol{g}(\boldsymbol{X})$ are independent random vectors.

1.3 Characteristic Function, MGF, CGF

Definition 1.31. The **moment generating function** (mgf) of a random variable Y is

$$m(t) = E[e^{tY}] \tag{1.17}$$

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 $m(t) = \int_{-\infty}^{\infty} e^{ty} f(y) dy$.

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

Definition 1.32. If the mgf exists, then the **cumulant generating function** (cgf) $k(t) = \log(m(t))$ for the values of t where the mgf is defined.

Definition 1.33. The **characteristic function** of a random variable Y is $c(t) = E[e^{itY}] = E[\cos(tY)] + iE[\sin(tY)]$ where the complex number $i = \sqrt{-1}$.

Moment generating functions do not necessarily exist in a neighborhood of zero, but a characteristic function always exists. 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 $k = 1, 2, 3, ...$ Let complex number $z = a + ib$. Then the modulus of z is $|z| = |a + ib| = \sqrt{a^2 + b^2}$.

Remark 1.3. a) Suppose that Y is a random variable with an mgf $m(t)$ that exists for $|t| < b$ for some constant $b > 0$. Then often the characteristic function of Y is i) $c(t) = m(it)$ while ii) $m(t) = c(-it)$. If Y has a pmf with $f(y_j) = P(Y = y_j) = p_j$), then the characteristic function of Y is $c(t) = c_Y(t) = \sum_j e^{ity_j} p_j$ while the mgf $m_Y(t) = \sum_j e^{ty_j} p_j$. Hence the two formulas i) and ii) "hold" if Y has a pmf, at least for t such that the mgf is defined. If Y is nonnegative then the mgf is a scaled Laplace transformation and $c(t)$ is a scaled Fourier transformation, and then the two formulas i) and ii) hold by Laplace and Fourier transformation theory, at least for t such that the mgf is defined. The Taylor series for the mgf is

$$m_Y(t) = \sum_{k=0}^{\infty} \frac{t^k}{k!} E[X^k]$$

for $|t| < b$ while the characteristic function

$$c_Y(t) = \sum_{k=0}^{\infty} \frac{(it)^k}{k!} E[X^k]$$

for all real t if Y has an mgf defined for all real t. Hence if $b = \infty$, the two formulas i) and ii) hold. See Billingsley (1986, pp. 285, 353).

b) If $E[Y^2]$ is finite, then

$$c_Y(t) = 1 + itE(Y) - \frac{1}{2}t^2 E[Y^2] + o(t^2) \text{ as } t \to 0.$$

In particular, if $E(Y) = 0$ and $E(Y^2) = V(Y) = \sigma^2$, then

$$c_Y(t) = 1 + \frac{t^2\sigma^2}{2} + o(t^2) \text{ as } t \to 0. \tag{1.18}$$

Here $a(t) = o(t^2)$ as $t \to 0$ if $\lim_{t \to 0} \frac{a(t)}{t^2} = 0$. See Billingsley (1986, p. 354).

c) Properties of $c(t)$: i) $c(0) = 1$, ii) the modulus $|c(t)| \leq 1$ for all real t, iii) $c(t)$ is a continuous function.

d) Let j and k be positive integers. If $E(Y^k)$ is finite, then $E(Y^j)$ is finite for $1 \leq j \leq k$.

e) If Y has mgf $m(t)$, then $E(Y^k)$ is finite for each positive integer k.

f) A complex random variable $Z = X + iY$ where X and Y are ordinary random variables. Then $E(Z) = E(X) + iE(Y)$, and $E(Z)$ exists if $E(|Z|) = E(\sqrt{X^2 + Y^2}) < \infty$. Linearity of expectation and key inequalities such as $|E(Z)| \leq E(|Z|)$ remain valid. Also, if $Z_1 \perp\!\!\!\perp Z_2$ and $g_i(Z_i)$ is a function of the complex random variable Z_i alone, then $E[g_1(Z_1)g_2(Z_2)] = E[g_1(Z_1)]E[g_2(Z_2)]$ if the expectations exist. $Z = e^{itY}$ is the main complex random variable in this book.

Remarks 1.3 and 1.4 are often used in proofs of the Central Limit Theorem. Note that by Remark 1.4a), $\lim_{n\to\infty} \left(1 - \frac{c \pm \epsilon}{n}\right)^n = e^{-[c \pm \epsilon]}$ where ϵ is a real number. Remark 1.4c), this result holds even if ϵ is complex valued. Letting positive $\epsilon \to 0$ proves Remark 1.4b). By

Remark 1.4. a) $\lim_{n\to\infty} \left(1 - \frac{c}{n}\right)^n = e^{-c}$.

b) If $c_n \to c$ as $n \to \infty$, then $\lim_{n\to\infty} \left(1 + \frac{-c_n}{n}\right)^n = e^{-c}$.

c) If c_n is a sequence of complex numbers such that $c_n \to c$ as $n \to \infty$ where c is real, then $\lim_{n\to\infty} \left(1 - \frac{c_n}{n}\right)^n = e^{-c}$.

In the following theorem, let the kth derivative of $g(t)$ be $g^{(k)}(t)$ with derivative $g^{(1)}(t) = g'(t)$ and second derivative $g^{(2)}(t) = g''(t)$.

Theorem 1.18. Suppose that the mgf $m(t)$ exists for $|t| < b$ for some constant $b > 0$, and suppose that the kth derivative $m^{(k)}(t)$ exists for $|t| < b$. Then $E[Y^k] = m^{(k)}(0)$ for positive integers k. In particular, $E[Y] = m'(0)$ and $E[Y^2] = m''(0)$. For the cumulant generating function $k(t)$, $E(Y) = k'(0)$ and $V(Y) = k''(0)$.

Remark 1.5. 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 the kth derivative

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

for integers $k \geq 2$. 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

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

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 1.34. The **characteristic function** (cf) of a random vector $\boldsymbol{Y}$ is

$$\phi_{\boldsymbol{Y}}(\boldsymbol{t}) = E(e^{i\boldsymbol{t}^T\boldsymbol{Y}})$$

$\forall \boldsymbol{t} \in \mathbb{R}^n$ where the complex number $i = \sqrt{-1}$.

Definition 1.35. The **moment generating function** (mgf) of a random vector $\boldsymbol{Y}$ is

$$m_{\boldsymbol{Y}}(\boldsymbol{t}) = E(e^{\boldsymbol{t}^T\boldsymbol{Y}})$$

provided that the expectation exists for all $\boldsymbol{t}$ in some neighborhood of the origin $\boldsymbol{0}$.

Theorem 1.19. If $Y_1, ..., Y_n$ have mgf $m(\boldsymbol{t})$, then moments of all orders exist and for any nonnegative integers $k_1, ..., k_j$,

$$E(Y_{i_1}^{k_1} \cdots Y_{i_j}^{k_j}) = \frac{\partial^{k_1+\cdots+k_j}}{\partial t_{i_1}^{k_1} \cdots \partial t_{i_j}^{k_j}} m(\boldsymbol{t})\Big|_{\boldsymbol{t}=\boldsymbol{0}}.$$

In particular,

$$E(Y_i) = \frac{\partial m(\boldsymbol{t})}{\partial t_i}\Big|_{\boldsymbol{t}=\boldsymbol{0}}$$

and

$$E(Y_iY_j) = \frac{\partial^2 m(\boldsymbol{t})}{\partial t_i \partial t_j}\Big|_{\boldsymbol{t}=\boldsymbol{0}}.$$

Theorem 1.20. If $Y_1, ..., Y_n$ have a cf $c_{\boldsymbol{Y}}(\boldsymbol{t})$ and mgf $m_{\boldsymbol{Y}}(\boldsymbol{t})$ then the marginal cf and mgf for $Y_{i_1}, ..., Y_{i_k}$ are found from the joint cf and mgf by replacing t_{i_j} by 0 for $j = k+1, ..., n$. In particular, if $\boldsymbol{Y} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2)^T$ and $\boldsymbol{t} = (\boldsymbol{t}_1, \boldsymbol{t}_2)^T$, then

$$c_{\boldsymbol{Y}_1}(\boldsymbol{t}_1) = c_{\boldsymbol{Y}}((\boldsymbol{t}_1^T, \boldsymbol{0}^T)^T) \text{ and } \mathrm{m}_{\boldsymbol{Y}_1}(\boldsymbol{t}_1) = \mathrm{m}_{\boldsymbol{Y}}((\boldsymbol{t}_1^{\mathrm{T}}, \boldsymbol{0}^{\mathrm{T}})^{\mathrm{T}}).$$

Proof. Use the definition of the cf and mgf. For example, if $\boldsymbol{Y}_1 = (Y_1, ..., Y_k)^T$ and $\boldsymbol{s} = \boldsymbol{t}_1$, then $m((\boldsymbol{t}_1^T, \boldsymbol{0}^T)^T) =$

$$E[\exp(t_1Y_1 + \cdots + t_kY_k + 0Y_{k+1} + \cdots + 0Y_n)] = E[\exp(t_1Y_1 + \cdots + t_kY_k)] =$$

$E[\exp(\boldsymbol{s}^T\boldsymbol{Y}_1)] = m_{\boldsymbol{Y}_1}(\boldsymbol{s})$, which is the mgf of $\boldsymbol{Y}_1$. □

Theorem 1.21. Partition the $1 \times n$ vectors $\boldsymbol{Y}$ and $\boldsymbol{t}$ as $\boldsymbol{Y} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2)^T$ and $\boldsymbol{t} = (\boldsymbol{t}_1, \boldsymbol{t}_2)$. Then the random vectors $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ are independent iff their joint cf factors into the product of their marginal cfs:

$$c_{\boldsymbol{Y}}(\boldsymbol{t}) = c_{\boldsymbol{Y}_1}(\boldsymbol{t}_1) c_{\boldsymbol{Y}_2}(\boldsymbol{t}_2) \ \ \forall \boldsymbol{t} \in \mathbb{R}^n.$$

If the joint mgf exists, then the random vectors $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ are independent iff their joint mgf factors into the product of their marginal mgfs:

$$m_{\boldsymbol{Y}}(\boldsymbol{t}) = m_{\boldsymbol{Y}_1}(\boldsymbol{t}_1) m_{\boldsymbol{Y}_2}(\boldsymbol{t}_2)$$

$\forall \boldsymbol{t}$ in some neighborhood of $\boldsymbol{0}$.

1.4 Sums of Random Variables

The assumption that the data are iid or a random sample is often used. The iid assumption is useful for finding the joint pdf or pmf, and the exact or large sample distribution of many important statistics.

Definition 1.36. $Y_1, ..., Y_n$ are a **random sample** or **iid** if $Y_1, ..., Y_n$ are independent and identically distributed (all of the Y_i have the same distribution).

An important statistic is $\sum_{i=1}^n Y_i$. Some properties of sums are given below.

Theorem 1.22. Assume that all relevant expectations exist. Let a, $a_1, ..., a_n$ and $b_1, ..., b_m$ be constants. Let $Y_1, ..., Y_n$, and $X_1, ..., X_m$ be random variables. Let $g_1, ..., g_k$ be functions of $Y_1, ..., Y_n$.

i) $E(a) = a$.

ii) $E[aY] = aE[Y]$

iii) $V(aY) = a^2 V(Y)$.

iv) $E[g_1(Y_1, ..., Y_n) + \cdots + g_k(Y_1, ..., Y_n)] = \sum_{i=1}^k E[g_i(Y_1, ..., Y_n)]$.

Let $W_1 = \sum_{i=1}^n a_i Y_i$ and $W_2 = \sum_{i=1}^m b_i X_i$.

v) $E(W_1) = \sum_{i=1}^n a_i E(Y_i)$.

vi) $V(W_1) = \text{Cov}(W_1, W_1) = \sum_{i=1}^n a_i^2 V(Y_i) + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^n a_i a_j \text{Cov}(Y_i, Y_j)$.

vii) $\text{Cov}(W_1, W_2) = \sum_{i=1}^n \sum_{j=1}^m a_i b_j \text{Cov}(Y_i, X_j)$.

viii) $E(\sum_{i=1}^n Y_i) = \sum_{i=1}^n E(Y_i)$.

ix) If $Y_1, ..., Y_n$ are independent, $V(\sum_{i=1}^n Y_i) = \sum_{i=1}^n V(Y_i)$.

Let $Y_1, ..., Y_n$ be iid random variables with $E(Y_i) = \mu$ and $V(Y_i) = \sigma^2$, then the

sample mean $\overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$. Then

x) $E(\overline{Y}) = \mu$ and
xi) $V(\overline{Y}) = \sigma^2/n$.

Hence the expected value of the sum is the sum of the expected values, the variance of the sum is the sum of the variances for independent random variables, and the covariance of two sums is the double sum of the covariances. Note that ix) follows from vi) with $a_i \equiv 1$, viii) follows from iv) with $g_i(\boldsymbol{Y}) = Y_i$ or from v) with $a_i \equiv 1$, x) follows from v) with $a_i \equiv 1/n$, and xi) can be shown using iii) and ix) using $\overline{Y} = \sum_{i=1}^n (Y_i/n)$.

Example 1.2. Let $Y_1, ..., Y_n$ be independent random variables with $E(Y_i) = \mu_i$ and $V(Y_i) = \sigma_i^2$. Let $W = \sum_{i=1}^n Y_i$. Then
a) $E(W) = E(\sum_{i=1}^n Y_i) = \sum_{i=1}^n E(Y_i) = \sum_{i=1}^n \mu_i$, and
b) $V(W) = V(\sum_{i=1}^n Y_i) = \sum_{i=1}^n V(Y_i) = \sum_{i=1}^n \sigma_i^2$.

A **statistic** is a function of the data (often a random sample) and known constants. A statistic is a random variable and the **sampling distribution** of a statistic is the distribution of the statistic. Important statistics are $\sum_{i=1}^n Y_i$, $\overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$ and $\sum_{i=1}^n a_i Y_i$ where $a_1, ..., a_n$ are constants. The following theorem shows how to find the mgf and characteristic function of such statistics.

Theorem 1.23. a) The characteristic function uniquely determines the distribution.

b) If the moment generating function exists, then it uniquely determines the distribution.

c) Assume that $Y_1, ..., Y_n$ are independent with characteristic functions $c_{Y_i}(t)$. Then the characteristic function of $W = \sum_{i=1}^n Y_i$ is

$$c_W(t) = \prod_{i=1}^n c_{Y_i}(t). \tag{1.19}$$

d) Assume that $Y_1, ..., Y_n$ are iid with characteristic functions $c_Y(t)$. Then the characteristic function of $W = \sum_{i=1}^n Y_i$ is

$$c_W(t) = [c_Y(t)]^n. \tag{1.20}$$

e) Assume that $Y_1, ..., Y_n$ are independent with mgfs $m_{Y_i}(t)$. Then the mgf of $W = \sum_{i=1}^n Y_i$ is

$$m_W(t) = \prod_{i=1}^{n} m_{Y_i}(t). \tag{1.21}$$

f) Assume that $Y_1, ..., Y_n$ are iid with mgf $m_Y(t)$. Then the mgf of $W = \sum_{i=1}^{n} Y_i$ is

$$m_W(t) = [m_Y(t)]^n. \tag{1.22}$$

g) Assume that $Y_1, ..., Y_n$ are independent with characteristic functions $c_{Y_i}(t)$. Then the characteristic function of $W = \sum_{j=1}^{n}(a_j + b_j Y_j)$ is

$$c_W(t) = \exp(it \sum_{j=1}^{n} a_j) \prod_{j=1}^{n} c_{Y_j}(b_j t). \tag{1.23}$$

h) Assume that $Y_1, ..., Y_n$ are independent with mgfs $m_{Y_i}(t)$. Then the mgf of $W = \sum_{i=1}^{n}(a_i + b_i Y_i)$ is

$$m_W(t) = \exp(t \sum_{i=1}^{n} a_i) \prod_{i=1}^{n} m_{Y_i}(b_i t). \tag{1.24}$$

Proof of g): Recall that $\exp(w) = e^w$ and $\exp(\sum_{j=1}^{n} d_j) = \prod_{j=1}^{n} \exp(d_j)$. It can be shown that for the purposes of this proof, that the complex constant i in the characteristic function (cf) can be treated in the same way as if it were a real constant. Now

$$c_W(t) = E(e^{itW}) = E(\exp[it \sum_{j=1}^{n}(a_j + b_j Y_j)])$$

$$= \exp(it \sum_{j=1}^{n} a_j) \; E(\exp[\sum_{j=1}^{n} itb_j Y_j)])$$

$$= \exp(it \sum_{j=1}^{n} a_j) \; E(\prod_{i=1}^{n} \exp[itb_j Y_j)])$$

$$= \exp(it \sum_{j=1}^{n} a_j) \prod_{i=1}^{n} E[\exp(itb_j Y_j)]$$

since by Theorem 1.12 the expected value of a product of independent random variables is the product of the expected values of the independent random variables. Now in the definition of a cf, the t is a dummy variable as long as t is real. Hence $c_Y(t) = E[\exp(itY)]$ and $c_Y(s) = E[\exp(isY)]$. Taking $s = tb_j$ gives $E[\exp(itb_j Y_j)] = \phi_{Y_j}(tb_j)$. Thus

$$c_W(t) = \exp(it \sum_{j=1}^{n} a_j) \prod_{i=1}^{n} c_{Y_j}(tb_j). \quad \square$$

The distribution of $W = \sum_{i=1}^{n} Y_i$ is known as the convolution of $Y_1, ..., Y_n$. Even for $n = 2$, convolution formulas tend to be hard; however, the following two theorems suggest that to find the distribution of $W = \sum_{i=1}^{n} Y_i$, first find the mgf or characteristic function of W using Theorem 1.19. If the mgf or cf is that of a brand name distribution, then W has that distribution. For example, if the mgf of W is a normal (ν, τ^2) mgf, then W has a normal (ν, τ^2) distribution, written $W \sim N(\nu, \tau^2)$. This technique is useful for several brand name distributions.

Theorem 1.24. a) If $Y_1, ..., Y_n$ are independent binomial BIN(k_i, ρ) random variables, then

$$\sum_{i=1}^{n} Y_i \sim \text{BIN}(\sum_{\text{i=1}}^{\text{n}} \text{k}_\text{i}, \rho).$$

Thus if $Y_1, ..., Y_n$ are iid BIN(k, ρ) random variables, then $\sum_{i=1}^{n} Y_i \sim \text{BIN}(\text{nk}, \rho)$.

b) Denote a chi–square χ^2_p random variable by $\chi^2(p)$. If $Y_1, ..., Y_n$ are independent chi–square $\chi^2_{p_i}$, then

$$\sum_{i=1}^{n} Y_i \sim \chi^2(\sum_{i=1}^{n} p_i).$$

Thus if $Y_1, ..., Y_n$ are iid χ^2_p, then

$$\sum_{i=1}^{n} Y_i \sim \chi^2_{np}.$$

c) If $Y_1, ..., Y_n$ are iid exponential EXP(λ), then

$$\sum_{i=1}^{n} Y_i \sim G(n, \lambda).$$

d) If $Y_1, ..., Y_n$ are independent Gamma $G(\nu_i, \lambda)$ then

$$\sum_{i=1}^{n} Y_i \sim G(\sum_{i=1}^{n} \nu_i, \lambda).$$

Thus if $Y_1, ..., Y_n$ are iid $G(\nu, \lambda)$, then

$$\sum_{i=1}^{n} Y_i \sim G(n\nu, \lambda).$$

e) If $Y_1, ..., Y_n$ are independent normal $N(\mu_i, \sigma_i^2)$, then

$$\sum_{i=1}^{n}(a_i + b_i Y_i) \sim N(\sum_{i=1}^{n}(a_i + b_i\mu_i), \sum_{i=1}^{n} b_i^2\sigma_i^2).$$

Here a_i and b_i are fixed constants. Thus if $Y_1, ..., Y_n$ are iid $N(\mu, \sigma^2)$, then $\overline{Y} \sim N(\mu, \sigma^2/n)$.

f) If $Y_1, ..., Y_n$ are independent Poisson POIS(θ_i), then

$$\sum_{i=1}^{n} Y_i \sim \text{POIS}(\sum_{\text{i=1}}^{\text{n}} \theta_{\text{i}}).$$

Thus if $Y_1, ..., Y_n$ are iid $POIS(\theta)$, then

$$\sum_{i=1}^{n} Y_i \sim \text{POIS}(\text{n}\theta).$$

Theorem 1.25. a) If $Y_1, ..., Y_n$ are independent Cauchy $C(\mu_i, \sigma_i)$, then

$$\sum_{i=1}^{n}(a_i + b_i Y_i) \sim C(\sum_{i=1}^{n}(a_i + b_i\mu_i), \sum_{i=1}^{n} |b_i|\sigma_i).$$

Thus if $Y_1, ..., Y_n$ are iid $C(\mu, \sigma)$, then $\overline{Y} \sim C(\mu, \sigma)$.

b) If $Y_1, ..., Y_n$ are iid geometric geom(p), then

$$\sum_{i=1}^{n} Y_i \sim \text{NB}(\text{n}, \text{p}).$$

c) If $Y_1, ..., Y_n$ are iid inverse Gaussian $IG(\theta, \lambda)$, then

$$\sum_{i=1}^{n} Y_i \sim IG(n\theta, n^2\lambda).$$

Also

$$\overline{Y} \sim IG(\theta, n\lambda).$$

d) If $Y_1, ..., Y_n$ are independent negative binomial NB(r_i, ρ), then

$$\sum_{i=1}^{n} Y_i \sim \text{NB}(\sum_{\text{i=1}}^{\text{n}} \text{r}_{\text{i}}, \rho).$$

Thus if $Y_1, ..., Y_n$ are iid $NB(r, \rho)$, then

$$\sum_{i=1}^{n} Y_i \sim NB(nr, \rho).$$

Example 1.3. Suppose $Y_1, ..., Y_n$ are iid $IG(\theta, \lambda)$ where the mgf

$$m_{Y_i}(t) = m(t) = \exp\left[\frac{\lambda}{\theta}\left(1 - \sqrt{1 - \frac{2\theta^2 t}{\lambda}}\right)\right]$$

for $t < \lambda/(2\theta^2)$. Then

$$m_{\sum_{i=1}^n Y_i}(t) = \prod_{i=1}^{n} m_{Y_i}(t) = [m(t)]^n = \exp\left[\frac{n\lambda}{\theta}\left(1 - \sqrt{1 - \frac{2\theta^2 t}{\lambda}}\right)\right]$$

$$= \exp\left[\frac{n^2\lambda}{n\ \theta}\left(1 - \sqrt{1 - \frac{2(n\theta)^2\ t}{n^2\lambda}}\right)\right]$$

which is the mgf of an $IG(n\theta, n^2\lambda)$ random variable. The last equality was obtained by multiplying $\frac{n\lambda}{\theta}$ by $1 = n/n$ and by multiplying $\frac{2\theta^2 t}{\lambda}$ by $1 = n^2/n^2$. Hence $\sum_{i=1}^n Y_i \sim IG(n\theta, n^2\lambda)$.

1.5 The Multivariate Normal Distribution

Definition 1.37: Rao (1965, p. 437). A $p \times 1$ random vector $\boldsymbol{X}$ has a $p-$dimensional *multivariate normal distribution* $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ iff $\boldsymbol{t}^T\boldsymbol{X}$ has a univariate normal distribution for any $p \times 1$ vector $\boldsymbol{t}$.

If $\boldsymbol{\Sigma}$ is positive definite, then $\boldsymbol{X}$ has a joint pdf

$$f(\boldsymbol{z}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-(1/2)(\boldsymbol{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})} \tag{1.25}$$

where $|\boldsymbol{\Sigma}|^{1/2}$ is the square root of the determinant of $\boldsymbol{\Sigma}$. Note that if $p = 1$, then the quadratic form in the exponent is $(z - \mu)(\sigma^2)^{-1}(z - \mu)$ and X has the univariate $N(\mu, \sigma^2)$ pdf. If $\boldsymbol{\Sigma}$ is positive semidefinite but not positive definite, then $\boldsymbol{X}$ has a degenerate distribution. For example, the univariate $N(0, 0^2)$ distribution is degenerate (the point mass at 0).

Some important properties of MVN distributions are given in the following three theorems. These theorems can be proved using results from Johnson and Wichern (1988, p. 127-132).

Theorem 1.26. a) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and

$$\text{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}.$$

b) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then any linear combination $\boldsymbol{t}^T\boldsymbol{X} = t_1X_1 + \cdots + t_pX_p \sim N_1(\boldsymbol{t}^T\boldsymbol{\mu}, \boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t})$. Conversely, if $\boldsymbol{t}^T\boldsymbol{X} \sim N_1(\boldsymbol{t}^T\boldsymbol{\mu}, \boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t})$ for every $p \times 1$ vector $\boldsymbol{t}$, then $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

c) **The joint distribution of independent normal random variables is MVN.** If $X_1, ..., X_p$ are independent univariate normal $N(\mu_i, \sigma_i^2)$ random vectors, then $\boldsymbol{X} = (X_1, ..., X_p)^T$ is $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} = (\mu_1, ..., \mu_p)^T$ and $\boldsymbol{\Sigma} = diag(\sigma_1^2, ..., \sigma_p^2)$ (so the off diagonal entries $\sigma_{i,j} = 0$ while the diagonal entries of $\boldsymbol{\Sigma}$ are $\sigma_{i,i} = \sigma_i^2$.)

d) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and if $\boldsymbol{A}$ is a $q \times p$ matrix, then $\boldsymbol{A}\boldsymbol{X} \sim N_q(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$. If $\boldsymbol{a}$ is a $p \times 1$ vector of constants, then $\boldsymbol{a} + \boldsymbol{X} \sim N_p(\boldsymbol{a} + \boldsymbol{\mu}, \boldsymbol{\Sigma})$.

It will be useful to partition $\boldsymbol{X}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$. Let $\boldsymbol{X}_1$ and $\boldsymbol{\mu}_1$ be $q \times 1$ vectors, let $\boldsymbol{X}_2$ and $\boldsymbol{\mu}_2$ be $(p-q) \times 1$ vectors, let $\boldsymbol{\Sigma}_{11}$ be a $q \times q$ matrix, let $\boldsymbol{\Sigma}_{12}$ be a $q \times (p-q)$ matrix, let $\boldsymbol{\Sigma}_{21}$ be a $(p-q) \times q$ matrix, and let $\boldsymbol{\Sigma}_{22}$ be a $(p-q) \times (p-q)$ matrix. Then

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix}, \ \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

Theorem 1.27. a) **All subsets of a MVN are MVN:** $(X_{k_1}, ..., X_{k_q})^T \sim N_q(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$ where $\tilde{\boldsymbol{\mu}}_i = E(X_{k_i})$ and $\tilde{\boldsymbol{\Sigma}}_{ij} = \text{Cov}(X_{k_i}, X_{k_j})$. In particular, $\boldsymbol{X}_1 \sim N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $\boldsymbol{X}_2 \sim N_{p-q}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$.

b) If $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent, then $\text{Cov}(\boldsymbol{X}_1, \boldsymbol{X}_2) = \boldsymbol{\Sigma}_{12} = E[(\boldsymbol{X}_1 - E(\boldsymbol{X}_1))(\boldsymbol{X}_2 - E(\boldsymbol{X}_2))^T] = \boldsymbol{0}$, a $q \times (p-q)$ matrix of zeroes.

c) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent iff $\boldsymbol{\Sigma}_{12} = \boldsymbol{0}$.

d) If $\boldsymbol{X}_1 \sim N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $\boldsymbol{X}_2 \sim N_{p-q}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$ are independent, then

$$\begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix} \sim N_p\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right).$$

Theorem 1.28. The conditional distribution of a MVN is MVN. If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the conditional distribution of $\boldsymbol{X}_1$ given that $\boldsymbol{X}_2 = \boldsymbol{x}_2$ is multivariate normal with mean $\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)$ and covariance matrix $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. That is,

$$\boldsymbol{X}_1 | \boldsymbol{X}_2 = \boldsymbol{x}_2 \sim N_q(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}).$$

Example 1.4. Let $p = 2$ and let $(Y, X)^T$ have a bivariate normal distribution. That is,

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \begin{pmatrix} \sigma_Y^2 & \text{Cov}(Y, X) \\ \text{Cov}(X, Y) & \sigma_X^2 \end{pmatrix} \right).$$

Also recall that the population correlation between X and Y is given by

$$\rho(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{VAR}(X)}\sqrt{\text{VAR}(Y)}} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$$

if $\sigma_X > 0$ and $\sigma_Y > 0$. Then $Y|X = x \sim N(E(Y|X = x), \text{VAR}(Y|X = x))$ where the conditional mean

$$E(Y|X = x) = \mu_Y + \text{Cov}(Y,X)\frac{1}{\sigma_X^2}(x - \mu_X) = \mu_Y + \rho(X,Y)\sqrt{\frac{\sigma_Y^2}{\sigma_X^2}}(x - \mu_X)$$

and the conditional variance

$$\text{VAR}(Y|X = x) = \sigma_Y^2 - \text{Cov}(X,Y)\frac{1}{\sigma_X^2}\text{Cov}(X,Y)$$

$$= \sigma_Y^2 - \rho(X,Y)\sqrt{\frac{\sigma_Y^2}{\sigma_X^2}}\rho(X,Y)\sqrt{\sigma_X^2}\sqrt{\sigma_Y^2}$$

$$= \sigma_Y^2 - \rho^2(X,Y)\sigma_Y^2 = \sigma_Y^2[1 - \rho^2(X,Y)].$$

Also $aX + bY$ is univariate normal with mean $a\mu_X + b\mu_Y$ and variance

$$a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\ \text{Cov}(X,Y).$$

Remark 1.6. There are several common misconceptions. First, **it is not true that every linear combination $\boldsymbol{t}^T\boldsymbol{X}$ of normal random variables is a normal random variable,** and **it is not true that all uncorrelated normal random variables are independent**. The key condition in Theorem 1.26b and Theorem 1.27c is that the joint distribution of $\boldsymbol{X}$ is MVN. It is possible that $X_1, X_2, ..., X_p$ each has a marginal distribution that is univariate normal, but the joint distribution of $\boldsymbol{X}$ is not MVN. Examine the following example from Rohatgi (1976, p. 229). Suppose that the joint pdf of X and Y is a mixture of two bivariate normal distributions both with $EX = EY = 0$ and $\text{VAR}(X) = \text{VAR}(Y) = 1$, but $\text{Cov}(X,Y) = \pm\rho$. Hence

$$f(x,y) = \frac{1}{2}\ \frac{1}{2\pi\sqrt{1-\rho^2}}\exp(\frac{-1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)) +$$

$$\frac{1}{2}\ \frac{1}{2\pi\sqrt{1-\rho^2}}\exp(\frac{-1}{2(1-\rho^2)}(x^2 + 2\rho xy + y^2)) \equiv \frac{1}{2}f_1(x,y) + \frac{1}{2}f_2(x,y)$$

where x and y are real and $0 < \rho < 1$. Since both marginal distributions of $f_i(x,y)$ are N(0,1) for $i = 1$ and 2 by Theorem 1.27a, the marginal distributions of X and Y are N(0,1). Since $\int\int xyf_i(x,y)dxdy = \rho$ for $i = 1$ and $-\rho$ for $i = 2$, X and Y are uncorrelated, but X and Y are not independent since $f(x,y) \neq f_X(x)f_Y(y)$.

Remark 1.7. In Theorem 1.28, suppose that $\boldsymbol{X} = (Y, X_2, ..., X_p)^T$. Let $X_1 = Y$ and $\boldsymbol{X}_2 = (X_2, ..., X_p)^T$. Then $E[Y|\boldsymbol{X}_2] = \beta_1 + \beta_2 X_2 + \cdots + \beta_p X_p$ and VAR$[Y|\boldsymbol{X}_2]$ is a constant that does not depend on $\boldsymbol{X}_2$. Hence $Y|\boldsymbol{X}_2 = \beta_1 + \beta_2 X_2 + \cdots + \beta_p X_p + e$ follows the multiple linear regression model.

Example 1.5. Severini (2005, p. 236): Let $W \sim N(\mu_W, \sigma_W^2)$ and let $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The characteristic function of W is

$$c_W(y) = E(e^{iyW}) = \exp\left(iy\mu_W - \frac{y^2}{2}\sigma_w^2\right).$$

Prove that the characteristic function of $\boldsymbol{X}$ is

$$c_{\boldsymbol{X}}(\boldsymbol{t}) = \exp\left(i\boldsymbol{t}^T\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}\right).$$

Proof. Let $W = \boldsymbol{t}^T\boldsymbol{X}$. Then $W \sim N(\mu_W, \sigma_W^2)$ where $\mu_W = E(\boldsymbol{t}^T\boldsymbol{X}) = \boldsymbol{t}^T\boldsymbol{\mu}$ and $\sigma_W^2 = V(\boldsymbol{t}^T\boldsymbol{X}) = \text{Cov}(\boldsymbol{t}^T\boldsymbol{X}) = \boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}$. Then

$$c_{\boldsymbol{X}}(\boldsymbol{t}) = E(e^{i\boldsymbol{t}^T\boldsymbol{X}}) = c_W(1) = \exp\left(i\mu_W - \frac{1}{2}\sigma_w^2\right) = \exp\left(i\boldsymbol{t}^T\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}\right).$$

1.6 Exponential Families

Suppose the data is a random sample from some parametric brand name distribution with parameters $\boldsymbol{\theta}$. This brand name distribution comes from a family of distributions parameterized by $\boldsymbol{\theta} \in \Theta$. Each different value of $\boldsymbol{\theta}$ in the parameter space Θ gives a distribution that is a member of the family of distributions. Often the brand name family of distributions is from an exponential family.

Often a "brand name distribution" such as the normal distribution will have three useful parameterizations: the *usual parameterization* with parameter space Θ_U is simply the formula for the probability distribution or mass function (pdf or pmf, respectively) given when the distribution is first defined. The *k-parameter exponential family parameterization* with parameter space Θ, given in Definition 1.38 below, provides a simple way to determine if the distribution is an exponential family while the *natural parameterization* with parameter space Ω, given in Definition 1.39 below, is used for *theory* that requires a complete sufficient statistic.

Definition 1.38. A *family* of joint pdfs or joint pmfs $\{f(\boldsymbol{y}|\boldsymbol{\theta}) : \boldsymbol{\theta} = (\theta_1, ..., \theta_j) \in \Theta\}$ for a random vector $\boldsymbol{Y}$ is an **exponential family** if

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = h(\boldsymbol{y})c(\boldsymbol{\theta}) \exp\left[\sum_{i=1}^{k} w_i(\boldsymbol{\theta})t_i(\boldsymbol{y})\right] \tag{1.26}$$

for all $\boldsymbol{y}$ where $c(\boldsymbol{\theta}) \geq 0$ and $h(\boldsymbol{y}) \geq 0$. The functions c, h, t_i, and w_i are real valued functions. The parameter $\boldsymbol{\theta}$ can be a scalar and $\boldsymbol{y}$ can be a scalar. It is crucial that $c, w_1, ..., w_k$ do not depend on $\boldsymbol{y}$ and that $h, t_1, ..., t_k$ do not depend on $\boldsymbol{\theta}$. The support of the distribution is $\mathcal{Y}$ and the parameter space is Θ. The family is a **k-parameter exponential family** if k is the smallest integer where (1.26) holds.

Notice that the distribution of Y is an exponential family if

$$f(y|\boldsymbol{\theta}) = h(y)c(\boldsymbol{\theta}) \exp\left[\sum_{i=1}^{k} w_i(\boldsymbol{\theta})t_i(y)\right] \tag{1.27}$$

and the distribution is a one parameter exponential family if

$$f(y|\theta) = h(y)c(\theta) \exp[w(\theta)t(y)]. \tag{1.28}$$

The parameterization is not unique since, for example, w_i could be multiplied by a nonzero constant a if t_i is divided by a. Many other parameterizations are possible. If $h(y) = g(y)I_{\mathcal{Y}}(y)$, then usually $c(\boldsymbol{\theta})$ and $g(y)$ are positive, so another parameterization is

$$f(y|\boldsymbol{\theta}) = \exp\left[\sum_{i=1}^{k} w_i(\boldsymbol{\theta})t_i(y) + d(\boldsymbol{\theta}) + S(y)\right] \; I_{\mathcal{Y}}(y) \tag{1.29}$$

where $S(y) = \log(g(y))$, $d(\boldsymbol{\theta}) = \log(c(\boldsymbol{\theta}))$, and $\mathcal{Y}$ does not depend on $\boldsymbol{\theta}$.

To demonstrate that $\{f(\boldsymbol{y}|\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ is an exponential family, find $h(\boldsymbol{y}), c(\boldsymbol{\theta}), w_i(\boldsymbol{\theta})$ and $t_i(\boldsymbol{y})$ such that (1.26), (1.27), (1.28) or (1.29) holds.

Theorem 1.29. Suppose that $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n$ are iid random vectors from an exponential family. Then the joint distribution of $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n$ follows an exponential family.

Proof. Suppose that $f_{\boldsymbol{Y}_i}(\boldsymbol{y}_i)$ has the form of (1.26). Then by independence,

$$f(\boldsymbol{y}_1, ..., \boldsymbol{y}_n) = \prod_{i=1}^{n} f_{\boldsymbol{Y}_i}(\boldsymbol{y}_i) = \prod_{i=1}^{n} h(\boldsymbol{y}_i)c(\boldsymbol{\theta}) \exp\left[\sum_{j=1}^{k} w_j(\boldsymbol{\theta})t_j(\boldsymbol{y}_i)\right]$$

$$= [\prod_{i=1}^{n} h(\boldsymbol{y}_i)][c(\boldsymbol{\theta})]^n \prod_{i=1}^{n} \exp\left[\sum_{j=1}^{k} w_j(\boldsymbol{\theta})t_j(\boldsymbol{y}_i)\right]$$

$$= [\prod_{i=1}^{n} h(\boldsymbol{y}_i)][c(\boldsymbol{\theta})]^n \exp\left(\sum_{i=1}^{n}\left[\sum_{j=1}^{k} w_j(\boldsymbol{\theta}) t_j(\boldsymbol{y}_i)\right]\right)$$

$$= [\prod_{i=1}^{n} h(\boldsymbol{y}_i)][c(\boldsymbol{\theta})]^n \exp\left[\sum_{j=1}^{k} w_j(\boldsymbol{\theta})\left(\sum_{i=1}^{n} t_j(\boldsymbol{y}_i)\right)\right].$$

To see that this has the form (1.26), take $h^*(\boldsymbol{y}_1, ..., \boldsymbol{y}_n) = \prod_{i=1}^{n} h(\boldsymbol{y}_i)$, $c^*(\boldsymbol{\theta}) = [c(\boldsymbol{\theta})]^n$, $w_j^*(\boldsymbol{\theta}) = w_j(\boldsymbol{\theta})$ and $t_j^*(\boldsymbol{y}_1, ..., \boldsymbol{y}_n) = \sum_{i=1}^{n} t_j(\boldsymbol{y}_i)$. □

The parameterization that uses the **natural parameter $\boldsymbol{\eta}$** is especially useful for theory. See Definition 1.40 for the natural parameter space Ω.

Definition 1.39. Let Ω be the natural parameter space for $\boldsymbol{\eta}$. The **natural parameterization for an exponential family** is

$$f(\boldsymbol{y}|\boldsymbol{\eta}) = h(\boldsymbol{y}) b(\boldsymbol{\eta}) \exp\left[\sum_{i=1}^{k} \eta_i t_i(\boldsymbol{y})\right] \tag{1.30}$$

where $h(\boldsymbol{y})$ and $t_i(\boldsymbol{y})$ are the same as in Equation (1.26) and $\boldsymbol{\eta} \in \Omega$. The natural parameterization for a random variable Y is

$$f(y|\boldsymbol{\eta}) = h(y) b(\boldsymbol{\eta}) \exp\left[\sum_{i=1}^{k} \eta_i t_i(y)\right] \tag{1.31}$$

where $h(y)$ and $t_i(y)$ are the same as in Equation (1.27) and $\boldsymbol{\eta} \in \Omega$. Again, the parameterization is not unique. If $a \neq 0$, then $a\eta_i$ and $t_i(y)/a$ would also work.

Notice that the natural parameterization (1.31) has the same form as (1.27) with $\boldsymbol{\theta}^* = \boldsymbol{\eta}$, $c^*(\boldsymbol{\theta}^*) = b(\boldsymbol{\eta})$ and $w_i(\boldsymbol{\theta}^*) = w_i(\boldsymbol{\eta}) = \eta_i$. In applications often $\boldsymbol{\eta}$ and Ω are of interest while $b(\boldsymbol{\eta})$ is not computed.

The next important idea is that of a regular exponential family (and of a full exponential family). Let $d_i(x)$ denote $t_i(y)$, $w_i(\boldsymbol{\theta})$ or η_i. A *linearity constraint* is satisfied by $d_1(x), ..., d_k(x)$ if $\sum_{i=1}^{k} a_i d_i(x) = c$ for some constants a_i and c and for all x (or η_i) in the sample or parameter space where not all of the $a_i = 0$. If $\sum_{i=1}^{k} a_i d_i(x) = c$ for all x only if $a_1 = \cdots = a_k = 0$, then the $d_i(x)$ do not satisfy a linearity constraint. In linear algebra, we would say that the $d_i(x)$ are *linearly independent* if they do not satisfy a linearity constraint.

For $k = 2$, a linearity constraint is satisfied if a plot of $d_1(x)$ versus $d_2(x)$ falls on a line as x varies. If the parameter space for the η_1 and η_2 is a nonempty open set, then the plot of η_1 versus η_2 is that nonempty open set, and the η_i can not satisfy a linearity constraint since the plot is not a line.

Let $\tilde{\Omega}$ be the set where the integral of the kernel function is finite:

$$\tilde{\Omega} = \{\boldsymbol{\eta} = (\eta_1, ..., \eta_k) : \frac{1}{b(\boldsymbol{\eta})} \equiv \int_{-\infty}^{\infty} h(y) \exp[\sum_{i=1}^{k} \eta_i t_i(y)] dy < \infty\}. \quad (1.32)$$

Replace the integral by a sum for a pmf. An interesting fact is that $\tilde{\Omega}$ is a convex set. If the parameter space Θ of the exponential family is not a convex set, then the exponential family can not be regular. Example 1.7 shows that the χ^2_p distribution is not regular since the set of positive integers is not convex.

Definition 1.40. Condition E1: the natural parameter space $\Omega = \tilde{\Omega}$.
Condition E2: assume that in the natural parameterization, neither the η_i nor the t_i satisfy a linearity constraint.
Condition E3: Ω is a k-dimensional nonempty open set.
If conditions E1), E2) and E3) hold then the exponential family is a k-parameter **regular exponential family** (REF).
If conditions E1) and E2) hold then the exponential family is a k-parameter *full exponential family.*

Notation. A kP–REF is a k-parameter regular exponential family. So a 1P–REF is a 1-parameter REF and a 2P–REF is a 2-parameter REF.

Notice that every REF is full. Any k–dimensional nonempty open set will contain a k–dimensional nonempty rectangle. A k–fold cross product of nonempty open intervals is a k–dimensional nonempty open set. For a one parameter exponential family, a one dimensional rectangle is just an interval, and the only type of function of one variable that satisfies a linearity constraint is a constant function. In the definition of an exponential family and in the usual parameterization, $\boldsymbol{\theta}$ is a $1 \times j$ vector. Typically $j = k$ if the family is a kP–REF. If $j < k$ and k is as small as possible, the family will usually not be regular. For example, a $N(\theta, \theta^2)$ family has $\boldsymbol{\theta} = \theta$ with $j = 1 < 2 = k$, and is not regular.

Some care has to be taken with the definitions of Θ and Ω since formulas (1.26) and (1.31) need to hold for every $\boldsymbol{\theta} \in \Theta$ and for every $\boldsymbol{\eta} \in \Omega$. Let Θ_U be the usual parameter space given for the distribution. For a continuous random variable or vector, the pdf needs to exist. Hence all degenerate distributions need to be deleted from Θ_U to form Θ and Ω. For continuous and discrete distributions, the natural parameter needs to exist (and often does not exist for discrete degenerate distributions). As a rule of thumb, remove values from Θ_U that cause the pmf to have the form 0^0. For example, for the binomial(k, ρ) distribution with k known, the natural parameter $\eta = \log(\rho/(1-\rho))$. Hence instead of using $\Theta_U = [0, 1]$, use $\rho \in \Theta = (0, 1)$, so that $\eta \in \Omega = (-\infty, \infty)$.

These conditions have some redundancy. If Ω contains a k-dimensional rectangle (e.g. if the family is a kP-REF, then Ω is a k-dimensional open set and contains a k-dimensional open ball which contains a k-dimensional rectangle), no η_i is completely determined by the remaining $\eta_j's$. In particular,

the η_i cannot satisfy a linearity constraint. If the η_i do satisfy a linearity constraint, then the η_i lie on a hyperplane of dimension at most k, and such a surface cannot contain a k-dimensional rectangle. For example, if $k = 2$, a line cannot contain an open box. If $k = 2$ and $\eta_2 = \eta_1^2$, then the parameter space is not a 2-dimensional open set and does not contain a 2-dimensional rectangle. Thus the family is not a 2P–REF although η_1 and η_2 do not satisfy a linearity constraint.

The most important 1P–REFs are the binomial (k, ρ) distribution with k known, the exponential (λ) distribution, and the Poisson (θ) distribution. A one parameter exponential family can often be obtained from a k–parameter exponential family by holding $k - 1$ of the parameters fixed. Hence a normal (μ, σ^2) distribution is a 1P–REF if σ^2 is known. When data is modeled with an exponential family, often the scale, location and shape parameters are unknown. For example, the mean and standard deviation are usually both unknown.

The most important 2P–REFs are the beta (δ, ν) distribution, the gamma (ν, λ) distribution and the normal (μ, σ^2) distribution. The chi (p, σ) distribution, the inverted gamma (ν, λ) distribution, the log-gamma (ν, λ) distribution and the lognormal (μ, σ^2) distribution are also 2P-REFs. Olive (2014) gives many other examples showing that a distribution is a 1P-REF or 2P-REF. The two parameter Cauchy distribution is not an exponential family because its pdf cannot be put into the form of Equation (1.26).

The natural parameterization can result in a family that is much larger than the family defined by the usual parameterization. See the definition of $\Omega = \tilde{\Omega}$ given by Equation (1.31). Casella and Berger (2002, p. 114) remarks that

$$\{\boldsymbol{\eta} : \boldsymbol{\eta} = (w_1(\boldsymbol{\theta}), ..., w_k(\boldsymbol{\theta})) | \boldsymbol{\theta} \in \Theta\} \subseteq \Omega, \tag{1.33}$$

but often Ω is a strictly larger set.

Remark 1.8. For the families in this text other than the χ_p^2 and inverse Gaussian distributions, make the following assumptions if $\dim(\Theta) = k = \dim(\Omega)$. Assume that $\eta_i = w_i(\boldsymbol{\theta})$. Assume the usual parameter space Θ_U is as big as possible (replace the integral by a sum for a pmf):

$$\Theta_U = \{\boldsymbol{\theta} \in \mathbb{R}^k : \int f(y|\boldsymbol{\theta}) dy = 1\},$$

and let

$$\Theta = \{\boldsymbol{\theta} \in \Theta_U : w_1(\boldsymbol{\theta}), ..., w_k(\boldsymbol{\theta}) \text{ are defined } \}.$$

Then assume that the natural parameter space satisfies condition E1) with

$$\Omega = \{(\eta_1, ..., \eta_k) : \eta_i = w_i(\boldsymbol{\theta}) \text{ for } \boldsymbol{\theta} \in \Theta\}.$$

In other words, simply define $\eta_i = w_i(\boldsymbol{\theta})$. For many common distributions, $\boldsymbol{\eta}$ is a one to one function of $\boldsymbol{\theta}$, and the above map is correct, especially if Θ_U is an open interval or cross product of open intervals.

Example 1.6. Let $f(x|\mu,\sigma)$ be the $N(\mu,\sigma^2)$ family of pdfs. Then $\boldsymbol{\theta} = (\mu,\sigma)$ where $-\infty < \mu < \infty$ and $\sigma > 0$. Recall that μ is the mean and σ is the standard deviation (SD) of the distribution. The usual parameterization is

$$f(x|\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma} \exp(\frac{-(x-\mu)^2}{2\sigma^2}) I_{\mathbb{R}}(x)$$

where $\mathbb{R} = (-\infty,\infty)$ and the indicator $I_A(x) = 1$ if $x \in A$ and $I_A(x) = 0$ otherwise. Notice that $I_{\mathbb{R}}(x) = 1 \ \ \forall x$. Since

$$f(x|\mu,\sigma) = \underbrace{\frac{1}{\sqrt{2\pi}\sigma} \exp(\frac{-\mu^2}{2\sigma^2})}_{c(\mu,\sigma)\geq 0} \exp(\underbrace{\frac{-1}{2\sigma^2}}_{w_1(\boldsymbol{\theta})} \underbrace{x^2}_{t_1(x)} + \underbrace{\frac{\mu}{\sigma^2}}_{w_2(\boldsymbol{\theta})} \underbrace{x}_{t_2(x)}) \underbrace{I_{\mathbb{R}}(x)}_{h(x)\geq 0},$$

this family is a 2-parameter exponential family. Hence $\eta_1 = -0.5/\sigma^2$ and $\eta_2 = \mu/\sigma^2$ if $\sigma > 0$, and $\Omega = (-\infty,0) \times (-\infty,\infty)$. Plotting η_1 on the horizontal axis and η_2 on the vertical axis yields the left half plane which certainly contains a 2-dimensional rectangle. Since t_1 and t_2 lie on a quadratic rather than a line, the family is a 2P–REF. Notice that if $X_1, ..., X_n$ are iid $N(\mu,\sigma^2)$ random variables, then the joint pdf $f(\boldsymbol{x}|\boldsymbol{\theta}) = f(x_1, ..., x_n|\mu,\sigma) =$

$$\underbrace{[\frac{1}{\sqrt{2\pi}\sigma} \exp(\frac{-\mu^2}{2\sigma^2})]^n}_{C(\mu,\sigma)\geq 0} \exp(\underbrace{\frac{-1}{2\sigma^2}}_{w_1(\boldsymbol{\theta})} \underbrace{\sum_{i=1}^n x_i^2}_{T_1(\boldsymbol{x})} + \underbrace{\frac{\mu}{\sigma^2}}_{w_2(\boldsymbol{\theta})} \underbrace{\sum_{i=1}^n x_i}_{T_2(\boldsymbol{x})}) \underbrace{1}_{h(\boldsymbol{x})\geq 0},$$

and is thus a 2P–REF.

Example 1.7. The χ^2_p distribution is not a 1P-REF since the usual parameter space Θ_U for the χ^2_p distribution is the set of positive integers, which is neither an open set nor a convex set. Nevertheless, the natural parameterization is the gamma$(\nu, \lambda = 2)$ family which is a 1P-REF. Note that this family has uncountably many members while the χ^2_p family does not.

Example 1.8. The binomial(k,ρ) pmf is

$$f(x|\rho) = \binom{k}{x} \rho^x (1-\rho)^{k-x} \ I_{\{0,...,k\}}(x)$$

$$= \underbrace{\binom{k}{x} I_{\{0,...,k\}}(x)}_{h(x)\geq 0} \ \underbrace{(1-\rho)^k}_{c(\rho)\geq 0} \ \exp[\underbrace{\log(\frac{\rho}{1-\rho})}_{w(\rho)} \underbrace{x}_{t(x)}]$$

where $\Theta_U = [0,1]$. Since the pmf and $\eta = \log(\rho/(1-\rho))$ is undefined for $\rho = 0$ and $\rho = 1$, we have $\Theta = (0,1)$. Notice that $\Omega = (-\infty, \infty)$.

Example 1.9. The uniform(0,θ) family is not an exponential family since the support $\mathcal{Y}_\theta = (0,\theta)$ depends on the unknown parameter θ.

1.6.1 Properties of $(t_1(Y), ..., t_k(Y))$

Write the *natural parameterization for the exponential family* as

$$f(y|\boldsymbol{\eta}) = h(y)b(\boldsymbol{\eta}) \exp\left[\sum_{i=1}^{k} \eta_i t_i(y)\right]$$

$$= h(y) \exp\left[\sum_{i=1}^{k} \eta_i t_i(y) \quad - \quad a(\boldsymbol{\eta})\right] \tag{1.34}$$

where $a(\boldsymbol{\eta}) = -\log(b(\boldsymbol{\eta}))$.

Theorem 1.30. Suppose that Y comes from an exponential family (1.34) and that $g(y)$ is any function with $E_{\boldsymbol{\eta}}[|g(Y)|] < \infty$. Then for any $\boldsymbol{\eta}$ in the interior of Ω, the integral $\int g(y)f(y|\boldsymbol{\eta})dy$ is continuous and has derivatives of all orders. These derivatives can be obtained by interchanging the derivative and integral operators. If f is a pmf, replace the integral by a sum.

Proof. See Lehmann (1986, p. 59).

Hence

$$\frac{\partial}{\partial \eta_i} \int g(y)f(y|\boldsymbol{\eta})dy = \int g(y) \frac{\partial}{\partial \eta_i} f(y|\boldsymbol{\eta})dy \tag{1.35}$$

if f is a pdf and

$$\frac{\partial}{\partial \eta_i} \sum g(y)f(y|\boldsymbol{\eta}) = \sum g(y) \frac{\partial}{\partial \eta_i} f(y|\boldsymbol{\eta}) \tag{1.36}$$

if f is a pmf.

Remark 1.9. If $\boldsymbol{Y}$ comes from an exponential family (1.26), then the derivative and integral (or sum) operators can be interchanged. Hence

$$\frac{\partial}{\partial \theta_i} \int ... \int g(\boldsymbol{y})f(\boldsymbol{y}|\boldsymbol{\theta})d\boldsymbol{y} = \int ... \int g(\boldsymbol{y}) \frac{\partial}{\partial \theta_i} f(\boldsymbol{y}|\boldsymbol{\theta})d\boldsymbol{y}$$

for any function $g(\boldsymbol{y})$ with $E_\theta|g(\boldsymbol{Y})| < \infty$.

The behavior of $(t_1(Y), ..., t_k(Y))$ will be of considerable interest in later chapters. The following result is in Lehmann (1983, p. 29-30). Also see Johnson, Ladella, and Liu (1979).

Theorem 1.31. Suppose that Y comes from an exponential family (1.34). Then a)

$$E(t_i(Y)) = \frac{\partial}{\partial \eta_i} a(\boldsymbol{\eta}) = - \frac{\partial}{\partial \eta_i} \log(b(\boldsymbol{\eta})) \tag{1.37}$$

and b)

$$\mathrm{Cov}(t_i(Y), t_j(Y)) = \frac{\partial^2}{\partial \eta_i \partial \eta_j} a(\boldsymbol{\eta}) = - \frac{\partial^2}{\partial \eta_i \partial \eta_j} \log(b(\boldsymbol{\eta})). \tag{1.38}$$

Notice that $i = j$ gives the formula for $\mathrm{VAR}(t_i(Y))$.

Proof. The proof will be for pdfs. For pmfs replace the integrals by sums. Use Theorem 1.30 with $g(y) = 1 \ \forall y$. a) Since $1 = \int f(y|\boldsymbol{\eta}) dy$,

$$0 = \frac{\partial}{\partial \eta_i} 1 = \frac{\partial}{\partial \eta_i} \int h(y) \exp\left[\sum_{m=1}^{k} \eta_m t_m(y) \quad - \quad a(\boldsymbol{\eta}) \right] dy$$

$$= \int h(y) \frac{\partial}{\partial \eta_i} \exp\left[\sum_{m=1}^{k} \eta_m t_m(y) \quad - \quad a(\boldsymbol{\eta}) \right] dy$$

$$= \int h(y) \exp\left[\sum_{m=1}^{k} \eta_m t_m(y) \quad - \quad a(\boldsymbol{\eta}) \right] (t_i(y) - \frac{\partial}{\partial \eta_i} a(\boldsymbol{\eta})) dy$$

$$= \int (t_i(y) - \frac{\partial}{\partial \eta_i} a(\boldsymbol{\eta})) f(y|\boldsymbol{\eta}) dy$$

$$= E(t_i(Y)) - \frac{\partial}{\partial \eta_i} a(\boldsymbol{\eta}).$$

b) Similarly,

$$0 = \int h(y) \frac{\partial^2}{\partial \eta_i \partial \eta_j} \exp\left[\sum_{m=1}^{k} \eta_m t_m(y) \quad - \quad a(\boldsymbol{\eta}) \right] dy.$$

From the proof of a),

$$0 = \int h(y) \frac{\partial}{\partial \eta_j} \left[\exp\left[\sum_{m=1}^{k} \eta_m t_m(y) \quad - \quad a(\boldsymbol{\eta}) \right] (t_i(y) - \frac{\partial}{\partial \eta_i} a(\boldsymbol{\eta})) \right] dy$$

$$= \int h(y) \exp\left[\sum_{m=1}^{k} \eta_m t_m(y) \quad - \quad a(\boldsymbol{\eta}) \right] (t_i(y) - \frac{\partial}{\partial \eta_i} a(\boldsymbol{\eta}))(t_j(y) - \frac{\partial}{\partial \eta_j} a(\boldsymbol{\eta})) dy$$

$$-\int h(y)\exp\left[\sum_{m=1}^{k}\eta_m t_m(y) \quad - \quad a(\boldsymbol{\eta})\right](\frac{\partial^2}{\partial\eta_i\partial\eta_j}a(\boldsymbol{\eta}))dy$$

$$= \mathrm{Cov}(t_i(Y), t_j(Y)) - \frac{\partial^2}{\partial\eta_i\partial\eta_j}a(\boldsymbol{\eta})$$

since $\frac{\partial}{\partial\eta_j}a(\boldsymbol{\eta}) = E(t_j(Y))$ by a). □

Theorem 1.32. Suppose that Y comes from an exponential family (1.34), and let $\boldsymbol{T} = (t_1(Y), ..., t_k(Y)) = (T_1, ..., T_k)$. Then the distribution of $\boldsymbol{T}$ is an exponential family with

$$f(\boldsymbol{t}|\boldsymbol{\eta}) = h^*(\boldsymbol{t})\exp\left[\sum_{i=1}^{k}\eta_i t_i \quad - \quad a(\boldsymbol{\eta})\right].$$

Proof. See Lehmann (1986, p. 58).

The main point of this section is that $\boldsymbol{T}$ is well behaved even if Y is not. For example, if Y follows a one sided stable distribution, then Y is from an exponential family, but $E(Y)$ does not exist. However the mgf of T exists, so all moments of T exist. If $Y_1, ..., Y_n$ are iid from a one parameter exponential family, then $T \equiv T_n = \sum_{i=1}^{n} t(Y_i)$ is from a one parameter exponential family. One way to find the distribution function of T is to find the distribution of $t(Y)$ using the transformation method, then find the distribution of $\sum_{i=1}^{n} t(Y_i)$ using moment generating functions or Theorems 1.24 and 1.25. This technique results in the following two theorems. Notice that T often has a gamma distribution.

1.7 MSE, Information Number, MLE, UMVUE

Definition 1.41. Let the sample $\boldsymbol{Y} = (Y_1, ..., Y_n)$ where $\boldsymbol{Y}$ has a pdf or pmf $f(\boldsymbol{y}|\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \Theta$. Assume all relevant expectations exist. Let $\tau(\boldsymbol{\theta})$ be a real valued function of $\boldsymbol{\theta}$, and let $T \equiv T(Y_1, ..., Y_n)$ be an estimator of $\tau(\boldsymbol{\theta})$. The **bias** of the estimator T for $\tau(\boldsymbol{\theta})$ is

$$B(T) \equiv B_{\tau(\boldsymbol{\theta})}(T) \equiv \mathrm{Bias(T)} \equiv \mathrm{Bias}_{\tau(\boldsymbol{\theta})}(\mathrm{T}) = \mathrm{E}_{\boldsymbol{\theta}}(\mathrm{T}) - \tau(\boldsymbol{\theta}). \tag{1.39}$$

The *mean squared error* (**MSE**) of an estimator T for $\tau(\boldsymbol{\theta})$ is

$$\mathrm{MSE(T)} \equiv \mathrm{MSE}_{\tau(\boldsymbol{\theta})}(\mathrm{T}) = \mathrm{E}_{\boldsymbol{\theta}}[(\mathrm{T} - \tau(\boldsymbol{\theta}))^2]$$

$$= Var_{\boldsymbol{\theta}}(T) + [\mathrm{Bias}_{\tau(\boldsymbol{\theta})}(\mathrm{T})]^2. \tag{1.40}$$

T is an *unbiased estimator* of $\tau(\boldsymbol{\theta})$ if

$$E_{\boldsymbol{\theta}}(T) = \tau(\boldsymbol{\theta}) \tag{1.41}$$

for all $\boldsymbol{\theta} \in \Theta$. Notice that $\text{Bias}_{\tau(\boldsymbol{\theta})}(\text{T}) = 0$ for all $\boldsymbol{\theta} \in \Theta$ if T is an unbiased estimator of $\tau(\boldsymbol{\theta})$.

Notice that the bias and MSE are functions of $\boldsymbol{\theta}$ for $\boldsymbol{\theta} \in \Theta$. If $MSE_{\tau(\boldsymbol{\theta})}(T_1) < MSE_{\tau(\boldsymbol{\theta})}(T_2)$ for all $\boldsymbol{\theta} \in \Theta$, then T_1 is "a better estimator" of $\tau(\boldsymbol{\theta})$ than T_2. So estimators with small MSE are judged to be better than ones with large MSE. Often T_1 has smaller MSE than T_2 for some $\boldsymbol{\theta}$ but larger MSE for other values of $\boldsymbol{\theta}$. Often θ is real valued.

Example 1.10. Find the bias and MSE (as a function of n and c) of an estimator $T = c\sum_{i=1}^{n} Y_i$ or $(T = b\overline{Y})$ of μ when $Y_1, ..., Y_n$ are iid with $E(Y_1) = \mu = \theta$ and $V(Y_i) = \sigma^2$.
Solution: $E(T) = c\sum_{i=1}^{n} E(Y_i) = nc\mu$, $V(T) = c^2\sum_{i=1}^{n} V(Y_i) = nc^2\sigma^2$, $B(T) = E(T) - \mu$ and $MSE(T) = V(T) + [B(T)]^2$. (For $T = b\overline{Y}$, use $c = b/n$.)

In the class of unbiased estimators, the UMVUE is best since the UMVUE has the smallest variance, hence the smallest MSE. Often the MLE and method of moments estimator are biased but have a smaller MSE than the UMVUE. MLEs and method of moments estimators are widely used because they often have good statistical properties and are relatively easy to compute. Sometimes the UMVUE, MLE and method of moments estimators for θ are the same for a 1P-REF when $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} t(Y_i)$ and $\theta = E(\hat{\theta}) = E[t(Y)]$.

Definition 1.42. Let the sample $\boldsymbol{Y} = (Y_1, ..., Y_n)$ where $\boldsymbol{Y}$ has a pdf or pmf $f(\boldsymbol{y}|\theta)$ for $\theta \in \Theta$. Assume all relevant expectations exist. Let $\tau(\theta)$ be a real valued function of θ, and let $U \equiv U(Y_1, ..., Y_n)$ be an estimator of $\tau(\theta)$. Then U is the *uniformly minimum variance unbiased estimator* (**UMVUE**) of $\tau(\theta)$ if U is an unbiased estimator of $\tau(\theta)$ and if $\text{Var}_\theta(U) \leq \text{Var}_\theta(\text{W})$ for all $\theta \in \Theta$ where W is any other unbiased estimator of $\tau(\theta)$.

Often students will be asked to compute a lower bound on the variance of unbiased estimators of $\eta = \tau(\theta)$ when θ is a scalar. Some preliminary results are needed to define the lower bound, known as the FCRLB. The Fisher information, defined below, is also useful for large sample theory in Chapter 2 since often the asymptotic variance of a good estimator of $\tau(\theta)$ is $1/I_n(\tau(\theta))$. Good estimators tend to have a variance $\geq c/n$, so the FCRLB should be c/n for some positive constant c that may depend on the parameters of the distribution. Often $c = [\tau'(\theta)]^2/I_1(\theta)$.

Definition 1.43. Let $\boldsymbol{Y} = (Y_1, ..., Y_n)$ have a pdf or pmf $f(\boldsymbol{y}|\theta)$. Then the **information number** or **Fisher Information** is

$$I_{\boldsymbol{Y}}(\theta) \equiv I_n(\theta) = E_\theta \left(\left[\frac{\partial}{\partial \theta} \log(f(\boldsymbol{Y}|\theta)) \right]^2 \right). \tag{1.42}$$

Let $\eta = \tau(\theta)$ where $\tau'(\theta) \neq 0$. Then

$$I_n(\eta) \equiv I_n(\tau(\theta)) = \frac{I_n(\theta)}{[\tau'(\theta)]^2.} \tag{1.43}$$

Theorem 1.33. a) Equations (1.42) and (1.43) agree if $\tau'(\theta)$ is continuous, $\tau'(\theta) \neq 0$, and $\tau(\theta)$ is one to one and onto so that an inverse function exists such that $\theta = \tau^{-1}(\eta)$.

b) If the $Y_1 \equiv Y$ is from a 1P–REF, then the Fisher information in a sample of size one is

$$I_1(\theta) = -E_\theta \left[\frac{\partial^2}{\partial \theta^2} \log(f(Y|\theta)) \right]. \tag{1.44}$$

c) If the $Y_1, ..., Y_n$ are iid from a 1P–REF, then

$$I_n(\theta) = nI_1(\theta). \tag{1.45}$$

Hence if $\tau'(\theta)$ exists and is continuous and if $\tau'(\theta) \neq 0$, then

$$I_n(\tau(\theta)) = \frac{nI_1(\theta)}{[\tau'(\theta)]^2}. \tag{1.46}$$

Proof. a) See Lehmann (1999, p. 467–468).

b) The proof will be for a pdf. For a pmf replace the integrals by sums. By Remark 1.9, the integral and differentiation operators of all orders can be interchanged. Note that

$$0 = E \left[\frac{\partial}{\partial \theta} \log(f(Y|\theta)) \right] \tag{1.47}$$

since

$$\frac{\partial}{\partial \theta} 1 = 0 = \frac{\partial}{\partial \theta} \int f(y|\theta) dy = \int \frac{\partial}{\partial \theta} f(y|\theta) dy = \int \frac{\frac{\partial}{\partial \theta} f(y|\theta)}{f(y|\theta)} f(y|\theta) dy$$

or

$$0 = \frac{\partial}{\partial \theta} \int f(y|\theta) dy = \int \left[\frac{\partial}{\partial \theta} \log(f(y|\theta)) \right] f(y|\theta) dy$$

which is (1.47). Taking 2nd derivatives of the above expression gives

$$0 = \frac{\partial^2}{\partial \theta^2} \int f(y|\theta) dy = \frac{\partial}{\partial \theta} \int \left[\frac{\partial}{\partial \theta} \log(f(y|\theta)) \right] f(y|\theta) dy =$$

$$\int \frac{\partial}{\partial\theta}\left(\left[\frac{\partial}{\partial\theta}\log(f(y|\theta))\right] f(y|\theta)\right) dy =$$

$$\int \left[\frac{\partial^2}{\partial\theta^2}\log(f(y|\theta))\right] f(y|\theta)dy + \int \left[\frac{\partial}{\partial\theta}\log(f(y|\theta))\right]\left[\frac{\partial}{\partial\theta}f(y|\theta)\right]\frac{f(y|\theta)}{f(y|\theta)}dy$$

$$= \int \left[\frac{\partial^2}{\partial\theta^2}\log(f(y|\theta))\right] f(y|\theta)dy + \int \left[\frac{\partial}{\partial\theta}\log(f(y|\theta))\right]^2 f(y|\theta)dy$$

or

$$I_1(\theta) = E_\theta[(\frac{\partial}{\partial\theta}\log f(Y|\theta))^2] = -E_\theta\left[\frac{\partial^2}{\partial\theta^2}\log(f(Y|\theta))\right].$$

c) By independence,

$$I_n(\theta) = E_\theta\left[\left(\frac{\partial}{\partial\theta}\log(\prod_{i=1}^n f(Y_i|\theta))\right)^2\right] = E_\theta\left[\left(\frac{\partial}{\partial\theta}\sum_{i=1}^n \log(f(Y_i|\theta))\right)^2\right] =$$

$$E_\theta\left[\left(\frac{\partial}{\partial\theta}\sum_{i=1}^n \log(f(Y_i|\theta))\right)\left(\frac{\partial}{\partial\theta}\sum_{j=1}^n \log(f(Y_j|\theta))\right)\right] =$$

$$E_\theta\left[\left(\sum_{i=1}^n \frac{\partial}{\partial\theta}\log(f(Y_i|\theta))\right)\left(\sum_{j=1}^n \frac{\partial}{\partial\theta}\log(f(Y_j|\theta))\right)\right] =$$

$$\sum_{i=1}^n E_\theta\left[\left(\frac{\partial}{\partial\theta}\log(f(Y_i|\theta))\right)^2\right] +$$

$$\sum_{i \neq j}\sum E_\theta\left[\left(\frac{\partial}{\partial\theta}\log(f(Y_i|\theta))\right)\left(\frac{\partial}{\partial\theta}\log(f(Y_j|\theta))\right)\right].$$

Hence

$$I_n(\theta) = nI_1(\theta) + \sum_{i \neq j}\sum E_\theta\left[\left(\frac{\partial}{\partial\theta}\log(f(Y_i|\theta))\right)\right] E_\theta\left[\left(\frac{\partial}{\partial\theta}\log(f(Y_j|\theta))\right)\right]$$

by independence. Hence

$$I_n(\theta) = nI_1(\theta) + n(n-1)\left[E_\theta\left(\frac{\partial}{\partial\theta}\log(f(Y_j|\theta))\right)\right]^2$$

since the Y_i are iid. Thus $I_n(\theta) = nI_1(\theta)$ by Equation (1.47) which holds since the Y_i are iid from a 1P–REF. □

Definition 1.44. Let $\boldsymbol{Y} = (Y_1, ..., Y_n)$ be the data, and consider $\tau(\theta)$ where $\tau'(\theta) \neq 0$. The quantity

$$\text{FCRLB}_{\text{n}}(\tau(\theta)) = \frac{[\tau'(\theta)]^2}{\text{I}_{\text{n}}(\theta)}$$

is called the **Fréchet Cramér Rao lower bound** (FCRLB) for the variance of unbiased estimators of $\tau(\theta)$. In particular, if $\tau(\theta) = \theta$, then $\text{FCRLB}_n(\theta) = \frac{1}{I_n(\theta)}$. The FCRLB is often called the Cramér Rao lower bound (CRLB).

Theorem 1.34, Fréchet Cramér Rao Lower Bound or Information Inequality. Let $Y_1, ..., Y_n$ be iid from a 1P–REF with pdf or pmf $f(y|\theta)$. Let $W(Y_1, ..., Y_n) = W(\boldsymbol{Y})$ be any unbiased estimator of $\tau(\theta) \equiv E_\theta W(\boldsymbol{Y})$. Then

$$\text{VAR}_\theta(W(\boldsymbol{Y})) \geq FCRLB_n(\tau(\theta)) = \frac{[\tau'(\theta)]^2}{I_n(\theta)} = \frac{[\tau'(\theta)]^2}{nI_1(\theta)} = \frac{1}{I_n(\tau(\theta))}.$$

Proof. See Olive (2014, pp. 164-166).

Definition 1.45. Let $f(\boldsymbol{y}|\boldsymbol{\theta})$ be the pmf or pdf of a sample $\boldsymbol{Y}$ with parameter space Θ. If $\boldsymbol{Y} = \boldsymbol{y}$ is observed, then the **likelihood function** is $L(\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta}|\boldsymbol{y}) = f(\boldsymbol{y}|\boldsymbol{\theta})$. For each sample point $\boldsymbol{y} = (y_1, ..., y_n)$, let $\hat{\boldsymbol{\theta}}(\boldsymbol{y}) \in \Theta$ be a parameter value at which $L(\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta}|\boldsymbol{y})$ attains its maximum as a function of $\boldsymbol{\theta}$ with $\boldsymbol{y}$ held fixed. Then a maximum likelihood estimator (**MLE**) of the parameter $\boldsymbol{\theta}$ based on the sample $\boldsymbol{Y}$ is $\hat{\boldsymbol{\theta}}(\boldsymbol{Y})$.

The following remarks are important. I) It is crucial to observe that the likelihood function is a function of $\boldsymbol{\theta}$ (and that $y_1, ..., y_n$ act as fixed constants). Note that the pdf or pmf $f(\boldsymbol{y}|\boldsymbol{\theta})$ is a function of n variables while $L(\boldsymbol{\theta})$ is a function of k variables if $\boldsymbol{\theta}$ is a $1 \times k$ vector. Often $k = 1$ or $k = 2$ while n could be in the hundreds or thousands.

II) If $Y_1, ..., Y_n$ is an independent sample from a population with pdf or pmf $f(y|\boldsymbol{\theta})$, then the likelihood function

$$L(\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta}|y_1, ..., y_n) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta}). \tag{1.48}$$

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f_i(y_i|\boldsymbol{\theta})$$

if the Y_i are independent but have different pdfs or pmfs.

III) If the MLE $\hat{\boldsymbol{\theta}}$ exists, then $\hat{\boldsymbol{\theta}} \in \Theta$. Hence if $\hat{\boldsymbol{\theta}}$ is not in the parameter space Θ, then $\hat{\boldsymbol{\theta}}$ is not the MLE of $\boldsymbol{\theta}$.

Theorem 1.35: Invariance Principle. If $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$, then $h(\hat{\boldsymbol{\theta}})$ is the MLE of $h(\boldsymbol{\theta})$ where h is a function with domain Θ.

Proof. When h is one to one, let $\boldsymbol{\eta} = h(\boldsymbol{\theta})$. Then the inverse function h^{-1} exists and $\boldsymbol{\theta} = h^{-1}(\boldsymbol{\eta})$. Hence

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = f(\boldsymbol{x}|h^{-1}(\boldsymbol{\eta})) \tag{1.49}$$

is the joint pdf or pmf of $\boldsymbol{x}$. So the likelihood function of $h(\boldsymbol{\theta}) = \boldsymbol{\eta}$ is

$$L^*(\boldsymbol{\eta}) \equiv K(\boldsymbol{\eta}) = L(h^{-1}(\boldsymbol{\eta})). \tag{1.50}$$

Also note that

$$\sup_{\boldsymbol{\eta}} K(\boldsymbol{\eta}|\boldsymbol{x}) = \sup_{\boldsymbol{\eta}} L(h^{-1}(\boldsymbol{\eta})|\boldsymbol{x}) = L(\hat{\boldsymbol{\theta}}|\boldsymbol{x}). \tag{1.51}$$

Thus

$$\hat{\boldsymbol{\eta}} = h(\hat{\boldsymbol{\theta}}) \tag{1.52}$$

is the MLE of $\boldsymbol{\eta} = h(\boldsymbol{\theta})$ when h is one to one.

If h is not one to one, then the new parameters $\boldsymbol{\eta} = h(\boldsymbol{\theta})$ do not give enough information to define $f(\boldsymbol{x}|\boldsymbol{\eta})$. Hence we cannot define the likelihood. That is, a $N(\mu, \sigma^2)$ density cannot be defined by the parameter μ/σ alone. Before concluding that the MLE does not exist if h is not one to one, note that if $X_1, ..., X_n$ are iid $N(\mu, \sigma^2)$ then $X_1, ..., X_n$ remain iid $N(\mu, \sigma^2)$ even though the investigator did not rename the parameters wisely or is interested in a function $h(\mu, \sigma) = \mu/\sigma$ that is not one to one. Berk (1967) said that if h is not one to one, define

$$w(\boldsymbol{\theta}) = (h(\boldsymbol{\theta}), u(\boldsymbol{\theta})) = (\boldsymbol{\eta}, \boldsymbol{\gamma}) = \boldsymbol{\xi} \tag{1.53}$$

such that $w(\boldsymbol{\theta})$ is one to one. Note that the choice

$$w(\boldsymbol{\theta}) = (h(\boldsymbol{\theta}), \boldsymbol{\theta})$$

works. In other words, we can always take u to be the identity function.

The choice of w is not unique, but the inverse function

$$w^{-1}(\boldsymbol{\xi}) = \boldsymbol{\theta}$$

is unique. Hence the likelihood is well defined, and $w(\hat{\boldsymbol{\theta}})$ is the MLE of $\boldsymbol{\xi}$. □

There are **four commonly used techniques** for finding the MLE.

- Potential candidates can be found by differentiating log $L(\boldsymbol{\theta})$, the log likelihood.
- Potential candidates can be found by differentiating the likelihood $L(\boldsymbol{\theta})$.
- The MLE can sometimes be found by direct maximization of the likelihood $L(\boldsymbol{\theta})$.

- **Invariance Principle**: If $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$, then $h(\hat{\boldsymbol{\theta}})$ is the MLE of $h(\boldsymbol{\theta})$.

The method of moments is another useful way for obtaining point estimators. Let $Y_1, ..., Y_n$ be an iid sample and let

$$\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} Y_i^j \text{ and } \mu_j \equiv \mu_j(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}(Y^j) \tag{1.54}$$

for $j = 1, ..., k$. So $\hat{\mu}_j$ is the jth sample moment and μ_j is the jth population moment. Fix k and assume that $\mu_j = \mu_j(\theta_1, ..., \theta_k)$. Solve the system

$$\begin{aligned} \hat{\mu}_1 &\stackrel{\text{set}}{=} \mu_1(\theta_1, ..., \theta_k) \\ &\vdots \qquad\qquad \vdots \\ \hat{\mu}_k &\stackrel{\text{set}}{=} \mu_k(\theta_1, ..., \theta_k) \end{aligned}$$

for $\tilde{\boldsymbol{\theta}}$.

Definition 1.46. The solution $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1, ..., \tilde{\theta}_k)^T$ is the **method of moments estimator** of $\boldsymbol{\theta}$. If g is a continuous function of the first k moments and $h(\boldsymbol{\theta}) = g(\mu_1(\boldsymbol{\theta}), ..., \mu_k(\boldsymbol{\theta}))$, then the method of moments estimator of $h(\boldsymbol{\theta})$ is

$$g(\hat{\mu}_1, ..., \hat{\mu}_k).$$

Definition 1.46 is similar to the invariance principle for the MLE, but note that g needs to be a continuous function, and the definition only applies to a function of $(\hat{\mu}_1, ..., \hat{\mu}_k)$ where $k \geq 1$. Hence $\overline{Y}$ is the method of moments estimator of the population mean μ, and $g(\overline{Y})$ is the method of moments estimator of $g(\mu)$ if g is a continuous function. Sometimes the notation $\hat{\theta}_{MLE}$ and $\hat{\theta}_{MM}$ will be used to denote the MLE and method of moments estimators of θ, respectively. As with maximum likelihood estimators, not all method of moments estimators exist in closed form, and then numerical techniques must be used.

1.8 Mixture Distributions

Mixture distributions are useful for model and variable selection since $\hat{\boldsymbol{\beta}}_{I_{min},0}$ is a mixture distribution of $\hat{\boldsymbol{\beta}}_{I_j,0}$, and the lasso estimator $\hat{\boldsymbol{\beta}}_L$ is a mixture distribution of $\hat{\boldsymbol{\beta}}_{L,\lambda_i}$ for $i = 1, ..., M$. See Chapter 6. A random vector $\boldsymbol{u}$ has a mixture distribution if $\boldsymbol{u}$ equals a random vector $\boldsymbol{u}_j$ with probability π_j

for $j = 1, ..., J$. See Definition 1.29 for the population mean and population covariance matrix of a random vector.

Definition 1.47. The distribution of a $g \times 1$ random vector $\boldsymbol{u}$ is a mixture distribution if the cumulative distribution function (cdf) of $\boldsymbol{u}$ is

$$F_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_{j=1}^{J} \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{t}) \tag{1.55}$$

where the probabilities π_j satisfy $0 \leq \pi_j \leq 1$ and $\sum_{j=1}^{J} \pi_j = 1$, $J \geq 2$, and $F_{\boldsymbol{u}_j}(\boldsymbol{t})$ is the cdf of a $g \times 1$ random vector $\boldsymbol{u}_j$. Then $\boldsymbol{u}$ has a mixture distribution of the $\boldsymbol{u}_j$ with probabilities π_j.

Theorem 1.36. Suppose $E(h(\boldsymbol{u}))$ and the $E(h(\boldsymbol{u}_j))$ exist. Then

$$E[h(\boldsymbol{u})] = \sum_{j=1}^{J} \pi_j E[h(\boldsymbol{u}_j)]. \tag{1.56}$$

Hence

$$E(\boldsymbol{u}) = \sum_{j=1}^{J} \pi_j E[\boldsymbol{u}_j], \tag{1.57}$$

and $Cov(\boldsymbol{u}) = E(\boldsymbol{u}\boldsymbol{u}^T) - E(\boldsymbol{u})E(\boldsymbol{u}^T) = E(\boldsymbol{u}\boldsymbol{u}^T) - E(\boldsymbol{u})[E(\boldsymbol{u})]^T =$
$\sum_{j=1}^{J} \pi_j E[\boldsymbol{u}_j \boldsymbol{u}_j^T] - E(\boldsymbol{u})[E(\boldsymbol{u})]^T =$

$$\sum_{j=1}^{J} \pi_j Cov(\boldsymbol{u}_j) + \sum_{j=1}^{J} \pi_j E(\boldsymbol{u}_j)[E(\boldsymbol{u}_j)]^T - E(\boldsymbol{u})[E(\boldsymbol{u})]^T. \tag{1.58}$$

If $E(\boldsymbol{u}_j) = \boldsymbol{\theta}$ for $j = 1, ..., J$, then $E(\boldsymbol{u}) = \boldsymbol{\theta}$ and

$$Cov(\boldsymbol{u}) = \sum_{j=1}^{J} \pi_j Cov(\boldsymbol{u}_j).$$

This theorem is easy to prove if the $\boldsymbol{u}_j$ are continuous random vectors with (joint) probability density functions (pdfs) $f_{\boldsymbol{u}_j}(\boldsymbol{t})$. Then $\boldsymbol{u}$ is a continuous random vector with pdf

$$f_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_{j=1}^{J} \pi_j f_{\boldsymbol{u}_j}(\boldsymbol{t}), \quad \text{and} \quad \mathrm{E[h(\boldsymbol{u})]} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathrm{h}(\boldsymbol{t}) \mathrm{f}_{\boldsymbol{u}}(\boldsymbol{t}) \mathrm{d}\boldsymbol{t}$$

$$= \sum_{j=1}^{J} \pi_j \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(\boldsymbol{t}) f_{\boldsymbol{u}_j}(\boldsymbol{t}) d\boldsymbol{t} = \sum_{j=1}^{J} \pi_j E[h(\boldsymbol{u}_j)]$$

where $E[h(\boldsymbol{u}_j)]$ is the expectation with respect to the random vector $\boldsymbol{u}_j$. Note that

$$E(\boldsymbol{u})[E(\boldsymbol{u})]^T = \sum_{j=1}^{J}\sum_{k=1}^{J} \pi_j \pi_k E(\boldsymbol{u}_j)[E(\boldsymbol{u}_k)]^T. \tag{1.59}$$

Alternatively, with respect to a Riemann Stieltjes integral, $E[h(\boldsymbol{u})] = \int h(\boldsymbol{t}) dF(\boldsymbol{t})$ provided the expected value exists, and the integral is a linear operator with respect to both h and F. Hence for a mixture distribution, $E[h(\boldsymbol{u})] = \int h(\boldsymbol{t}) dF(\boldsymbol{t}) =$

$$\int h(\boldsymbol{t}) \; d\left[\sum_{j=1}^{J} \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{t})\right] = \sum_{j=1}^{J} \pi_j \int h(\boldsymbol{t}) dF_{\boldsymbol{u}_j}(\boldsymbol{t}) = \sum_{j=1}^{J} \pi_j E[h(\boldsymbol{u}_j)].$$

1.9 Elliptically Contoured Distributions

Definition 1.48: Johnson (1987, pp. 107-108). A $p \times 1$ random vector $\boldsymbol{X}$ has an *elliptically contoured distribution,* also called an *elliptically symmetric distribution,* if $\boldsymbol{X}$ has joint pdf

$$f(\boldsymbol{z}) = k_p |\boldsymbol{\Sigma}|^{-1/2} g[(\boldsymbol{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{z} - \boldsymbol{\mu})], \tag{1.60}$$

and we say $\boldsymbol{X}$ has an elliptically contoured $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution.

If $\boldsymbol{X}$ has an elliptically contoured (EC) distribution, then the characteristic function of $\boldsymbol{X}$ is

$$\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \exp(i\boldsymbol{t}^T \boldsymbol{\mu}) \psi(\boldsymbol{t}^T \boldsymbol{\Sigma} \boldsymbol{t}) \tag{1.61}$$

for some function ψ. If the second moments exist, then

$$E(\boldsymbol{X}) = \boldsymbol{\mu} \tag{1.62}$$

and

$$\text{Cov}(\boldsymbol{X}) = c_X \boldsymbol{\Sigma} \tag{1.63}$$

where

$$c_X = -2\psi'(0).$$

Definition 1.49. The *population squared Mahalanobis distance*

$$U \equiv D^2 = D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{X} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{X} - \boldsymbol{\mu}). \tag{1.64}$$

For elliptically contoured distributions, U has pdf

$$h(u) = \frac{\pi^{p/2}}{\Gamma(p/2)} k_p u^{p/2-1} g(u). \tag{1.65}$$

For $c > 0$, an $EC_p(\boldsymbol{\mu}, c\boldsymbol{I}, g)$ distribution is *spherical about* $\boldsymbol{\mu}$ where $\boldsymbol{I}$ is the $p \times p$ identity matrix. The *multivariate normal distribution* $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ has $k_p = (2\pi)^{-p/2}$, $\psi(u) = g(u) = \exp(-u/2)$, and $h(u)$ is the χ^2_p pdf.

1.10 Some Useful Distributions

Let the population quantile be y_δ. Then $P(Y \leq y_\delta) = \delta$ if Y has a pdf that is positive at y_δ.

Definition 1.50. The *gamma function* $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$ for $x > 0$.

Some properties of the gamma function follow. i) $\Gamma(k) = (k-1)!$ for integer $k \geq 1$. ii) $\Gamma(x+1) = x\ \Gamma(x)$ for $x > 0$. iii) $\Gamma(x) = (x-1)\ \Gamma(x-1)$ for $x > 1$. iv) $\Gamma(0.5) = \sqrt{\pi}$.

1) $Y \sim \text{beta}(\delta, \nu)$

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

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

$$E(Y) = \frac{\delta}{\delta+\nu}, \quad V(Y) = \frac{\delta\nu}{(\delta+\nu)^2(\delta+\nu+1)}.$$

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

$$m(t) = [(1-\rho) + \rho e^t], \quad c(t) = [(1-\rho) + \rho e^{it}].$$

3) binomial(k, ρ), $Y \sim BIN(k, \rho)$,

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

for $y = 0, 1, \ldots, k$ where $0 < \rho < 1$. $E(Y) = k\rho$, $\quad V(Y) = k\rho(1-\rho)$. 1P-REF is k is known, and $I_1(\rho) = \dfrac{k}{\rho(1-\rho)}$. $m(t) = [(1-\rho)+\rho e^t]^k$, $\ c(t) = [(1-\rho) + \rho e^{it}]^k$. If $Y_1, ..., Y_n$ are independent binomial BIN(k_i, ρ) random variables, then

$$\sum_{i=1}^{n} Y_i \sim \text{BIN}\left(\sum_{\text{i=1}}^{\text{n}} \text{k}_\text{i}, \rho\right).$$

Thus if $Y_1, ..., Y_n$ are iid BIN(k, ρ) random variables, then $\sum_{i=1}^{n} Y_i \sim \text{BIN}(\text{nk}, \rho)$.

4) $Y \sim \text{Cauchy}(\mu, \sigma)$,

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

where y and μ are real numbers and $\sigma > 0$. $E(Y) = \infty = \text{VAR}(Y)$. $E(Y)$ and $V(Y)$ do not exist. $c(t) = \exp(it\mu - |t|\sigma)$.

$$F(y) = \frac{1}{\pi}[\arctan(\frac{y-\mu}{\sigma}) + \pi/2].$$

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

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

where $y > 0$ and p is a positive integer. $E(Y) = p$, $V(Y) = 2p$.

$$m(t) = \left(\frac{1}{1-2t}\right)^{p/2} = (1-2t)^{-p/2} \text{ for } \text{t} < 1/2, \quad \text{c(t)} = \left(\frac{1}{1-\text{i2t}}\right)^{\text{p/2}}.$$

If $Y_1, ..., Y_n$ are independent chi–square $\chi^2_{p_i}$, then

$$\sum_{i=1}^{n} Y_i \sim \chi^2(\sum_{i=1}^{n} p_i).$$

Thus if $Y_1, ..., Y_n$ are iid χ^2_p, then

$$\sum_{i=1}^{n} Y_i \sim \chi^2_{np}.$$

6) exponential(λ)= gamma$(\nu = 1, \lambda)$, $Y \sim \text{EXP}(\lambda)$

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

where $\lambda > 0$. $E(Y) = \lambda$, $V(Y) = \lambda^2$, and $y_\delta = -\lambda \ln(1-\delta)$. 1P-REF and $I_1(\lambda) = 1/\lambda^2$.

$$m(t) = 1/(1-\lambda t) \text{ for } \text{t} < 1/\lambda, \quad \text{c(t)} = 1/(1-\text{i}\lambda\text{t}).$$

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

If $Y_1, ..., Y_n$ are iid exponential EXP(λ), then

$$\sum_{i=1}^{n} Y_i \sim G(n, \lambda).$$

7) gamma(ν, λ), $Y \sim G(\nu, \lambda)$,

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

where ν, λ, and y are positive. $E(Y) = \nu\lambda$, $V(Y) = \nu\lambda^2$. 2P-REF and if ν is known, then $I_1(\lambda) = \nu/\lambda^2$.

$$m(t) = \left(\frac{1}{1-\lambda t}\right)^\nu \quad \text{for} \quad \mathrm{t} < 1/\lambda, \quad \mathrm{c(t)} = \left(\frac{1}{1-\mathrm{i}\lambda\mathrm{t}}\right)^\nu.$$

If $Y_1, ..., Y_n$ are independent Gamma $G(\nu_i, \lambda)$ then

$$\sum_{i=1}^{n} Y_i \sim G(\sum_{i=1}^{n} \nu_i, \lambda).$$

Thus if $Y_1, ..., Y_n$ are iid $G(\nu, \lambda)$, then $\sum_{i=1}^{n} Y_i \sim G(n\nu, \lambda)$.

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

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

where $\sigma > 0$ and μ and y are real. $E(Y) = \mu$, $V(Y) = \sigma^2$, and $y_\delta = \mu + \sigma z_\delta$. 2P-REF. If σ^2 is known, then $I_1(\mu) = 1/\sigma^2$. If μ is known, then $I_1(\sigma^2) = \frac{1}{2\sigma^4}$.

$$I_1(\mu, \sigma) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{pmatrix}, \quad I_1(\mu, \sigma^2) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}.$$

$$m(t) = \exp(t\mu + t^2\sigma^2/2), \quad c(t) = \exp(it\mu - t^2\sigma^2/2).$$

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

If $Y_1, ..., Y_n$ are independent normal $N(\mu_i, \sigma_i^2)$, then

$$\sum_{i=1}^{n} (a_i + b_i Y_i) \sim N(\sum_{i=1}^{n} (a_i + b_i\mu_i), \sum_{i=1}^{n} b_i^2\sigma_i^2).$$

Here a_i and b_i are fixed constants. Thus if $Y_1, ..., Y_n$ are iid $N(\mu, \sigma^2)$, then $\overline{Y} \sim N(\mu, \sigma^2/n)$.

9) Poisson(θ), $Y \sim$ POIS(θ)

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

for $y = 0, 1, \ldots$, where $\theta > 0$. $E(Y) = \theta = V(Y)$. 1P-REF and $I_1(\theta) = 1/\theta$.

$$m(t) = \exp(\theta(e^t - 1)), \quad c(t) = \exp(\theta(e^{it} - 1)).$$

If $Y_1, ..., Y_n$ are independent POIS(θ_i), then

$$\sum_{i=1}^{n} Y_i \sim \text{POIS}(\sum_{\text{i=1}}^{\text{n}} \theta_{\text{i}}).$$

Thus if $Y_1, ..., Y_n$ are iid $POIS(\theta)$, then

$$\sum_{i=1}^{n} Y_i \sim \text{POIS}(\text{n}\theta).$$

10) uniform(θ_1, θ_2), $Y \sim U(\theta_1, \theta_2)$.

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

$F(y) = (y - \theta_1)/(\theta_2 - \theta_1)$ for $\theta_1 \leq y \leq \theta_2$. $E(Y) = (\theta_1 + \theta_2)/2$. $V(Y) = (\theta_2 - \theta_1)^2/12$, and $y_\delta = (\theta_2 - \theta_1)\delta + \theta_1$. By definition, $m(0) = c(0) = 1$. For $t \neq 0$,

$$m(t) = \frac{e^{t\theta_2} - e^{t\theta_1}}{(\theta_2 - \theta_1)t}, \quad \text{and} \quad \text{c(t)} = \frac{\text{e}^{\text{it}\theta_2} - \text{e}^{\text{it}\theta_1}}{(\theta_2 - \theta_1)\text{it}}.$$

11) point mass at c: The distribution of Y is a point mass at c (or Y is degenerate at c) if $P(Y = c) = 1$ with pmf $f(c) = 1$. Hence $Y \sim N(c, 0)$, $E(Y) = c$, $V(Y) = 0$. $m(t) = e^{tc}$. $c(t) = e^{itc}$.

More Distributions:

12) If Y has a geometric distribution, $Y \sim$ geom(ρ) then the pmf of Y is

$$f(y) = P(Y = y) = \rho(1 - \rho)^y$$

for $y = 0, 1, 2, ...$ and $0 < \rho < 1$. $E(Y) = (1 - \rho)/\rho$. $V(Y) = (1 - \rho)/\rho^2$. $Y \sim NB(1, \rho)$. Hence the mgf of Y is

$$m(t) = \frac{\rho}{1 - (1 - \rho)e^t}$$

for $t < -\log(1 - \rho)$.

13) If Y has an inverse Gaussian distribution, $Y \sim \text{IG}(\theta, \lambda)$, then the pdf of Y is

$$f(y) = \sqrt{\frac{\lambda}{2\pi y^3}} \exp\left[\frac{-\lambda(y-\theta)^2}{2\theta^2 y}\right]$$

where $y, \theta, \lambda > 0$. $E(Y) = \theta$ and

$$V(Y) = \frac{\theta^3}{\lambda}.$$

The mgf is

$$m(t) = \exp\left[\frac{\lambda}{\theta}\left(1 - \sqrt{1 - \frac{2\theta^2 t}{\lambda}}\right)\right] \quad t < \lambda/(2\theta^2), \quad c(t) = \exp\left[\frac{\lambda}{\theta}\left(1 - \sqrt{1 - \frac{2\theta^2 it}{\lambda}}\right)\right].$$

14) If Y has a negative binomial distribution, $Y \sim \text{NB}(\text{r}, \rho)$, then the pmf of Y is

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

for $y = 0, 1, \ldots$ where $0 < \rho < 1$. $E(Y) = r(1-\rho)/\rho$, and

$$V(Y) = \frac{r(1-\rho)}{\rho^2}.$$

The moment generating function

$$m(t) = \left[\frac{\rho}{1 - (1-\rho)e^t}\right]^r$$

for $t < -\log(1-\rho)$.

15) If Y has an F distribution, $Y \sim F(\nu_1, \nu_2)$, then the pdf of Y is

$$f(y) = \frac{\Gamma(\frac{\nu_1+\nu_2}{2})}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} \frac{y^{(\nu_1-2)/2}}{\left(1 + (\frac{\nu_1}{\nu_2})y\right)^{(\nu_1+\nu_2)/2}}$$

where $y > 0$ and ν_1 and ν_2 are positive integers.

$$E(Y) = \frac{\nu_2}{\nu_2 - 2}, \quad \nu_2 > 2$$

and

$$V(Y) = 2\left(\frac{\nu_2}{\nu_2 - 2}\right)^2 \frac{(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 4)}, \quad \nu_2 > 4.$$

16) If Y has a Student's t distribution, $Y \sim t_p$, then the pdf of Y is

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

where p is a positive integer and y is real. This family is symmetric about 0. The t_1 distribution is the Cauchy$(0,1)$ distribution. If Z is $N(0,1)$ and is independent of $W \sim \chi^2_p$, then

$$\frac{Z}{(\frac{W}{p})^{1/2}}$$

is t_p. $E(Y) = 0$ for $p \geq 2$. $V(Y) = p/(p-2)$ for $p \geq 3$.

Two Multivariate Distributions:

17) point mass at $\boldsymbol{c}$: The distribution of the $p \times 1$ random vector $\boldsymbol{Y}$ is a point mass at $\boldsymbol{c}$ (or $\boldsymbol{Y}$ is degenerate at $\boldsymbol{c}$) if $P(\boldsymbol{Y} = \boldsymbol{c}) = 1$ with pmf $f(\boldsymbol{c}) = 1$. Hence $\boldsymbol{Y} \sim N_p(\boldsymbol{c}, \boldsymbol{0})$, $E(\boldsymbol{Y}) = \boldsymbol{c}$, $\mathrm{Cov}(\boldsymbol{Y}) = \boldsymbol{0}$, $m(\boldsymbol{t}) = e^{\boldsymbol{t}^T\boldsymbol{c}}$, $c(\boldsymbol{t}) = e^{i\boldsymbol{t}^T\boldsymbol{c}}$.

18) MVN distribution: If $\boldsymbol{Y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $E(\boldsymbol{Y}) = \boldsymbol{\mu}$ and $\mathrm{Cov}(\boldsymbol{Y}) = \boldsymbol{\Sigma}$.

$$m(\boldsymbol{t}) = \exp\left(\boldsymbol{t}^T\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}\right), \quad c(\boldsymbol{t}) = \exp\left(i\boldsymbol{t}^T\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}\right).$$

If $\boldsymbol{Y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and if $\boldsymbol{A}$ is a $q \times p$ matrix, then $\boldsymbol{A}\boldsymbol{Y} \sim N_q(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$. If $\boldsymbol{a}$ is a $p \times 1$ vector of constants, then $\boldsymbol{Y} + \boldsymbol{a} \sim N_p(\boldsymbol{\mu} + \boldsymbol{a}, \boldsymbol{\Sigma})$.

$$\text{Let } \boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{pmatrix}, \ \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \text{and } \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

All subsets of a MVN are MVN: $(Y_{k_1}, ..., Y_{k_q})^T \sim N_q(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$ where $\tilde{\boldsymbol{\mu}}_i = E(Y_{k_i})$ and $\tilde{\boldsymbol{\Sigma}}_{ij} = \mathrm{Cov}(Y_{k_i}, Y_{k_j})$. In particular, $\boldsymbol{Y}_1 \sim N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $\boldsymbol{Y}_2 \sim N_{p-q}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$. If $\boldsymbol{Y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ are independent iff $\boldsymbol{\Sigma}_{12} = \boldsymbol{0}$.

1.11 Summary

1) See Section 1.10 for some useful distributions.

1.12 Complements

1.13 Problems