Otto-Friedrich-Universität Bamberg
Lehrstuhl für Statistik und Ökonometrie

Formulary for the courses

Statistics 3 - Introduction to Probability Theory
Statistics 4 - Estimation and Inference

Formelsammlung zu den Lehrveranstaltungen[1]

Statistik 3 - Grundlagen der Wahrscheinlichkeitstheorie
Statistik 4 - Schätz- und Testtheorie

Dr. Christian Aßmann
M.Sc. Doris Stingl

[1] Die aufgelisteten Definitionen und Theoreme stellen eine Auswahl aus dem Lehrbuch von Mittelhammer 1996 (Mathematical Statistics for Econometrics and Business, Springer-Verlag New York Inc.) dar und bauen auf einer Formelsammlung von Prof. Dr. Roman Liesenfeld auf.

1 Probability Theory

Definition 1.1 [Sample Space]
A set that contains all possible outcomes of a given experiment.

Definition 1.2 [Event]
An event is a subset of the sample space.

Definition 1.3 [Classical Probability]
Let S be the finite sample space for an experiment having $N(S)$ equally likely outcomes, and let $A \subset S$ be an event containing $N(A)$ elements. Then the probability of the event A, denoted by $P(A)$, is given by $P(A) = \frac{N(A)}{N(S)}$.

Definition 1.4 [Relative Frequency Probability]
Let n be the number of times that an experiment is repeatedly performed under identical conditions. Let A be an event in the sample space S, and define n_A to be the number of times in n repetitions of the experiment that the event A occurs. Then the probability of the event A is given by the limit of the relative frequency $\frac{n_A}{n}$, as $P(A) = \lim_{n \to \infty} \frac{n_A}{n}$.

Definition 1.5 [Subjective Probability]
The subjective probability of an event A is a real number, $P(A)$, in the interval $[0,1]$, chosen to express the degree of personal belief in the likelihood of occurrence or validity of event A, the number 1 being associated with certainty.

Definition 1.6 [Event Space]
The set of all events in the sample space S is called the event space.

Probability Axioms

Axiom 1: For any event $A \subset S$, $P(A) \geq 0$.

Axiom 2: $P(S) = 1$.

Axiom 3: Let I be a finite or countably infinite index set of positive integers, and let $\{A_i : i \in I\}$ be a collection of disjoint events contained in S. Then, $P(U_{i \in I} A_i) = \sum_{i \in I} P(A_i)$.

Definition 1.7 [Probability Space]
A probability space is the 3-tuple $\{S, \Upsilon, P\}$, where S is the sample space of an experiment, Υ is the event space, and P is a probability set function having domain Υ.

Probability Theorems

Theorem 1: Let A be an event in the sample space S. Then $P(A) = 1 - P(\bar{A})$.

Theorem 2: $P(\emptyset) = 0$.

Theorem 3: Let A and B be two events in a sample space such that $A \subset B$. Then $P(A) \leq P(B)$ and $P(B - A) = P(B) - P(A)$.

Theorem 4: Let A and B be two events in a sample space S. Then $P(A) = P(A \cap B) + P(A \cap \bar{B})$.

Theorem 5: Let A and B be two events in a sample space S. Then $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Corollary 1: (Booles's Inequality)[2] $P(A \cup B) \leq P(A) + P(B)$. (This follows directly from Theorem since $P(A \cap B) \geq 0$.)

[2] Named after the English mathematician and logician George Boole.

Theorem 6: Let A be an event in a sample space S. Then $P(A) \in [0,1]$.

Theorem 7: (Bonferroni's Inequality (2-event case))[3] Let A and B be two events in a sample space S. Then $P(A \cap B) \geq 1 - P(\bar{A}) - P(\bar{B})$.

Theorem 8: (Bonferroni's Inequalite (general)) Let $A_1, \ldots, A_n$ be events in a sample space S. Then

$$P\left(\bigcap_{i=1}^{n} A_i\right) \geq 1 - \sum_{i=1}^{n} P(\bar{A}_i).$$

Theorem 1.1 [Classical Probability]
Let S be the finite sample space for an experiment having $N(S)$ equally likely outcomes, and let $A \subset S$ be an event containing $N(A)$ elements. Then the probability of the event A is given by $\frac{N(A)}{N(S)}$.

Definition 1.8 [Rectangles in R^n]
Rectangles in R^n are sets of points in R^n defined as[4]

a. closed rectangle: $\{(x_1, \ldots x_n): \quad a_i \leq x_i \leq b_i, \quad i = 1, \ldots, n\}$,

b. open rectangle: $\{(x_1, \ldots x_n): \quad a_i < x_i < b_i, \quad i = 1, \ldots, n\}$,

c. half-open/half-closed rectangle:

$$\{(x_1, \ldots x_n): \quad a_i < x_i \leq b_i, \quad i = 1, \ldots, n\} \quad \text{or} \quad \{(x_1, \ldots x_n): \quad a_i \leq x_i < b_i, \quad i = 1, \ldots, n\},$$

where the a_i's and b_i's are real numbers, with $-\infty$ or ∞ being admissible for strong inequalities. Clearly, rectangles are intervals when $n = 1$.

Definition 1.9 [Borel Sets in S]
Let $S \subset R^n$. The collection of Borel sets in S consists of all closed, open, and half-open/half-closed rectangles contained in S, as well as any other set that can be defined by applying a countable number of union, intersection, and/or complement operations to these rectangles.[5]

Definition 1.10 [Conditional Probability]
Let A and B be any two events in a sample space S. If $P(B) \neq 0$, then the conditional probability of event A, given event B, is given by $P(A|B) = \frac{P(A \cap B)}{P(B)}$.

Theorem 1.2
Given a probability space $\{S, \Upsilon, P\}$ and an event B for which $P(B) \neq 0$, $P(A|B) = \frac{P(A \cap B)}{P(B)}$ defines a probability set function with domain Υ.

Theorem 1.3 [Multiplication Rule]
Let A and B be any two events in the sample space for which $P(B) \neq 0$. Then $P(A \cap B) = P(A|B)P(B)$.

[3] Named for the Italian mathematician C. E. Bonferroni.

[4] One can also define rectangles that are represented as Cartesian products of any collection of closed, open, *and/or* half-open/half-closed intervals, rather than as Cartesian products of only closed intervals, *or* open intervals, *or* half-open/half-closed intervals as in the definition. These might also be referred to as nonopen/nonclosed rectangles.

[5] The collection of Borel sets in S is an example of what is known in the literature as a sigma-field (σ-field), or a sigma-algebra (σ-algebra). A σ-field is a nonempty set of sets that is closed under countable union, intersection, and complement operations. The use of the word "'closed" here means that if A_i, $i \in I$, all belong to the σ-field, any set formed by applying a countable number of unions, intersections, and/or complement operations to the A_i's is also a set that belongs to the σ-field, where I is any countable index set.

Theorem 1.4 [Extended Multiplication Rule]
Let $A_1, A_2, \ldots, A_n$, $n \geq 2$, be events in the sample space. Then if all of the conditional probabilities exist,

$$P\left(\bigcap_{i=1}^{n} A_i\right) = P(A_1) \prod_{i=2}^{n} P\left(A_i | \bigcap_{j=1}^{i-1} A_j\right).$$

Definition 1.11 [Independence of Events (2-event case)]
Let A and B be two events in a sample space S. Then A and B are independent iff $P(A \cap B) = P(A)P(B)$. If A and B are not independent, A and B are said to be dependent events.

Theorem 1.5
If events A and B are independent, then events A and $\bar{B}$, $\bar{A}$ and B, and $\bar{A}$ and $\bar{B}$ are also independent.

Theorem 1.6 [Independence and Disjointness]

1. $P(A) > 0$, $P(B) > 0$, $A \cap B = \emptyset \Rightarrow A$ and B are dependent.
2. $P(A)$ and/or $P(B) = 0$, $A \cap B = \emptyset \Rightarrow A$ and B are independent.
3. $P(A)$ and/or $P(B) = 0$, $A \cap B \neq \emptyset \Rightarrow A$ and B are dependent.

Definition 1.12 [Independence of Events (n-event case)]
Let $A_1, A_2, \ldots, A_n$ be events in the sample space S. The events $A_1, A_2, \ldots, A_n$ are independent iff

$$P\left(\bigcap_{j \in J} A_j\right) = \prod_{j \in J} P(A_j)$$

for all subsets $J \subset \{1, 2, \ldots, n\}$ for which $N(J) \geq 2$. If the events $A_1, A_2, \ldots, A_n$ are not independent, they are said to be dependent events.

Theorem 1.7 [Theorem of Total Probability]
Let the events B_i, $i \in I$, be a finite or countably infinite partion of the sample space, S, so that $B_j \cap B_k = \emptyset$ for $j \neq k$, and $\cup_{i \in I} B_i = S$. Let $P(B_i) > 0\ \forall i$. Then, $P(A) = \sum_{i \in I} P(A \mid B_i) P(B_i)$.

Corollary 1.1 [Bayes's Rule]
Let the events B_i, $i \in I$, be a finite or countable infinite partition of the sample space, S, so that $B_j \cap B_k = \emptyset$ for $j \neq k$ and $\cup_{i \in I} B_i = S$. Let $P(B_i) > 0\ \forall i \in I$. Then, provided $P(A) \neq 0$,

$$P(B_j \mid A) = \frac{P(A \mid B_j) P(B_j)}{\sum_{i \in I} P(A \mid B_i) P(B_i)}, \quad \forall j \in I.$$

2 Random Variables

Definition 2.1 [Univariate Random Variable]
Let $\{S, \Upsilon, P\}$ be a probability space. If $X : \quad S \mapsto R$ (or simply, X) is a real-valued function having as its domain the elements of S, then $X : \quad S \mapsto R$ (or X) is called a random variable.

Definition 2.2 [Discrete Random Variable]
A random variable is called discrete if its range consists of a countable number of elements.

Definition 2.3 [Discrete Probability Density Function]
The discrete probability density function, f, for a discrete random variable X is defined as $f(x) =$ probability of x, $\forall X \in R(X)$, and $f(x) = 0$, $\forall x \notin R(X)$.

Definition 2.4 [Continuous Random Variables and Continuous Probability Density Functions]
A random variable is called continuous if its range is uncountably infinite, and if there exists a nonnegative-valued function $f(x)$, defined for all $x \in]-\infty, \infty[$, such that for any event $A \subset R(X)$, $P_X(A) = \int_{x \in A} f(x)dx$, and $f(x) = 0 \quad \forall x \notin R(X)$. The function $f(x)$ is called a continuous probability density function.

Definition 2.5 [The Classes of Discrete and Continuous Probability Density Functions (univariate case)]

a. *Class of Discrete Density Functions*
The function $f : \quad R \mapsto R$ is a member of the class of discrete density functions iff (1) the set $C = \{x : \quad f(x) > 0, x \in R\}$ (i. e., the subset of points in R having a positive image under f) is countable; (2) $f(x) = 0$ for $x \in \bar{C}$; and (3) $\sum_{x \in C} f(x) = 1$.

b. *Class of Continuous Density Functions*
The function $f : \quad R \mapsto R$ is a member of the class of continuous density functions iff (1) $f(x) \geq 0$ for $x \in (-\infty, \infty)$, and (2) $\int_{-\infty}^{\infty} f(x)dx = 1$.

Definition 2.6 [Univariate Cumulative Distribution Function]
The cumulative distribution function of a random variable X is defined by $F(b) = P(x \leq b) \quad \forall b \in$ (-$\infty, \infty)$. The functional representation of $F(b)$ in particular cases is as follows:

a. *Discrete X:*

$$F(b) = \sum_{x<b, f(x)>0} f(x) \,, \quad \text{for } b \in (-\infty, \infty);$$

b. *Continuous X:*

$$F(b) = \int_{-\infty}^{b} f(x)dx \,, \quad \text{for } b \in (-\infty, \infty);$$

c. *Mixed discrete-continuous X:*

$$F(b) = \sum_{x<b, f_d(x)>0} f_d(x) + \int_{-\infty}^{b} f_c(x)dx \,, \quad \text{for } b \in (-\infty, \infty).$$

Theorem 2.1 [Discrete PDFs from CDFs]
Let $x_1 < x_2 < x_3 < \ldots$ be the countable collection of outcomes in the range of the discrete random variable X. Then the discrete probability density function for X can be defined as

$$\begin{aligned} f(x_1) &= F(x_1) \\ f(x_i) &= F(x_i) - F(x_{i-1}), \quad i = 2, 3, \ldots \\ f(x) &= 0 \quad \text{for } x \notin R(X). \end{aligned}$$

Theorem 2.2 [Continuous PDFs from CDFs]
Let $f(x)$ and $F(x)$ represent the probability density function and CDF, respectively, for the continuous random variable X. The density function for X can be defined as $f(x) = \frac{d}{dx}F(x)$ wherever $f(x)$ is continuous, and $f(x) = 0$ (or any nonnegative number) elsewhere.

Definition 2.7 [Real-Valued Vector Function]
Let $f_i: \quad A \mapsto R$, $i = 1, \ldots, n$, be a collection of n real-valued functions, each function being defined on the domain A. Then the function $f: \quad A \mapsto R^n$ defined by

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} f_1(x) \\ \vdots \\ f_n(x) \end{bmatrix} = f(x), \text{ for } x \in A,$$

is called an (n-dimensional) real-valued vector function. The real-valued vector functions $f_1, \ldots, f_n$ are called coordinate functions of the vector function f.

Definition 2.8 [Multivariate (n-Variate) Random Variable]
Let $\{S, \Upsilon, P\}$ be a probability space. If $X: \quad S \mapsto R^n$ (or simply X) is a real-valued vector function having as its domain the elements of S, then $X: \quad S \mapsto R^n$ (or X) is called a multivariate (n-variate) random variable.

Definition 2.9 [Discrete Multivariate Random Variables and Discrete Joint Probability Density Functions]
A multivariate random variable is called discrete if its range consists of a countable number of elements. The discrete joint probability density function, f, for a discrete multivariate random variable $X = (X_1, \ldots, X_n)$ is defined as $f(x_1, \ldots, x_n) =$ probability of $(x_i, \ldots, x_n)$ if $(x_1, \ldots, x_n) \in R(X)$, and 0 otherwise.

Definition 2.10 [Continuous Multivariate Random Variables and Continuous Joint Probability Density Functions]
A multivariate random variable is called continuous if its range is uncountably infinite and if there exists a nonnegative-valued function $f(x_1, \ldots, x_n)$, defined for all $(x_1, \ldots, x_n) \in R^n$, such that for any event $A \subset R(X)$,

$$P(A) = \int_{(x_1,\ldots,x_n)\in A} \cdots \int f(x_1, \ldots, x_n) dx_1 \ldots dx_n$$

and $f(x_1, \ldots, x_n) = 0 \quad \forall (x_1, \ldots, x_n) \notin R(X)$. The function $f(x_1, \ldots, x_n)$ is called a continuous joint probability density function.

Definition 2.11 [The Classes of Discrete and Continuous Joint Probability Density Functions]

a. *Class of Discrete Joint Density Functions:* A function $F: \quad R^n \mapsto R$ is a member of the class of discrete joint density functions iff:

   1. the set $C = \{(x_1, \ldots, x_n): \quad f(x_1, \ldots, x_n) > 0, (x_1, \ldots, x_n) \in R^n\}$ is countable,
   2. $f(x_1, \ldots, x_n) = 0$ for $x \in \bar{C}$, and
   3. $\sum \cdots \sum_{(x_1,\ldots,x_n)\in C} f(x_1, \ldots, x_n) = 1$.

b. *Class of Continuous Joint Density Functions:* A function $f: R^n \mapsto R$ ist a member of the class of continuous joint density functions iff:

   1. $f(x_1, \ldots, x_n) \geq 0 \quad \forall \quad (x_1, \ldots, x_n) \in R^n$ and
   2. $\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_n)\, dx_1 \ldots dx_n = 1$.

Definition 2.12 [Joint Cumulative Distribution Function]
The joint cumulative distribution function of an n-dimensional random variable X is defined by $F(b_1,\ldots,b_n) = P(x_i \le b_i, i = 1,\ldots,n) \quad \forall \quad (b_1,\ldots,b_n) \in R^n$. The algebraic representation of $F(b_1,\ldots,b_n)$ in the discrete and continuous cases can be given as follows:

a. *Discrete X:*

$$F(b_1,\ldots,b_n) = \sum_{\substack{x1 \le b_1 \\ f(x_1,\ldots,x_n)>0}} \cdots \sum_{x_n \le b_n} f(x_1,\ldots,x_n) \text{ for } (b_1,\ldots,b_n) \in R^n;$$

b. *Continuous X:*

$$F(b_1,\ldots,b_n) = \int_{-\infty}^{b_n} \cdots \int_{-\infty}^{b_1} f(x_1,\ldots,x_n)\, dx_1 \ldots dx_n \text{ for } (b_1,\ldots,b_n) \in R^n.$$

Theorem 2.3 [Discrete Bivariate PDFs from Joint CDFs]
Let (X,Y) be a discrete bivariate random variable with joint cumulative distribution function $F(x,y)$, ald let $x_1 < x_2 < x_3 < \ldots$ and $y_1 < y_2 < y_3 < \ldots$ represent the possible outcomes of X and Y. Then

$$\begin{aligned} f(x_1,y_1) &= F(x_1,y_1), \\ f(x_1,y_j) &= F(x_1,y_j) - F(x_1,y_{j-1}), \quad j \ge 2, \\ f(x_i,y_1) &= F(x_i,y_1) - F(x_{i-1},y_1), \quad i \ge 2, \\ f(x_i,y_j) &= F(x_i,y_j) - F(x_i,y_{j-1}) - F(x_{i-1},y_j) + F(x_{i-1},y_{j-1}), \quad i \text{ and } j \ge 2. \end{aligned}$$

Theorem 2.4 [Continuous Joint PDFs from Joint CDFs]
Let $F(x_1,\ldots,x_n)$ and $f(x_1,\ldots,x_n)$ represent the joint CDF and PDF for the continuous multivariate random variable $X = (X_1,\ldots,X_n)$. The joint PDF of X can be defined as

$$f(x_1,\ldots,x_n) = \begin{cases} \frac{\partial^n F(x_1,\ldots,x_n)}{\partial x_1 \ldots x_n} \text{ where } f(\cdot) \text{ is continuous} \\ 0 \text{ (or any nonnegative number) elsewhere.} \end{cases}$$

Definition 2.13 [Discrete Marginal Probability Density Functions]
Let $f(x_1,\ldots,x_n)$ be the joint discrete probability density function for the n-dimensional random variable $(X_1,\ldots,X_n)$. Let $J = \{j_1, j_2, \ldots, j_m\}$, $1 \le m < n$, be a set of indices selected from the index set $I = \{1,2,\ldots,n\}$. Then the marginal density function for the m-dimensional discrete random variable $(X_{j_1},\ldots,X_{j_m})$ is given by

$$f_{j_1,\ldots,j_m}(x_{j_1},\ldots,x_{j_m}) = \sum_{(x_i \in R(X_i), i \in I-J)} \cdots \sum f(x_1,\ldots,x_n).$$

Definition 2.14 [Continuous Marginal Probability Density Functions]
Let $f(x_1,\ldots,x_n)$ be the joint continuous probability density function for the n-variate random variable $(X_1,\ldots,X_n)$. Let $J = \{j_1, j_2, \ldots, j_m\}$, $1 \le m < n$, be a set of indices selected from the index set $I = \{1,2,\ldots,n\}$. Then the marginal density function for the m-variate continuous random variable $(X_{j_1},\ldots,X_{j_m})$ is given by

$$f_{j_1 \ldots j_m}(x_{j_1},\ldots,x_{j_m}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1,\ldots,x_n) \prod_{i \in I-J} dx_i.$$

Definition 2.15 [Conditional Probability Density Functions]
Let $f(x_1,\ldots,x_n)$ be the joint density function for the n-dimensional random variable $(X_1,\ldots,X_n)$. Let $J_1 = \{j_1,\ldots,j_m\}$ and $J_2 = \{j_{m+1},\ldots,j_n\}$ be two mutually exclusive index sets whose union is equal to the index set $\{1,2,\ldots,n\}$. Then the conditional density function for the m-dimensional random variable $(X_{j_1},\ldots,X_{j_m})$, given that $(X_{j_{m+1}},\ldots,X_{j_n}) \in D$ and $P_{x_{j_{m+1}} \ldots x_{j_n}}(D) > 0$, is as follows:

Discrete case:

$$f(x_{j_1},\ldots,x_{j_m}|(x_{j_{m+1}},\ldots,x_{j_n})\in D)=\frac{\sum\cdots\sum_{(x_{j_{m+1}},\ldots,x_{j_n})\in D} f(x_1,\ldots,x_n)}{\sum\cdots\sum_{(x_{j_{m+1}},\ldots,x_{j_n})\in D} f_{j_{m+1}\cdots j_n}(x_{j_{m+1}},\ldots,x_{j_n})};$$

Continuous case:

$$f(x_{j_1},\ldots,x_{j_m}|(x_{j_{m+1}},\ldots,x_{j_n})\in D)=\frac{\int\cdots\int_{(x_{j_{m+1}},\ldots,x_{j_n})\in D} f(x_1,\ldots,x_n)dx_{j_{m+1}}\ldots dx_{j_n}}{\int\cdots\int_{(x_{j_{m+1}},\ldots,x_{j_n})\in D} f_{j_{m+1}\cdots j_n}(x_{j_{m+1}},\ldots,x_{j_n})dx_{j_{m+1}}\ldots dx_{j_n}}.$$

If D is equal to the elementary event $(d_{m+1},\ldots,d_n)$, then the definition of the conditional density in both the discrete and continuous cases can be represented as

$$f(x_{j_1},\ldots,x_{j_m}|x_{j_i}=d_i, i=m+1,\ldots,n)=\frac{f(x_1,\ldots,x_n)}{f_{j_{m+1}\cdots j_n}(d_{m+1},\ldots,d_n)},$$

where $x_{j_i}=d_i$ for $j_i\in J_2$, and if the marginal density in the denominator is positive valued.[6]

Definition 2.16 [Independence of Random Variables]
The random variables X_1 and X_2 are said to be independent iff $P(x_1\in A_1, x_2\in A_2)=P(x_1\in A_1)P(x_2\in A_2)$ for all events A_1, A_2.

Definition 2.17 [Independence of Random Variables (n-Variate)]
The random variables $X_1,X_2,\ldots,X_n$ are said to be independent iff $P(x_i\in A_i, i=1,\ldots,n)=\prod_{i=1}^n P(x_i\in A_i)$ for all choices of the events $A_1,\ldots,A_n$.

Theorem 2.5 [Joint Density Factorization for Independence of Random Variables (n)-Variate Case]
The random variables $X_1,X_2,\ldots,X_n$ with joint probability density function $f(x_1,\ldots,x_n)$ and marginal probability density functions $f_i(x_i)$, $i=1,\ldots,n$, are independent iff the joint density can be factored into the product of the marginal densities as

$$f(x_1,\ldots,x_n)=\prod_{i=1}^n f_i(x_i)\quad\forall(x_1,\ldots,x_n)$$

except, possibly, at points of discontinuity for the joint density function as a continuous random variable.

Theorem 2.6 [Independence of Functions of Random Variables, Bivariate]
If X_1 and X_2 are independent random variables, and if the random variables Y_1 and Y_2 are defined by $y_1=Y_1(x_1)$ and $y_2=Y_2(x_2)$, then Y_1 and Y_2 are independent random variables.

Theorem 2.7 [Independence of Functions of Random Variables, n-Variate]
Let $X_1,\ldots,X_n$ be a collection of n independent random vectors, and let the random vectors $Y_1,\ldots,Y_n$ be defined by $y_i=Y_i(x_i)$, $i=1,\ldots,n$. Then the random vectors $Y_1,\ldots,Y_n$ are independent.

[6] In the continuous case, it is also presumed that f and $f_{j_{m+1}\cdots j_n}$ are continuous in $(x_{j_{m+1}\cdots j_n})$ within some neighborhood of points around the point where the conditional density is evaluated in order to justify the conditional density definition via a limiting argument analogous to the bivariate case. Motivation for the conditional density expression when conditioning on an elementary event in the continuous case can then be provided by extending the mean value theorem argument used in the bivariate case. See R. G. Bartle, *Real Analysis*, p. 429, for a statement of the general mean value theorem for integrals.

3 Moments

Definition 3.1 [Expectation of a Random Variable; Discrete Case]
The expected value of a discrete random variable exists, and is defined by $\mathrm{E}X = \sum_{x \in R(X)} x f(x)$, iff $\sum_{x \in R(X)} |x| f(x) < \infty$.

Definition 3.2 [Expected Value of a Random Variable; Continuous Case]
The expected value of a continuous random variable X exists, and is defined by $\mathrm{E}X = \int_{-\infty}^{\infty} x f(x)\, dx$, iff $\int_{-\infty}^{\infty} |x| f(x)\, dx < \infty$.

Theorem 3.1 [Existence of EX for Bounded $R(X)$]
If $|x| < c \quad \forall \quad x \in R(X)$, for some choice of $c \in (0, \infty)$, then EX exists.

Theorem 3.2 [Expectation of a Function of a Univariate Random Variable]
Let X be a random variable having density function $f(x)$. Then the expectation of $Y = g(x)$ is given by[7]

(discrete)

$$\mathrm{E}g(x) = \sum_{x \in R(X)} g(x) f(x),$$

(continuous)

$$\mathrm{E}g(x) = \int_{-\infty}^{\infty} g(x) f(x)\, dx.$$

Lemma 3.1
For any continuous random variable Y, the expectation of Y, if it exists, can be written as

$$\mathrm{E}Y = \int_0^{\infty} P(y > z) \; - \int_0^{\infty} P(y \leq -z)\, dz.$$

Theorem 3.3 [Expectation of an Indicator Function]
Let X be a random variable with density function $f(x)$, and suppose A is an event for X. Then $\mathrm{E}(I_A(X)) = P_X A$.

Theorem 3.4 [Jensen's Inequality]
Let X be a random variable with expectation EX, and let g be a continuous function on an open interval I containing $R(X)$. Then

a. $\mathrm{E}g(X) \geq g(\mathrm{E}X)$ if g is convex on I, and $\mathrm{E}g(X) > g(\mathrm{E}X)$ if g is strictly convex on I and X is not degenerate;[8]

b. $\mathrm{E}g(X) \leq g(\mathrm{E}X)$ if g is concave on I, and $\mathrm{E}g(X) < g(\mathrm{E}X)$ if g is strictly concave on I and X is not degenerate.

Expectations of Functions

Theorem 1: If c is a constant, then $\mathrm{E}(c) = c$.

Theorem 2: If c is a constant, then $\mathrm{E}(cX) = c\mathrm{E}X$.

Theorem 3: $\mathrm{E}\sum_{i=1}^{k} g_i(X) = \sum_{i=1}^{k} \mathrm{E}g_i(X)$.

Corollary 1: Let $Y = a + bX$ for real constants a and b, and let EX exist. Then $\mathrm{E}Y = a + b\mathrm{E}X$.

[7] It is tacitly assumed that the sum and integral are absolutely convergent for the expectation to exist.
[8] A degenerate random variable is a random variable that has one outcome that is assigned a probability of 1.

Theorem 4: Expectation of a Function of a Multivariate Random Variable Let $(X_1,\ldots,X_n)$ be a multivariate random variable with joint density function $f(x_1,\ldots,x_n)$. Then the expectation of $Y = g(X_1,\ldots,X_n)$ is given by[9]

discrete: $\mathrm{E}Y = \sum\ldots\sum_{(x_1,\ldots,x_n)\in R(X)} g(x_1,\ldots,x_n)f(x_1,\ldots,x_n)$,
continuous: $\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} g(x_1,\ldots,x_n)f(x_1,\ldots,x_n) \quad dx_1\ldots dx_n$.

We remind the reader that since $f(x_1,\ldots,x_n) = 0\forall(x_1,\ldots,x_n) \notin R(X)$, one could also sum over the points $(x_1,\ldots,x_n) \in \times_{i=1}^{n} R(X_i)$ to define $\mathrm{E}Y$ in the discrete case.

Theorem 5: $\mathrm{E}\sum_{i=1}^{k} g_i(X_1,\ldots,X_n) = \sum_{i=1}^{k} \mathrm{E}g_i(X_1,\ldots,X_n)$.

Corollary 2: $\mathrm{E}\sum_{i=1}^{k} X_i = \sum_{i=1}^{k} \mathrm{E}X_i$.

Theorem 6: Let $(X_1,\ldots,X_n)$ be independent random variables. Then $\mathrm{E}\prod_{i=1}^{n} X_i = \prod_{i=1}^{n} \mathrm{E}X_i$.

Definition 3.3 [Expectation of a Matrix of Random Variables]
Let $\mathbf{W}$ be an $n \times k$ matrix of random variables whose (i,j)th element is $\mathbf{W}_{ij}$. Then $\mathrm{E}\mathbf{W}$, the expectation of the matrix $\mathbf{W}$, is the matrix of expectations of the elements of $\mathbf{W}$, where the (i,j)th element of $\mathrm{E}\mathbf{W}$ is equal to $\mathrm{E}\mathbf{W}_{ij}$.

Definition 3.4 [Conditional Expectation; Bivariate]
Let X and Y be random variables with joint density function $f(x,y)$. Let the conditional density of Y, given $x \in B$, be $f(y|x \in B)$. Let $g(Y)$ be a real-valued function of Y. Then the conditional expectation of $g(Y)$, given $x \in B$, is defined as

discrete: $\mathrm{E}(g(Y)|x \in B) = \sum_{y\in R(Y)} g(y)f(y|x \in B)$,

continuous: $\mathrm{E}(g(Y)|x \in B) = \int_{-\infty}^{\infty} g(y)f(y|x \in B) \quad dy$.

Theorem 3.5 [Double Expectation Theorem]
$\mathrm{E}(\mathrm{E}(g(Y)|X)) = \mathrm{E}g(Y)$.

Theorem 3.6 [Substitution Theorem]
$\mathrm{E}(g(X,Y)|x = b) = \mathrm{E}(g(g,Y)|x = b)$.

Theorem 3.7 [Generalized Double Expectation Theorem]
$\mathrm{EE}(g(X,Y)|X) = \mathrm{E}(g(X,Y))$.

Definition 3.5 [Conditional Expectation (General)]
Let $(X_1,\ldots,X_n)$ and $(Y_1,\ldots,Y_m)$ be random vectors having a joint density function $f(x_1,\ldots,x_n,y_1,\ldots,y_m)$. Let $g(Y_1,\ldots,Y_m)$ be a real-valued function of $(Y_1,\ldots,Y_m)$. Then the conditional expectation of $g(Y_1,\ldots,Y_m)$, given $(x_1,\ldots,x_n) \in B$, is defined as for discrete random variables as

$$\mathrm{E}(g(Y_1,\ldots,Y_m)|(x_1,\ldots,x_n) \in B) = \sum\ldots\sum_{(y_1,\ldots,y_m)\in R(Y)} g(y_1,\ldots,y_m)f(y_1,\ldots,y_m|(x_1,\ldots,x_n) \in B),$$ [10]

and for continuous random variables as

$$\mathrm{E}(g(Y_1,\ldots,Y_m)|(x_1,\ldots,x_n) \in B) = \int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} g(y_1,\ldots,y_m)f(y_1,\ldots,y_m|(x_1,\ldots,x_n) \in B) \quad dy_1\ldots dy_m.$$

Conditional Expectations of Functions

[9] It is tacitly assumed that the sum and integral are absolutely convergent for the expectation to exist.

[10] One can equivalently sum over the points $(y_1,\ldots,y_m) \in \times_{i=1}^{m} R(Y_i)$ in defining the expectation in the discrete case.

Theorem 1: *Substitution Theorem for Multivariate Random Variables*

$$\mathrm{E}(g(X_1,\ldots,X_n,Y_1,\ldots,Y_m)|x=b) = \mathrm{E}(g(b_1,\ldots,b_n,Y_1,\ldots,Y_m)|x=b.)$$

Theorem 2: *Double Expectation Theorem for Multivariate Random Variables*

$$\mathrm{E}(\mathrm{E}(g(Y_1,\ldots,Y_m)|X_1,\ldots,X_n)) = \mathrm{E}[g(Y_1,\ldots,Y_m)], \quad \text{and}$$

$$\mathrm{E}(\mathrm{E}(g(X_1,\ldots,X_n,Y_1,\ldots,Y_m)|X_1,\ldots,X_n)) = \mathrm{E}[g(X_1,\ldots,X_n,Y_1,\ldots,Y_m)].$$

Theorem 3: $\mathrm{E}(c|(x_1,\ldots,x_n)\in B) = c.$

Theorem 4: $\mathrm{E}(cY|(x_1,\ldots,x_n)\in B) = c\mathrm{E}(Y|(x_1,\ldots,x_n)\in B).$

Theorem 5:

$$\mathrm{E}\left(\sum_{i=1}^{k} g_i(Y_1,\ldots,Y_m)|(x_1,\ldots,x_n)\in B\right) = \sum_{i=1}^{k}\mathrm{E}(g_i(Y_1,\ldots,Y_m)|(x_1,\ldots,x_n)\in B).$$

Definition 3.6 [rth Moment about the Origin]
Let X be a random variable with density function $f(x)$. Then the rth moment of X about the origin, denoted by μ_r', is defined as

discrete: $\mu_r' = \mathrm{E}(X^r) = \sum_{x\in R(X)} x^r f(x)$,

continuous: $\mu_r' = \mathrm{E}(X^r) = \int_{-\infty}^{\infty} x^r f(x) \quad dx.$

Definition 3.7 [Mean of a Random Variable (or Mean of a Density Function)]
The first moment about the origin of a random variable, X, is called the mean of the random variable X (or mean of the density functio of X) and will be denoted by the symbol μ.

Definition 3.8 [rth Central Moment (or rth Moment about the Mean)]
Let X be a random variable with density function $f(x)$. Then the rth central moment of X (or the rth moment of X about the mean), denoted by μ_r, is defined as

discrete: $\mu_r = \mathrm{E}(X-\mu)^r = \sum_{x\in R(X)}(x-\mu)^r f(x)$,

continuous: $\mu_r = \mathrm{E}(X-\mu)^r = \int_{-\infty}^{\infty}(x-\mu)^r f(x) \quad dx.$

Definition 3.9 [Variance of a Random Variable (or Variance of a Density Function)]
The second central moment, $\mathrm{E}(X-\mu)^2$, of a random variable, X, is called the variance of the random variable X (or the variance of the density function of X) and will be denoted by the symbol σ^2, or by $\mathrm{var}(X)$.

Definition 3.10 [Standard Deviation of a Random Variable (or Standard Deviation of a Density Function)]
The nonnegative square root of the variance of a random variable, X, (i. e. $\sqrt{\sigma^2}$), is called the standard deviation of the random variable X (or standard deviation of the density function of X) and will be denoted by the symbol σ, or by $\mathrm{std}(X)$.

Theorem 3.8 [Markov's Inequality]
Let X be a random variable with density function $f(x)$, and let g be a nonnegative-valued function of X. Then $\Pr(g(x)\geq\alpha)\leq \mathrm{E}\frac{g(X)}{\alpha}$ for any value $\alpha>0$.

Corollary 3.1 [Chebyshev' Inequality]

$$\Pr(|x-\mu| \geq k\sigma) \leq \frac{1}{k^2} \quad \text{for} \quad k > 0.$$

Corollary 3.2 [Chebyshev' Inequality]

$$\Pr(|x-\mu| < k\sigma) \geq 1 - \frac{1}{k^2} k > 0.$$

Lemma 3.2 [Binomial Theorem]
Let a and b be real numbers. Then $(a+b)^n = \sum_{j=0}^{n} \left(\begin{array}{c} n \\ j \end{array} \right) a^j b^{n-j}$.

Theorem 3.9 [Moments about the Origin as Functions of Central Moments]
If μ_r' exists and r is a positive integer, then $\mu_r' = \sum_{j=1}^{r} \left(\begin{array}{c} r \\ j \end{array} \right) \mu_{r-j} \mu^j$.

Existence Conditions for Moments

Theorem 1: If $\mathrm{E}X^r$ exists for a given $r > 0$, then $\mathrm{E}X^s$ exists for all $s \in [0, r]$.

Theorem 2: If $\mathrm{E}(Y-\mu)^r$ exists for a given $r > 0$, then $\mathrm{E}(Y-\mu)^s$ exists for all $s \in [0, r]$.

Theorem 3: If $\mathrm{E}X^r$ (or $\mathrm{E}(Y-\mu)^r$) exists for a given integral $r > 0$, then $\mathrm{E}X^s$ (or $\mathrm{E}(Y-\mu)^s$) exists $\forall s \in [0, r]$.

Definition 3.11 [Median of X]
Any number, b, satisfying $\Pr(x \leq b) \geq \frac{1}{2}$ and $\Pr(x \geq b) \geq \frac{1}{2}$ is called a median of X and is denoted by $\mathrm{med}(X)$.

Definition 3.12 [Quantile of X]
Any number b satisfying $\Pr(x \leq b) \geq p$ and $\Pr(x \geq b) \geq 1-p$ for $p \in (0,1)$ is called a quantile of X of order p (or the $(100p)$th percentile of the distribution of X).

Definition 3.13 [Mode of $f(x)$]
Let X be a random variable with density function $f(x)$. Then any point b at which $f(x)$ exhibits a maximum is called a mode of X, or a mode of the distribution of X, and is denoted by $\mathrm{mode}(X)$.

Definition 3.14 [Moment Generating Function (MGF)]
The expected value of e^{tX} is defined to be the moment-generating function of X if the expected value exists for every value of t in some interval $t \in (-h, h)$, $h > 0$. The moment-generating function of X will be denoted by $M_X(t)$. Thus,

discrete: $M_X(T) = \mathrm{E}e^{tX} = \sum_{x \in R(X)} e^{tx} f(x)$,

continuous: $M_X(T) = \mathrm{E}e^{tX} = \int_{-\infty}^{\infty} e^{tx} f(x) \quad dx$.

Theorem 3.10 [Moments from MGF]
Let X be a random variable for which the MGF, $M_X(t)$, exists. Then

$$\mu_r' = \mathrm{E}X^r = \frac{d^r M_X(0)}{dt^r}.$$

Lemma 3.3 [Differentiating under the Integral Sign]
If the function $g(t)$ defined by $g(t) = \sum_{x \in R(X)} e^{tx} f(x)$ or $\int_{-\infty}^{\infty} e^{tx} f(x)\, dx$ converges for $t \in (-h,h)$, $h > 0$, then $\frac{d^r g(t)}{dt^r}$ exists $\forall t \in (-h,h)$ and for all positive integers r, and the derivate can be found by differentiating under the summation sign or differentiating under the integral sign, respectively, as

$$\frac{d^r g(t)}{dt^r} = \sum_{x \in R(X)} \frac{d^r e^{tx}}{dt^r} f(x) \text{ or } \int_{-\infty}^{\infty} \frac{d^r e^{tx}}{dt^r} f(x)\, dx$$

(see D. V. Widder (1961), *Advanced Calculus*, 2nd ed., Englewood Cliffs, N. J.: Prentice-Hall, pp. 442-447).

Theorem 3.11 [Properties of MGFs]
Let $(X_1, \ldots, X_n)$ be independent random variables having respective MGFs $M_{X\,i}(t)$, $i = 1, \ldots, n$.

a. If $Y_i = aX_i + b$, then $M_{Y\,i}(t) = e^{bt} M_{X\,i}(at)$.

b. If $Y = \sum_{i=1}^{n} X_i$, then $M_Y(t) = \prod_{i=1}^{n} M_{X\,i}(t)$.

c. If $Y = \sum_{i=1}^{n} a_i X_i + b$, then $M_Y(t) = e^{bt} \prod_{i=1}^{n} M_{X\,i}(a_i t)$.

Theorem 3.12 [MGF Uniqueness Theorem]
If a moment-generating function exists for a random variable X having density function $f(x)$, then the moment generating function is unique. Conversely, the moment generating function determines the density function of X uniquely, at least up to a set of points having probability zero.

Definition 3.15 [Cumulant-Generating Function and Cumulants]
The cumulant-generating function of X is defined as $\Psi(t) = \ln(M_X(t))$. The rth cumulant of X is given by $\kappa_r = \frac{d^r \Psi(0)}{dt^r}$. The first four cumulants are related to moments as follows: $\kappa_1 = \mu_1'$; $\kappa_2 = \sigma^2$; $\kappa_3 = \mu_3$; and $\kappa_4 = \mu_4 - 3\sigma^4$.

Definition 3.16 [MGF and Cumulant Generating Function; Multivariate]
The expected value of $\exp\left(\sum_{j=1}^{n} t_j X_j\right)$ is defined to be the MGF of the n-variate random variable $X = (X_1, \ldots, X_n)$ if the expected value exists for all $t_i \in (-h,h)$ for some $h > 0$, $i = 1, \ldots, n$. The MGF will be denoted by $M_X(t)$, where $t = (t_1, \ldots, t_n)$. Thus,

discrete: $M_X(t) = \sum \cdots \sum_{(x_1,\ldots,x_n) \in R(X)} \exp\left(\sum_{j=1}^{n} t_j x_j\right) f(x_1, \ldots, x_n)$,

continuous: $M_X(t) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left(\sum_{j=1}^{n} t_j x_j\right) f(x_1, \ldots, x_n) dx_1 \ldots dx_n$.

The cumulant generating function of X is defined as $\Psi_X(t) = \ln M_X(t)$.

Theorem 3.13 [Marginal MGFs from Multivariate MGFs]
Let $(X_1, \ldots, X_n)$ have MGF $M_X(t)$, and let $X_{(m)} = (X_j, j \in J)$ be any m-element subset of random variables in X, where $J \subset \{1, 2, \ldots, n\}$, $N(J) = m < n$. Define $t_{(m)} = (t_j, j \in J)$. Then the MGF of $X_{(m)}$, referred to as the marginal MGF of $X_{(m)}$, can be represented as $M_{X\,(m)}(t_{(m)}) = M_X(t^*)$, where the elements in t^* are defined by $t_j^* = t_j I_J(j)$.

Definition 3.17 [Joint Moments about the Mean (or Central Joint Moments)]
Let X and Y be two random variables having joint density function $f(x,y)$. Then the (r,s)th joint moment of (X,Y) (or of $f(x,y)$) about the mean is defined by

discrete: $\mu_{r,s} = \sum_{x \in R(X)} \sum_{y \in R(Y)} (x - \mathrm{E}X)^r (y - \mathrm{E}y)^s f(x,y)$, *continuous:* $\mu_{r,s} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mathrm{E}X)^r(y-\mathrm{E}y)^s f(x,y)\,dxdy$.

Definition 3.18 [Covariance]
The central joint moment $\mu_{1,1} = \mathrm{E}(X-\mathrm{E}X)(Y-\mathrm{E}Y)$ is called the covariance between X and Y and is denoted by the symbol σ_{XY}, or by $\mathrm{cov}(X,Y)$.

Theorem 3.14 [Cauchy-Schwarz Inequality]

$$(\mathrm{E}WZ)^2 \leq \mathrm{E}W^2 EZ^2.$$

Theorem 3.15 [Covariance Bound]

$$|\sigma_{XY}| \leq \sigma_X\sigma_Y.$$

Definition 3.19 [Correlation]
The correlation between two random variables X and Y is defined by

$$\mathrm{corr}(X,Y) = \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X\sigma_y}.$$

Theorem 3.16 [Correlation Bound]

$$-1 \leq \rho_{XY} \leq 1.$$

Theorem 3.17 [Independence and Correlation]
If X and Y are independent, then $\sigma_{XY} = 0$ (assuming the covariance exists).

Theorem 3.18 [Degenerate Random Variable]
Let Z be a random variable for which $\sigma_Z^2 = 0$. Then $P(z = \mathrm{E}Z) = 1$.

Theorem 3.19 [Correlation Bounds and Linearity]
If $\rho_{XY} = +1$ or -1, then $P(y = a_1 + bx) = 1$ or $P(y = a_2 - bx) = 1$, respectively, where $a_1 = \mathrm{E}Y - \frac{\sigma_Y}{\sigma_X}\mathrm{E}X$, $a_2 = \mathrm{E}Y + \frac{\sigma_Y}{\sigma_X}\mathrm{E}X$, and $b = \frac{\sigma_Y}{\sigma_X}$.

Theorem 3.20 [Best Linear Prediction of Y Outcomes]
Let (X,Y) have moments of at least second order, and let $\hat{Y} = a + bX$. Then the choices of a and b that minimize $\mathrm{E}d^2(Y,\hat{Y}) = \mathrm{E}(Y-(a+bX))^2$ are given by $a = \mathrm{E}Y - \frac{\sigma_{XY}}{\sigma_X^2}\mathrm{E}X$ and $b = \frac{\sigma_{XY}}{\sigma_X^2}$.

Definition 3.20 [Covariance Matrix]
The covariance matrix of an n-variate random variable $\mathbf{X} = [X_1, \ldots, X_n]'$ is the $n \times n$ symmetric matrix $\mathbf{Cov}(\mathbf{X}) = \mathrm{E}(\mathbf{X} - \mathrm{E}\mathbf{X})(\mathbf{X} - \mathrm{E}\mathbf{X})'$.

Moments of Linear Combinations

Theorem 1: Let $Y = \sum_{i=1}^n a_i X_i$ where the a_i's are real constants. Then $\mathrm{E}Y = \sum_{i=1}^n a_i \mathrm{E}X_i$.

Theorem 2: Let $Y = \sum_{i=1}^n a_i X_i$ where the a_i's are real constants. Then

$$\sigma_Y^2 = \sum_{i=1}^n a_i^2 \sigma_{X\,i}^2 + 2\sum\sum_{i<j} a_i a_j \sigma_{X\,i}\sigma_{X\,j}.$$

Theorem 3 Let $\mathbf{Y} = \mathbf{AX}$, where $\mathbf{A}$ is a $k \times n$ matrix of real constants, and $\mathbf{X}$ is an $n \times 1$ vector of random variables. Then $\mathrm{E}Y = \mathrm{E}\mathbf{AX} = \mathbf{A}\mathrm{E}\mathbf{X}$.

Corollary 1: Let $\mathbf{Y} = \mathbf{AX}$, where $\mathbf{A}$ is a $k \times n$ matrix of real constants and $\mathbf{X}$ is an $n \times l$ matrix of random variables. Then $\mathrm{E}Y = \mathbf{A}\mathrm{E}\mathbf{X}$.

Corollary 2: Let $\mathbf{Y} = \mathbf{XB}$, where $\mathbf{X}$ is a $n \times l$ matrix of random variables and $\mathbf{B}$ is an $l \times m$ matrix of real constants. Then $\mathrm{E}Y = (\mathrm{E}\mathbf{X})\mathbf{B}$.

Corollary 3: Let $\mathbf{A}$ be a $k \times n$ matrix of real constants, and let $\mathbf{X}$ be an $n \times l$ matrix of random variables, and let $\mathbf{B}$ be an $l \times m$ matrix of real constants. Then $\mathrm{E}\mathbf{AXB} = \mathbf{A}(\mathrm{E}\mathbf{X})\mathbf{B}$.

Theorem 4: Let $\mathbf{Y} = \mathbf{AX}$, where $\mathbf{A}$ is a $k \times n$ matrix of real constants and $\mathbf{X}$ is an $n \times 1$ vector of random variables. Then $\mathbf{Cov}(\mathbf{Y}) = \mathbf{Cov}(\mathbf{AX}) = \mathbf{A}\mathbf{Cov}(\mathbf{X})\mathbf{A}'$.

4 Parametric Functions of Densities

4.1 Parametric Families of Discrete Density Functions

Uniform

Parameterization	$N \in \Omega = \{N : N \text{ is a positive integer}\}$
Density Definition	$f(x;N) = \frac{1}{N} I_{\{1,2,\ldots,N\}}(x)$
Moments	$\mu = \frac{(N+1)}{2}$, $\sigma^2 = \frac{(N^2-1)}{12}$, $\mu_3 = 0$
MGF	$M_X(t) = \frac{\sum_{j=1}^{N} e^{jt}}{N}$

Bernoulli

Parameterization	$p \in \Omega = \{p : 0 \leq p \leq 1\}$
Density Definition	$f(x;p) = p^x (1-p)^{1-x} I_{\{0,1\}}(x)$
Moments	$\mu = p$, $\sigma^2 = p(1-p)$, $\mu_3 = 2p^3 - 3p^2 + p$
MGF	$M_X(t) = pe^t + (1-p)$

Binomial

Parameterization	$(n,p) \in \Omega = \{(n,p) : n \text{ is a positive integer, } 0 \leq p \leq 1\}$
Density Definition	$f(x;n,p) = \begin{cases} \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}, \text{ for } x = 0,1,2,\ldots,n \\ 0 \text{ otherwise} \end{cases}$
Moments	$\mu = np$, $\sigma^2 = np(1-p)$, $\mu_3 = np(1-p)(1-2p)$
MGF	$M_X(t) = (1 - p + pe^t)^n$

Multinomial

Parameterization	$(n,p_1,\ldots,p_m) \in \Omega = \{(n,p_1,\ldots,p_m) : n \text{ is a positive integer,}$ $0 \leq p_i \leq 1, \forall i, \sum_{i=1}^{m} p_i = 1\}$
Density Definition	$f(x_1,\ldots,x_m;n,p_1,\ldots,p_m) =$ $= \begin{cases} \frac{n!}{\prod_{i=1}^{m} x_i!} \prod_{i=1}^{m} p_i^{x_i} \text{ for } x_i = 0,1,2,\ldots,n \forall i, \sum_{i=1}^{m} x_i = n \\ 0 \text{ otherwise} \end{cases}$
Moments	$\mu_i = np_i$, $\sigma_i^2 = np_i(1-p_i)$, $\mu_{3,i} = np_i(1-p_i)(1-2p_i)$, $Cov(X_i,X_j) = -np_i p_j$
MGF	$M_X(t) = (\sum_{i=1}^{m} p_i e^{t_i})^n$

Negative Binomial and Geometric

Parameterization	(for the geometric density family, $r = 1$) $(r,p) \in \Omega = \{(r,p); r \text{ is a positive integer, } 0 \leq p \leq 1\}$
Density Definition	$f(x;r,p) = \begin{cases} \frac{(x-1)!}{(r-1)!(x-r)!} p^r (1-p)^{x-r} \text{ for } x = r, r+1, r+2, \ldots \\ 0 \text{ otherwise} \end{cases}$
Moments	$\mu = \frac{r}{p}$, $\sigma^2 = \frac{(r(1-p))}{p^2}$, $\mu_3 = \frac{(r((1-p)+(1-p)^2))}{p^3}$
MGF	$M_X(t) = e^{rt} p^r (1 - (1-p)e^t)^{-r}$ for $t < -\ln(1-p)$

Poisson

Parameterization	$\lambda \in \Omega = \{\lambda : \lambda > 0\}$
Density Definition	$f(x;\lambda) = \begin{cases} \frac{e^{-\lambda} \lambda^x}{x!} \text{ for } x = 0,1,2,\ldots \\ 0 \text{ otherwise} \end{cases}$
Moments	$\mu = \lambda$, $\sigma^2 = \lambda$, $\mu_3 = \lambda$
MGF	$M_X(t) = e^{\lambda(e^t - 1)}$

Hypergeometric

Parameterization	$(M,K,n) \in \Omega = \{(M,K,n) : M = 1,2,3,\ldots; K = 0,1,\ldots,M; n = 1,2,\ldots,M\}$
Density Definition	$f(x;M,K,n) = \begin{cases} \dfrac{\begin{pmatrix} K \\ x \end{pmatrix}\begin{pmatrix} M-K \\ n-x \end{pmatrix}}{\begin{pmatrix} M \\ n \end{pmatrix}} \text{ for integer values} \\ \quad \max[0, n-(M-K)] \le x \le \min(n,K) \\ 0 \text{ otherwise} \end{cases}$
Moments	$\mu = \frac{(nk)}{M}$, $\sigma^2 = n\left(\frac{K}{M}\right)\left(\frac{M-K}{M}\right)\left(\frac{M-n}{M-1}\right)$, $\mu_3 = n\left(\frac{K}{M}\right)\left(\frac{MK}{M}\right)\left(\frac{M-2K}{M}\right)\left(\frac{M-n}{M-1}\right)\left(\frac{M-2n}{M-2}\right)$
MGF	$M_X(t) = [\frac{((M-n)!(M-K)!)}{M!}]H(-n,-K,M-K-n+1,e^t)$, where $H(\cdot)$ is the hypergeometric function $H(\alpha,\beta,r,Z) = 1 + \frac{\alpha\beta}{r}\frac{Z}{1!} + \frac{\alpha\beta(\alpha+1)(\beta+1)}{r(r+1)}\frac{Z^2}{2!} + \ldots$

Multivariate Hypergeometric

Parameterization	$(M,K_1,\ldots,K_m,n) \in \Omega\{(M,K_1,\ldots,K_m,n) : M = 1,2,\ldots; K_i = 0,1,\ldots,M \text{ for } i = 1,\ldots,m; \sum_{i=1}^m K_i = M; n = 1,2,\ldots,M\}$
Density Definition	$f(x_1,\ldots,x_m;M,K_1,\ldots,K_m,n) = \begin{cases} \dfrac{\prod_{j=1}^{m}\begin{pmatrix} K_j \\ x_j \end{pmatrix}}{\begin{pmatrix} M \\ n \end{pmatrix}} \text{ for } x_i \in \{0,1,2,\ldots,n\} \forall i, \sum_{i=1}^m x_i = n \\ 0 \text{ otherwise} \end{cases}$
Moments	$\mu_i = \frac{nK_i}{M}$, $\sigma_i^2 = n\left(\frac{K_i}{M}\right)\left(\frac{M-K_i}{M}\right)\left(\frac{M-n}{M-1}\right)$, $\mu_{3,i} = n\left(\frac{K_i}{M}\right)\left(\frac{MK_i}{M}\right)\left(\frac{M-2K_i}{M}\right)\left(\frac{M-n}{M-1}\right)\left(\frac{M-2n}{M-2}\right)$
MGF	not useful

4.2 Parametric Families of Continuous Density Functions

Uniform

Parameterization	$(a,b) \in \Omega\{(a,b) : -\infty < a < b < \infty\}$
Density Definition	$f(x;a,b) = \frac{1}{(b-a)}I_{[a,b]}(x)$
Moments	$\mu = \frac{a+b}{2}$, $\sigma^2 = \frac{(b-a)^2}{12}$, $\mu_3 = 0$
MGF	$M_X(t) = \begin{cases} \frac{e^{bt}-e^{at}}{(b-a)t} \text{ for } t \neq 0 \\ 1 \text{ for } t = 0 \end{cases}$

Gamma

Parameterization	$(\alpha,\beta)\in\Omega=\{(\alpha,\beta):\alpha>0,\beta>0\}$
Density Definition	$f(x;\alpha,\beta)=\frac{1}{\beta^\alpha\Gamma(\alpha)}x^{\alpha-1}e^{-\frac{x}{\beta}}I_{(0,\infty)}(x)$, where $\Gamma(\alpha)=\int_0^\infty y^{\alpha-1}e^{-y}dy$ is called the gamma function, having the property that if α is a positive integer, $\Gamma(\alpha)$ has values $\Gamma(\alpha)=(\alpha-1)!$, and if $\alpha=\frac{1}{2}$, then $\Gamma(\frac{1}{2})=\pi^{\frac{1}{2}}$. Also, for any real $\alpha>0$, $\Gamma(\alpha+1)=\alpha\Gamma(\alpha)$.
Moments	$\mu=\alpha\beta$, $\sigma^2=\alpha\beta^2$, $\mu_3=2\alpha\beta^3$
MGF	$M_X(t)=(1-\beta t)^{-\alpha}$ for $t<\beta^{-1}$

Exponential

Parameterization	$\theta\in\Omega=\{\theta:\theta>0\}$
Density Definition	The gamma density, with $\alpha=1$ and $\beta=\theta$ $f(x;\theta)=\frac{1}{\theta}e^{-\frac{x}{\theta}}I_{(0,\infty)}(x)$
Moments	$\mu=\theta$, $\sigma^2=\theta^2$, $\mu_3=2\theta^3$
MGF	$M_X(t)=(1-\theta t)^{-1}$ for $t<\theta^{-1}$

Chi-Square

Parameterization	$\nu\in\Omega=\{\nu:\nu \text{is a positive integer}\}$
Density Definition	The gamma density, with $\alpha=\frac{\nu}{2}$ and $\beta=2$. $f(x;\nu)=\frac{1}{2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})}x^{\frac{\nu}{2}-1}e^{-\frac{x}{2}}I_{(0,\infty)}(x)$
Moments	$\mu=\nu$, $\sigma^2=2\nu$, $\mu_3=8\nu$
MGF	$M_X(t)=(1-2t)^{-\frac{\nu}{2}}$ for $t<\frac{1}{2}$

Beta

Parameterization	$(\alpha,\beta)\in\Omega=\{(\alpha,\beta):\alpha>0,\beta>0\}$
Density Definition	$f(x;\alpha,\beta)=\frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}I_{(0,1)}(x)$ where $B(\alpha,\beta)=\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$ is called the beta function. Some useful properties of the beta function include the fact that $B(\alpha,\beta)=B(\beta,\alpha)$ and $B(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ so that the beta function can be evaluated in terms of the gamma function.
Moments	$\mu=\frac{\alpha}{\alpha+\beta}$, $\sigma^2=\frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$, $\mu_3=\frac{2(\beta-\alpha)(\alpha\beta)}{(\alpha+\beta+2)(\alpha+\beta+1)(\alpha+\beta)^3}$
MGF	$M_X(t)=\sum_{t=1}^\infty\left[\frac{B(r+\alpha,\beta)}{B(\alpha,\beta)}\right]\frac{t^r}{r!}$

Univariate Normal

Parameterization	$(a,b)\in\Omega=\{(a,b):a\in(-\infty,\infty),b>0\}$
Density Definition	$f(x;a,b)=\frac{1}{\sqrt{2\pi}b}\exp\{(-\frac{1}{2})(\frac{x-a}{b})^2\}$
Moments	$\mu=a$, $\sigma^2=b^2$, $\mu_3=0$
MGF	$M_X(t)=\exp\{at+\frac{1}{2}b^2t^2\}$

Multivariate Normal

Parameterization	$\mathbf{a} = (a_1, \ldots, a_n)'$ and $\mathbf{B} = \begin{pmatrix} b_{11} & \ldots & b_{1n} \\ \vdots & \ldots & \vdots \\ b_{n1} & \ldots & b_{nn} \end{pmatrix}$
	$(\mathbf{a}, \mathbf{B}) \in \Omega = \{(\mathbf{a}, \mathbf{B}) : \mathbf{a} \in R^n, \mathbf{B}$ is a symmetric, $(n \times n)$, positive semidefinite matrix$\}$.
Density Definition	$f(\mathbf{x}; \mathbf{a}, \mathbf{B}) = \frac{1}{(2\pi)^{\frac{n}{2}} \lvert\mathbf{B}\rvert^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\mathbf{x}-\mathbf{a})'\mathbf{B}^{-1}(\mathbf{x}-\mathbf{a})\}$
Moments	$\boldsymbol{\mu}_{(n\times 1)} = \mathbf{a}$, $\mathrm{Cov}(\mathbf{X})_{(n\times n)} = \mathbf{B}$, $\mu_{3(n\times 1)} = [\mathbf{0}]$
MGF	$M_X(t) = \exp\{\mathbf{a}'\mathbf{t} + \frac{1}{2}\mathbf{t}'\mathbf{B}\mathbf{t}\}$, where $\mathbf{t} = (t_1, \ldots, t_n)'$

Univariate t-distribution

Parameterization	$\nu \in \Omega = \{\nu : \nu$ is a positive integer$\}$
Density Definition	$f(x; \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}}(1 + \frac{x^2}{\nu})^{-(\frac{\nu+1}{2})}$
Moments	$\mu = 0$ for $\nu > 1$, $\sigma^2 = \frac{\nu}{\nu-2}$ for $\nu > 2$, $\mu_3 = 0$ for $\nu > 3$
MGF	Does not exist

Univariate F-distribution

Parameterization	$(\nu_1, \nu_2) \in \Omega = \{(\nu_1, \nu_2) : \nu_1$ and ν_2 are positive integers$\}$
Density Definition	$f(x; \nu_1, \nu_2) = \frac{\Gamma(\frac{\nu_1+\nu_2}{2})}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_1}{2})}(\frac{\nu_1}{\nu_2})^{\frac{\nu_1}{2}} x(\frac{\nu_1}{2}) - 1(1 + \frac{\nu_1}{\nu_2}x)^{-\frac{1}{2}(\nu_1+\nu_2)} I_{(0,\infty)}(x)$
Moments	$\mu = \frac{\nu_2}{\nu_2-2}$ for $\nu_2 > 2$, $\sigma^2 = \frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)}$ for $\nu_2 > 4$,
	$\mu_3 = (\frac{\nu_2}{\nu_1})\frac{8\nu_1(\nu_1+\nu_2-2)(2\nu_1+\nu_2-2)}{(\nu_2-2)^3(\nu_2-4)(\nu_2-6)} > 0$ for $\nu_2 > 6$
MGF	Does not exist

Definition 4.1 [Poisson Process]
Let an experiment consist of observing the number of type A outcomes that occur over a fixed interval of time, say $[0,t]$. The experiment is said to follow the Poisson process if:

1. the probability that precisely one type A outcome will occur in a small time interval of length Δt is approximately proportional to the length of the interval, as $\gamma[\Delta t] + \mathrm{o}(\Delta t)$, where $\gamma > 0$ is the proportionality factor,[11]

2. the probability of two or more type A outcomes occuring in a small time interval of length Δt is negligible relative to the probability that one type A outcome occurs, the negligible probability being of order of magnitude $\mathrm{o}(\Delta t)$,

3. the numbers of type A outcomes that occur in nonoverlapping time intervals are independent events.

Theorem 4.1 [Poisson Process $\Rightarrow$ Poisson Density]
Let X represent the number of times event A occurs in an interval of time $[0,t]$. If the experiment underlying X follows the Poisson process, then the density of X is the Poisson density.

[11] $\mathrm{o}(t)$ is a generic notation applied to any function of Δt whose values approach zero faster than Δt, so that $\lim_{\Delta t \to 0} \frac{\mathrm{o}(\Delta t)}{\Delta t} = 0$. The " $\mathrm{o}(\Delta t)$"' stands for "of smaller order of magnitude than Δt". For example, $h(\Delta t) = (\Delta t)^2$ is a function to which we could affix the label $\mathrm{o}(\Delta t)$, while $h(\Delta t) = (\Delta t)^{\frac{1}{2}}$ is not.

Theorem 4.2 [Gamma Additivity]
Let $X_1, \ldots, X_n$ be independent random variables with respective gamma densities Gamma(α_i, β), $i = 1 \ldots, n$. Then $Y = \sum_{i=1}^n X_i$ has the gamma density Gamma$(\sum_{i=1}^n \alpha_i, \beta)$.

Theorem 4.3 [Scaling of Gamma Random Variables]
Let X have a gamma density Gamma (α, β), and let $c > 0$. Then $Y = cX$ has a gamma density Gamma $(\alpha, \beta c)$.

Theorem 4.4 [Gamma Inverse Additivity]
Let $Y = X_1 + X_2$, where Y has the gamma density Gamma (α, β), X_1 has the gamma density Gamma α_1, β, $\alpha > \alpha_1$, and X_1 and X_2 are independent. Then X_2 has the gamma density Gamma $(\alpha - \alpha_1, \beta)$.

Theorem 4.5 [Memoryless Property of Exponential Density]
If X has an exponential density, then $\Pr(x > s + t | x > s) = \Pr(x > t) \quad \text{for all} \quad t$ and $s > 0$.

Corollary 4.1 [Chi-Square Additivity]
Let $X_1, \ldots, X_k$ be independent random variables having χ^2-square densities with $\nu_1, \ldots, \nu_k$ degrees of freedom, respectively. Then $Y = \sum_{i=1}^k X_i$ has a χ^2-square density with degrees of freedom $\nu = \sum_{i=1}^k \nu_i$.

Theorem 4.6 [Standardized Normal]
Let X have the density $\mathscr{N}(x; \mu, \sigma^2)$. Then $Z = \frac{X - \mu}{\sigma}$ has the density $\mathscr{N}(z; 0, 1)$.

Theorem 4.7 [Squared Standard Normal]
If X has the density $\mathscr{N}(0, 1)$, then $Y = X^2$ has a χ^2 density with 1 degree of freedom.

Theorem 4.8 [Sums of Squares of Independent Standard Normal Random Variables]
Let $(X_1, \ldots, X_n)$ be independent random variables, each having the density $\mathscr{N}(0, 1)$. Then $Y = \sum_{i=1}^n X_i^2$ has a χ^2 density with n degrees of freedom.

Theorem 4.9 [PDF of Linear Combinations of Normal Random Variables]
Let X be an n-variate random variable having the density function $\mathscr{N}(x; \mu, \Sigma)$. Let A be any $(k \times n)$ matrix of real constants with rank k, and let b be any $(k \times 1)$ vector of real constants. Then the $(k \times 1)$ random vector $Y = AX + b$ has the density $\mathscr{N}(y; A\mu + b, A\Sigma A')$.

Theorem 4.10 [Marginal Densities for $\mathscr{N}(\mu, \Sigma)$]
Let Z have the density $\mathscr{N}(z; \mu, \Sigma)$, where

$$Z = \left[\begin{array}{c} Z_{(1)} \\ (m \times 1) \\ \hline Z_{(2)} \\ (n-m) \times 1 \end{array}\right], \mu = \left[\begin{array}{c} \mu_{(1)} \\ (m \times 1) \\ \hline \mu_{(2)} \\ (n-m) \times 1 \end{array}\right], \text{ and } \Sigma = \left[\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ (m \times m) & (m \times (n-m)) \\ \hline \Sigma_{21} & \Sigma_{22} \\ ((n-m) \times m) & ((n-m) \times (n-m)) \end{array}\right].$$

Then the marginal PDF of $Z_{(1)}$ ist $\mathscr{N}(\mu_1, \Sigma_{11})$, and the marginal PDF of $Z_{(2)}$ is $\mathscr{N}(\mu_2, \Sigma_{22})$.

Theorem 4.11 [Conditional Densities for $\mathscr{N}(\mu, \Sigma)$]
Let Z be as defined in Theorem (**Marginal Densities for $\mathscr{N}(\mu, \Sigma)$**), and

$$z^0_{(n \times 1)} = \left[\begin{array}{c} z^0_{(1)} \\ (m \times 1) \\ \hline z^0_{(2)} \\ (n-m) \times 1 \end{array}\right]$$

be a vector of constants. Then

$$\begin{aligned} f(z_{(1)} | z_{(2)} = z^0_{(2)}) &= \mathscr{N}(\mu_{(1)} + \Sigma_{12}\Sigma_{22}^{-1}(z^0_{(2)} - \mu_{(2)}), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}), \\ f(z_{(2)} | z_{(1)} = z^0_{(1)}) &= \mathscr{N}(\mu_{(2)} + \Sigma_{21}\Sigma_{11}^{-1}(z^0_{(1)} - \mu_{(1)}), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}). \end{aligned}$$

Lemma 4.1 [Partitioned Inversion and Partitioned Determinants]
Partition the $(n \times n)$ matrix Σ as

$$\Sigma = \left[\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ (m \times m) & (m \times (n-m)) \\ \hline \Sigma_{21} & \Sigma_{22} \\ ((n-m) \times m) & ((n-m) \times (n-m)) \end{array}\right].$$

a. If Σ_{11} is nonsingular, then $|\Sigma| = |\Sigma_{11}| \cdot |\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}|$.

b. If Σ_{22} is nonsingular, then $|\Sigma| = |\Sigma_{22}| \cdot |\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}|$.

c. If $|\Sigma| \neq 0$, $|\Sigma_{11}| \neq 0$, and $|\Sigma_{22}| \neq 0$, then

$$\Sigma^{-1} = \left[\begin{array}{c|c} (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} & -(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1} \\ \hline -\Sigma_{22}^{-1}\Sigma_{21}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} & (\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1} \end{array}\right].$$

d. The diagonal blocks in the partitioned matrix of part (c) can also be expressed as

$$\begin{aligned} (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} &= \Sigma_{11}^{-1} + \Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1}\Sigma_{21}\Sigma_{11}^{-1} \\ (\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1} &= \Sigma_{11}^{-1} + \Sigma_{22}^{-1}\Sigma_{21}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1} \end{aligned}$$

(see F. A. Graybill (1983). Matrices with Applications in Statistics, 2nd ed., Belmon, CA: Wadsworth, pp. 183-186, for further discussion and proofs).

Theorem 4.12 [$Cov(X) = 0 \Rightarrow$ Independence when X Has PDF $\mathcal{N}(\mu, \Sigma)$]
Let $X = (X_1, \ldots, X_n)'$ have the density $\mathcal{N}(\mu, \Sigma)$. Then $(X_1, \ldots, X_n)$ are independent iff Σ is a diagonal matrix.

Theorem 4.13 [$Z_{(1)}$ and $Z_{(2)}$ Independent $\Leftrightarrow \Sigma_{12} = [0]$ for $\mathcal{N}(\mu, \Sigma)$]
Let

$$Z_{(n \times 1)} = \left[\begin{array}{c} Z_{(1)} \\ (m \times 1) \\ \hline Z_{(2)} \\ (n-m) \times 1 \end{array}\right]$$

have the multivariate normal density identified in Theorem (**Marginal Densities for $\mathcal{N}(\mu, \Sigma)$**). Then the vectors $Z_{(1)}$ and $Z_{(2)}$ are independent iff $\Sigma_{12} = \Sigma_{21}' = [0]$.

Definition 4.2 [Exponential Class of Densities]
The density function $f(x; \theta)$ is a member of the exponential class of density functions iff

$$f(x; \theta) = \begin{cases} \exp\left(\sum_{i=1}^{k} c_i(\theta) g_i(x) + d(\theta) + z(x)\right) & \text{for } x \in A \\ 0 & \text{otherwise} \end{cases}$$

where $x = (x_1, \ldots, x_n)'$, $\theta = (\theta_1, \ldots, \theta_k)'$; $c_i(\theta)$, $i = 1, \ldots, k$, and $d(\theta)$ are real-valued functions of θ that do not depend on x; $g_i(x)$, $i = 1, \ldots, k$, and $z(x)$ are real-valued functions of x that do not depend on θ; and $A \in R^n$ is a set of n-tuples contained in n-dimensional real space whose definition does not depend on the parameter vector θ.

5 Basic Asymptotics

Definition 5.1 [Sequence]
Let A be any set. A sequence in A is a function having the natural numbers, N, for its domain, and its range contained in A, i. e., $f : N \mapsto A$, is a sequence in A.

Definition 5.2 [Limit of a Real Number Sequence]
Let $\{y_n\}$ be a sequence whose elements are (scalar) real numbers. Suppose there exists a real number, y, such that for every real $\varepsilon > 0$ there exists an integer $N(\varepsilon)$ for which $n \geq N(\varepsilon) \Rightarrow |y_n - y| < \varepsilon$. Then y is the limit of the sequence $\{y_n\}$, and the sequence $\{y_n\}$ is said to converge to y as $n \to \infty$. The existence of the limit is denoted by $y_n \to y$ or $\lim_{n\to\infty} y_n = y$. If the limit does not exist, the sequence is said to be divergent.

Definition 5.3 [Bounded Sequence of Real Numbers]
The sequence of real numbers $\{y_n\}$ is bounded iff there exists a finite number $m > 0$ such that $|y_n| \leq m \quad \forall \quad n \in N$; otherwise the sequence is said to be unbounded.

Definition 5.4 [Limit of a Real-Valued Matrix Sequence]
Let $\{\mathbf{Y}_n\}$ be a sequence whose elements are $(q \times k)$ real-valued matrices. Suppose there exists an $(q \times k)$ matrix of real numbers $\mathbf{Y}$ such that $Y_n[i, j] \to Y[i, j]$ for $i = 1, \ldots, q$ and $j = 1, \ldots, k$. Then the matrix $\mathbf{Y}$ is the limit of the matrix sequence $\{\mathbf{Y}_n\}$, and the sequence $\{\mathbf{Y}_n\}$ is said to converge to $\mathbf{Y}$ as $n \to \infty$. The existence of the limit is denoted by $\mathbf{Y}_n \to Y$, or by $\lim_{n\to\infty} \mathbf{Y}_n = \mathbf{Y}$. If the limit does not exist, the sequence is said to be divergent.

Definition 5.5 [Adding, Subtracting, and Multiplying Sequences]
Let $\{\mathbf{X}_n\}$ and $\{\mathbf{Z}_n\}$ be sequences of conformable, real-valued matrices.

a. Summation: The summation of $\{\mathbf{X}_n\}$ and $\{\mathbf{Z}_n\}$, $\{\mathbf{X}_n\} + \{\mathbf{Z}_n\}$, is a sequence $\{\mathbf{Y}_n\}$ defined by $\mathbf{Y}_n = \mathbf{X}_n + \mathbf{Z}_n \forall n$.

b. Difference: The difference between $\{\mathbf{X}_n\}$ and $\{\mathbf{Z}_n\}$, $\{\mathbf{X}_n\} - \{\mathbf{Z}_n\}$, is a sequence $\{\mathbf{Y}_n\}$ defined by $\mathbf{Y}_n = \mathbf{X}_n - \mathbf{Z}_n \forall n$.

c. Product: The product of $\{\mathbf{X}_n\}$ and $\{\mathbf{Z}_n\}$, $\{\mathbf{X}_n\}\{\mathbf{Z}_n\}$, is a sequence $\{\mathbf{Y}_n\}$ defined by $\mathbf{Y}_n = \mathbf{X}_n\mathbf{Z}_n \forall n$.

Lemma 5.1 [Combinations of Sequences]
Let $\{\mathbf{X}_n\}$ and $\{\mathbf{Z}_n\}$ be convergent sequences of conformable, real-valued matrices such that $\mathbf{X}_n \to \mathbf{X}$ and $\mathbf{Z}_n \to \mathbf{Z}$. Then

a. $\mathbf{X}_n + \mathbf{Z}_n \to \mathbf{X} + \mathbf{Z}$,

b. $\mathbf{X}_n - \mathbf{Z}_n \to \mathbf{X} - \mathbf{Z}$,

c. $\mathbf{X}_n\mathbf{Z}_n \to \mathbf{X}\mathbf{Z}$,

d. if $\{a_n\}$ is a sequence in R that converges to a, then $a_n\mathbf{X}_n \to a\mathbf{X}$,

e. if $\{b_n\}$ is a sequence of nonzero numbers in R that converges to $b \neq 0$, then $b_n^{-1}\mathbf{X}_n \to b^{-1}\mathbf{X}$,

f. $\sum_{i=1}^{k} \mathbf{X}_n[\cdot, i] \to \sum_{i=1}^{k} \mathbf{X}[\cdot, i]$,

g. if $\{\mathbf{Z}_n\}$ is a sequence of nonsingular matrices that converges to the nonsingular matrix $\mathbf{Z}$, then $\mathbf{Z}_n^{-1} \to \mathbf{Z}^{-1}$ and $\mathbf{Z}_n^{-1}\mathbf{X}_n \to \mathbf{Z}^{-1}\mathbf{X}$.

Definition 5.6 [Continuous Function]
[12]The function $g : A \mapsto R$, for $A \subset R^m$, is continuous at the point $x \in A$ iff either

a. $\forall \varepsilon > 0$, $\exists \delta(\varepsilon) > 0$ such that $\omega \in A$ and $d(x,\omega) < \delta(\varepsilon)$ implies $|g(\omega) - g(x)| < \varepsilon$, or

b. $\forall$ sequence $\{x_n\}$ in A for which $x_n \to x$, it is true that $g(x_n) \to g(x)$.

The vector function $g : A \mapsto R^k$ is continuous at the point $x \in A$ iff each coordinate function $g_j(x)$ is continuous at the point x, $j = 1,\ldots,k$. The function g is said to be continuous on (the set) $B \subset A$ if the function is continous at every point in B.

Definition 5.7 [Convergence of a Function Sequence]
Let $\{f_n\}$ be a sequence of functions, $f_n : D \mapsto R^\ell$, having common domain $D \subset R^m$. Let $f : D_0 \mapsto R^\ell$ be a function with domain $D_0 \subset D$. The function sequence $\{f_n\}$ is said to converge on D_0 to f if $f_n(x) \to f(x) \forall x \in D_0$. If $\{f_n\}$ converges to f on D_0, f is called the limiting function of $\{f_n\}$ on D_0, and $\{f_n\}$ is said to be convergent on D_0.

Definition 5.8 [Order of Magnitude of a Sequence]
Let $\{x_n\}$ be a real number sequence, and let $\{\mathbf{W}_n\}$ be a real-valued matrix sequence.

a. $O(n^k)$: The sequence $\{x_n\}$ is said to be at most of order n^k, denoted by $O(n^k)$, if there exists a finite real number c such that $|n^{-k}x_n| \leq c \forall n \in N$.

b. $o(n^k)$: The sequence $\{x_n\}$ is said to be of order smaller than n^k, denoted by $o(n^k)$, if $n^{-k}x_n \to 0$.

c. If $\{\mathbf{W}_n[i,j]\}$ is $O(n^k)$ (of $o(n^k)$) $\forall i$ and j, then the matrix sequence $\{\mathbf{W}\}$ is said to be $O(n^k)$ (or $o(n^k)$).

Lemma 5.2 [Relationships between Orders of Magnitudes]
Let $\{x_n\}$ and $\{z_n\}$ be real number sequences. The following relationship between orders of magnitude hold:

IF		THEN	
$\{x_n\}$	$\{z_n\}$	$\{x_n + z_n\}$	$\{x_n z_n\}$
$O(n^k)$	$O(n^m)$	$O\left(n^{\max(k,m)}\right)$	$O\left(n^{k+m}\right)$
$o(n^k)$	$o(n^m)$	$o\left(n^{\max(k,m)}\right)$	$o\left(n^{k+m}\right)$
$O(n^k)$	$o(n^m)$	$O\left(n^{\max(k,m)}\right)$	$o\left(n^{k+m}\right)$

Definition 5.9 [Convergence in Distribution (CDFs)]
Let $\{Y_n\}$ be a sequence of random variables, and let $\{F_n\}$ be the associated sequence of cumulative distribution functions corresponding to the random variables. If there exists a cumulative distribution function F such that $F_n(y) \to F(y) \forall y$ at which F is continuous, then F is called the limiting CDF of $\{Y_n\}$. Letting Y have the distribution F, i. e., $Y \sim F$, we then say that Y_n converges in distribution (or converges in law) to the random variable Y, and we denote this convergence by $Y_n \xrightarrow{d} Y$ (or $Y_n \xrightarrow{L} Y$). We also write $Y_n \xrightarrow{d} F$ as a shorthand notation for $Y_n \xrightarrow{d} Y \sim F$, which is read "$Y_n$ converges in distribution to F."

[12]This definition can be altered to provide definitions for continuity from the right and continuity from the left. For continuity from the right, the condition $\omega \geq x$ is added in part (a). The condition $x_n \geq x \forall n$ is added to part (b). For continuity from the left, the conditions become $\omega \leq x$ and $x_n \leq x$, respectively, $\forall n$.

Theorem 5.1 [Convergence in Distribution (Densities)]
Let $\{Y_n\}$ be a sequence of either continuous or nonnegative, integer-values, discrete random variables, and let $\{f_n\}$ be the associated sequence of probability density functions corresponding to the random variables. Let there exist a density function f such that $f_n(y) \to f(y) \forall y$, except perhaps on a set of points A such that $P_Y(A) = 0$ in the continuous case, where $Y \sim f$. It follows that $Y_n \xrightarrow{d} Y$ (or $Y_n \xrightarrow{L} Y$).

Theorem 5.2 [Convergence in Distribution (MGFs)]
Let $\{Y_n\}$ be a sequence of random variables having an associated sequence of moment generating functions $\{M_{Y_n}(t)\}$. Let Y have the moment-generating function $M_Y(t)$. Then $Y_n \xrightarrow{d} Y$ iff $M_{Y_n}(t) \to M_Y(t) \forall t \in (-h, h)$, for some $h > 0$.

Definition 5.10 [Asymptotic Distribution for $g(X_n, \theta_n)$ when $X_n \xrightarrow{d} X$]
Let the sequence of random variables $\{Z_n\}$ be defined by $Z_n = g(X_n, \theta_n)$, where $X_n \xrightarrow{d} X$ for nondegenerate X, and $\{\theta_n\}$ is a sequence of real numbers, matrices, and/or parameters. Then an asymptotic distribution for Z_n is given by the distribution of $g(X, \theta_n)$, denoted by $Z_n \overset{a}{\sim} g(X, \theta_n)$ and meaning "Z_n is asymptotically distributed as $g(X, \theta_n)$."

Theorem 5.3 [Convergence in Distribution for Continuous Functions]
Let $X_n \xrightarrow{d} X$, and let the random variable $g(X)$ be defined by a function $g(x)$ that is continuous, except perhaps on a set of points assigned probability zero by the probability distribution of X. Then $g(X_n) \xrightarrow{d} g(X)$.

Definition 5.11 [Convergence in Probability]
The sequence of random variables $\{Y_n\}$ converges in probability to the random variable Y iff

a. Scalar case: $\lim_{n \to \infty} P(|y_n - y| < \varepsilon) = 1 \ \forall \varepsilon > 0$,

b. Matrix case: $\lim_{n \to \infty} P(|y_n[i,j] - y[i,j]| < \varepsilon) = 1 \ \forall \varepsilon > 0$, $\forall i$ and j.

Convergence in probability will be denoted by $Y_n \xrightarrow{p} Y$, or $\operatorname{plim} Y_n = Y$, the latter notation meaning the probability limit of Y_n is Y.

Definition 5.12 [Probability Limits of Matrices (and Vectors for $k = 1$)]
Let $\{Y_n\}$ be a sequence of $(m \times k)$ random matrices. Then

$$\operatorname{plim} \begin{bmatrix} Y_n[1,1] & \dots & Y_n[1,k] \\ \vdots & \ddots & \vdots \\ Y_n[m,1] & \dots & Y_n[m,k] \end{bmatrix} = \begin{bmatrix} \operatorname{plim} Y_n[1,1] & \dots & \operatorname{plim} Y_n[1,k] \\ \vdots & \ddots & \vdots \\ \operatorname{plim} Y_n[m,1] & \dots & \operatorname{plim} Y_n[m,k] \end{bmatrix}.$$

Theorem 5.4 [Convergence in Probability for Continuous Functions]
Let $X_n \xrightarrow{p} X$, and let the random variable $g(X)$ be defined by a function $g(x)$ that is continuous, except perhaps on a set of points assigned probability zero by the probability distribution of X. Then $g(X_n) \xrightarrow{p} g(X)$, or equivalently, $\operatorname{plim} g(X_n) = g(\operatorname{plim} X_n)$.

Theorem 5.5 [plim **Properties]**
For conformable $\mathbf{X}_n \mathbf{Y}_n$, and constant matrix $\mathbf{A}$,

a. $\operatorname{plim} \mathbf{A}\mathbf{X}_n = \mathbf{A}(\operatorname{plim} \mathbf{X}_n)$;

b. $\operatorname{plim} \sum_{i=1}^{m} X_n[i] = \sum_{i=1}^{m} \operatorname{plim} X_n[i]$ (the plim of a sum = the sum of the plims);

c. $\operatorname{plim} \prod_{i=1}^{m} X_n[i] = \prod_{i=1}^{m} \operatorname{plim} X_n[i]$ (the plim of a product = the product of the plims);

d. $\operatorname{plim} \mathbf{X}_n \mathbf{Y}_n = (\operatorname{plim} \mathbf{X}_n)(\operatorname{plim} \mathbf{Y}_n)$;

e. $\operatorname{plim} \mathbf{X}_n^{-1} \mathbf{Y}_n = (\operatorname{plim} \mathbf{X}_n)^{-1} \operatorname{plim} \mathbf{Y}_n$ ($\operatorname{plim} X_n$ being nonsingular).

Corollary 5.1 [Convergence in Probability and in Distribution]

$$Y_n \xrightarrow{p} Y \Rightarrow Y_n \xrightarrow{d} Y.$$

Theorem 5.6 [Convergence in Probability and in Distribution, 2]

$$Y_n \xrightarrow{d} c \Rightarrow Y_n \xrightarrow{p} c.$$

Theorem 5.7 [Convergence in Probability and in Distribution, 3]
Let $\{\mathbf{X}_n\}$, $\{\mathbf{Y}_n\}$, and $\{\mathbf{a}_n\}$ be such that

$$\mathbf{X}_{n(k\times m)} \xrightarrow{d} \mathbf{X}_{(k\times m)},\ \mathbf{Y}_{n(l\times q)} \xrightarrow{p} \mathbf{y}_{(l\times q)}, \text{ and } \mathbf{a}_{n(j\times p)} \to \mathbf{a}_{(j\times p)}.$$

Let the set B be such that the probability distribution of $\mathbf{X}$ assigns $P(x \in B) = 1$, and let the random variable $g(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{a}_n)$ be defined by a (possibly vector) function g that is continuous at every point in $B \times \mathbf{y} \times \mathbf{a}$. Then $g(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{a}_n) \xrightarrow{d} g(\mathbf{X}, \mathbf{y}, \mathbf{a})$.

Theorem 5.8 [Slutsky's Theorems]
Let $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$ and $\mathbf{Y}_n \xrightarrow{p} \mathbf{c}$. Then, for conformable $\mathbf{X}_n$ and $\mathbf{Y}_n$,

a. $\mathbf{X}_n + \mathbf{Y}_n \xrightarrow{d} \mathbf{X} + \mathbf{c}$,

b. $\mathbf{Y}_n \mathbf{X}_n \xrightarrow{d} \mathbf{c}\mathbf{X}$,

c. $\mathbf{Y}_n^{-1} \mathbf{X}_n \xrightarrow{d} \mathbf{c}^{-1}\mathbf{X}$ (if $\mathbf{c}^{-1}$ exists).

Definition 5.13 [Order of Magnitude in Probability]
Let (X_n) be a sequence of random scalars, and let $\{\mathbf{W}_n\}$ be a real-valued, random matrix sequence.

a. $O_p(n^k)$: The sequence $\{X_n\}$ is said to be at most of order n^k in probability, denoted by $O_p(n^k)$, iff for every $\varepsilon > 0$ there exists a corresponding positive constant $c(\varepsilon) < \infty$ such that $P\left(n^{-k}|X_n| \le c(\varepsilon)\right) \ge 1 - \varepsilon$, $\forall n$.

b. $o_p(n^k)$: The sequence $\{X_n\}$ is said to be of order smaller than n^k in probability, denoted by $o_p(n^k)$, iff $n^{-k} X_n \xrightarrow{p} 0$.

c. If $\{\mathbf{W}_n[i,j]\}$ is $O_p(n^k)$ (or $o_p(n^k)$) $\forall i$ and j, then the random matrix sequence $\{\mathbf{W}_n\}$ is said to be $O_p(n^k)$ (or $o_p(n^k)$).

Definition 5.14 [Convergence in Mean Square (or Convergence in Quadratic Mean)]
The sequence of random variables $\{Y_n\}$ converges in mean square to the random variable Y, iff

a. Scalar case: $\lim_{n\to\infty} E(Y_n - Y)^2 = 0$,

b. Matrix case: $\lim_{n\to\infty} E(Y_n[i,j] - Y[i,j])^2 = 0$, $\forall i$ and j.

Convergence in mean square will be denoted by $Y_n \xrightarrow{m} Y$.

Theorem 5.9 [Conditions for Mean Square Convergence]
$Y_n \xrightarrow{m} Y$ iff $\forall i$ and j

a. $\mathrm{E}Y_n[i,j] \to \mathrm{E}Y[i,j]$,

b. $var(Y_n[i,j]) \to var(Y[i,j])$,

c. $cov(Y_n[i,j], Y[i,j]) \to var(Y[i,j])$.

Corollary 5.2 [Conditions for Mean Square Convergence]
$\mathbf{Y}_n \xrightarrow{m} \mathbf{c}$ iff $\mathrm{E}Y_n[i,j] \to c[i,j]$ and $var(Y_n[i,j] \to 0)$ $\forall i$ and j.

Theorem 5.10 [Convergence in Quadratic Mean, in Probability and in Distribution]

$$Y_n \xrightarrow{m} Y \Rightarrow Y_n \xrightarrow{p} Y \Rightarrow Y_n \xrightarrow{d} Y.$$

Definition 5.15 [Almost Sure Convergence (or Convergence with Probability 1)]
The sequence of random variables $\{Y_n\}$ converges almost surely to the random variable Y iff

a. Scalar case: $P(y_n \to y) = P(\lim_{n\to\infty} y_n = y) = 1$,

b. Matrix case: $P(y_n[i,j] \to y[i,j]) = P(\lim_{n\to\infty} y_n[i,j] = y[i,j]) = 1$, $\forall i$ and j.

Almost-sure convergence will be denoted by $Y_n \xrightarrow{as} Y$, or by $\text{aslim}\, Y_n = Y$, the latter notation meaning the almost-sure limit of Y_n is Y.

Theorem 5.11 [Khinchin's WLLN]
Let $\{X_n\}$ be a sequence of iid random variables, and suppose $\mathrm{E}X_i = \mu < \infty$, $\forall i$. Then $\bar{X}_n \xrightarrow{p} \mu$.

Theorem 5.12 [Convergence in Probability of Relative Frequency]
Let $\{S, \Upsilon, P\}$ be the probability space of an experiment, and let A be any event contained in S. Let an outcome of N_A be the number of times that event A occurs in n independent and identical repetitions of the experiment. Then the relative frequency of event A occuring is such that $\frac{N_A}{n} \xrightarrow{p} P(A)$.

Theorem 5.13 [Necessary and Sufficient Conditions for WLLN]
Let $\{X_n\}$ be a sequence of random variables with finite variances (not necessarily independent), and let $\{\mu_n\}$ be the corresponding sequence of their expectations. Then

$$\lim_{n\to\infty} P(|\bar{x}_n - \bar{\mu}_n| < \varepsilon) = 1, \forall \varepsilon > 0 \text{ iff } \mathrm{E}\left[\frac{(\bar{X}_n - \bar{\mu}_n)^2}{1 + (\bar{X}_n - \bar{\mu}_n)^2}\right] \to 0.$$

Theorem 5.14 [WLLN for Non-IID Case]
Let $\{X_n\}$ be a sequence of random variables with respective means given by $\{\mu_n\}$. If $var(\bar{X}_n) \to 0$, then $(\bar{X}_n - \bar{\mu}_n) \xrightarrow{p} 0$.

Theorem 5.15 [Convergence in Probability, Different Means]
$\bar{X}_n - \bar{\mu}_n \xrightarrow{p} 0$ and $\bar{\mu}_n \to c \Rightarrow \bar{X}_n \xrightarrow{p} c$.

Theorem 5.16 [Lindberg-Levy CLT]
Let $\{X_n\}$ be a sequence of iid random variables with $\mathrm{E}X_i = \mu$ and $\mathrm{var}(X_i) = \sigma^2 \in (0,\infty) \forall i$. Then,

$$(n^{\frac{1}{2}}\sigma)^{-1}\left(\sum_{i=1}^{n} X_i - n\mu\right) = \frac{n^{\frac{1}{2}}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} \mathscr{N}(0,1).$$

Theorem 5.17 [Lindberg's CLT]
Let $\{X_n\}$ be a sequence of independent random variables with $\mathrm{E}X_i = \mu_i$ and $\mathrm{var}(X_i = \sigma_i^2 < \infty) \forall i$. Define $b_n^2 = \sum_{i=1}^n \sigma_i^2$, $\bar{\sigma}_n^2 = n^{-1}\sum_{i=1}^n \sigma_i^2$, $\bar{\mu}_n = n^{-1}\sum_{i=1}^n \mu_i$, and let f_i be the PDF of X_i. If $\forall \varepsilon > 0$,

(continuous case:) $\lim_{n\to\infty} \frac{1}{b_n^2} \sum_{i=1}^n \int_{(x_i-\mu_i)^2 \geq \varepsilon b_n^2} (x_i - \mu_i)^2 f_i(x_i)\, dx_i = 0$,

(discrete case:) $\lim_{n\to\infty} \frac{1}{b_n^2} \sum_{i=1}^n \sum_{(x_i-\mu_i)^2 \geq \varepsilon b_n^2, f_i(x_i)>0} (x_i - \mu_i)^2 f_i(x_i) = 0$.

then

$$\frac{\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i}{(\sum_{i=1}^n \sigma_i^2)^{\frac{1}{2}}} = \frac{n^{\frac{1}{2}}(\bar{X}_n.\bar{\mu}_n)}{\bar{\sigma}_n} \xrightarrow{d} \mathscr{N}(0,1).$$

Theorem 5.18 [CLT for Bounded Random Variables]
Let $\{X_n\}$ be a sequence of independent random variables such that $P(|x_i| \leq m) = 1 \forall i$ for some $m \in (0,\infty)$, and suppose $\mathrm{E}X_i = \mu_i$ and $\mathrm{var}(X_i) = \sigma_i^2 < \infty \forall i$. If $\sum_{i=1}^n \mathrm{var}(X_i) = \sum_{i=1}^n \sigma_i^2 \to \infty$ as $n \to \infty$, then $\frac{n^{\frac{1}{2}}(\bar{X}_n - \bar{\mu}_n)}{\bar{\sigma}_n} \xrightarrow{d} \mathscr{N}(0,1)$.

Theorem 5.19 [Liapounov's CLT]
Let $\{X_n\}$ be a sequence of independent random variables such that $\mathrm{E}X_i = \mu_i$ and $\mathrm{var}(X_i) = \sigma_i^2 < \infty \forall i$. If, for some $\delta > 0$,

$$\lim_{n\to\infty} \frac{\sum_{i=1}^n \mathrm{E}|X_i - \mu_i|^{2+\delta}}{(\sum_{i=1}^n \sigma_i^2)^{1+\frac{\delta}{2}}} = 0,$$

then

$$\frac{\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i}{(\sum_{i=1}^n \sigma_i^2)^{\frac{1}{2}}} = \frac{n^{\frac{1}{2}}(\bar{X}_n - \bar{\mu}_n)}{\bar{\sigma}_n} = \xrightarrow{d} \mathscr{N}(0,1).$$

Theorem 5.20 [Liapounov's CLT: Triangular Arrays]
Let $\{X_{nn}\}$ be a triangular array of random variables with independent random variables within rows. Let $\mathrm{E}X_{ij} = \mu_{ij}$ and $\mathrm{var}(X_{ij}) = \sigma_{ij}^2 < \infty \forall i,j$. If, for some $\delta > 0$,

$$\lim_{n\to\infty} \frac{\sum_{i=1}^n \mathrm{E}|X_{ni} - \mu_{ni}|^{2+\delta}}{(\sum_{i=1}^n \sigma_{ni}^2)^{1+\frac{\delta}{2}}} = 0$$

then

$$\frac{\sum_{i=1}^n X_{ni} - \sum_{i=1}^n \mu_{ni}}{(\sum_{i=1}^n \sigma_{ni}^2)^{\frac{1}{2}}} = \frac{n^{\frac{1}{2}}(\bar{X}(n) - \bar{\mu}(n))}{\bar{\sigma}(n)} = \xrightarrow{d} \mathscr{N}(0,1).$$

Definition 5.16 [M-Dependence]
The sequence $\{X_n\}$ is said to exhibit m-dependence (or is said to be m-dependent) if, for $a_1 < a_2 < \ldots < a_k < b_1 < b_2 < \ldots < b_r$, $(X_{a1}, X_{a2}, \ldots, X_{ak})$ is independent of $(X_{b1}, X_{b2}, \ldots, X_{br})$ whenever $b_1 - a_k < m$.

Theorem 5.21 [CLT for Bounded M-Dependent Sequences]
Let $\{X_n\}$ be an m-dependent sequence of random scalars for which $\mathrm{E}X_i = \mu_i$ and $P(|x_i| \leq c) = 1$ for some $c < \infty \forall i$. Let $\sigma_{*n}^2 = \mathrm{var}(\sum_{i=1}^n X_i)$. If $n^{-\frac{2}{3}} \sigma_{*n}^2 \to \infty$, then

$$(\sigma_{*n}^2)^{-\frac{1}{2}} \left(\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i \right) \xrightarrow{d} \mathscr{N}(0,1).$$

Theorem 5.22 [Cramr-Wold's Device]
The sequence of $(k \times 1)$ random vectors $(\mathbf{X}_n)$ converges in distribution to the random $(k \times 1)$ vector $\mathbf{X}$ iff $\ell' \mathbf{X}_n \xrightarrow{d} \ell' \mathbf{X} \forall \ell \in R^k$.

Corollary 5.3 [Cramr-Wold Device for Normal Limiting Distributions]
$\mathbf{X}_n \xrightarrow{d} \mathscr{N}(\mu, \Sigma)$ iff $\ell' \mathbf{X}_n \xrightarrow{d} \mathscr{N}(\ell' \mu, \ell' \Sigma \ell) \forall \ell \in R^k$.

Theorem 5.23 [Multivariate Lindberg-Levy CLT]
Let $\{\mathbf{X}_n\}$ be a sequence of iid $(k \times 1)$ random vectors with $\mathrm{E}\mathbf{X}_i = \boldsymbol{\mu}$ and $\mathbf{Cov}(\mathbf{X}_i) = \Sigma \forall i$, where $\boldsymbol{\Sigma}$ is a $(k \times k)$ positive definite matrix. Then

$$n^{\frac{1}{2}} \left(n^{-1} \sum_{i=1}^{n} \mathbf{X}_i - \boldsymbol{\mu} \right) \xrightarrow{d} \mathcal{N}([\mathbf{0}], \boldsymbol{\Sigma}).$$

Theorem 5.24 [Multivariate CLT: Independent Bounded Random Vectors]
Let $\{\mathbf{X}_n\}$ be a sequence of independent $(k \times 1)$ random vectors such that $P(|x_{1i}| \le m, x_{2i} \le m, \ldots, x_{ki} \le m) = 1 \forall i$, where $m \in (0, \infty)$. Let $\mathrm{E}\mathbf{X}_i = \boldsymbol{\mu}_i$, $\mathrm{Cov}(\mathbf{X}_i) = \boldsymbol{\psi}_i$, and suppose that $\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n}$ $boldsymbolpsi_i = \boldsymbol{\psi}$, a finite, positive definite $(k \times k)$ matrix. Then

$$n^{-\frac{1}{2}} \sum_{i=1}^{n} (\mathbf{X}_i - \boldsymbol{\mu}_i) \xrightarrow{d} \mathcal{N}([\mathbf{0}], \boldsymbol{\psi}).$$

Lemma 5.3
First-Order Taylor Series Expansion and Remainder (Young's Form)] Let $g : D \mapsto R$ be a function having partial derivates in a a neighborhood of the point $\boldsymbol{\mu} \in D$ that are continuous at $\boldsymbol{\mu}$. Let $\mathbf{G} = \left[\frac{\partial g(\boldsymbol{\mu})}{\partial x_1}, \ldots, \frac{\partial g(\boldsymbol{\mu})}{\partial x_k} \right]$ be the gradient vector of $g(\mathbf{x})$ evaluated at the point $\mathbf{x} = \boldsymbol{\mu}$. For $\mathbf{x} \in D$, define the remainder term $R(\mathbf{x})$ via $g(\mathbf{x}) = g(\boldsymbol{\mu}) + G(\mathbf{x} - \boldsymbol{\mu}) + d(\mathbf{x}, \boldsymbol{\mu}) R(\mathbf{x})$, with $R(\boldsymbol{\mu}) = 0$. Then $R(\mathbf{x})$ is continuous at $x = \boldsymbol{\mu}$ and $\lim_{n\to\infty} R(\mathbf{x}) = R(\boldsymbol{\mu}) = 0$.

Theorem 5.25 [Asymptotic Distribution of $g(\mathbf{X}_n)$ – Scalar Function Case]
Let $\{\mathbf{X}_n\}$ be a sequence of $(k \times 1)$ random vectors such that $n^{\frac{1}{2}}(X_n - \boldsymbol{\mu}) \xrightarrow{d} Z \sim \mathcal{N}([\mathbf{0}], \boldsymbol{\Sigma})$. Let $g(\mathbf{x})$ have first-order partial derivates in a neighborhood of the point $\mathbf{x} = \boldsymbol{\mu}$ that are continuous at $\boldsymbol{\mu}$, and suppose the gradient vector of $g(x)$ evaluated at $x = \boldsymbol{\mu}$, $\mathbf{G}_{(1\times k)} = \left[\frac{\partial g(\boldsymbol{\mu})}{\partial x_1}, \ldots, \frac{\partial g(\boldsymbol{\mu})}{\partial x_k} \right]'$, is not the zero vector. Then

$$n^{\frac{1}{2}} \left(g(\mathbf{X}_n) - g(\boldsymbol{\mu}) \right) \xrightarrow{d} \mathcal{N}(0, \mathbf{G}\boldsymbol{\Sigma}\mathbf{G}') \text{ and } g(\mathbf{X}_n) \overset{a}{\sim} \mathcal{N}\left(g(\boldsymbol{\mu}), n^{-1}\mathbf{G}\boldsymbol{\Sigma}\mathbf{G}' \right).$$

Theorem 5.26 [Asymptotic Distribution of $\mathbf{g}(\mathbf{X}_n)$ – Vector Function Case]
Let $\{\mathbf{X}_n\}$ be a sequence of $(k \times 1)$ random vectors such that $n^{\frac{1}{2}}(X_n - \boldsymbol{\mu}) \xrightarrow{d} Z \sim \mathcal{N}([\mathbf{0}], \boldsymbol{\Sigma})$. Let $\mathbf{g}(\mathbf{x}) = (g_1(\mathbf{x}), \ldots, g_m(\mathbf{x}))'$ be an $(m \times 1)$ vector function $(m \le k)$ having first-order partial derivates in a neighborhood of the point $\mathbf{x} = \boldsymbol{\mu}$ that are continuous at μ. Let the Jacobian matrix of $\mathbf{g}(\mathbf{x})$ evaluated at $\mathbf{x} = \boldsymbol{\mu}$,

$$\mathbf{G}_{m\times k} = \begin{bmatrix} \frac{\partial g_1(\boldsymbol{\mu})}{\partial \mathbf{x}'} \\ \vdots \\ \frac{\partial g_m(\boldsymbol{\mu})}{\partial \mathbf{x}'} \end{bmatrix} = \begin{bmatrix} \frac{\partial g_1(\boldsymbol{\mu})}{\partial x_1} & \cdots & \frac{\partial g_1(\boldsymbol{\mu})}{\partial x_k} \\ \vdots & \cdots & \vdots \\ \frac{\partial g_m(\boldsymbol{\mu})}{\partial x_1} & \cdots & \frac{\partial g_m(\boldsymbol{\mu})}{\partial x_m} \end{bmatrix},$$

have full rank. Then

$$n^{\frac{1}{2}} \left(\mathbf{g}(\mathbf{X}_n) - \mathbf{g}(\boldsymbol{\mu}) \right) \xrightarrow{d} \mathcal{N}([0], \mathbf{G}\boldsymbol{\Sigma}\mathbf{G}') \text{ and } \mathbf{g}(\mathbf{X}_n) \overset{a}{\sim} \mathcal{N}(\mathbf{g}(\boldsymbol{\mu}), n^{-1}\mathbf{G}\boldsymbol{\Sigma}\mathbf{G}').$$

Theorem 5.27 [Asymptotic Distribution of $g(\mathbf{X}_n)$ – Generalized]
Let $\{\mathbf{X}_n\}$ be a sequence of $(k \times 1)$ random vectors such that $\mathbf{V}^{-\frac{1}{2}}(\mathbf{X}_n - \boldsymbol{\mu}) \xrightarrow{d} \mathcal{N}([\mathbf{0}], \mathbf{I})$, where $(\mathbf{V}_n)$ is a sequence of $(m \times m)$ positive definite matrices for which $\mathbf{V}_n \to [\mathbf{0}]$. Let $\mathbf{g}(\mathbf{x})$ be an $(m \times 1)$ vector function satisfying the conditions of Theorem **Asymptotic Distributions of $\mathbf{g}(\mathbf{X}_n)$ – Vector Function Case**. If there exists a sequence of positive real numbers $\{a_n\}$ such that $\left\{ [a_n \mathbf{G}\mathbf{V}_n\mathbf{G}']^{-\frac{1}{2}} \right\}$ is $O(1)$ and $a_n^{\frac{1}{2}}(\mathbf{X}_n - \boldsymbol{\mu})$ is $O_p(1)$, then

$$(\mathbf{G}\mathbf{V}_n\mathbf{G}')^{-\frac{1}{2}} [\mathbf{g}(\mathbf{X}_n) - \mathbf{g}(\boldsymbol{\mu})] \xrightarrow{d} \mathcal{N}([\mathbf{0}], \mathbf{I}) \text{ and } \mathbf{g}(\mathbf{X}_n) \overset{a}{\sim} \mathcal{N}(\mathbf{g}(\boldsymbol{\mu}), \mathbf{G}\mathbf{V}_n\mathbf{G}').$$

6 Sampling, Sample Moments, Sampling Distributions

Definition 6.1 [Random Sampling without Replacement]

1. The first object is selected from the population in a way that gives all objects in the population an equal chance of being selected.
2. The characteristics level of the object is observed, but the object is not returned to the population.
3. An object is selected from the remaining objects in the population in a way that gives all remaining objects an equal chance of being selected, and step (2) is repeated. For a sample of size n, step (3) is performed $(n-1)$ times.

Definition 6.2 [Statistic]
A real-valued function of observable random variables that is itself an observable random variable not depending on any unknown parameters.

Definition 6.3 [Empirical Distribution Function, Scalar Case]
Let the scalar random variables $X_1, \ldots, X_n$ denote a random sample from some population distribution. Then the empirical distribution function is defined, for $t \in (-\infty, \infty)$, by

$$F_n(t) = n^{-1} \sum_{i=1}^{n} I_{(-\infty,t]}(X_i),$$

an outcome of which is defined by

$$\hat{F}_n(t) = n^{-1} \sum_{i=1}^{n} I_{(-\infty,t]}(x_i).$$

Theorem 6.1 [PDF of EDF]
Let $F_n(t)$ be the EDF corresponding to a random sample of size n from a population characterized by the CDF $F(t)$. Then the PDF of $F_n(t)$ is defined by

$$P\left(\hat{F}_n(t) = \frac{j}{n}\right) = \begin{cases} \begin{pmatrix} n \\ j \end{pmatrix} [F(t)]^j [1-F(t)]^{n-j} & \text{for } j \in \{0,1,2,\ldots,n\}, \\ 0 & \text{otherwise.} \end{cases}$$

Theorem 6.2 [Properties of EDF, 1]
Let $F_n(t)$ be the EDF defined in Theorem **PDF of EDF**. Then, $\forall t \in (-\infty, \infty)$,

a. $\mathrm{E} F_n(t) = F(t)$,

b. $\mathrm{var}(F_n(t)) = n^{-1}[F(t)(1-F(t))]$,

c. $\mathrm{plim}\, F_n(t) = F(t)$,

d. $F_n(t) \overset{a}{\sim} \mathcal{N}\left(F(t), n^{-1}[F(t)(1-F(t))]\right)$.

Theorem 6.3 [Properties of EDF, 2]
Let $F_n(t)$ be the EDF defined in Theorem **PDF of EDF**. Then $\forall t \in (-\infty, \infty)$, and for $a < b$,

a. $\mathrm{E}\left[F_n(b) - F_n(a)\right] = F(b) - F(a)$,

b. $\mathrm{var}\left(F_n(b) - F_n(a)\right) = n^{-1}[F(b) - F(a)][1 - F(b) + F(a)]$,

c. $\operatorname{plim}[F_n(b) - F_n(a)] = F(b) - F(a)$,

d. $F_n(b) - F_n(a) \overset{a}{\sim} \mathscr{N}\left(F(b) - F(a), n^{-1}[F(b) - F(a)][1 - F(b) + F(a)]\right)$.

Theorem 6.4 [Glivenko-Cantelli's Theorem]
Let $D_n = \sup_t |F_n(t) - F(t)|$. Then

$$P(\lim_{n\to\infty} D_n = 0) = 1.$$

Definition 6.4 [Empirical Distribution Function, Multivariate Case]
Let the $(k \times 1)$ random vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$ denote a random sample from some population distribution. Then the empirical distribution function is defined for $\mathbf{t} = [t_1, \ldots, t_k]' \in R^k$ and $A(\mathbf{t}) = \times_{i=1}^{k}(-\infty, t_i]$ as

$$F_n(t) = n^{-1} \sum_{i=1}^{n} I_{A(\mathbf{t})}(\mathbf{X}_i),$$

an outcome defined by

$$\hat{F}_n(t) = n^{-1} \sum_{i=1}^{n} I_{A(\mathbf{t})}(x_i).$$

Definition 6.5 [Sample Moments about the Origin and Mean]
Let the scalar random variables $X_1, \ldots, X_n$ be a random sample. Then outcomes of the rth order sample moments about the origin and mean are defined as:

Sample moments about the origin: $m'_r = n^{-1} \sum_{i=1}^{n} x_i^r$,

Sample moments about the mean: $m_r = n^{-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)^r$,

where $\bar{x}_n = m'_1 = n^{-1} \sum_{i=1}^{n} x_i$.

Theorem 6.5 [Properties of M'_r]
Let $M'_r = n^{-1} \sum_{i=1}^{n} X_i^r$ be the rth sample moment about the origin for a random sample $(X_1, \ldots, X_n)$ from a population distribution. Then, assuming the appropriate population moments exist,

a. $\mathrm{E}M'_r = \mu'_r$,

b. $\operatorname{var}(M'_r) = n^{-1}\left[\mu'_{2r} - (\mu'_r)^2\right]$,

c. $\operatorname{plim} M'_r = \mu'_r$,

d. $\frac{(M'_r - \mu'_r)}{[\operatorname{var}(M'_r)]^{\frac{1}{2}}} \xrightarrow{d} \mathscr{N}(0, 1)$,

e. $M'_r \overset{d}{\sim} \mathscr{N}(\mu'_r, \operatorname{var}(M'_r))$.

Definition 6.6 [Sample Mean]
Let $(X_1, \ldots, X_n)$ be a random sample. The sample mean is defined by

$$\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i = M'_1.$$

Theorem 6.6 [Multivariate, Asymptotic Normality of Sample Moments about the Origin]

$$n^{\frac{1}{2}} \begin{bmatrix} M'_1 - \mu'_1 \\ \vdots \\ M'_r - \mu'_r \end{bmatrix} \xrightarrow{d} \mathscr{N}(|0|_{r\times 1}, \Sigma_{r\times r}) \text{ and } \begin{bmatrix} M'_1 \\ \vdots \\ M'_r \end{bmatrix} \overset{a}{\sim} \mathscr{N}\left((\mu'_1, \ldots, \mu'_r)', n^{-1}\Sigma\right),$$

where the nonsingular covariance matrix Σ has a typical (j, k) entry $\sigma_{jk} = \mu'_{j+k} - \mu'_j \mu'_k$.

Definition 6.7 [Sample Variance]
Let $X_1,\ldots,X_n$ be a random sample of size n. The sample variance is defined as[13]

$$S_n^2 = n^{-1}\sum_{i=1}^{n}(X_i-\bar{X}_n)^2 = M_2.$$

Theorem 6.7 [Properties of S_n^2]
Let S_n^2 be the sample variance for a random sample $(X_1,\ldots,X_n)$ from a population distribution. Then, assuming the appropriate population moments exist,

a. $\mathrm{E}S_n^2 = \frac{(n-1)}{n}\sigma^2$,

b. $\mathrm{var}S_n^2 = n^{-1}\left[\left(\frac{(n-1)}{n}\right)^2\mu_4 - \left(\frac{(n-1)(n-3)}{n^2}\right)\sigma^4\right]$,

c. $\mathrm{plim}\,S_n^2 = \sigma^2$,

d. $n^{\frac{1}{2}}\left(S_n^2-\sigma^2\right)\xrightarrow{d}\mathscr{N}\left(0,\mu_4-\sigma^4\right)$,

e. $S_n^2 \overset{a}{\sim} \mathscr{N}\left(\sigma^2, n^{-1}(\mu_4-\sigma^4)\right)$.

Definition 6.8 [Sample Moments, Multivariate Case]
Let the $(k\times 1)$ vector of random variables $\mathbf{X}_1,\ldots,\mathbf{X}_n$ be a random sample from a population distribution. Then the following outcomes of sample moments can be defined for j and $\ell\in\{1,2,\ldots,k\}$:

Sample moments about the origin: $m_r'[j] = n^{-1}\sum_{i=1}^{n} x_i[j]^r$;

Sample means: $\bar{x}[j] = m_1'[j] = n^{-1}\sum_{i=1}^{n} x_i[j]$;

Sample moments about the mean: $m_r[j] = n^{-1}\sum_{i=1}^{n}(x_i[j]-\bar{x}[j])^r$;

Sample variances: $s^2[j] = m_2[j] = n^{-1}\sum_{i=1}^{n}(x_i[j]-\bar{x}[j])^2$;

Sample covariance: $s_{j\ell} = n^{-1}\sum_{i=1}^{n}(x_i[j]-\bar{x}[j])(x_i[\ell]-\bar{x}[\ell])$.

Theorem 6.8 [Properties of Sample Covariance]
Let $((X_1,Y_1),\ldots,(X_n,Y_n))$ be a random sample from a population distribution, and let S_{XY} be the sample covariance between X and Y. Then, assuming the appropriate population moments exist,

a. $\mathrm{E}S_{XY} = \frac{(n-1)}{n}\sigma_{XY}$,

b. $\mathrm{var}(S_{XY}) = n^{-1}\left(\mu_{2,2}-(\mu_{1,1})^2\right)+\mathrm{o}\left(n^{-1}\right)$,

c. $\mathrm{plim}\,S_{XY} = \sigma_{XY}$,

d. $S_{XY}\xrightarrow{a}\mathscr{N}\left(\sigma_{XY}, n^{-1}\left(\mu_{2,2}-(\mu_{1,1})^2\right)\right)$.

Theorem 6.9 [Sample Correlation]
Let $((X_1,Y_1),\ldots,(X_n,Y_n))$ be a random sample from a population distribution. Then the sample correlation between X and Y is given by

$$R_{XY} = \frac{S_{XY}}{S_X S_Y},$$

where $S_X = (S_X^2)^{\frac{1}{2}}$ and $S_Y = (S_Y^2)^{\frac{1}{2}}$ are the sample standard deviations of X and Y, respectively.

[13] Some authors define the sample variance as $S_n^2 = \frac{n}{(n-1)}M_2$, so that $\mathrm{E}S_n^2 = \sigma^2$, which identifies S_n^2 as an unbiased estimator of σ^2. However, this definition would be inconsistent with the aforementioned fact that M_2, and not $\frac{n}{(n-1)}M_2$, is the second moment about the mean, and thus the variance, of the sample (empirical) distribution function, $\hat{F}_n$.

Theorem 6.10 [Properties of Sample Correlation]
Let (X_i, Y_i), $i = 1, \ldots, n$, be a random sample from a population distribution, and let R_{XY} be the sample correlation between X and Y. Then

a. $\operatorname{plim} R_{XY} = \rho_{XY}$,

b. $R_{XY} \overset{a}{\sim} \mathscr{N}\left(\rho_{XY}, n^{-1}\tau'\Sigma\tau\right)$,

with τ and Σ defined ahead.

Theorem 6.11 [Independence of Linear and Quadratic Forms]
Let $\mathbf{B}$ be a $(q \times n)$ matrix of real numbers, let $\mathbf{A}$ be an $(n \times n)$ symmetric matrix of real numbers having rank p, and let $\mathbf{X}$ be an $(n \times 1)$ random vector such that $\mathbf{X} \sim \mathscr{N}\left(\mu_X, \sigma^2 I\right)$. Then $\mathbf{BX}$ and $\mathbf{X}'\mathbf{AX}$ are independent if $\mathbf{BA} = [0]$.[14]

Theorem 6.12 [Properties of S_n^2]
If $\bar{X}_n$ and S_n^2 are the sample mean and sample variance, respectively, of a random sample of size n from a normal distribution with mean μ and variance σ^2, then

a. $\bar{X}_n$ and S_n^2 are independent,

b. $\left(nS_n^2/\sigma^2\right) \sim \chi^2_{n-1}$.

Theorem 6.13 [(Change of Variables Technique (Univariate and Invertible)]
Suppose the continuous random variable X has PDF $f(x)$. Let $g(x)$ be continuously differentiable with $dg/dx \neq 0\ \forall x$ in some open interval, Δ, containing the support of $f(x)$, Ξ. Also, let the inverse function $x = g^{-1}(y)$ be defined $\forall y \in \Psi = \{y : y = g(x),\ x \in \Xi\}$. Then the PDF of $Y = g(X)$ is given by $h(y) = f\left(g^{-1}(y)\right)\left|\frac{dg^{-1}(y)}{dy}\right|$ for $y \in \Psi$, with $h(y) = 0$ elsewhere.

Theorem 6.14 [Change of Variables Technique (Univariate and Piecewise Invertible)]
Suppose the continuous random variable X has PDF $f(x)$. Let $g(x)$ be continuously differentiable with $dg(x)/dx \neq 0$ for all but perhaps a finite number of x's in an open interval Δ containing the support of $f(x)$, Ξ. Let Ξ be partitioned into a collection of disjoint intervals $D_1, \ldots, D_n$ for which $g : D_i \to R_i$ has an inverse function $g_i^{-1} : R_i \to D_i\ \forall i$.[15] Then the probability density of $Y = g(X)$ is given by

$$h(y) = \begin{cases} \sum\limits_{i \in I(y)} f\left(g_i^{-1}(y)\right)\left|\frac{dg^{-1}(y)}{dy}\right| & \text{for } y \in \Psi = \{y : y = g(x),\ x \in \Xi\}, \\ 0 & \text{elsewhere,} \end{cases}$$

where $I(y) = \{i : \exists x \in D_i \text{ such that } y = g(x),\ i = 1, \ldots, n\}$ and $\left(dg_i^{-1}(y)/dy\right) \equiv 0$ whenever it would otherwise be undefined.[16]

Theorem 6.15 [Change of Variables Technique (Multivariate and Invertible)]
Suppose the continuous $(n \times 1)$ random vector $\mathbf{X}$ has joint PDF $f(\mathbf{x})$. Let $\mathbf{g}(\mathbf{x})$ be a $(n \times 1)$ real valued vector function that is continuously differentiable $\forall \mathbf{x}$ vector in some open rectangle of points, Δ, containing the support of $f(\mathbf{x})$, Ξ. Assume the inverse vector function $\mathbf{x} = \mathbf{g}^{-1}(\mathbf{y})$ exists, $\forall \mathbf{y} \in \Psi = \{\mathbf{y} : \mathbf{y} = \mathbf{g}(\mathbf{x}),\ \mathbf{x} \in \Xi\}$. Furthermore, let

$$\mathbf{J} = \begin{bmatrix} \frac{\partial g_1^{-1}(\boldsymbol{y})}{\partial y_1} & \cdots & \frac{\partial g_1^{-1}(\boldsymbol{y})}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_n^{-1}(\boldsymbol{y})}{\partial y_1} & \cdots & \frac{\partial g_n^{-1}(\boldsymbol{y})}{\partial y_n} \end{bmatrix},$$

[14] The theorem can be extended to the case where $\mathbf{X} \sim \mathscr{N}(\mu_X, \Sigma)$, in which case the condition for independence is that $\mathbf{B}\Sigma\mathbf{A} = [0]$.

[15] These properties define a function that is piecewise invertible on the domain $\bigcup_{i=1}^n D_i$.

[16] Note that $I(y)$ is an index set containing the indices of all of the D_i sets that have an element whose image under the function g is the value y.

called the Jacobian matrix, be such that $\det(\mathbf{J}) \neq 0$ with all partial derivatives in $\mathbf{J}$ being continuous $\forall \mathbf{y} \in \Psi$. Then the joint density of $\mathbf{Y} = \mathbf{g}(\mathbf{X})$ is given by

$$h(\mathbf{y}) = \begin{cases} f\left(g_1^{-1}(\mathbf{y}), \ldots, g_n^{-1}(\mathbf{y})\right) |\det(\mathbf{J})| & \text{for } \mathbf{y} \in \Psi, \\ 0 & \text{otherwise,} \end{cases}$$

where $|\det(\mathbf{J})|$ denotes the absolute value of the determinant of the Jacobian.

Theorem 6.16 [t-density for a ratio of $\mathcal{N}(0,1)$ and χ^2_ν]
Let $Z \sim \mathcal{N}(0,1)$, let $Y \sim \chi^2_\nu$, and let Z and Y be independent random variables. Then $T = Z/|Y/\nu|^{1/2}$ has the t-density with ν degrees of freedom, defined as

$$f(t;\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma(\nu/2)\sqrt{\pi\nu}} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}.$$

Theorem 6.17 [t-distribution for the standardized sample mean]
Under the assumptions of Theorem Properties of S_n^2 and defining $\hat{\sigma}_n = (n/(n-1))^{1/2} S_n$,

$$T = \frac{n^{1/2}(\bar{X}_n - \mu)}{\hat{\sigma}_n} \sim \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma(\nu/2)(\pi\nu)^{1/2}} \left[1 + \frac{t^2}{\nu}\right]^{-(\nu+1)/2},$$

where $\nu = n - 1$.

Theorem 6.18 [F-density for a ratio of $\chi^2_{\nu_1}$ and $\chi^2_{\nu_2}$]
Let $Y_1 \sim \chi^2_{\nu_1}$, let $Y_2 \sim \chi^2_{\nu_2}$, and let Y_1 and Y_2 be independent. Then $F = (Y_1/\nu_1)/(Y_2/\nu 2)$ has the F-density with ν_1 numerator and ν_2 denominator degrees of freedom, defined as

$$m(f;\nu_1,\nu_2) = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} f^{(\nu_1/2)-1} \left(1 + \frac{\nu_1}{\nu_2} f\right)^{-(1/2)(\nu_1+\nu_2)} I_{(0,\infty)}(f).$$

7 Order statistics

Definition 7.1 [Order statistics]
Given a data generation process described in form of a probability density $f_X(x)$, and given a sample of size n implying n independent and identical draws of the random variable X, let $\mathbf{y} = SORT(x_1, \ldots, x_n)$ be the $n \times 1$ vector function whose value is the $n \times 1$ vector $[x_1, \ldots, x_n]'$ sorted from the lowest to the highest value. Then the order statistics $\mathbf{X}_o = [X_{[1]}, \ldots, X_{[n]}]'$ corresponding to the random sample $X = [X_1, \ldots, X_n]'$ are defined as $\mathbf{X}_o = SORT(X)$ with PDF

$$f_{X_{[1]},X_{[2]},\ldots,X_{[n]}}(x_o) = n! f(x_{[1]}) f(x_{[2]}) \ldots f(x_{[n]}) \cdot I_{(-\infty,\infty)}(x_{[1]}) I_{(x_{[1]},\infty)}(x_{[2]}) \ldots I_{(x_{[n-1]},\infty)}(x_{[n]}).$$

The random variable $X_{[k]}$ is called the kth order statistic.

Theorem 7.1 [Sampling distribution of $X_{[k]}$]
Let $(X_1, \ldots, X_n)$ be a random sample from a population distribution with CDF F, and let $X_{[k]}$ be the kth order statistic corresponding to the random sample. Then the CDF of $X_{[k]}$ is given by

$$F_{X_{[k]}}(b) = \sum_{j=k}^{n} \binom{n}{j} F(b)^j [1 - F(b)]^{n-j},$$

while the corresponding PDF is given as

$$f_{X_{[k]}}(x) = \frac{n!}{(k-1)!(n-k)!} F(x)^{k-1} (1 - F(x))^{n-k} f(x).$$

Corollary 7.1 [Sampling Distributions of $X_{[1]}$ and $X_{[n]}$]
Assume the conditions of Theorem Sampling distribution of $X_{[k]}$ (Thm. 7.1). Then

$$F_{X_{[1]}}(b) = 1 - [1 - F(b)]^n \quad \text{and} \quad F_{X_{[n]}}(b) = F(b)^n,$$

and correspondingly,

$$f_{X_{[1]}}(x) = n f_X(x) [1 - F_X(x)]^{n-1} \quad \text{and} \quad f_{X_{[n]}}(x) = n f_X(x) [F_X(x)]^{n-1}.$$

Theorem 7.2 [(Sampling Distributions of $\left(X_{[k]}, X_{[\ell]}\right)$]
Let $X_{[k]}$ and $X_{[\ell]}$, $k < \ell$, be the kth and ℓth order statistics corresponding to the random sample $X = (X_1, \ldots, X_n)$ from a population distribution with CDF F and PDF f. Then the joint CDF of $X_{[k]}, X_{[\ell]}$ is given by

$$F_{X_{[k]}X_{[\ell]}}(b_k, b_\ell) = \begin{cases} F_{X_{[\ell]}}(b_\ell) & \text{for} \quad b_k \geq b_\ell, \\ \sum_{i=k}^{n} \sum_{j=\max\{0,\ell-i\}}^{n-i} \frac{n!}{i!j!(n-i-j)!} \times & \\ \qquad F(b_k)^i [F(b_\ell) - F(b_k)]^j [1 - F(b_\ell)]^{n-i-j} & \text{for} \quad b_k < b_\ell. \end{cases}$$

Correspondingly, the PDF is given as

$$f_{X_{[k]},X_{[l]}} = \begin{cases} \frac{n!}{(k-1)!(l-1-k)!(n-l)!} (F(x_k))^{k-1} f(x_k) (F(x_l) - F(x_k))^{l-1-k} f(x_l) (1 - F(x_l))^{n-l} & \text{if } x_k < x_l, \\ 0 \text{ else.} \end{cases}$$

Corollary 7.2 [Sampling Distributions of $X_{[1]}$ and $X_{[n]}$]
Let $k = 1$ and $\ell = n$ in Theorem Sampling Distributions of $(X_{[k]}, X_{[\ell]})$ (Thm. 7.2). Then by the binomial theorem,

$$\mathrm{E}_{X_{[1]},X_{[n]}}(b_1, b_n) = \begin{cases} F(b_n)^n & \text{for} \quad b_1 \geq b_n, \\ F(b_n)^n - [F(b_n) - F(b_1)]^n & \text{for} \quad b_1 < b_n. \end{cases}$$

8 Point Estimation

8.1 Stochastic Models

Definition 8.1 [Statistical model]
A statistical model for a random sample $\boldsymbol{X}$ consists of

1. a parametric functional form, $f(\boldsymbol{x};\boldsymbol{\Theta})$, for the joint pdf of $\boldsymbol{X}$ indexed by the parameters $\boldsymbol{\Theta}$,

2. together with a parameter space, Ω, that defines the set of potential candidates for the true joint pdf of $\boldsymbol{X}$ as $\{f(\boldsymbol{x};\boldsymbol{\Theta}),\ \boldsymbol{\Theta} \in \Omega\}$.

Assumption 1. Ω contains the true parameter value so that $\boldsymbol{\Theta}_0 \in \Omega$.

Assumption 2. Ω is such that the parameter vector $\boldsymbol{\Theta}$ is identified.

Definition 8.2 [Parameter identifiability]
Let $\{f(\boldsymbol{x};\boldsymbol{\Theta}),\ \boldsymbol{\Theta} \in \Omega\}$ be a statistical model for the random sample $\boldsymbol{X}$. The parameter vector $\boldsymbol{\Theta}$ is said to be identified iff $\forall\ \boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2 \in \Omega,\ f(\boldsymbol{x};\boldsymbol{\Theta}_1)$ and $f(\boldsymbol{x};\boldsymbol{\Theta}_2)$ are distinct if $\boldsymbol{\Theta}_1 \neq \boldsymbol{\Theta}_2$.

8.2 Estimators and Their Properties

Definition 8.3 [Point estimator]
A statistic, $\boldsymbol{T} = \boldsymbol{t}(\boldsymbol{X})$, whose outcomes are used to estimate the value of a scalar or vector function, $\boldsymbol{q}(\boldsymbol{\Theta})$, of the parameter vector, $\boldsymbol{\Theta}$, is called a **point estimator**. An observed outcome of an estimator is called a **point estimate**.

8.2.1 Finite-Sample Properties

Definition 8.4 [Mean square error (scalar case)]
The mean square error (MSE) of an estimator $T = t(\boldsymbol{X})$ of $q(\boldsymbol{\Theta})$ is defined as

$$MSE_{\Theta}(T) = \mathrm{E}_{\Theta}[T - q(\boldsymbol{\Theta})]^2\ \forall\ \boldsymbol{\Theta} \in \Omega.$$

Definition 8.5 [Relative Efficiency (scalar case)]
Let T and $T^\star$ be two estimators of a scalar $q(\boldsymbol{\Theta})$. The relative efficiency of T w.r.t. $T^\star$ is given by

$$RE_{\Theta}(T,T^\star) = \frac{MSE_{\Theta}(T^\star)}{MSE_{\Theta}(T)},\ \forall\ \boldsymbol{\Theta} \in \Omega.$$

T is relatively more efficient than $T^\star$ if

$$RE_{\Theta}(T,T^\star) \geq 1\ \forall\ \boldsymbol{\Theta} \in \Omega \quad \text{and} \quad RE_{\Theta}(T,T^\star) > 1 \ \text{ for at least one } \boldsymbol{\Theta} \in \Omega.$$

Definition 8.6 [Unbiased estimator]
An estimator $\boldsymbol{T}$ is said to be an unbiased estimator of $\boldsymbol{q}(\boldsymbol{\Theta})$ iff

$$\mathrm{E}_{\Theta}\boldsymbol{T} = \boldsymbol{q}(\boldsymbol{\Theta})\ \forall\ \boldsymbol{\Theta} \in \Omega.$$

Otherwise, the estimation is said to be biased.

Definition 8.7 [Minimum Variance unbiased estimator (MVUE) (scalar case)]
An estimator T is said to be a minimum-variance unbiased estimator of $q(\boldsymbol{\Theta})$ iff

1. $\mathrm{E}_{\Theta}T = q(\boldsymbol{\Theta})\ \forall\ \boldsymbol{\Theta} \in \Omega$, that is, T is unbiased, and

2. $\mathrm{Var}_\Theta(T) \leq \mathrm{Var}_\Theta(T^\star)\ \forall\ \Theta \in \Omega$ for any other unbiased estimator $T^\star$.

Definition 8.8 [Best linear unbiased estimator (BLUE) (scalar case)]
An estimator T is said to be a BLUE of $q(\Theta)$ iff

1. T is a linear function of the random sample $\boldsymbol{X} = (X_1, \dots, X_n)'$, i.e.,

$$T = \boldsymbol{a}'\boldsymbol{X} = a_0 + a_1 X_1 + \ldots + a_n X_n,$$

2. $\mathrm{E}_\Theta T = q(\Theta)\ \forall\ \Theta \in \Omega$, that is, T is unbiased, and

3. $\mathrm{Var}_\Theta(T) \leq \mathrm{Var}_\Theta(T^\star)\ \forall\ \Theta \in \Omega$ for any other linear and unbiased estimator $T^\star$.

8.2.2 Asymptotic Properties

Definition 8.9 [Consistent estimator]
An estimator $\boldsymbol{T}_n$ is said to be a consistent estimator of $\boldsymbol{q}(\Theta)$ iff $\mathrm{plim}_\Theta \boldsymbol{T}_n = \boldsymbol{q}(\Theta)\ \forall\ \Theta \in \Omega$.

Definition 8.10 [Consistent asymptotically normal (CAN) estimator]
An estimator $\boldsymbol{T}_n$ is said to be a CAN estimator of $\boldsymbol{q}(\Theta)$ iff

$$\sqrt{n}(\boldsymbol{T}_n - \boldsymbol{q}(\Theta)) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma}$ is a positive-definite covariance matrix.

Asymptotic versions of MSE, bias and variance can be defined w.r.t. the unique asymptotic distribution of CAN estimators.
The **asymptotic MSE for a CAN estimator** T_n for the scalar $q(\Theta)$ with $T_n \overset{a}{\sim} \mathcal{N}(q(\Theta), \frac{1}{n}\sigma^2)$ is

$$AMSE_\Theta(T_n) = \mathrm{E}_A\left[(T_n - q(\Theta))^2\right] = \mathrm{Var}_A[T_n] + (\underbrace{E_A\left[T_n - q(\Theta)\right]}_{\text{Asymtotic Bias} = 0})^2 = \mathrm{Var}_A(T_n) = \frac{1}{n}\sigma^2,$$

where E_A is the expectation w.r.t. the asym. distribution and Var_A is the variance of the asym. distribution.

Definition 8.11 [Asymptotic relative efficiency (scalar case)]
Let T_n and $T_n^\star$ be CAN estimators of $q(\Theta)$ such that

$$\sqrt{n}(T_n - q(\Theta)) \xrightarrow{d} \mathcal{N}(0, \sigma_T^2) \quad \text{and} \quad \sqrt{n}(T_n^\star - q(\Theta)) \xrightarrow{d} \mathcal{N}(0, \sigma_{T^\star}^2).$$

The asymptotic relative efficiency (ARE) of T_n with respect to $T_n^\star$ is given by

$$ARE_\Theta(T_n, T_n^\star) = \frac{AMSE_\Theta(T_n^\star)}{AMSE_\Theta(T_n)} = \frac{\sigma_{T^\star}^2}{\sigma_T^2}\ \forall\ \Theta \in \Omega.$$

T_n is asymptotically relatively more efficient than $T_n^\star$ if

$$ARE_\Theta(T, T^\star) \geq 1\ \forall\ \Theta \in \Omega \quad \text{and} \quad ARE_\Theta(T, T^\star) > 1 \ \text{ for at least one } \Theta \in \Omega.$$

Definition 8.12 [Asymptotic efficiency (scalar case)]
If T_n is a CAN estimator of $q(\Theta)$ having the smallest asymptotic variance among all CAN estimators $\forall\ \Theta \in \Omega$, except on a set of Lebesque measure zero, T_n is said to be asymptotically efficient.

8.3 Sufficient Statistics

Definition 8.13 [Sufficient statistics]
Let $(X_1,\ldots,X_n) \sim f(x_1,\ldots,x_n;\Theta)$ be a random sample, and let $S_1 = s_1(X_1,\ldots,X_n),\ldots,S_r = s_r(X_1,\ldots,X_n)$ be r statistics. The r statistics are said to be sufficient statistics for $f(x;\Theta)$ iff

$$f(x_1,\ldots,x_n;\Theta|s_1,\ldots,s_r) = h(x_1,\ldots,x_n),$$

i.e., the conditional density of $\boldsymbol{X}$, given $\boldsymbol{s} = [s_1,\ldots,s_r]'$, does not depend on the parameter Θ.

Theorem 8.1 [Neyman's Factorization Theorem]
Let $f(\boldsymbol{x};\Theta)$ be the pdf of the random sample $(X_1,\ldots,X_n)$. The statistics $S_1,\ldots,S_r$ are sufficient statistics for $f(\boldsymbol{x};\Theta)$ iff $f(\boldsymbol{x};\boldsymbol{\Theta})$ can be factored as

$$f(\boldsymbol{x};\Theta) = g(s_1(\boldsymbol{x}),\ldots,s_r(\boldsymbol{x});\Theta)\cdot h(\boldsymbol{x}),$$

where g is a function of only $s_1(\boldsymbol{x}),\ldots,s_r(\boldsymbol{x})$ and Θ, and $h(\boldsymbol{x})$ does not depend on Θ.

8.4 Minimal Sufficient Statistics

Definition 8.14 [Minimal sufficient statistics]
A sufficient statistic $\boldsymbol{S} = \boldsymbol{s}(\boldsymbol{X})$ for $f(\boldsymbol{x};\Theta)$ is said to be a minimal sufficient statistic if, for every other sufficient statistic $\boldsymbol{T} = \boldsymbol{t}(\boldsymbol{X}), \exists$ a function[17]

$$\boldsymbol{h}_T(\cdot) \text{ such that } \boldsymbol{s}(\boldsymbol{x}) = \boldsymbol{h}_T(\boldsymbol{t}(\boldsymbol{x}))\ \forall\ \boldsymbol{x} \in R(\boldsymbol{X}).$$

Corollary 8.1 [Minimal Sufficiency when $R(\boldsymbol{X})$ in independent of Θ]
Let $\boldsymbol{X} \sim f(\boldsymbol{x};\Theta)$, and suppose that $R(\boldsymbol{X})$ does not depend on Θ. If the statistic $\boldsymbol{S} = \boldsymbol{s}(\boldsymbol{X})$ is such that

$$\frac{f(\boldsymbol{x};\Theta)}{f(\boldsymbol{y};\Theta)} \text{ does not depend on } \Theta \text{ iff } (\boldsymbol{x},\boldsymbol{y}) \text{ satisfies } \boldsymbol{s}(\boldsymbol{x}) = \boldsymbol{s}(\boldsymbol{y}),$$

then $\boldsymbol{S} = \boldsymbol{s}(\boldsymbol{X})$ is a minimal sufficient statistic.

Theorem 8.2 [Exponential class and sufficient statistics]
Let $f(\boldsymbol{x};\boldsymbol{\Theta})$ be a member of the exponential class of density functions

$$f(\boldsymbol{x};\boldsymbol{\Theta}) = \exp\left[\sum_{i=1}^{k} c_i(\boldsymbol{\Theta})g_i(\boldsymbol{x}) + d(\boldsymbol{\Theta}) + z(\boldsymbol{x})\right].$$

Then $\boldsymbol{s}(\boldsymbol{X}) = [g_1(\boldsymbol{X}),\ldots,g_k(\boldsymbol{X})]$ is a k-variate sufficient statistic, and if $c_1(\boldsymbol{\Theta}),\ldots,c_k(\boldsymbol{\Theta})$, are linearly independent, the sufficient statistic is a minimal sufficient statistic.

Theorem 8.3 [Sufficiency of invertible functions of sufficient statistics]
Let $\boldsymbol{S} = \boldsymbol{s}(\boldsymbol{X})$ be an r-dimensional sufficient statistic for $f(\boldsymbol{x};\Theta)$. If $\boldsymbol{\tau}[\boldsymbol{s}(\boldsymbol{X})]$ is an r-dimensional invertible function of $\boldsymbol{s}(\boldsymbol{X})$, then

1. $\boldsymbol{\tau}[\boldsymbol{s}(\boldsymbol{X})]$ is an r-dimensional sufficient statistic for $f(\boldsymbol{x};\Theta)$;
2. if $\boldsymbol{s}(\boldsymbol{X})$ is a minimal sufficient statistic, then $\boldsymbol{\tau}[\boldsymbol{s}(\boldsymbol{X})]$ is a minimal sufficient statistic.

[17] The notation for the sample space $R_\Omega(\boldsymbol{X})$ indicates that the range of $\boldsymbol{X}$ is taken over all Θs in the parameter space Ω. If the support of the pdf does not change with Θ (e.g. Normal, Gamma, etc.) then $R_\Omega(\boldsymbol{X}) = R(\boldsymbol{X})$.

8.5 Minimum Variance Unbiased Estimation

8.5.1 Cramér-Rao Lower Bound (CRLB)

Definition 8.15 [CRLB regularity conditions (scalar case)]

1. The parameter space Ω for the parameter θ indexing the pdf $f(\boldsymbol{x};\theta)$ is an open interval with $\theta \in \Omega \subset \mathbb{R}^1$.

2. The support of $f(\boldsymbol{x};\theta)$, say A, is the same $\forall\ \theta \in \Omega$.

3. $\partial \ln f(\boldsymbol{x};\theta)/\partial\theta$ exists and is finite $\forall\ \boldsymbol{x} \in A$, and $\forall\ \theta \in \Omega$.

4. We can differentiate under the integral as follows

$$\frac{\partial}{\partial\theta}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(\boldsymbol{x};\theta)\,dx_1\cdots dx_n = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{\partial f(\boldsymbol{x};\theta)}{\partial\theta}\,dx_1\cdots dx_n.$$

5. For all unbiased estimators $t(\boldsymbol{X})$ for $q(\theta)$ with finite variance, we can differentiate under the integral as follows

$$\frac{\partial}{\partial\theta}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} t(\boldsymbol{x})\cdot f(\boldsymbol{x};\theta)\,dx_1\cdots dx_n = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} t(\boldsymbol{x})\cdot\frac{\partial f(\boldsymbol{x};\theta)}{\partial\theta}\,dx_1\cdots dx_n.$$

6. $0 < \mathrm{E}\left[\left(\frac{\partial \ln f(\boldsymbol{X};\theta)}{\partial\theta}\right)^2\right] < \infty\ \forall\ \theta \in \Omega$.

Theorem 8.4 [Cramér-Rao Lower Bound (scalar case)]
Let $X_1,\ldots,X_n$ be a random sample from a population with pdf $f(x;\theta)$ and let $T = t(\boldsymbol{X})$ be an unbiased estimator for $q(\theta)$. Then under the CRLB regularity conditions for the joint pdf $f(\boldsymbol{x},\theta)$ given above

$$\mathrm{Var}_\theta(T) \geq \frac{\left[\frac{\partial q(\theta)}{\partial\theta}\right]^2}{n\,\mathrm{E}_\theta\left[\left\{\frac{\partial}{\partial\theta}\ln f(X;\theta)\right\}^2\right]}.$$

Equality prevails iff there exists a function, say $K(\theta,n)$, such that

$$\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\ln f(x_i;\theta) = K(\theta,n)[t(\boldsymbol{x}) - q(\theta)].$$

Definition 8.16 [Information Equality]
For the joint density function $f(\boldsymbol{X};\theta)$ with $\theta \in \Theta \subset \mathbb{R}^n$ the information equality holds if

$$\mathrm{E}_\theta\left[\left\{\frac{\partial}{\partial\theta}\ln f(\boldsymbol{X},\theta)\right\}^2\right] = -\,\mathrm{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\ln f(\boldsymbol{X},\theta)\right].$$

Proposition [Exponential class and CRLB]
If T is an unbiased estimator of some $q(\theta)$ whose variance coincides with the CRLB, then the pdf $f(\boldsymbol{x};\theta)$ belongs to the exponential class; and, conversely, if $f(\boldsymbol{x};\theta)$ belongs to the exponential class, then there exists an unbiased estimator T of some $q(\theta)$ whose variance coincides with the CRLB.

8.5.2 Sufficiency and Completeness

Theorem 8.5 [Rao-Blackwell's Theorem (scalar case)]
Let $\boldsymbol{S} = (S_1, \ldots, S_r)'$ be an r-dimensional sufficient statistic for $f(\boldsymbol{x};\Theta)$, and let $T = t(\boldsymbol{X})$ be any unbiased estimator for the scalar $q(\Theta)$. Define

$$T' = t'(\boldsymbol{X}) = \mathrm{E}[T(\boldsymbol{X})|S_1, \ldots, S_r].$$

Then

1. T' is a statistic and it is a function of $S_1, \ldots, S_r$,
2. $\mathrm{E}T' = q(\Theta)$, that is T' is an unbiased estimator of $q(\Theta)$, and
3. $\mathrm{Var}(T') \leq \mathrm{Var}(T)\ \forall\ \Theta \in \Omega$, where the equality is attained only if $P(T' = T) = 1$.

Definition 8.17 [Complete sufficient statistics]
Let $\boldsymbol{S} = [S_1, \ldots, S_r]'$ be a sufficient statistic for $f(\boldsymbol{x};\Theta)$. The sufficient statistic $\boldsymbol{S}$ is said to be complete iff

$$\mathrm{E}_\Theta[z(\boldsymbol{S})] = 0\ \forall\ \Theta \in \Omega \quad \text{implies that} \quad P_\Theta[z(\boldsymbol{s}) = 0] = 1\ \forall\ \Theta \in \Omega,$$

where $z(\boldsymbol{S})$ is a statistic.

Theorem 8.6 [Completeness in the exponential class]
Let the joint density, $f(\boldsymbol{x};\boldsymbol{\Theta})$, of the random sample $(X_1, \ldots, X_n)$ be a member of a parametric family of densities belonging to the exponential class of densities with pdf

$$f(\boldsymbol{x};\boldsymbol{\Theta}) = \exp\left[\sum_{i=1}^{k} c_i(\boldsymbol{\Theta}) g_i(\boldsymbol{x}) + d(\boldsymbol{\Theta}) + z(\boldsymbol{x})\right].$$

If the range of $[c_1(\boldsymbol{\Theta}), \ldots, c_k(\boldsymbol{\Theta})]'$, $\boldsymbol{\Theta} \in \Omega$, contains an open k-dimensional rectangle[18], then $s(\boldsymbol{X}) = [g_1(\boldsymbol{X}), \ldots, g_k(\boldsymbol{X})]'$ is a complete sufficient statistic for $f(\boldsymbol{x};\boldsymbol{\Theta})$, $\boldsymbol{\Theta} \in \Omega$.

Theorem 8.7 [Lehmann-Scheffé's completeness Theorem]
Let $\boldsymbol{S} = (S_1, ..., S_r)'$ be a complete sufficient statistics for $f(\boldsymbol{x};\Theta)$. Let $T = t(\boldsymbol{S})$ be an unbiased estimator for the function $q(\Theta)$. Then $T = t(\boldsymbol{S})$ is the MVUE of $q(\Theta)$.

[18] The condition that the range of $[c_1(\boldsymbol{\Theta}), \ldots, c_k(\boldsymbol{\Theta})]'$ contains an open k-dimensional rectangle excludes cases where the $c_i(\boldsymbol{\Theta})$s are linearly dependent. For a random sample from a $\mathscr{N}(\mu, \sigma^2)$ distribution with $(\mu, \sigma^2) \in \mathbb{R}^1 \times \mathbb{R}^1_+$, for example, the range of $[c_1(\cdot), c_2(\cdot)]' = \left[\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right]$ is the set $\mathbb{R}^1 \times \mathbb{R}^1_-$ and contains an open 2-dimensional rectangle.

9 Point Estimation Methods

9.1 Least-Squares Estimators for Linear Regression Models

9.1.1 The classical LRM assumptions

Assumption 1. $\mathrm{E}[\boldsymbol{Y}] = \boldsymbol{x} \cdot \boldsymbol{\beta}$ and $\mathrm{E}[\varepsilon] = 0$.

Assumption 2. $\boldsymbol{x}$ is a non-random $n \times k$ matrix with rank $\mathrm{rk}(\boldsymbol{x}) = k$ (full column rank).

Assumption 3. $\mathrm{Cov}(\varepsilon) = \mathrm{E}[\varepsilon\varepsilon'] = \sigma^2 \boldsymbol{I}$.

9.1.2 The Least-Squares estimator for β in the classical LRM

The LS estimate of $\boldsymbol{\beta}$ denoted by $\boldsymbol{b}$ solves for given observations of the dependent and the vector of regressors $\{y_i, \boldsymbol{x}_{i.}\}_{i=1}^n$ the minimization problem

$$\boldsymbol{b} = \arg\min_{\boldsymbol{\beta}} S(\boldsymbol{\beta}), \qquad \text{where}$$

$$S(\beta) = \sum_{i=1}^{n} (y_i - \boldsymbol{x}_{i.}\boldsymbol{\beta})^2 = (\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}) = \boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{\beta}'\boldsymbol{x}'\boldsymbol{y} + \boldsymbol{\beta}'\boldsymbol{x}'\boldsymbol{x}\boldsymbol{\beta}.$$

The k first-order conditions for a minimum can be represented as

$$\frac{\partial S(\boldsymbol{b})}{\partial \boldsymbol{\beta}} = -2\boldsymbol{x}'\boldsymbol{y} + 2\boldsymbol{x}'\boldsymbol{x}\boldsymbol{b} \quad = \quad 0,$$

such that

$$\boldsymbol{x}'\boldsymbol{x}\boldsymbol{b} \quad = \quad \boldsymbol{x}'\boldsymbol{y}$$

form the k **LS normal equations** resulting in the least-squares estimator

$$\boldsymbol{b} \quad = \quad (\boldsymbol{x}'\boldsymbol{x})^{-1}\boldsymbol{x}'\boldsymbol{y}.$$

Note that $\boldsymbol{x}$ is assumed to have full rank (Assumption 2). This implies that the $(k \times k)$ matrix $\boldsymbol{x}'\boldsymbol{x}$ has full rank and is thus invertible.

9.1.3 Properties of the LS estimator in the classical LRM

Theorem 9.1 [Gauss-Markov Theorem]
Under the classical assumptions of the LRM, $\hat{\boldsymbol{\beta}} = (\boldsymbol{x}'\boldsymbol{x})^{-1}\boldsymbol{x}'\boldsymbol{Y}$ is the best linear unbiased estimator of $\boldsymbol{\beta}$.

Theorem 9.2 [Consistency of $\hat{\beta}$]
Under the classical assumptions of the LRM, if

$$(\boldsymbol{x}'\boldsymbol{x})^{-1} \to 0 \quad \text{as} \quad n \to \infty,$$

then $\hat{\boldsymbol{\beta}} = (\boldsymbol{x}'\boldsymbol{x})^{-1}\boldsymbol{x}'\boldsymbol{Y} \xrightarrow{p} \boldsymbol{\beta}$, so that $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$.

Theorem 9.3 [Consistency of S^2 - iid case]
Under the classical assumptions of the LRM, if the disturbances ε_i are iid, then $\hat{S}^2 \xrightarrow{p} \sigma^2$, so that $\hat{S}^2$ is a consistent estimator of σ^2.

Theorem 9.4 [Asymptotic Normality of $\hat{\beta}$ - iid case]
Assume the classical assumptions of the LRM. In addition, assume that

1. the ε_is are iid with $P(|\varepsilon_i| < m) = 1$ for $m < \infty$ and $\forall i$,
2. the regressors are such that $|x_{ij}| < \xi < \infty$ $\forall i$ and j, and
3. $\lim_{n\to\infty} n^{-1}\boldsymbol{x}'\boldsymbol{x} = \boldsymbol{Q}$, where $\boldsymbol{Q}$ is a finite, positive-definite matrix.

Then

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(0, \sigma^2\boldsymbol{Q}^{-1}) \quad \text{and} \quad \hat{\boldsymbol{\beta}} \overset{a}{\sim} \mathcal{N}(\boldsymbol{\beta}, n^{-1}\sigma^2\boldsymbol{Q}^{-1}).$$

Theorem 9.5 [Asymptotic Normality of $\hat{S}^2$ - iid case]
Under the classical assumptions of the LRM, if the ε_is are iid, and if $\mathrm{E}[\varepsilon_i^4] = \mu_4' \leq \tau < \infty$, then

$$\sqrt{n}(\hat{S}^2 - \sigma^2) \xrightarrow{d} \mathcal{N}(0, \mu_4' - \sigma^4) \quad \text{and} \quad \hat{S}^2 \overset{a}{\sim} \mathcal{N}(\sigma^2, n^{-1}[\mu_4' - \sigma^4]).$$

Theorem 9.6
Under the classical assumptions of the LRM, if $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2\boldsymbol{I})$, then

1. $\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{x}'\boldsymbol{x})^{-1})$,
2. $(n-k)\hat{S}^2/\sigma^2 \sim \chi^2_{(n-k)}$,
3. $\hat{\boldsymbol{\beta}}$ and $\hat{S}^2$ are independent.

Theorem 9.7 [MVUE Property of $(\hat{\boldsymbol{\beta}}, \hat{S}^2)$ Under Normality]
Assume the classical assumptions of the LRM, and assume that $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2\boldsymbol{I})$. Then $(\hat{\boldsymbol{\beta}}, \hat{S}^2)$ is the MVUE for $(\boldsymbol{\beta}, \sigma^2)$.

9.2 Method of Maximum Likelihood

9.2.1 The Likelihood function and the ML estimator

The ML method leads to an estimate of the parameter Θ or $\boldsymbol{q}(\Theta)$ by maximizing the **likelihood function** of the parameters, given the outcome of the random sample.

The likelihood function is identical in functional form to the joint pdf of the random sample.

In particular, let $f(\boldsymbol{x};\Theta)$ denote the joint pdf of the random sample variables $\boldsymbol{X} = (X_1, \ldots, X_n)$ indexed by the unknown parameter $\Theta \in \Omega$, then the likelihood function is defined as

$$L(\Theta;\boldsymbol{x}) \equiv f(\boldsymbol{x};\Theta).$$

Note that we write the joint pdf as a function in the data $\boldsymbol{x}$ indexed/conditioned on the parameter Θ, whereas when we form the likelihood, we write this function in reverse, as a function in the parameters Θ for given values of the data $\boldsymbol{x}$.

The **maximum likelihood (ML) estimator** $\hat{\boldsymbol{\theta}}$ is obtained as the value of Θ that maximizes the likelihood function. Thus

$$\hat{\boldsymbol{\theta}} = \arg\max_{\Theta\in\Omega} L(\Theta;\boldsymbol{x}).$$

The ML-method can be interpreted

> as choosing, from all candidates, the value of Θ indexing the joint pdf $f(\boldsymbol{x};\Theta)$ that assigns the **highest probability** (discrete case) or **highest density weighting** (continuous case) to the random sample outcome, $\boldsymbol{x}$, actually observed.
>
> Put another way, the ML estimate $\hat{\boldsymbol{\theta}}$ defines a particular member of a parametric family of pdfs $\{f(\boldsymbol{x};\Theta), \Theta \in \Omega\}$ that assigns the highest 'likelihood' to generating the data actually observed.

9.2.2 Finite sample properties of the ML estimator

Theorem 9.8 [MLE Attainment of the CRLB]
If there exists an unbiased estimator, $\boldsymbol{T} = \boldsymbol{t}(\boldsymbol{X})$, of Θ that has a covariance matrix equal to the CRLB, and if the MLE can be defined by solving the f.o.c. for maximizing the likelihood function, then the MLE is equal to $\boldsymbol{T} = \boldsymbol{t}(\boldsymbol{X})$ with probability 1.

Theorem 9.9 [Unique MLEs are Functions of Any Sufficient Statistics for $f(\boldsymbol{x};\Theta)$]
Assume that the MLE of Θ, say $\hat{\boldsymbol{\theta}}$, is uniquely defined in terms of $\boldsymbol{X}$. If $\boldsymbol{S} = [S_1,\ldots,S_r]'$ is any vector of sufficient statistics for $f(\boldsymbol{x};\Theta) \equiv L(\Theta;\boldsymbol{x})$, then there exists a function of $\boldsymbol{S}$, say $\boldsymbol{\tau}(\boldsymbol{S})$, such that $\hat{\boldsymbol{\theta}} = \boldsymbol{\tau}(\boldsymbol{S})$.

9.2.3 Large sample properties of the ML estimator

Theorem 9.10 [MLE Consistency - iid and scalar case]
Let $X_1,\ldots,X_n$ be an iid random sample from a population with pdf $f(x,\theta)$, where $\theta \in \Omega$ is a scalar. Assume that

R1. the set of joint pdfs $\{f(\boldsymbol{x};\theta), \theta \in \Omega\}$, have common support, Ξ,

R2. the parameter space, Ω, is an open interval,

R3. $\ln L(\theta;\boldsymbol{x})$ is continuously differentiable w.r.t. $\theta \in \Omega \; \forall \boldsymbol{x} \in \Xi$, and

R4. $\partial \ln L(\theta;\boldsymbol{x})/\partial\theta = 0$ has a unique solution for $\theta \in \Omega$, and the solution defines the unique ML estimate, $\hat{\theta}(\boldsymbol{x}), \forall \boldsymbol{x} \in \Xi$.

Then $\hat{\Theta} \xrightarrow{p} \theta_0$ (true value of θ), and the MLE is thus consistent for θ.

Note that the fairly restrictive iid assumption is not needed if we add to the list of the 4 regularity conditions in Theorem 9.10 the (fairly weak) condition

R5. $\lim_{n\to\infty} \mathrm{P}[\ln L(\theta_0;\boldsymbol{x}) > \ln L(\theta;\boldsymbol{x})] = 1$ for $\theta \neq \theta_0$.

This condition essentially requires that the likelihood is such that as $n \to \infty$ the true value θ_0 maximizes the likelihood (and hence satisfies the definition of the ML-estimate) with probability 1. For further details see Mittelhammer (1996, Theorem 8.14.).

Theorem 9.11 [MLE Consistency - Sufficient Conditions]
Let $\{f(\boldsymbol{x};\Theta), \Theta \in \Omega\}$ be the statistical model for the random sample $\boldsymbol{X}$. Let $N(\varepsilon) = \{\Theta : d(\Theta,\Theta_0) < \varepsilon\}$ be an open ε-neighbourhood of Θ_0 (true value of Θ).[19] Assume

M1. the pdfs $f(\boldsymbol{x};\Theta)$, $\Theta \in \Omega$, have common support, Ξ,

M2. $\ln L(\Theta;\boldsymbol{x})$ has continuous first-order partial derivatives w.r.t. $\Theta \in \Omega \; \forall \boldsymbol{x} \in \Xi$,

M3. $\partial \ln L(\Theta;\boldsymbol{x})/\partial\Theta = 0$ has a unique solution that defines the unique ML estimate $\hat{\Theta} = \operatorname{argmax}_{\Theta\in\Omega} L(\Theta;\boldsymbol{x}) \; \forall \boldsymbol{x} \in$ Ξ, and

M4. $\lim_{n\to\infty} P[\ln L(\Theta_0;\boldsymbol{x}) > \max_{\Theta\in\overline{N(\varepsilon)}} \ln L(\Theta)] = 1 \; \forall \varepsilon > 0$ with Ω being an open rectangle containing Θ_0.

Then the MLE, $\hat{\boldsymbol{\Theta}}$ is such that

$$\hat{\boldsymbol{\Theta}} \xrightarrow{p} \boldsymbol{\theta}_0.$$

[19] $N(\varepsilon)$ is an open interval, the interior of a circle, the interior of a sphere, and the interior of a hypersphere in 1, 2, 3, and ≥ 4 dimensions, respectively.

Theorem 9.12 [MLE Asymptotic Normality - iid and scalar case]
In addition to conditions (R1)-(R4) of Theorem 9.10 assume that

R6. $\partial^2 \ln L(\theta;\boldsymbol{x})/\partial\theta^2$ exists and is continuous in θ $\forall\theta \in \Omega$ and $\forall \boldsymbol{x} \in \Xi$, and

R7. $\operatorname{plim}[\frac{1}{n}(\partial^2 \ln L(\Theta^\star;\boldsymbol{X})/\partial\theta^2 = H(\theta_0) \neq 0$ for any sequence of random variables $\{\Theta_n^\star\}$ such that $\operatorname{plim}\Theta_n^\star = \theta_0$.

Then the MLE, $\hat{\Theta}$ is such that

$$\sqrt{n}(\hat{\Theta}-\theta_0) \xrightarrow{d} \mathscr{N}\left(0,\ \frac{\mathrm{E}\left[(\partial \ln f(X_i;\theta_0)/\partial\theta)^2\right]}{H(\theta_0)^2}\right),$$
$$\hat{\Theta} \overset{a}{\sim} \mathscr{N}\left(\theta_0,\ \frac{1}{n}\frac{\mathrm{E}\left[(\partial \ln f(X_i;\theta_0)/\partial\theta)^2\right]}{H(\theta_0)^2}\right).$$

Theorem 9.13 [MLE Asymptotic Normality - Sufficient Conditions]
In addition to conditions (M1)-(M4) of Theorem 9.11, assume

M5. $\partial^2 \ln L(\Theta;\boldsymbol{x})/\partial\Theta\partial\Theta'$ exists and is continuous in Θ $\forall\Theta \in \Omega$ and $\forall \boldsymbol{x} \in \Xi$;

M6. $\operatorname{plim}\left[\frac{1}{n}\left(\partial^2 \ln L(\Theta^\star;\boldsymbol{X})/\partial\Theta\partial\Theta'\right)\right] = \boldsymbol{H}(\Theta_0)$ is a nonsingular matrix for any sequence of random variables $\{\Theta_n^\star\}$ such that $\operatorname{plim}\Theta_n^\star = \Theta_0$;

M7. $n^{-1/2}[\partial \ln L(\Theta_0;\boldsymbol{X})]/\partial\Theta \xrightarrow{d} \mathscr{N}(0,\boldsymbol{M}(\Theta_0))$ where $\boldsymbol{M}(\Theta_0)$ is a symmetric, p.d. matrix.

Then the MLE, $\hat{\boldsymbol{\Theta}}$, is such that

$$\sqrt{n}\left(\hat{\boldsymbol{\Theta}}-\Theta_0\right) \xrightarrow{d} \mathscr{N}\left(0,\boldsymbol{H}(\Theta_0)^{-1}\boldsymbol{M}(\Theta_0)\boldsymbol{H}(\Theta_0)^{-1}\right).$$

9.2.4 MLE invariance principle

Theorem 9.14 [MLE Invariance Property - scalar case[20]]
Let $\hat{\Theta}$ be a MLE of the scalar θ, and let $q(\theta)$ be a real-valued function of θ. Then $q(\hat{\Theta})$ is a MLE of $q(\theta)$.

9.3 The Method of Moments

9.3.1 Moment Conditions and Method of Moments Estimator

Definition 9.1 [Moment Conditions]
Let $\boldsymbol{Y} = (Y_1,\ldots,Y_n)$ be a random sample from a statistical model $\{f(\boldsymbol{y};\Theta), \Theta \in \Omega \subseteq \mathbb{R}^k\}$ with true parameter value Θ_0. Let $\boldsymbol{g}(Y_t;\Theta)$ be a continuous ℓ-dimensional vector function of Θ with $\ell \geq k$ such that

$$\mathrm{E}\boldsymbol{g}(Y_t,\Theta_0) = 0, \quad t = 1,\ldots,n.$$

This set of ℓ equations are called moment conditions and the vector function $\boldsymbol{g}$ is called moment function.

[20] For a multivariate version of this theorem see Mittelhammer (1996, Theorem 8.20).

9.3.2 Properties of the MM Estimator

Theorem 9.15 [Consistency of MM Estimator]
Let the MM estimator $\hat{\Theta}_{(k\times 1)} = \boldsymbol{h}^{-1}(M_1', \ldots, M_k')$ be such that $\boldsymbol{h}^{-1}(\mu_1', \ldots, \mu_k')$ is continuous $\forall(\mu_1', \ldots, \mu_k') \in \Gamma = \{(\mu_1', \ldots, \mu_k') : \mu_i' = \boldsymbol{h}_i(\Theta),\ i = 1, \ldots, k,\ \Theta \in \Omega\}$. Then $\hat{\Theta} \xrightarrow{p} \Theta$.

Theorem 9.16 [Asymptotic Normality of MM Estimator]
Let the MM estimator $\hat{\Theta} = \boldsymbol{h}^{-1}(M_1', \ldots, M_k')$ such that $\boldsymbol{h}^{-1}(\mu_1', \ldots, \mu_k')$ is differentiable $\forall(\mu_1', \ldots, \mu_k') \in \Gamma = \{(\mu_1', \ldots, \mu_k' : \mu_i' = \boldsymbol{h}_i(\Theta),\ i = 1, \ldots, k,\ \Theta \in \Omega\}$, and let the elements of the Jakobian

$$\boldsymbol{A}(\mu_1', \ldots, \mu_k') = \begin{bmatrix} \frac{\partial h_1^{-1}(\mu_1', \ldots, \mu_k')}{\partial \mu_1'} & \cdots & \frac{\partial h_1^{-1}(\mu_1', \ldots, \mu_k')}{\partial \mu_k'} \\ \vdots & \ddots & \vdots \\ \frac{\partial h_k^{-1}(\mu_1', \ldots, \mu_k')}{\partial \mu_1'} & \cdots & \frac{\partial h_k^{-1}(\mu_1', \ldots, \mu_k')}{\partial \mu_k'} \end{bmatrix}$$

be continuous functions with $\boldsymbol{A}(\mu_1', \ldots, \mu_k')$ having full rank $\forall(\mu_1', \ldots, \mu_k') \in \Gamma$. Then

$$\sqrt{n}(\hat{\Theta} - \Theta) \xrightarrow{d} \mathscr{N}(0, \boldsymbol{A}\Sigma\boldsymbol{A}'), \quad \text{and} \quad \hat{\Theta} \overset{a}{\sim} \mathscr{N}(\Theta, \frac{1}{n}\boldsymbol{A}\Sigma\boldsymbol{A}'),$$

where $\underset{k\times k}{\Sigma} = \mathrm{Cov}(M_1', \ldots, M_k')$.

9.3.3 Generalized Method of Moments (GMM) Estimator

The GMM estimator is used when the k-dimensional parameter vector Θ is over-identified by the $\ell > k$ moment conditions $\mathrm{E}\boldsymbol{g}(Y_t; \Theta) = 0$. In this case the corresponding sample moment conditions

$$\boldsymbol{g}_n(\boldsymbol{y}; \Theta) = \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{g}(Y_t; \Theta) = 0$$

is a system with more equations than unknowns, such that we cannot find a vector $\hat{\theta}$ that exactly satisfies the sample moment conditions. In this over-identified case the GMM estimate is defined as the value of Θ that satisfies the sample moment conditions as closely as possible.

9.3.4 GMM Properties

Definition 9.2 [GMM Consistency Conditions]

C1. The expectation of the moment function $\boldsymbol{g}(Y_t, \Theta)$ used to define the moment conditions

$$\mathrm{E}\boldsymbol{g}(Y_t, \Theta) \overset{\text{(say)}}{=} \boldsymbol{h}(\Theta) \quad \text{exists and is finite} \quad \forall \Theta \in \Omega.$$

C2. There exists a $\Theta_0 \in \Omega$ such that

$$\mathrm{E}\boldsymbol{g}(Y_t, \Theta) = 0 \quad \Leftrightarrow \quad \Theta = \Theta_0.$$

C3. Let $g_{n,j}(\boldsymbol{Y}, \Theta)$ be the jth sample moment in $\boldsymbol{g}_n(\boldsymbol{Y}, \Theta)$ and $h_j(\Theta)$ the corresponding population moment in $\boldsymbol{h}(\Theta)$. Then

$$\sup_{\Theta \in \Omega} |g_{n,j}(\boldsymbol{Y}, \Theta) - h_j(\Theta)| \xrightarrow{p} 0 \quad \text{for} \quad j = 1, \ldots, \ell.$$

Theorem 9.17 [Consistency of GMM Estimator]
The GMM estimator of Θ defined as

$$\hat{\Theta} = \arg\min_{\Theta \in \Omega} Q_n(\Theta; \boldsymbol{Y}),$$

where

$$Q_n(\Theta; \boldsymbol{Y}) = \boldsymbol{g}_n(\boldsymbol{Y}; \Theta)' \cdot \boldsymbol{W}_n \cdot \boldsymbol{g}_n(\boldsymbol{Y}; \Theta),$$

and where $\boldsymbol{W}_n \xrightarrow{p} \boldsymbol{w}$, with $\boldsymbol{w}$ being a nonrandom, symmetric, and p.d. matrix. Then under the conditions (C1) to (C3), $\hat{\Theta} \xrightarrow{p} \Theta_0$.

Definition 9.3 [GMM Asymptotic Normality Conditions]

C4. The sample moment $\boldsymbol{g}_n(\boldsymbol{y}; \Theta) = \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{g}(y_t; \Theta)$ is continuously differentiable w.r.t. Θ $\forall \Theta \in \Omega$, with a Jakobian matrix

$$\boldsymbol{G}_n(\boldsymbol{y}; \Theta) = \frac{\partial \boldsymbol{g}_n(\boldsymbol{y}; \Theta)}{\partial \Theta'} = \frac{1}{n}\sum_{t=1}^{n} \frac{\partial \boldsymbol{g}(y_t; \Theta)}{\partial \Theta'}.$$

C5. For any sequence $\{\Theta_n^\star\}$ such that $\Theta_n^\star \xrightarrow{p} \Theta_0$,

$$\boldsymbol{G}_n(\boldsymbol{Y}; \Theta_n^\star) \xrightarrow{p} \boldsymbol{G}(\Theta_0),$$

where $\boldsymbol{G}(\Theta_0)$ is a nonrandom $\ell \times k$ matrix.

C6. The sequence of moment functions $\{\boldsymbol{g}(Y_t; \Theta)\}$ satisfies a CLT, so that

$$\sqrt{n}\boldsymbol{g}_n(\boldsymbol{Y}; \Theta) \xrightarrow{d} \boldsymbol{Z} \sim \mathcal{N}[0, \boldsymbol{V}(\Theta_0)],$$

where $\boldsymbol{V}(\Theta_0) = n \cdot \mathrm{Cov}[\boldsymbol{g}_n(\boldsymbol{Y}; \Theta_0)]$.

Theorem 9.18 [Asymptotic Normality of GMM Estimator]
Under the conditions (C1) to (C6), the GMM estimator $\hat{\Theta}$ of Θ defined in Theorem 9.17 is such that

$$\begin{aligned}\sqrt{n}(\hat{\Theta} - \Theta_0) \xrightarrow{d} \mathcal{N}\,(\,0,\ & [\boldsymbol{G}(\Theta_0)'\boldsymbol{w}\boldsymbol{G}(\Theta_0)]^{-1} \\ & \times \boldsymbol{G}(\Theta_0)'\boldsymbol{w}\boldsymbol{V}(\Theta_0)\boldsymbol{w}\boldsymbol{G}(\Theta_0)[\boldsymbol{G}(\Theta_0)'\boldsymbol{w}\boldsymbol{G}(\Theta_0)]^{-1}\,)\end{aligned}$$

9.4 Bayesian Estimation

9.4.1 Prior and Posterior Distribution

Definition 9.4 [Posterior Bayes Estimate]
Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ be a random sample from the joint pdf $f(\boldsymbol{y}|\Theta)$ and $f(\Theta)$ the prior density for Θ. The posterior Bayes estimator of $\boldsymbol{q}(\Theta)$ is defined to be

$$\mathrm{E}[\boldsymbol{q}(\Theta)|\boldsymbol{y}] = \int \boldsymbol{q}(\Theta) f(\Theta|\boldsymbol{y}) d\Theta = \frac{\int \boldsymbol{q}(\Theta) f(\boldsymbol{y}|\Theta) f(\Theta) d\Theta}{\int f(\boldsymbol{y}|\Theta) f(\Theta) d\Theta}.$$

9.4.2 Loss-Function Approach

Definition 9.5 [Bayes Estimator]
The Bayes estimator of $q(\Theta)$ is that estimator which has for a given loss function and a given prior distribution for Θ the smallest Bayes risk.

Theorem 9.19
Let $\ell(t;\Theta) \geq 0$ be the loss function for estimating $q(\Theta)$ and let $f(\Theta|\boldsymbol{y})$ denote the posterior density obtained from the prior $f(\Theta)$ with domain Ω and the likelihood $f(\boldsymbol{y}|\Theta)$ with domain Ξ. Then, the Bayes estimator is that estimator $T_\star$ which minimizes the posterior risk

$$\mathrm{E}[\ell(T;\Theta)|\boldsymbol{y}] = \int_\Omega \ell(t;\Theta) f(\Theta|\boldsymbol{y}) d\Theta$$

Corollary 9.1
Under a quadratic loss function $\ell(t;\Theta)$, the Bayes estimator of $q(\Theta)$ is given by the posterior expectation of $q(\Theta)$, i.e.

$$\mathrm{E}\left[q(\Theta)|\boldsymbol{y}\right] = \int q(\Theta) f(\Theta|\boldsymbol{y}) d\Theta.$$

10 Hypothesis Testing

10.1 Fundamental Notations and Terminology of Hypotheses Testing

Definition 10.1 [Statistical hypothesis]
A set of potential probability distributions for a random sample from a population is called a *statistical hypothesis*.
If the statistical hypothesis completely and uniquely identifies the probability distribution, the hypothesis is called *simple*.
If the statistical hypothesis contains two or more potential probability distributions, the hypothesis is called *composite*.

Definition 10.2 [Statistical hypothesis test]
A statistical hypothesis test is a rule, based on a random sample outcome $\boldsymbol{x}$, used to decide whether or not to reject a hypothesis H.

Definition 10.3 [Critical region]
A subset C_r of the sample range such that if $\boldsymbol{x} \in C_r$, then the hypothesis H is rejected is called the critical region or rejection region. (The complement of C_r is the acceptance region C_a with $C_r \cap C_a = \emptyset$).

Definition 10.4 [Type I and type II error]
The **type I error** of a test for the hypotheses H is the random event that H is rejected when H is true, i.e. the event

$$\{\boldsymbol{x} \in C_r \quad \text{and} \quad H \text{ is true}\}.$$

The **type II error** is the random event that H is accepted when H is false, i.e.

$$\{\boldsymbol{x} \notin C_r \quad \text{and} \quad H \text{ is not true}\}.$$

Definition 10.5 [Test statistic]
Let C_r define the critical region associated with a test of the hypothesis H versus $\bar{H}$. If $T = t(\boldsymbol{X})$ is a scalar statistic such that $C_r = \{\boldsymbol{x} : t(\boldsymbol{x}) \in C_r^T\}$, i.e., the critical region can be defined in terms of outcomes, C_r^T, of the statistic T, then T is referred to as a test statistic for the hypothesis H versus $\bar{H}$. The set C_r^T will be referred to as the critical (or rejection) region of the test statistic, T.

10.2 Parametric Tests and Test Properties

Definition 10.6 [Power function]
Let C_r be the critical region of a test of $H : \boldsymbol{\Theta} \in \Omega_H \subseteq \Omega$, where $\boldsymbol{\Theta}$ indexes a parametric family of densities $\{f(\boldsymbol{x};\boldsymbol{\Theta}), \boldsymbol{\Theta} \in \Omega = \Omega_H \cup \Omega_{\bar{H}}\}$. Then the **power function** of the test is defined by

$$\begin{aligned} \pi(\boldsymbol{\Theta}) = P(\boldsymbol{x} \in C_r;\boldsymbol{\Theta}) \quad &\equiv \quad \underset{\boldsymbol{x}\in C_r}{\int\cdots\int} f(\boldsymbol{x};\boldsymbol{\Theta})\,dx \quad \text{(continuous case)} \\ &\equiv \quad \underset{\boldsymbol{x}\in C_r}{\sum\cdots\sum} f(\boldsymbol{x};\boldsymbol{\Theta}) \qquad \text{(discrete case)} \end{aligned}$$

Definition 10.7 [Size of test]
Let $\pi(\boldsymbol{\Theta})$ be the power function of a test C_r for the hypothesis H. Then

$$\alpha = \sup_{\Theta \in H} \pi(\boldsymbol{\Theta}) = \sup_{\Theta \in H} P(\boldsymbol{x} \in C_r;\boldsymbol{\Theta})$$

is called the size of the test C_r .

Definition 10.8 [Significance level of test]
A test of significance level α is any test for which

$$P(\text{type I error}) = P(\boldsymbol{x} \in C_r | \Theta \in H) \leq \alpha.$$

Definition 10.9 [Unbiasedness of a test]
Let $\pi(\Theta)$ be the power function of a test for the hypothesis H. The test is called unbiased iff

$$\sup_{\Theta \in H} \pi(\Theta) \leq \inf_{\Theta \in \bar{H}} \pi(\Theta).$$

Definition 10.10 [Uniformly most powerful (UMP) size-α test]
Let $\Xi = \{C_r : \sup_{\Theta \in H} \pi(\Theta) \leq \alpha\}$ be the set of all critical regions with a size of at most α for the hypothesis H. A test with critical region $C_r^\star$ and with a power function $\pi c_r^\star(\Theta)$ is called **uniformly most powerful of size α** iff[21]

$$\begin{array}{rrcl} & \sup_{\Theta \in H} \pi c_r^\star & = & \alpha, \\ \text{and} & \pi c_r^\star(\Theta) & \geq & \pi c_r(\Theta) \quad \forall\, \Theta \in \bar{H} \text{ and } \quad \forall\, C_r \in \Xi. \end{array}$$

Definition 10.11 [Admissibility of test]
Let C_r be a test of H. If there exists an alternative test $C_r^\star$ such that

$$\pi c_r^\star \begin{Bmatrix} \geq \\ \leq \end{Bmatrix} \pi c_r(\Theta) \quad \forall\, \Theta \in \begin{Bmatrix} \bar{H} \\ H \end{Bmatrix},$$

with strict inequality holding for some $\Theta \in H \cup \bar{H}$, then C_r is **inadmissible**.

Definition 10.12 [Consistency of test]
Let C_{rn} be a sequence of tests of H based on a random sample $(X_1, \ldots, X_n)$. Let the significance level of the test C_{rn} be $\alpha\ \forall\, n$. Then the sequence of tests C_{rn} is said to be a **consistent sequence of significance level-α** tests iff

$$\lim_{n \to \infty} \pi c_{rn}(\Theta) = 1 \quad \forall\, \Theta \in \bar{H}.$$

10.3 Construction of UMP Tests

Theorem 10.1 [Neyman-Pearson Lemma]
Let $\boldsymbol{X}$ be a random sample from $f(\boldsymbol{x};\theta)$. Furthermore, let $k > 0$ be a positive constant and C_r a critical region which satisfy

(1) $p(\boldsymbol{x} \in C_r; \theta_0) = \alpha, \qquad 0 < \alpha < 1;$

(2) $\frac{f(x;\theta_0)}{f(x;\theta_1)} \leq k \qquad \forall \boldsymbol{x} \in C_r;$

(3) $\frac{f(x;\theta_0)}{f(x;\theta_1)} > k \qquad \forall \boldsymbol{x} \notin C_r.$

Then C_r is the most powerful critical region of size α for testing the hypothesis $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$.

Theorem 10.2 [UMP Test of a simple H_0 versus a composite H_1]
Let $\boldsymbol{X}$ be a random sample from $f(\boldsymbol{x};\theta)$. Furthermore, let $\{k(\theta_1) > 0\}$ be a sequence of positive constants with $\theta_1 \in \Omega_1$ and C_r a critical region which satisfy

[21] Mittelhammer (1996, p.534) defines the uniformly most powerful of **level** α instead of **size** α. This amounts to replacing the first condition $\sup_{\Theta \in H} \pi c_r^\star(\Theta) = \alpha$ by $\sup_{\Theta \in H} \pi c_r^\star(\Theta) \leq \alpha$.

(1') $p(\boldsymbol{x} \in C_r; \theta_0) = \alpha, \qquad 0 < \alpha < 1;$

(2') $\frac{f(\boldsymbol{x};\theta_0)}{f(\boldsymbol{x};\theta_1)} \leq k(\theta_1) \qquad \forall \boldsymbol{x} \in C_r$ and $\forall \theta_1 \in \Omega_1$;

(3') $\frac{f(\boldsymbol{x};\theta_0)}{f(\boldsymbol{x};\theta_1)} > k(\theta_1) \qquad \forall \boldsymbol{x} \notin C_r$ and $\forall \theta_1 \in \Omega_1$.

Then C_r is the uniformly most powerful critical region of size α for testing the hypothesis $H_0 : \theta = \theta_0$ versus $H_1 : \theta \in \Omega_1$.

Definition 10.13 [Monotone likelihood ratio]
A family of density functions $\{f(\boldsymbol{x};\theta), \theta \in \Omega\}$ is said to have a monotone likelihood ratio in the statistic $T = t(\boldsymbol{X})$ iff

$$\forall\ \theta_1 > \theta_2 \quad \text{the likelihood ratio} \quad \frac{L(\theta_1;\boldsymbol{x})}{L(\theta_2;\boldsymbol{x})} = \frac{f(\boldsymbol{x};\theta_1)}{f(\boldsymbol{x};\theta_2)}$$

can be expressed as a **nondecreasing function** of $t(\boldsymbol{x})\ \forall \boldsymbol{x}$.

Theorem 10.3 [Monotone likelihood ratio and the exponential class]
Let $\{f(\boldsymbol{x};\theta), \theta \in \Omega\}$, be a density family belonging to the exponential class of densities, as

$$f(\boldsymbol{x};\theta) = \exp\{c(\theta)g(\boldsymbol{x}) + d(\theta) + z(\boldsymbol{x})\}.$$

If $c(\theta)$ is a nondecreasing function of θ, then $\{f(\boldsymbol{x};\theta),\ \theta \in \Omega\}$, has a monotone likelihood ratio in the statistic $g(\boldsymbol{X})$.

Theorem 10.4 [Monotone likelihood ratios and UMP size-α tests]
Let $\{f(\boldsymbol{x};\theta),\ \theta \in \Omega\}$, be a density family having a monotone likelihood ratio in the statistic $t(\boldsymbol{X})$. Then

(1.)
$$C_r = \{\boldsymbol{x} : t(\boldsymbol{x}) \geq k\}, \text{ where } k \text{ is such that } P(t(\boldsymbol{x}) \geq k; \theta_0) = \alpha$$

is the a UMP size-α test for $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$;

(2.)
$$C_r = \{\boldsymbol{x} : t(\boldsymbol{x}) \leq k\}, \text{ where } k \text{ is such that } P(t(\boldsymbol{x}) \leq k; \theta_0) = \alpha$$

is the a UMP size-α test for $H_0 : \theta \geq \theta_0$ versus $H_1 : \theta < \theta_0$.

10.4 Hypothesis-Testing Methods

10.4.1 Likelihood Ratio Tests

Definition 10.14 [Likelihood ratio test]
Let $L(\Theta;\boldsymbol{x})$ be the likelihood function for a sample $\boldsymbol{X} = (X_1, \ldots, X_n)$. The **generalized likelihood ratio** (GLR) is defined as

$$\lambda(\boldsymbol{x}) = \frac{\sup_{\Theta \in H_0} L(\Theta;\boldsymbol{x})}{\sup_{\Theta \in H_0 \cup H_1} L(\Theta;\boldsymbol{x})}.$$

A likelihood ratio test for testing H_0 versus H_1 is given by the critical region

$$C_r = \{x : \lambda(\boldsymbol{x}) \leq c\}.$$

For a size α test, the constant c is chosen to satisfy

$$\sup_{\Theta \in H_0} \pi(\Theta) = \sup_{\Theta \in H_0} P(\lambda(\boldsymbol{x}) \leq c; \Theta) = \alpha.$$

Theorem 10.5 [Equivalence of LR and Neyman-Pearson MP test when H_0 and H_1 are simple]
Suppose a size-α LR test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ exists with critical region

$$C_r^{LR} = \{\boldsymbol{x} : \lambda(\boldsymbol{x}) \leq c\}, \quad \text{where} \quad P(\boldsymbol{x} \in C_r^{LR}; \theta_0) = \alpha \in (0,1).$$

Furthermore, suppose a Neyman-Pearson most powerful size-α test also exists with critical region

$$C_r = \{\boldsymbol{x} : L(\theta_0; \boldsymbol{x})/L(\theta_1; \boldsymbol{x}) \leq k\}, \quad \text{where} \quad P(\boldsymbol{x} \in C_r; \theta_0) = \alpha.$$

Then the LR test and the Neyman-Pearson most powerful test are equivalent.

Theorem 10.6 [Equivalence of LR and Neyman-Pearson MP test when H_0 is simple and H_1 is composite]
Consider the hypotheses $H_0 : \theta = \theta_0$ versus $H_1 : \theta \in \Omega_1$ and suppose the given critical region,

$$C_r^{LR} = \{\boldsymbol{x} : \lambda(\boldsymbol{x}) \leq c\}$$

of the LR test defines a size $\alpha \in (0,1)$ test.
Furthermore, suppose that $\forall \theta_1 \in H_1 \; \exists \; c(\theta_1) \geq 0$ such that

$$C_r^{LR} = \{\boldsymbol{x} : \lambda_{\theta_1}(\boldsymbol{x}) \leq c(\theta_1)\},$$

where

$$\lambda_{\theta_1}(\boldsymbol{x}) = \frac{L(\theta_0; \boldsymbol{x})}{\max_{\theta \in \{\theta_0, \theta_1\}} L(\theta; \boldsymbol{x})} \quad \text{and} \quad P(\lambda_{\theta_1}(\boldsymbol{x}) \leq c(\theta_1); \theta_0) = \alpha.$$

Finally, suppose a Neyman-Pearson UMP test C_r of H_0 versus H_1 having size α exists. Then the size-α LR test C_r^{LR} and the size-α Neyman-Pearson UMP test C_r are equivalent.

Theorem 10.7 [Asymptotic distribution of the GLR when H_0 is true]
Assume that the MLE of the $(k \times 1)$ vector Θ is consistent, asymptotically normal and asymptotically efficient. Let

$$\lambda(\boldsymbol{x}) = \frac{\sup_{\Theta \in H_0} L(\Theta; \boldsymbol{x})}{\sup_{\Theta \in H_0 \cup H_1} L(\Theta; \boldsymbol{x})}$$

be the GLR statistic for testing $H_0 : \mathrm{R}(\Theta) = \boldsymbol{r}$ versus $H_1 : \mathrm{R}(\Theta) \neq \boldsymbol{r}$, where $\mathrm{R}(\Theta)$ is a $(q \times 1)$ continuously differentiable vector function having nonredundant coordinate functions and $q \leq k$. Then, when H_0 is true,

$$-2 \ln \lambda(\boldsymbol{X}) \xrightarrow{d} \chi_q^2.$$

10.4.2 Lagrange Multiplier (LM) Tests

Theorem 10.8 [Asymptotic distribution of the LM test statistic when H_0 is true]
Assume that the MLE of the $(k \times 1)$ vector Θ is consistent, asymptotically normal and asymptotically efficient. Let $\hat{\Theta}_r$ and Λ_r denote the restricted MLE and the LM that solve

$$\max_{\Theta, \lambda} \ln L(\Theta; \boldsymbol{x}) - \boldsymbol{\lambda}'[\mathrm{R}(\Theta) - \boldsymbol{r}]$$

where $\mathrm{R}(\Theta)$ is a continuously differentiable $(q \times 1)$ vector function that contains no redundant coordinate functions. If $\partial \mathrm{R}(\Theta_0)/\partial \Theta'$ has full row rank, then under $H_0 : \mathrm{R}(\Theta) = r$ it follows that:

$$\begin{aligned} G &= \Lambda_r' \frac{\partial \mathrm{R}(\hat{\Theta}_r)'}{\partial \Theta} \left[-\frac{\partial^2 \ln L(\hat{\Theta}_r; \boldsymbol{X})}{\partial \Theta \partial \Theta'} \right]^{-1} \frac{\partial \mathrm{R}(\hat{\Theta}_r)}{\partial \Theta} \Lambda_r \\ &= \frac{\partial \ln L(\hat{\Theta}_r; \boldsymbol{X})'}{\partial \Theta} \left[-\frac{\partial^2 \ln L(\hat{\Theta}_r; \boldsymbol{X})}{\partial \Theta \partial \Theta'} \right]^{-1} \frac{\partial \ln L(\hat{\Theta}_r; \boldsymbol{X})}{\partial \Theta} \xrightarrow{d} \chi_q^2. \end{aligned}$$

10.4.3 Wald Tests

Theorem 10.9 [Asymptotic distribution of the Wald test statistic when H_0 is true]
Let the random sample $\boldsymbol{X}$ of size n have the joint probability density function $f(\boldsymbol{x};\Theta_0)$, let $\hat{\Theta}$ be a consistent estimator for Θ_0 such that $\sqrt{n}(\hat{\Theta}-\Theta_0) \xrightarrow{d} \mathscr{N}(0,\Sigma)$, and let $n\hat{\Sigma}_n$ be a consistent estimator of Σ.
Furthermore, consider the hypotheses $H_0 : \mathrm{R}(\Theta) = r$ versus $H_1 : \mathrm{R}(\Theta) \neq r$, where $\mathrm{R}(\Theta)$ is a $(q \times 1)$ continuously differentiable vector function of Θ for which $q \leq k$ and $\mathrm{R}(\Theta)$ contains no redundant coordinate functions.
Finally, let $\partial\,\mathrm{R}(\Theta_0)/\partial\Theta'$ have full row rank. Then under H_0 it follows that:

$$W = \left[\mathrm{R}(\hat{\Theta}) - \boldsymbol{r}\right]' \left[\frac{\partial\,\mathrm{R}(\hat{\Theta})'}{\partial\Theta}\hat{\Sigma}_n\frac{\partial\,\mathrm{R}(\hat{\Theta})}{\partial\Theta}\right]^{-1} \left[\mathrm{R}(\hat{\Theta}) - \boldsymbol{r}\right] \xrightarrow{d} \chi^2_q.$$