IE 230 Concise Notes,* v2.4.3†

Text in blue will be given on a test; ALL OTHER MATERIAL SHOULD BE COMMITTED TO MEMORY.

---

*The content in this document is based in part on Prof. Bruce Schmeiser's original IE 230 Concise Notes, as well as materials from Profs. Larry Leemis, Joel Nachlas, Raghu Pasupathy, Michael Taaffe, and the class text, Montgomery & Runger (M&R).

†**This document has been proofread, but may have typos.** If you find a typo, email susanhunter[AT]purdue[DOT]edu.

1 Review

1.1 Set Theory

Definition 1.1 (Set)**.** A set is a collection of items.

> **Notation.** Sets are usually denoted by capital letters, such as A, B, C, E, F, G, S.
> $A = \{x, y, z\}$ denotes that A contains elements (or members) x, y, and z.
> If a set has members defined by a condition, $A = \{x\colon x \text{ satisfies the condition}\}$, where the colon is read as "such that."
> M&R (book) uses the notation $A = \{x \mid x \text{ satisfies the condition}\}$, where " $\mid$ " is read as "such that."
> $x \in A$ denotes that x "is an element of" A.

Definition 1.2 (Empty set)**.** The empty set, denoted $\emptyset$, contains no items. It is the smallest set.

Definition 1.3 (Universe)**.** The set containing all relevant items is called the universe. It is the largest set.

Definition 1.4 (Cardinality)**.** The cardinality of a set A, written $|A|$, is the number of elements in the set.

Definition 1.5 (Finite)**.** The set A is finite if $|A|$ is finite.

Definition 1.6 (Countably Infinite)**.** The set A is countably infinite if $|A|$ is infinite, but its members can be counted, that is, a unique integer can be assigned to each member.

Definition 1.7 (Uncountably Infinite)**.** The set A is uncountably infinite if $|A|$ is infinite and its members *cannot* be counted. (E.g., $\mathbb{R}$ is uncountably infinite.)

Definition 1.8 (Subset)**.** If all members of a set A are contained in a set B, then A is a subset of B, $A \subseteq B$.

Definition 1.9 (Superset)**.** If all members of a set A are contained in a set B, then B is a superset of A, written $B \supseteq A$.

Definition 1.10 (Set Equality)**.** Two sets A and B are equal, and we write $A = B$, if they contain the same elements. (That is, $A \subseteq B$ and $A \supseteq B$.)

> **Notation.** The symbols $\subset$ and $\supset$ are "strict" versions of $\subseteq$ and $\supseteq$, just like $<$ and $>$ are "strict" versions of $\leq$ and $\geq$. That is, $A \subset B$ if all elements of A are contained in B *and* A *and* B *are not equal*. Notice that not all texts make this distinction.

Definition 1.11 (Union)**.** The union of sets A and B is the set of items contained in *at least one* of the sets.

> **Notation.** $A \cup B = \{x\colon x \in A \text{ or } x \in B\}$. Remember $A \cup B$ means "A or B or both." Union can be extended to a collection of sets. Consider sets $A_1, A_2, \ldots, A_n$, all subsets of the universe S. Then $\cup_{i=1}^{n} A_i = \{x \in S\colon x \in A_i \text{ for some } i\}$.

Definition 1.12 (Intersection)**.** The intersection of sets A and B is the set of items contained in *both* sets.

> **Notation.** $A \cap B = \{x\colon x \in A \text{ and } x \in B\}$. Remember $A \cap B$ means "A and B." Intersection can be extended to a collection of sets. Consider sets $A_1, A_2, \ldots, A_n$, all subsets of the universe S. Then $\cap_{i=1}^{n} A_i = \{x \in S\colon x \in A_i \text{ for all } i\}$.

Definition 1.13 (Complement)**.** The complement of a set A is the set of items not in A.

> **Notation.** $A^c = \{x\colon x \notin A\}$ or $A' = \{x\colon x \notin A\}$

Remark 1 (Set Operators)**.** *Union, intersection, and complement are operations that are defined for sets.*

Result 1.14 (Distributive Laws)**.** For any sets A, B, and C, it holds that $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ and $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.

Result 1.15 (DeMorgan's Laws)**.** For any sets A and B, $(A \cup B)^c = A^c \cap B^c$ and $(A \cap B)^c = A^c \cup B^c$.

Definition 1.16 (Mutually Exclusive / Disjoint)**.** Two sets A and B are mutually exclusive (or disjoint) if they contain no elements in common, that is, $A \cap B = \emptyset$. More generally, $A_1, A_2, \ldots, A_n$ are mutually exclusive if $A_i \cap A_j = \emptyset$ for every pair of sets where $i \neq j$.

Definition 1.17 (Partition)**.** The sets $B_1, B_2, \ldots, B_n$ partition the set A if $\cup_{i=1}^{n} B_i = A$ and $B_i \cap B_j = \emptyset$ for all $i \neq j$. That is, together, all the B_i's contain all the elements of A, and each member of A lies in exactly one of the B_i's. (The B_i's are "mutually exclusive and collectively exhaustive" with respect to the set A.)

1.2 Functions

Definition 1.18 (Function, not precise)**.** A function assigns a single value to each argument. The set of possible arguments is called the domain, and the set of values is called the range.

Example. $f(x) = x^2$ has domain $\mathbb{R}$ and range $[0, \infty)$.

Definition 1.19 (Undefined function)**.** A function is said to be undefined, or "not defined" at points outside its domain.

Example. The natural logarithm of x, $\ln(x)$ or $\log_e(x)$, is not defined for negative numbers.

2 Probability Basics

2.1 Sample Spaces and Events

Definition 2.1 (Random Experiment)**.** A random experiment is a procedure that can result in a different *outcome* each time it is performed.

Definition 2.2 (Replication)**.** A replication is one instance of the random experiment, which results in exactly one outcome.

Definition 2.3 (Sample Space)**.** The set S of all possible outcomes of a particular random experiment is called the sample space. (Choose the simplest such space to answer the question at hand.)

Definition 2.4 (Discrete Sample Space)**.** A sample space S is discrete if it is finite or countably infinite.

Definition 2.5 (Event)**.** An event is a *subset* of the sample space S. (That is, $E \subseteq S$). For a given replication of the experiment, the event E *occurs* if it contains the outcome; otherwise it does not occur.

Remark 2. *An event is a set. All of the set operators, such as union, intersection, and complement, operate on events because events are sets. Further, the distributive laws and DeMorgan's laws apply to events.*

Definition 2.6 (Mutually Exclusive / Disjoint)**.** Two events, say E_1 and E_2, are mutually exclusive (or disjoint) if they cannot occur together in the same replication of the experiment, that is, $E_1 \cap E_2 = \emptyset$. More generally, $E_1, E_2, \ldots, E_n$ are mutually exclusive if only one can occur in the same replication, that is, $E_i \cap E_j = \emptyset$ for every pair of events.

Definition 2.7 (Partition of the Sample Space)**.** If $\cup_{i=1}^{n} E_i = S$ and $E_1, E_2, \ldots, E_n$ are mutually exclusive, then $E_1, E_2, \ldots, E_n$ are said to partition the sample space.

2.2 Event Probability

Definition 2.8 (Probability)**.** The probability of an event E, denoted $P\{E\}$, is a numerical measure of how likely the event E is to occur when the experiment is performed. $P\{\cdot\}$ is a function that maps a set to a real number in $[0, 1]$.

Remark 3. *A common interpretation of probability is as a relative frequency: if the experiment were repeated infinitely often, $P\{E\}$ is the fraction of the replications in which E occurs.*

Definition 2.9 (Axiom)**.** An axiom is a statement that is assumed and requires no proof.

Definition 2.10 (The Three Axioms of Probability)**.** Consider an experiment with sample space S. For each event E of the sample space S, we assume that a number $P\{E\}$ is defined and satisfies the following three axioms:

1. $P\{S\} = 1$. (With probability 1, the outcome will be a point in the sample space S.)
2. $0 \leq P\{E\} \leq 1$. (The probability that the outcome of the experiment is an outcome in E is a number between 0 and 1.)
3. For all mutually exclusive events E_1 and E_2, (that is, $E_1 \cap E_2 = \emptyset$), we have

$$P\{E_1 \cup E_2\} = P\{E_1\} + P\{E_2\}.$$

(If E_1 and E_2 have no events in common, then the relative frequency of outcomes in $E_1 \cup E_2$ is the sum of the relative frequencies of the outcomes in E_1 and E_2)

Result 2.11 (Axioms Result 1: Complement)**.** For every event E, $P\{E^c\} = 1 - P\{E\}$. (In particular, $P\{\emptyset\} = 1 - P\{S\} = 0$, which implies that the "impossible" event has probability zero.)

Result 2.12 (Axioms Result 2: Dominance)**.** If $E_1 \subset E_2$, then $P\{E_1\} \leq P\{E_2\}$.

Result 2.13 (Axioms Result 3: Axiom 3 for n events)**.** If events $E_1, E_2, \ldots, E_n$ are mutually exclusive, then

$$P\{\cup_{i=1}^{n} E_i\} = \sum_{i=1}^{n} P\{E_i\}$$

Result 2.14 (Axioms Result 4: Equally likely events)**.** If equally likely events $E_1, E_2, \ldots, E_n$ *partition* the sample space, then $P(E_i) = 1/n$ for $i = 1, 2, \ldots, n$.

Result 2.15 (Axioms Result 5: Always true)**.** For any two events E_1 and E_2,

$$P\{E_1 \cup E_2\} = P\{E_1\} + P\{E_2\} - P\{E_1 \cap E_2\}.$$

Remark 4. *Notice that this result can be generalized. For any three events, E_1, E_2, and E_3,*

$$\begin{aligned} P\{E_1 \cup E_2 \cup E_3\} =& P\{E_1\} + P\{E_2\} + P\{E_3\} \\ & - P\{E_1 \cap E_2\} - P\{E_2 \cap E_3\} - P\{E_1 \cap E_3\} \\ & + P\{E_1 \cap E_2 \cap E_3\} \end{aligned}$$

For n events, continue the pattern and alternate signs.

2.3 Conditional Probability and Independence

Definition 2.16 (Conditional Probability)**.** For $P\{B\} > 0$, the conditional probability of A *given* B is

$$P\{A \mid B\} = \frac{P\{A \cap B\}}{P\{B\}}$$

Definition 2.17 (Unconditional Probability)**.** With respect to a sample space S,

$$P\{A \mid S\} = \frac{P\{A \cap S\}}{P\{S\}} := P\{A\}$$

is the *unconditional* or *marginal* probability of A.

Result 2.18 (Multiplication Rule)**.** $P\{A \cap B\} = P\{A \mid B\}P\{B\}$.

Result 2.19 (Baby Bayes' Theorem)**.** For $P\{B\} > 0$,

$$P\{A \mid B\} = \frac{P\{B \mid A\}P\{A\}}{P\{B\}}.$$

Result 2.20 (The Law of Total Probability)**.** Let $E_1, E_2, \ldots, E_n$ be mutually exclusive events in S such that $\cup_{i=1}^{n} E_i = S$. (That is, $E_1, E_2, \ldots, E_n$ partition the sample space S.) Then for an event $A \subseteq S$,

$$P\{A\} = \sum_{i=1}^{n} P\{A \cap E_i\} = \sum_{i=1}^{n} P\{A \mid E_i\}P\{E_i\}.$$

Definition 2.21 (Independence)**.** Two events A and B are independent if $P\{A \cap B\} = P\{A\}P\{B\}$.

Remark 5. *For independent events A and B,*

$$P\{A \mid B\} = \frac{P\{A \cap B\}}{P\{B\}} = \frac{P\{A\}P\{B\}}{P\{B\}} = P\{A\}.$$

Result 2.22. The following four statements are equivalent:

- Events A and B are independent.
- $P\{A \cap B\} = P\{A\}P\{B\}$
- $P\{A \mid B\} = P\{A\}$
- $P\{B \mid A\} = P\{B\}$

Result 2.23. The following four statements are equivalent:

- Events A and B are independent.
- Events A^c and B are independent.
- Events A and B^c are independent.
- Events A^c and B^c are independent.

Definition 2.24 (Mutually Independent Events)**.** The n events $A_1, A_2, \ldots, A_n$ are mutually independent if and only if for every subset $A_{i_1}, A_{i_2}, \ldots, A_{i_k}$ of the n events,

$$P\{A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}\} = P\{A_{i_1}\}P\{A_{i_2}\} \cdots P\{A_{i_k}\}$$

for $k = 2, 3, \ldots, n$. That is, $P\left\{\cap_{j=1}^{k} A_{i_j}\right\} = \prod_{j=1}^{k} P\{A_{i_j}\}$.

Remark 6 (Pairwise Independence)**.** *Pairwise independence is a weaker form of independence than mutual independence, which requires only that every pair of events be independent ($k = 2$ in the Definition 2.24).*

2.4 Random Variables

Definition 2.25 (Random Variable)**.** A random variable is a function that assigns a real number to each outcome in the sample space of an experiment.

> **Notation.** We will denote random variables with upper case letters (usually at the end of the alphabet) such as X, Y, or Z. We reserve lower case letters such as x, y, and z to represent *constants*. Since random variables are functions, for a sample space S, we write $X \colon S \to \mathbb{R}$.

Definition 2.26 (Probability Distribution)**.** The probability distribution of a random variable X is a description, in whatever form, of the likelihoods associated with the values of X.

Remark 7. *Events can be constructed from random variables. Consider the case of rolling two dice, and define X as the sum of the numbers showing on the two dice. "$X = 2$" denotes the event $(1, 1)$ is rolled. "$X < 3$" denotes the event $(1, 1)$ is rolled. "$X \leq 3$" denotes the event $(1, 1)$ or $(1, 2)$ or $(2, 1)$ is rolled. Therefore stating $P\{X = 2\}$ makes sense because $P\{\cdot\}$ is a set function and "$X = 2$" is an event, and events are sets. The statement $P\{X\}$ is meaningless because X is a random variable, not a set.*

3 Discrete Random Variables

Definition 3.1 (Discrete Random Variable, Defn 1)**.** Let S be a sample space. A discrete random variable is a function $X\colon S \to \mathbb{R}$ that takes on a finite number of values or a countably infinite number of values.

Remark 8. *Discrete random variables often arise from counting.*

3.1 Probability Mass Function and Cumulative Distribution Function

Definition 3.2 (Probability Mass Function (pmf))**.** For a discrete random variable, the probability mass function (pmf) is $f_X(x) = P\{X = x\}$ for every real number $-\infty < x < \infty$. That is, the <u>domain</u> of $f_X(x)$ is $\mathbb{R}$, and hence $f_X(x)$ is defined for all $x \in \mathbb{R}$.

Definition 3.3 (Support[‡])**.** The support of a distribution is the set of all $x \in \mathbb{R}$ such that $f_X(x) > 0$. That is, $\mathcal{X} = \{x \in \mathbb{R}\colon f_X(x) > 0\}$. The support $\mathcal{X}$ is written with the distribution function, and the function is assumed to be zero elsewhere. In this class, we usually state explicitly that the function is zero elsewhere.

Definition 3.4 (Cumulative Distribution Function (cdf)[‡])**.** The cumulative distribution function (cdf) of any random variable X is

$$F_X(x) = P\{X \le x\} \text{ for every real number } -\infty < x < \infty.$$

Result 3.5. For a discrete random variable X having possible values $x_1, x_2, \ldots, x_m$, the cdf is

$$P\{X \le x\} = F_X(x) = \sum_{\text{all } x_i \le x} f_X(x_i) \text{ for every real number } -\infty < x < \infty.$$

Result 3.6. [‡]If $a \le b$, then $F_X(a) \le F_X(b)$.

Result 3.7. [‡]For every random variable X, if $a \le b$, then $P\{a < X \le b\} = F_X(b) - F_X(a)$.

Result 3.8 (Properties of a cdf[‡])**.** The function $F_X(x)$ is a cdf if and only if the following conditions hold:

1. $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$.
2. $F_X(x)$ is a nondecreasing function of x.
3. $F_X(x)$ is right-continuous, that is, for every number x_0, $\lim_{x\downarrow x_0} F_X(x) = F_X(x_0)$.

Definition 3.9 (Discrete Random Variable, Defn 2)**.** A random variable X is discrete if its cdf $F_X(x)$ is a step function of x.

3.2 Mean and Variance of a Discrete Random Variable

Definition 3.10 (Expected Value)**.** For a discrete random variable X having possible values $x_1, x_2, \ldots, x_m$, the mean or expected value is the *constant*

$$\mathrm{E}[X] = \sum_{i=1}^{m} x_i f_X(x_i) = \sum_{i=1}^{m} x_i P\{X = x_i\}.$$

Notation. Traditionally, the mean $\mathrm{E}[X]$ is also denoted as μ or μ_X.

Remark 9. *The mean is also called the first moment. The mean can be considered the "balance point" or center of mass of the pmf.*

Definition 3.11 (Expected Value of a Function of X)**.** If X is a discrete random variable having possible values $x_1, x_2, \ldots, x_m$, and $h(\cdot)$ is a function of X, then

$$\mathrm{E}[h(X)] = \sum_{i=1}^{m} h(x_i) f_X(x_i).$$

[‡]Although this content is part of the section on Discrete Random Variables, it holds generally.

Result 3.12. In particular, for constants a and b, $\mathrm{E}[aX+b] = a\mathrm{E}[X]+b$, and $\mathrm{E}[X^2] = \sum_{i=1}^{m} x_i^2 f_X(x_i)$.

Definition 3.13 (Variance[‡])**.** The variance of a random variable X is the constant $\mathrm{Var}(X) = \mathrm{E}[(X-\mathrm{E}[X])^2]$.

Notation. Traditionally, the variance is denoted by σ^2 or σ_X^2.

Remark 10. *The variance is the second moment about the mean, or the second centralized moment. It can be considered a measure of average "penalty" for deviating from the mean.*

Result 3.14. [‡]$\mathrm{Var}(X) = \mathrm{E}[(X-\mathrm{E}[X])^2] = \mathrm{E}[X^2] - \mathrm{E}[X]^2$

Result 3.15. For a discrete random variable X having possible values $x_1, x_2, \ldots, x_m$,

$$\mathrm{Var}(X) = \sum_{i=1}^{m} (x_i - \mathrm{E}[X])^2 f_X(x_i) = \left[\sum_{i=1}^{m} x_i^2 f_X(x_i)\right] - \mathrm{E}[X]^2 = \mathrm{E}[X^2] - \mathrm{E}[X]^2.$$

Definition 3.16 (Standard Deviation[‡])**.** The standard deviation of X is the constant $\sigma_X = \sqrt{\mathrm{Var}(X)}$.

3.3 Common Discrete Distributions

3.3.1 Review of Permutations and Combinations

Definition 3.17 (Permutation)**.** An ordering of r elements from a set of n elements is called a permutation.

Remark 11. *Note that a set of n elements has $n!$ possible permutations, where 0! is defined as 1.*

$$n! = n \times (n-1) \times (n-2) \times \ldots \times 1 \tag{1}$$

Result 3.18. The number of permutations (order matters) of subsets of r elements selected from a set of n elements is

$$P_r^n = \frac{n!}{(n-r)!} \text{ for } r = 0, 1, \ldots, n.$$

Definition 3.19 (Combination)**.** A selection of r elements from a set of n elements *without regard to order* is called a combination.

Result 3.20. The number of combinations (order does not matter) of r elements from a set of n elements is

$$\binom{n}{r} = \frac{n!}{r!(n-r)!} = \frac{P_r^n}{r!} \text{ for } r = 0, 1, \ldots, n.$$

3.3.2 Bernoulli

Definition 3.21 (Bernoulli Trial)**.** A single experiment with only two outcomes is a Bernoulli Trial.

Definition 3.22 (Bernoulli Distribution)**.** A discrete random variable Y has a Bernoulli distribution with parameter p, where $0 \le p \le 1$, if its pmf is

$$f_Y(y) = \begin{cases} p & \text{if } y = 1 \\ 1-p & \text{if } y = 0 \\ 0 & \text{otherwise} \end{cases}$$

Notation. If Y has a Bernoulli distribution, we write $Y \sim \mathrm{Bernoulli}(p)$.

Result 3.23 (Mean of Bernoulli)**.** $\mathrm{E}[Y] = p$.

Result 3.24 (Variance of Bernoulli)**.** $\mathrm{Var}(Y) = p(1-p)$.

Definition 3.25 (Sequence of Bernoulli Trials)**.** A sequence of Bernoulli Trials has three properties:

1. Each trial has exactly two outcomes ("success" and "failure").
2. Each trial has P(success)=p, where p is a constant.
3. Each trial is independent of every other trial.

Remark 12. *An example of a sequence of Bernoulli trials is the repeated, independent tossing of a coin.*

3.3.3 Binomial

Definition 3.26 (Binomial Distribution)**.** Let X be a random variable indicating the number of successes in n Bernoulli trials. Then X has a binomial distribution with *parameters* n and p, where $n \in \{1, 2, 3, \ldots\}$, $0 \leq p \leq 1$, and

$$f_X(x) = P(X = x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & x = 0, 1, \ldots, n; \\ 0 & \text{elsewhere.} \end{cases}$$

Notation. If X has a binomial distribution, we write $X \sim \text{binomial}(n, p)$.

Result 3.27 (Mean of Binomial)**.** $\mathrm{E}[X] = np$

Result 3.28 (Variance of Binomial)**.** $\mathrm{Var}(X) = np(1-p)$.

3.3.4 Geometric

Definition 3.29 (Geometric Distribution)**.** Let X be a random variable indicating the number of Bernoulli trials until the first success. Then X has a geometric distribution with parameter p, where $0 < p < 1$, and

$$f_X(x) = P\{X = x\} = \begin{cases} (1-p)^{x-1} p, & x = 1, 2, \ldots, \infty \\ 0 & \text{elsewhere.} \end{cases}$$

Notation. If X has a geometric distribution, we write $X \sim \text{geometric}(p)$.

Result 3.30 (Mean of Geometric)**.** $\mathrm{E}[X] = 1/p$

Result 3.31 (Variance of Geometric)**.** $\mathrm{Var}(X) = (1-p)/p^2$

Property. The geometric distribution is the only discrete *memoryless* distribution, that is, if $X \sim \text{geometric}(p)$,

$$P\{X > x + c \mid X > x\} = P\{X > c\}.$$

3.3.5 Negative Binomial

Definition 3.32 (Negative Binomial Distribution)**.** Let X be a random variable indicating the number of Bernoulli trials until the rth success. Then X has a negative binomial distribution with parameters r and p, where $0 < p < 1$ and

$$f_X(x) = \begin{cases} \binom{x-1}{r-1} p^r (1-p)^{x-r}, & x = r, r+1, r+2, \ldots \\ 0 & \text{elsewhere.} \end{cases}$$

Notation. If X has a negative binomial distribution, we write $X \sim \text{negative binomial}(r, p)$.

Result 3.33 (Mean of Negative Binomial)**.** $\mathrm{E}[X] = r/p$

Result 3.34 (Variance of Negative Binomial)**.** $\mathrm{Var}(X) = r(1-p)/p^2$

3.3.6 Hypergeometric

Remark 13. *We have related distributions to Bernoulli trials. Now consider an urn containing n marbles, m of which are red and $n - m$ of which are green. Suppose drawing a red marble from the urn is considered a "success." Then successive independent draws from the urn with replacement is a form of Bernoulli trial. Successive draws from the urn without replacement induce dependence and cannot be considered Bernoulli trials (the probability of a success changes across draws). The number of successes in a fixed number of trials when sampling k items from n items <u>without</u> replacement has a hypergeometric distribution.*

Definition 3.35 (Hypergeometric Distribution)**.** Let X be a random variable indicating the number of successes when sampling k items from n items without replacement, m of which are a "success" and $n - m$ of which are a "failure." Then X has a hypergeometric distribution with parameters n, m, and k, where

$$f_X(x) = \begin{cases} \dfrac{\binom{m}{x}\binom{n-m}{k-x}}{\binom{n}{k}}, & x = 0, 1, \ldots, k \text{ and } m - (n-k) \leq x \leq m \\ 0 & \text{elsewhere.} \end{cases}$$

Notation. If X has a hypergeometric distribution, we write $X \sim \text{hypergeometric}(n, m, k)$.

Result 3.36 (Mean of Hypergeometric). $\mathrm{E}[X] = k(m/n)$

Result 3.37 (Variance of Hypergeometric). $\mathrm{Var}(X) = k\left(\frac{m}{n}\right)\left(\frac{(n-m)(n-k)}{n(n-1)}\right)$

Table 1: Categorization of Distributions with respect to the Urn Model

	With Replacement	Without Replacement
# Successes in Fixed # Trials	Binomial	Hypergeometric
# Trials until Fixed # Successes	Negative Binomial (1 success: Geometric)	(Negative Hypergeometric; not covered)

3.3.7 Poisson

Definition 3.38 (Poisson Distribution). A random variable X is said to have a Poisson distribution with parameter $\lambda > 0$ if

$$f_X(x) = \begin{cases} \dfrac{e^{-\lambda}\lambda^x}{x!}, & x = 0, 1, 2, \ldots, \infty \\ 0 & \text{elsewhere.} \end{cases}$$

Result 3.39 (Mean and Variance of Poisson). $\mathrm{E}[X] = \mathrm{Var}(X) = \lambda$.

Remark 14. *The parameter λ is called the rate, and is expressed in units — often the units are units of time, or units of length or area for spatial modeling. For example, the number of people who enter a sub shop in an hour might be well-modeled by a Poisson random variable with parameter $\lambda = 10$ people per hour. The units of λ must be correct for the question at hand. That is, if a sub shop employee is asked about the probability that 15 people will enter the shop in two hours, then the number of people entering the sub shop in two hours is a Poisson random variable with parameter $\lambda = 10 * 2 = 20$ people per <u>two</u> hours.*

Result 3.40 (Poisson Approximation to the Binomial). Poisson($\lambda = np$) is a good approximation to binomial(n, p) when n is large and p is small.

3.4 Summary of Named Univariate Discrete Distributions

Table 2: Summary of Named Univariate Discrete Distributions

Distribution	$f_X(x)$	Support	Mean	Variance
Bernoulli(p)	$p^x(1-p)^{1-x}$	$x = 0, 1$	p	$p(1-p)$
binomial(n, p)	$\binom{n}{x}p^x(1-p)^{n-x}$	$x = 0, 1, \ldots, n$	np	$np(1-p)$
geometric(p)	$p(1-p)^{x-1}$	$x = 1, 2, \ldots$	$\dfrac{1}{p}$	$\dfrac{1-p}{p^2}$
negative binomial(r, p)	$\binom{x-1}{r-1}p^r(1-p)^{x-r}$	$x = r, r+1, \ldots$	$\dfrac{r}{p}$	$\dfrac{r(1-p)}{p^2}$
hypergeometric(n, m, k)	$\dfrac{\binom{m}{x}\binom{n-m}{k-x}}{\binom{n}{k}}$	$x = 0, 1, \ldots, k;$ $m - (n-k) \le x \le m$	$k(m/n)$	$\dfrac{km(n-m)(n-k)}{n^2(n-1)}$
Poisson(λ)	$\dfrac{\lambda^x e^{-\lambda}}{x!}$	$x = 0, 1, 2, \ldots$	λ	λ

4 Continuous Random Variables

Remark 15. *Continuous random variables take on an uncountably infinite number of values. They often arise from measuring (height, weight, volume).*

4.1 Probability Density Function and Cumulative Distribution Function

Definition 4.1 (Continuous Random Variable)**.** A random variable X is continuous if its cdf $F_X(x)$ is a continuous function of x.

Remark 16. *Given a continuous random variable X, for any $x \in \mathbb{R}$, $P\{X = x\} = 0$.*

Definition 4.2 (Probability Density Function (pdf))**.** For a continuous random variable X, a probability density function (pdf) is a function such that

1. $f_X(x) \geq 0$ for all $x \in \mathbb{R}$
2. $\int_{-\infty}^{\infty} f_X(x)dx = 1$
3. $P\{a < X \leq b\} = \int_a^b f_X(x)dx = F_X(b) - F_X(a)$ for all real numbers $a \leq b$.

The domain of $f_X(x)$ is $\mathbb{R}$, and hence $f_X(x)$ is defined for all $x \in \mathbb{R}$. Recall from Definition 3.3 that the support of $f_X(x)$ is $\mathcal{X} = \{x \in \mathbb{R} \colon f_X(x) > 0\}$.

Remark 17. *The pdf is analogous to the pmf, except that the pmf directly provides probabilities while the pdf must be integrated to obtain a probability. Notice that while a pmf cannot have a value greater than one, the pdf can assume values greater than one!*

Result 4.3. For a continuous random variable X, the cumulative distribution function (cdf) is

$$F_X(x) = P\{X \leq x\} = \int_{-\infty}^{x} f_X(t)dt \quad \text{for all } -\infty < x < \infty.$$

Result 4.4. By the fundamental theorem of calculus, given the cdf $F_X(x)$, $f_X(x) = \frac{d}{dx}F_X(x)$.

4.2 Mean and Variance of a Continuous Random Variable

Definition 4.5 (Expected Value)**.** The mean or expected value of a continuous random variable X is the constant

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x f_X(x)dx.$$

Definition 4.6 (Expected Value of a Function of X)**.** For a continuous random variable X and function $h(\cdot)$,

$$\mathrm{E}[h(X)] = \int_{-\infty}^{\infty} h(x) f_X(x)dx.$$

Definition 4.7 (Variance)**.** The variance of a continuous random variable X is the constant

$$\mathrm{Var}(X) = \mathrm{E}[(X - \mathrm{E}[X])^2] = \mathrm{E}[X^2] - \mathrm{E}[X]^2.$$

Thus the definition is unchanged from the discrete case, but the expected values are calculated using integrals:

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} (x - \mathrm{E}[X])^2 f_X(x)dx = \int_{-\infty}^{\infty} x^2 f_X(x)dx - \mathrm{E}[X]^2.$$

Definition 4.8 (Standard Deviation)**.** The standard deviation is $\sigma_X = \sqrt{\mathrm{Var}(X)}$.

4.3 Common Continuous Distributions

4.3.1 Uniform

Definition 4.9 (Uniform Distribution)**.** A random variable X is said to have a uniform distribution with parameters a and b where $a \leq b$ if

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0 & \text{elsewhere.} \end{cases}$$

Notation. If X has a uniform distribution, we write $X \sim \text{uniform}(a, b)$.

Result 4.10 (Uniform cdf)**.** The cdf of a uniform random variable X is

$$F_X(x) = P\{X \leq x\} = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \leq x < b \\ 1 & x \geq b \end{cases}$$

Result 4.11 (Mean of Uniform)**.** $\text{E}[X] = \frac{a+b}{2}$

Result 4.12 (Variance of Uniform)**.** $\text{Var}(X) = \frac{(b-a)^2}{12}$

4.3.2 Triangular

Definition 4.13 (Mode of a Distribution)**.** The mode of a distribution is the location(s) at which the pmf or pdf achieves a maximum.

Definition 4.14 (Triangular Distribution)**.** A random variable X has a triangular distribution if its pdf forms a triangle with base $[a, b]$ and mode at m, where $a \leq m \leq b$, and

$$f_X(x) = \begin{cases} \frac{2(x-a)}{(b-a)(m-a)}, & a \leq x \leq m \\ \frac{2(b-x)}{(b-a)(b-m)}, & m < x \leq b \\ 0 & \text{elsewhere.} \end{cases}$$

Notation. If X has a triangular distribution, we write $X \sim \text{triangular}(a, b, m)$.

Result 4.15 (Mean of Triangular)**.** $\text{E}[X] = \frac{a+m+b}{3}$

Result 4.16 (Variance of Triangular)**.** $\text{Var}(X) = \frac{(b-a)^2-(m-a)(b-m)}{18}$

4.3.3 Normal

Definition 4.17 (Normal Distribution)**.** A random variable X is said to have a normal distribution with location parameter μ and scale parameter σ if

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty.$$

Notation. If X has a normal distribution, we write $X \sim \text{N}(\mu, \sigma^2)$.

Result 4.18 (Mean of Normal)**.** $\text{E}[X] = \mu$.

Result 4.19 (Variance of Normal)**.** $\text{Var}(X) = \sigma^2$.

Result 4.20 (Points of Inflection)**.** The two points of inflection in the normal pdf occur at $\mu - \sigma$ and $\mu + \sigma$.

Result 4.21 (68-95-99.7 Rule)**.** For any normally distributed random variable X, the probability that X is within 1 standard deviation of the mean is 0.68, within 2 standard deviations of the mean is 0.95, and within 3 standard deviations of the mean is 0.997.

Result 4.22 (Normal cdf)**.** The normal cdf,

$$F_X(x) = P\{X \le x\} = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt,$$

has no closed-form expression. Values of the cdf may be found by numerical integration.

Result 4.23 (Standard Normal)**.** If $X \sim \mathrm{N}(\mu, \sigma^2)$ and

$$Z = \frac{X - \mu}{\sigma},$$

then $Z \sim \mathrm{N}(0, 1)$. Likewise, if $Z \sim \mathrm{N}(0, 1)$ and $X = Z\sigma + \mu$, then $X \sim \mathrm{N}(\mu, \sigma^2)$.

Notation. If $Z \sim \mathrm{N}(0, 1)$, then Z is called a standard normal random variable. The letter Z is (nearly) universally used for standard normal random variables.

Remark 18. *Probabilities corresponding to a random variable $X \sim N(\mu, \sigma^2)$ can be found by standardizing X to find a corresponding probability in terms of Z. The probability corresponding to Z can be found in a Z table. The cdf for the random variable Z is often denoted $\Phi(\cdot)$, and the pdf for Z is often denoted $\phi(\cdot)$.*

4.3.4 Exponential

Remark 19. *The exponential distribution is a continuous analogue of the geometric distribution. It is often used to model the time between events, e.g. as in a Poisson Process, or used to model lifetimes.*

Definition 4.24 (Exponential Distribution)**.** A random variable X is said to have an exponential distribution with parameter $\lambda > 0$ if

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x \ge 0 \\ 0 & \text{elsewhere.} \end{cases}$$

Notation. If X has an exponential distribution, we write $X \sim \exp(\lambda)$.

Result 4.25 (Mean of Exponential)**.** $\mathrm{E}[X] = 1/\lambda$.

Result 4.26 (Variance of Exponential)**.** $\mathrm{Var}(X) = 1/\lambda^2$.

Result 4.27 (Exponential cdf)**.** The cdf for an exponential random variable X is

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x}, & x \ge 0 \\ 0 & x < 0. \end{cases}$$

Definition 4.28 (Memoryless)**.** A random variable X is said to be memoryless if

$$P\{X > s + t \mid X > t\} = P\{X > s\} \text{ for all } s, t \ge 0.$$

Result 4.29. Exponential is the only memoryless continuous distribution.

4.3.5 Gamma / Erlang

Definition 4.30 (Gamma function)**.** The gamma function is a normalizing constant in the pdf of a gamma random variable and is defined as $\Gamma(r) = \int_0^\infty x^{r-1} e^{-x} dx,\ r > 0$.

Result 4.31. If r is an integer, then $\Gamma(r) = (r-1)!$.

Definition 4.32 (Gamma Distribution)**.** A random variable X is said to have a gamma distribution with parameters $r > 0$ and $\lambda > 0$ if

$$f_X(x) = \begin{cases} \frac{1}{\Gamma(r)} \lambda^r x^{r-1} e^{-\lambda x} & x \ge 0 \\ 0 & \text{elsewhere.} \end{cases}$$

If r is an integer, then the distribution is called Erlang.

Notation. If X has an gamma distribution, we write $X \sim \text{gamma}(r, \lambda)$.

Remark 20. *The Erlang distribution is a continuous analogue of the negative binomial distribution.*

Result 4.33 (Mean of Gamma)**.** $\mathrm{E}[X] = r/\lambda$.

Result 4.34 (Variance of Gamma)**.** $\mathrm{Var}(X) = r/\lambda^2$.

4.4 Summary of Named Univariate Continuous Distributions

Table 3: Summary of Named Univariate Continuous Distributions

Distribution	$f_X(x)$, with support	Mean	Variance
uniform(a, b)	$\frac{1}{b-a},\ a < x < b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
triangular(a, b, m)	$\begin{cases} \frac{2(x-a)}{(b-a)(m-a)}, & a \le x \le m \\ \frac{2(b-x)}{(b-a)(b-m)}, & m < x \le b \end{cases}$	$\frac{a+b+m}{3}$	$\frac{(b-a)^2-(m-a)(b-m)}{18}$
$N(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},\ -\infty < x < \infty$	μ	σ^2
exponential(λ)	$\lambda e^{-\lambda x},\ x \ge 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
gamma(r, λ)	$\frac{1}{\Gamma(r)} \lambda^r x^{r-1} e^{-\lambda x},\ x \ge 0$	$\frac{r}{\lambda}$	$\frac{r}{\lambda^2}$

Definition 4.35 (Distribution Parameter: Location Parameter)**.** A location parameter is additive: $Y = a + X$. The distribution of Y is identical to that of X except that its location is shifted a units to the right when $a > 0$.

Definition 4.36 (Distribution Parameter: Scale Parameter)**.** A scale parameter is multiplicative: $Y = bX$. The distribution of Y is identical to that of X, except that each unit of Y is b units of X; the location of zero is unchanged.

Definition 4.37 (Distribution Parameter: Shape Parameter)**.** A shape parameter is nonlinear: $Y = g(X; c)$, where the function g is nonlinear in c. The distributions of Y and X have different shapes.

Remark 21. *The normal distribution has location parameter μ and scale parameter σ. The normal distribution does not have a shape parameter.*

5 Summary of General Univariate Distributions

Remark 22. *When finding a pmf/pdf from a cdf or finding a cdf from a pmf/pdf, you must include complete support (see Definition 3.3).*

Table 4: Summary of Univariate Distributions

	Discrete Random Variable X	Continuous Random Variable X
possible values	finite or countably infinite: $x_1, x_2, \ldots, x_m$	uncountably infinite: for example, x can assume values in an interval
pmf / pdf	pmf $f_X(x)$ satisfies (i) $f_X(x) \geq 0$, all $x \in \mathbb{R}$ (ii) $\sum_{\text{all } x_i:\ f_X(x_i)>0} f_X(x_i) = 1$ (iii) $f_X(x) = P\{X = x\}$	pdf $f_X(x)$ satisfies (i) $f_X(x) \geq 0$, all $x \in \mathbb{R}$ (ii) $\int_{-\infty}^{\infty} f_X(x) = 1$ (iii) For any a and b such that $a < b$, $P\{a < X \leq b\} = \int_a^b f_X(x)dx$
cdf	$F_X(x) = P\{X \leq x\}$	$F_X(x) = P\{X \leq x\}$
obtaining probabilities	directly given by pmf $f_X(x)$; $P\{X = a\} = f_X(a)$ for all $a \in \mathbb{R}$	must integrate pdf $f_X(x)$; $P\{X = a\} = 0$ for all $a \in \mathbb{R}$
obtaining cdf from pmf/pdf	$F_X(x) = P\{X \leq x\} = \sum_{\text{all } x_i \leq x} f_X(x_i)$	$F_X(x) = P\{X \leq x\} = \int_{-\infty}^{x} f_X(t)dt$
obtaining pr./pmf/pdf from cdf	$P\{a < X \leq b\} = F_X(b) - F_X(a)$	$P\{a < X \leq b\} = F_X(b) - F_X(a)$ $f_X(x) = \frac{d}{dx}F_X(x)$
expected value of X	$\mathrm{E}[X] = \sum_{i=1}^{m} x_i f_X(x_i)$	$\mathrm{E}[X] = \int_{-\infty}^{\infty} x f_X(x)dx$
expected value of a function of X	$\mathrm{E}[h(X)] = \sum_{i=1}^{m} h(x_i) f_X(x_i)$	$\mathrm{E}[h(X)] = \int_{-\infty}^{\infty} h(x) f_X(x)dx$
variance of X	$\mathrm{Var}(X) = \mathrm{E}[(X - \mathrm{E}[X])^2]$	$\mathrm{Var}(X) = \mathrm{E}[(X - \mathrm{E}[X])^2]$
standard deviation of X	$\sigma_X = \sqrt{\mathrm{Var}(X)}$	$\sigma_X = \sqrt{\mathrm{Var}(X)}$
common named distributions	binomial, geometric, negative binomial hypergeometric, Poisson	uniform, triangular, normal, exponential, gamma / Erlang

6 Bivariate Distribution Functions

6.1 Summary of General Bivariate Distributions

Table 5: Summary of Joint Distribution Functions for Two Random Variables (i.e., Bivariate Distributions)

	Discrete Random Variables X, Y	Continuous Random Variables X, Y
pmf / pdf	$f_{X,Y}(x,y)$ satisfies (i) $f_{X,Y}(x,y) \geq 0$, all $(x,y) \in \mathbb{R}^2$ (ii) $\sum_x \sum_y f_{X,Y}(x,y) = 1$ (iii) $f_{X,Y}(x,y) = P\{X = x, Y = y\}$	$f_{X,Y}(x,y)$ satisfies (i) $f_{X,Y}(x,y) \geq 0$, all $(x,y) \in \mathbb{R}^2$ (ii) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y) = 1$ (iii) For any $R \in \mathbb{R}^2$, $P\{(X,Y) \in R\} = \iint_R f_{X,Y}(x,y)dxdy$
cdf	$F_{X,Y}(x,y) = P\{X \leq x, Y \leq y\}$	$F_{X,Y}(x,y) = P\{X \leq x, Y \leq y\}$
expected value of fn of X, Y	$\mathrm{E}[h(X,Y)] = \sum_x \sum_y h(x,y) f_{X,Y}(x,y)$	$\mathrm{E}[h(X,Y)] = \iint_{\mathbb{R}^2} h(x,y) f_{X,Y}(x,y)dxdy$
marginals	$f_X(x) = \sum_y f_{X,Y}(x,y)$ $f_Y(y) = \sum_x f_{X,Y}(x,y)$	$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$ $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx$
conditional pmf/pdf	$f_{Y\|X=x}(y) = \dfrac{f_{X,Y}(x,y)}{f_X(x)}$ for $f_X(x) > 0$ Interpretation: $P\{Y = y \mid X = x\} = \dfrac{P\{Y = y, X = x\}}{P\{X = x\}}$	$f_{Y\|X=x}(y) = \dfrac{f_{X,Y}(x,y)}{f_X(x)}$ for $f_X(x) > 0$
conditional expectation	$\mathrm{E}[Y \mid X = x] = \sum_y y f_{Y\|X=x}(y)$	$\mathrm{E}[Y \mid X = x] = \int_{-\infty}^{\infty} y f_{Y\|X=x}(y)dy$
common named dist'ns	(multinomial)	bivariate normal

Definition 6.1 (Support)**.** The support of a bivariate distribution is the set of all $(x,y) \in \mathbb{R}^2$ such that $f_{X,Y}(x,y) > 0$. That is, $\mathcal{X} = \{(x,y) \in \mathbb{R}^2 \colon f_{X,Y}(x,y) > 0\}$. The support $\mathcal{X}$ is written with the distribution function, and the function is assumed to be zero elsewhere.

Remark 23. *Going from a joint pmf/pdf to a collection of marginals results in a loss of information about how the random variables occur together.*

Remark 24. *Conditional pmfs/pdfs have all the usual properties of regular pmfs/pdfs.*

6.2 Independence

Definition 6.2 (Independence)**.** Two random variables X and Y are independent if

$$f_{X,Y}(x,y) = f_X(x) f_Y(y) \text{ for all } x \text{ and } y.$$

Result 6.3 (Independence: Equivalent Statements)**.** For random variables X and Y, if any one of the following properties is true, the others are also true, and X and Y are independent:

1. $f_{X,Y}(x,y) = f_X(x) f_Y(y)$ for all x and y
2. $f_{Y|X=x}(y) = f_Y(y)$ for all x and y with $f_X(x) > 0$
3. $f_{X|Y=y}(x) = f_X(x)$ for all x and y with $f_Y(y) > 0$
4. $P\{X \in A, Y \in B\} = P\{X \in A\} P\{Y \in B\}$ for any sets A and B in the range of X and Y, respectively.

Result 6.4. If X and Y are independent random variables, then

$$F_{X,Y}(x,y) = F_X(x)F_Y(y) \text{ for all } x \text{ and } y.$$

6.3 Expected Values

Definition 6.5 (See Table 5)**.** The expected value of a function of two random variables is

$$\mathrm{E}[h(X,Y)] = \begin{cases} \sum_x \sum_y h(x,y) f_{X,Y}(x,y) & \text{if } X,Y \text{ discrete} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x,y) f_{X,Y}(x,y) dx dy & \text{if } X,Y \text{ continuous.} \end{cases}$$

Result 6.6. $\mathrm{E}[g(X) + h(Y)] = \mathrm{E}[g(X)] + \mathrm{E}[h(Y)]$

Result 6.7. If X and Y are independent random variables, $\mathrm{E}[g(X)h(Y)] = \mathrm{E}[g(X)]\mathrm{E}[h(Y)]$

Result 6.8. If X and Y are random variables, then $\mathrm{E}[X] = \mathrm{E}[\mathrm{E}[X \mid Y]]$ when the expectations exist.

6.4 Covariance and Correlation

Definition 6.9 (Covariance)**.** The covariance between the random variables X and Y, denoted $\mathrm{Cov}(X,Y)$, is

$$\mathrm{Cov}(X,Y) = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])] = \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y].$$

Remark 25. *Covariance measures the linear relationship between two random variables.*

Result 6.10. For random variables X and Y, $\mathrm{Cov}(aX + c, bY + d) = ab\mathrm{Cov}(X,Y)$.

Remark 26. *Note that for $Y = X$ and $a = b, c = d$, the previous result implies $Var(aX + c) = a^2\, Var(X)$.*

Result 6.11. $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y)$

Result 6.12. If X and Y are independent random variables, then $\mathrm{Cov}(X,Y) = 0$.

Definition 6.13 (Correlation)**.** The correlation between random variables X and Y, denoted $\rho_{X,Y}$, is

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

Remark 27. *Correlation is a dimensionless measure of the linear relationship between two random variables.*

Result 6.14. If X and Y are independent random variables, then $\rho_{X,Y} = 0$.

Result 6.15. For any two random variables X and Y, $-1 \le \rho_{X,Y} \le 1$.

Remark 28. *If $Cov(X,Y) = \rho_{X,Y} = 0$, then it is not necessarily true that X and Y are independent! X and Y may have nonlinear dependence.*

6.5 Common Bivariate Distribution: Bivariate Normal

Remark 29. *The bivariate normal can also be extended to the case of many random variables, resulting in a multivariate normal.*

Definition 6.16 (Bivariate Normal Distribution)**.** The probability density function of a bivariate normal distribution with parameters $\mu_X, \mu_Y, \sigma_X > 0, \sigma_Y > 0$, and $-1 < \rho < 1$ is

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]}, \quad -\infty < x < \infty, -\infty < y < \infty.$$

Result 6.17. If X and Y have a bivariate normal distribution with joint pdf $f_{X,Y}(x,y)$, then the marginal probability distributions of X and Y are normal with means μ_X and μ_Y and standard deviations σ_X and σ_Y, respectively.

Result 6.18. If X and Y have a bivariate normal distribution, the correlation between X and Y is ρ.

Result 6.19. If X and Y have a bivariate normal distribution with $\rho = 0$, then X and Y are independent.

7 More than Two Random Variables

7.1 Multivariate Distributions

Definition 7.1. A joint probability mass function for discrete random variables $X_1, X_2, \ldots, X_p$, denoted as $f_{X_1,X_2,\ldots,X_p}(x_1, x_2, \ldots, x_p)$, satisfies the following properties:

1. $f_{X_1,X_2,\ldots,X_p}(x_1, x_2, \ldots, x_p) \geq 0$ for all $x_1, x_2, \ldots, x_p$.
2. $\sum_{x_1}\sum_{x_2}\cdots\sum_{x_p} f_{X_1,X_2,\ldots,X_p}(x_1, x_2, \ldots, x_p) = 1$.
3. $f_{X_1,X_2,\ldots,X_p}(x_1, x_2, \ldots, x_p) = P\{X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p\}$.

Definition 7.2. A joint probability density function for continuous random variables $X_1, X_2, \ldots, X_p$, denoted as $f_{X_1,X_2,\ldots,X_p}(x_1, x_2, \ldots, x_p)$, satisfies the following properties:

1. $f_{X_1,X_2,\ldots,X_p}(x_1, x_2, \ldots, x_p) \geq 0$ for all $x_1, x_2, \ldots, x_p$.
2. $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f_{X_1,X_2,\ldots,X_p}(x_1, x_2, \ldots, x_p)dx_1dx_2\ldots dx_p = 1$.
3. For any region B of p-dimensional space,
   $P\{(X_1, X_2, \ldots, X_p) \in B\} = \iint\ldots\int_B f_{X_1,X_2,\ldots,X_p}(x_1, x_2, \ldots, x_p)dx_1dx_2\ldots dx_p$

Remark 30. *Marginal distributions of random variables and conditional distributions are found in methods analogous to the case of two random variables.*

Definition 7.3 (Multiple Independent Random Variables)**.** Random variables $X_1, X_2, \ldots, X_p$ are independent if and only if

$$f_{X_1,X_2,\ldots,X_p}(x_1, x_2, \ldots, x_p) = f_{X_1}(x_1)f_{X_2}(x_2)\cdots f_{X_p}(x_p).$$

Result 7.4. If $X_1, X_2, \ldots, X_p$ are independent, then

$$P\{X_1 \in A_1, X_2 \in A_2, \ldots, X_p \in A_p\} = P\{X_1 \in A_1\}P\{X_2 \in A_2\}\cdots P\{X_p \in A_p\}$$

for any regions $A_1, A_2, \ldots, A_p$ in the range of $X_1, X_2, \ldots, X_p$, respectively.

7.2 Expected Values

Definition 7.5. The expected value of a function of p random variables $X_1, X_2, \ldots, X_p$ having joint pmf or pdf $f_{X_1,X_2,\ldots,X_p}(x_1, x_2, \ldots, x_p)$ is

$$\mathrm{E}[h(X_1, X_2, \ldots, X_p)] = \begin{cases} \sum_{x_1}\sum_{x_2}\cdots\sum_{x_p} h(x_1, x_2, \ldots, x_p)f_{X_1,X_2,\ldots,X_p}(x_1, x_2, \ldots, x_p) \\ \qquad\qquad \text{if } X_1, X_2, \ldots, X_p \text{ discrete} \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} h(x_1, x_2, \ldots, x_p)f_{X_1,X_2,\ldots,X_p}(x_1, x_2, \ldots, x_p)dx_1dx_2\cdots dx_p \\ \qquad\qquad \text{if } X_1, X_2, \ldots, X_p \text{ continuous.} \end{cases}$$

Result 7.6. If $X_1, X_2, \ldots, X_p$ are random variables, then

$$\mathrm{E}\left[\sum_{i=1}^{p} g_i(X_i)\right] = \sum_{i=1}^{p}\mathrm{E}[g_i(X_i)].$$

Proof. This proof is for the continuous case, but a similar proof holds in the discrete case.

$$\begin{aligned} \mathrm{E}\left[\sum_{i=1}^{p} g_i(X_i)\right] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left(\sum_{i=1}^{p} g_i(X_i)\right)f_{X_1,X_2,\ldots,X_p}(x_1, x_2, \ldots, x_p)dx_1dx_2\ldots dx_p \\ &= \sum_{i=1}^{p}\left(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} g_i(X_i)f_{X_1,X_2,\ldots,X_p}(x_1, x_2, \ldots, x_p)dx_1dx_2\ldots dx_p\right) \\ &= \sum_{i=1}^{p}\mathrm{E}[g_i(X_i)] \end{aligned}$$

□

Result 7.7. If $X_1, X_2, \ldots, X_p$ are random variables and $c_1, c_2, \ldots, c_p$ are constants, then

$$\text{Var}\Big(\sum_{i=1}^{p} c_i X_i\Big) = \sum_{i=1}^{p}\sum_{j=1}^{p} \text{Cov}(c_i X_i, c_j X_j) = \sum_{i=1}^{p}\sum_{j=1}^{p} c_i c_j \text{Cov}(X_i, X_j).$$

Then in succinct form, $\text{Var}\Big(\sum_{i=1}^{p} c_i X_i\Big) = \sum_{i=1}^{p} c_i^2 \text{Var}(X_i) + 2\sum_{1\le i<j\le p} c_i c_j \text{Cov}(X_i, X_j)$. In the specific case that all constants $c_1, c_2, \ldots, c_p$ equal one, then $\text{Var}\Big(\sum_{i=1}^{p} X_i\Big) = \sum_{i=1}^{p} \text{Var}(X_i) + 2\sum_{1\le i<j\le p} \text{Cov}(X_i, X_j)$.

Definition 7.8. The variance-covariance matrix of $X_1, X_2, \ldots, X_p$ is the $p \times p$ symmetric, positive semidefinite matrix

$$\Sigma = \begin{pmatrix} \text{Cov}(X_1,X_1) & \text{Cov}(X_1,X_2) & \cdots & \text{Cov}(X_1,X_p) \\ \text{Cov}(X_2,X_1) & \text{Cov}(X_2,X_2) & \cdots & \text{Cov}(X_2,X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_p,X_1) & \text{Cov}(X_p,X_2) & \cdots & \text{Cov}(X_p,X_p) \end{pmatrix} = \begin{pmatrix} \text{Var}(X_1) & \text{Cov}(X_1,X_2) & \cdots & \text{Cov}(X_1,X_p) \\ \text{Cov}(X_2,X_1) & \text{Var}(X_2) & \cdots & \text{Cov}(X_2,X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_p,X_1) & \text{Cov}(X_p,X_2) & \cdots & \text{Var}(X_p) \end{pmatrix}.$$

Remark 31. *Notice that in Result 7.7, we are merely summing all entries in the variance-covariance matrix.*

7.3 Common Multivariate Distribution: Multinomial

Remark 32. *Our example of a sequence of Bernoulli trials was a sequence of independent tosses of a coin. Now instead of tossing a coin, think of tossing a die, where instead of just two outcomes, there can be six different outcomes. A random vector $(X_1, X_2, X_3, X_4, X_5, X_6)$ will count the number of times each face of the die is shown in n (fixed) successive tosses. Then $(X_1, X_2, X_3, X_4, X_5, X_6)$ has a multinomial distribution. The multinomial distribution is a multi-dimensional version of the binomial distribution.*
Or, think of an urn model where there are k colors of marbles and each successive draw is made independently, with replacement. A multinomial random variable will count the number of each color of marble drawn in n (fixed) trials.

Definition 7.9 (Multinomial Distribution)**.** Suppose a random experiment consists of n trials, and assume

1. The result of each trial is one of k potential "classes."
2. The probability of a trial generating a result in the ith class is equal to p_i for all $i = 1, \ldots, k$, and each p_i is constant across the trials.
3. The trials are independent.

Let X_i denote the number of trials that resulted in the ith class for $i = 1, \ldots, k$, respectively. Then $X_1, X_2, \ldots, X_k$ have a multinomial distribution with joint pdf

$$f_{X_1,X_2,\ldots,X_k}(x_1, x_2, \ldots, x_k) = \frac{n!}{x_1! x_2! \cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

for $x_1 + x_2 + \ldots + x_k = n$ and $p_1 + p_2 + \ldots + p_k = 1$ (0 elsewhere).

Result 7.10. If $X_1, X_2, \ldots, X_k$ have a multinomial distribution, the marginal probability distribution of X_i is binomial(n, p_i) with mean np_i and variance $np_i(1 - p_i)$ for all $i = 1, \ldots, k$.

Remark 33. *Suppose X_1, X_2, X_3, X_4 have a multinomial distribution. To find the marginal probability distribution of some subset of these random variables, for example, the joint marginal distribution of X_1, X_2, we can consider a new multinomial experiment in which we combine the classes corresponding to X_3 and X_4 into a single class. Then*

$$f_{X_1,X_2}(x_1, x_2) = P\{X_1 = x_1, X_2 = x_2\} = \frac{n!}{x_1! x_2! (n - x_1 - x_2)!} p_1^{x_1} p_2^{x_2} (1 - (p_1 + p_2))^{n - x_1 - x_2}$$

for $x_1 + x_2 \le n$ (0 elsewhere).

8 Estimation

Remark 34. *We now suppose that there exists a system we want to study. We further suppose that there is some underlying distribution (say with cdf F_X) governing data from this system. Since the underlying distribution F_X is unknown, we wish to draw conclusions about the system by sampling and constructing estimators for quantities of interest, such as $E[X]$, $Var(X)$, or other parameters of the distribution F_X.*

Definition 8.1 (Independent and Identically Distributed)**.** Random variables $X_1, X_2, \ldots, X_n$ are independent and identically distributed (iid) if they have the same distribution and they are mutually independent.

Remark 35. *We sometimes say $X_1, X_2, \ldots, X_n$ are iid copies of a random variable X, where X has cdf F_X with $E[X] = \mu$ and $Var(X) = \sigma^2$.*

Definition 8.2 (Random Sample)**.** If a random variable X has distribution F_X and $X_1, X_2, \ldots, X_n$ are iid copies of X, then $X_1, X_2, \ldots, X_n$ is a random sample from F_X.

8.1 Estimation of the Mean

Remark 36. *We first consider the case in which the parameter of interest is the mean of the distribution F_X, which we will call μ.*

Result 8.3. If $X_1, X_2, \ldots, X_n$ are iid random variables with mean μ and variance σ^2, and

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

then

$$\mathrm{E}[\overline{X}] = \mu \quad \text{and} \quad \mathrm{Var}(\overline{X}) = \sigma^2/n \quad \text{and} \quad s.e.(\overline{X}) = \sqrt{\mathrm{Var}(\overline{X})} = \sigma/\sqrt{n},$$

where $\overline{X}$ is called the "sample mean" and *s.e.* stands for "standard error."

Remark 37. *That is, $\overline{X}$ is a random variable and it has a distribution. We care about the standard error of $\overline{X}$ because it tells us how much variability there is in the estimator $\overline{X}$.*

Result 8.4 (Strong Law of Large Numbers (SLLN))**.** Let $X_1, X_2, \ldots, X_n$ be iid random variables with mean μ and finite variance σ^2. Then as n tends to infinity,

$$\overline{X} \to \mu$$

with probability one.

Remark 38 (For the curious)**.** *It is broadly beyond the scope of the class, but the with probability one convergence of $\overline{X}$ in the SLLN means that for all $\epsilon > 0$, $P\{\lim_{n\to\infty} |\overline{X} - \mu| < \epsilon\} = 1$.*

Corollary 8.5. If $X_1, X_2, \ldots, X_n$ are iid random variables with mean μ and variance σ^2, then $\sum_{i=1}^{n} X_i$ has mean $n\mu$ and variance $n\sigma^2$.

Theorem 8.6 (Central Limit Theorem (CLT))**.** If $X_1, X_2, \ldots, X_n$ are iid random variables with mean μ and finite variance σ^2, then as n tends to infinity,

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

tends to the standard normal distribution.

Remark 39. *How large should n be such that the CLT provides a reasonable approximation to the distribution of $\overline{X}$? A good "rule of thumb" is that n should be larger than about 30.*

Result 8.7. If $X_1, X_2, \ldots, X_n$ are iid **normal** random variables with mean μ and variance σ^2, then $\overline{X}$ is **normally distributed** with mean μ and variance σ^2/n.

Result 8.8. If $X_1, X_2, \ldots, X_n$ are iid **normal** random variables with mean μ and variance σ^2, then $\sum_{i=1}^{n} X_i$ is **normally distributed** with mean $n\mu$ and variance $n\sigma^2$.

8.2 Estimation of a Difference of Means

Remark 40. *Now consider the case in which we have two independent populations, the first governed by some underlying distribution with cdf F_X, mean μ_X, and finite variance σ_X^2, and the second governed by some underlying distribution with cdf F_Y, mean μ_Y, and finite variance σ_Y^2. We wish to draw conclusions about the difference in means between these two distributions by sampling.*

Result 8.9. $\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X, Y)$.

Result 8.10. If $X_1, X_2, \ldots, X_{n_X}$ and $Y_1, Y_2, \ldots, Y_{n_Y}$ are each iid samples drawn independently from their respective populations, with

$$\overline{X} = \frac{1}{n_X}\sum_{i=1}^{n_X} X_i, \quad \text{and} \quad \overline{Y} = \frac{1}{n_Y}\sum_{i=1}^{n_Y} Y_i,$$

then

$$\text{Var}(\overline{X} - \overline{Y}) = \text{Var}(\overline{X}) + \text{Var}(\overline{Y}) = \sigma_X^2/n_X + \sigma_Y^2/n_Y.$$

Further,

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\sigma_X^2/n_X + \sigma_Y^2/n_y}}$$

has a distribution that is approximately standard normal, if the conditions of the central limit theorem apply. If the two populations are exactly normal, then the distribution of Z is exactly normal.

8.3 Estimation of the Variance

Remark 41. *We now consider the case in which the parameter of interest is the variance of the distribution F_X, which we will call σ^2. Estimating the population variance σ^2 also enables estimation of s.e.$(\overline{X})$.*

Result 8.11. Let $X_1, X_2, \ldots, X_n$ be iid random variables with mean μ and variance σ^2. An estimator of σ^2 is

$$\hat{\sigma}^2 = \widehat{\text{Var}}(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Remark 42. *One can also use the estimator $\hat{\sigma}^2 = \widehat{Var}(X) = \frac{1}{n}\sum_{i=1}^n X_i^2 - \left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2$ for easier calculations. In this class, either is acceptable. M&R tends to use $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$, which is also okay. Note that the estimator $\hat{\sigma}^2$ is biased, and S^2 is unbiased (see Definition 8.15).*
From the definition of standard deviation, we can estimate the standard deviation of X using the estimator

$$\hat{\sigma} = \widehat{s.d.}(X) = \sqrt{\widehat{Var}(X)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}.$$

Remark 43. *Recall that s.e.$(\overline{X}) = \sigma/\sqrt{n}$ is the variance of the distribution of $\overline{X}$. When σ is unknown and must be estimated, we may estimate the standard error of $\overline{X}$ as $\widehat{s.e.}(\overline{X}) = \hat{\sigma}/\sqrt{n}$.*

8.4 Summary of Fixed and Random Quantities

Remark 44. *We use the terms "fixed" and "constant" interchangeably to refer to quantities that are numbers and are not random variables. Random variables are "random."*

Remark 45. *A good test for whether a quantity is fixed or random is to ask the question, if I repeat the experiment, will the quantity change?*

Remark 46 (Data)**.** *M&R uses the notation $x_1, x_2, \ldots, x_n$ to denote an observed sample of size n. These quantities are numbers and are fixed, because they have already been observed as data. If we were to perform the experiment again, we would collect a new data set $x_1', x_2', \ldots, x_n'$. The old data would remain unchanged by observing the new data.*

Table 6: Let $X_1, X_2, \ldots, X_n$ be iid copies of a random variable X having mean μ and variance σ^2. The table presents a summary of fixed and random quantities in this scenario.

Quantity	Expression	Random/ Fixed	As n ↑, tends to	
μ	$\mathrm{E}[X]$	fixed	μ	← unknown
$\overline{X}$	$\frac{1}{n}\sum_{i=1}^n X_i$	random	μ	
$\sigma = s.d.(X)$	$\sqrt{\mathrm{Var}(X)}$	fixed	σ	← inherent variability
$\hat{\sigma} = \widehat{s.d.}(X)$	$\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2 - \left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2}$	random	σ	
$s.e.(\overline{X})$	$\sigma/\sqrt{n}$	fixed	0	← experimental variability
$\widehat{s.e.}(\overline{X})$	$\hat{\sigma}/\sqrt{n}$	random	0	

NOTE: The notation "$s.d.$" stands for "standard deviation." The notation "$s.e.$" stands for "standard error," which, in this scenario, is a measure of the variability of the estimator $\overline{X}$.

8.5 General Parameter Estimation

Remark 47. *We may want to estimate something other than a mean or variance. For example, we may want to estimate the median or some other distributional parameter, which we call θ.*

Remark 48 (Distributional Family)**.** *Distributional families are characterized by location, scale, and/or shape parameters. The named distributions we have studied are distributional families. Within a distributional family, one must select parameters of the distributional family to specify a distribution. For example, exponential, normal, binomial are distributional families. Exponential($\lambda = 1$), $N(\mu = 5, \sigma^2 = 4)$, and binomial($n = 10, p = 1/2$) are distributions; λ, (μ, σ^2), and (n, p) are the parameters of their respective distributional families.*

Remark 49. *Suppose that we assume our dataset is modeled as a realization of a random sample $X_1, X_2, \ldots, X_n$ from a named distribution. Then we need to estimate the parameters for the distribution from the assumed family. The parameters that determine the model distribution are the model parameters.*

Definition 8.12 (Statistic)**.** A statistic $\hat{\theta}$ is a function of the random sample $X_1, X_2, \ldots, X_n$; that is, there is some function h so that $\hat{\theta} = h(X_1, X_2, \ldots, X_n)$. (Statistics are random since functions of random variables are random. $\overline{X}$ and $\hat{\sigma}$ are statistics.)

Definition 8.13 (Point Estimator)**.** A statistic $\hat{\theta}$ is a point estimator of the parameter θ if its purpose is to guess the value of the parameter θ. ($\overline{X}$ is a point estimator for $\mathrm{E}[X] = \mu$; $\hat{\sigma}^2$ is a point estimator for $\mathrm{Var}(X) = \sigma^2$.)

Definition 8.14 (Consistent Estimator, not precise)**.** A point estimator $\hat{\theta}$ is a consistent estimator for θ if it is guaranteed to be arbitrarily close to θ for large sample sizes.

Definition 8.15 (Bias)**.** The bias of the point estimator $\hat{\theta}$ is $\mathrm{bias}(\hat{\theta}, \theta) = \mathrm{E}[\hat{\theta}] - \theta$.

Definition 8.16 (Unbiased)**.** The point estimator $\hat{\theta}$ is said to be unbiased if $\mathrm{E}[\hat{\theta}] = \theta$.

Example. The sample mean $\overline{X}$ is an unbiased estimator of μ. However, $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$ is a biased estimator of σ^2 because

$$\mathrm{E}[\hat{\sigma}^2] = \frac{n-1}{n}\sigma^2 \neq \sigma^2.$$

The bias of $\hat{\sigma}^2$ is $\mathrm{bias}(\hat{\sigma}^2, \sigma^2) = \mathrm{E}[\hat{\sigma}^2] - \sigma^2 = -\sigma^2/n < 0$. Thus $\hat{\sigma}^2$ will tend to underestimate σ^2. The statistic $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$ is an unbiased estimator of σ^2 since $\mathrm{E}[S^2] = \sigma^2$.

Definition 8.17 (Standard Error)**.** The standard error of a point estimator $\hat{\theta}$ is its standard deviation, $\sigma_{\hat{\theta}} = \sqrt{\mathrm{Var}(\hat{\theta})}$. (Notice that $\sigma_{\hat{\theta}}$ is a fixed constant.)

Definition 8.18 (Mean Squared Error (MSE))**.** The mean squared error of a point estimator $\hat{\theta}$ for the parameter θ is

$$MSE(\hat{\theta}, \theta) = \mathrm{E}[(\hat{\theta} - \theta)^2].$$

Result 8.19. The mean squared error is variance plus bias squared:

$$MSE(\hat{\theta}, \theta) = \mathrm{E}[(\hat{\theta} - \theta)^2] = \mathrm{E}[(\hat{\theta} - \mathrm{E}[\hat{\theta}]^2)] + (\mathrm{E}[\hat{\theta}] - \theta)^2 = \mathrm{Var}(\hat{\theta}) + \mathrm{bias}(\hat{\theta}, \theta)^2.$$

Definition 8.20 (Root MSE)**.** The root mean squared error (RMSE) is the square root of the MSE. (MSE is in [units]^2, while RMSE is in [units].)

Remark 50. *Some biased point estimators are good in the sense that they have a small MSE.*

Remark 51. *Method of Maximum Likelihood is a method of deriving estimators for parameters of a distributional family.*

8.5.1 Method of Moments Estimator (MOME)

Definition 8.21 (rth Moment)**.** The rth moment of a random variable X is $\mathrm{E}[X^r]$.

Definition 8.22 (rth Sample Moment)**.** For iid random variables $X_1, X_2, \ldots, X_n$, the rth sample moment is

$$\frac{1}{n}\sum_{i=1}^{n} X_i^r.$$

Remark 52. *A method of moments estimator (MOME) is derived by setting the first m population moments to the first m sample moments and solving for the unknown parameters.*

Example. For an exponential random variable, the first moment is $\mathrm{E}[X] = 1/\lambda$ and the first sample moment is $\overline{X}$. Then setting $\mathrm{E}[X] = 1/\lambda = \overline{X}$ and solving for the unknown quantity λ implies that the MOME estimator of λ is $\hat{\lambda}_{MOME} = 1/\overline{X}$.

8.5.2 Maximum Likelihood Estimator (MLE)

Definition 8.23 (Likelihood)**.** The likelihood L of a sample $(x_1, x_2, \ldots, x_n)$ is $f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n)$, the n-dimensional joint pmf (if discrete) or joint pdf (if continuous) of $(X_1, X_2, \ldots, X_n)$ evaluated at the observed data points $(x_1, x_2, \ldots, x_n)$. (In the discrete case, the likelihood can be interpreted as the probability of observing the data.)

Definition 8.24 (Likelihood Function)**.** The likelihood function of an observed sample $(x_1, x_2, \ldots, x_n)$ is

$$L(\theta \mid x_1, x_2, \ldots, x_n) = f_{X_1,X_2,\ldots,X_n}(\theta \mid x_1, x_2, \ldots, x_n),$$

where θ is a single unknown distributional parameter (the distributional family must be assumed by the analyst).

Result 8.25. If $X_1, X_2, \ldots, X_n$ are iid and the observed data from this random sample is $x_1, x_2, \ldots, x_n$, then the sample's likelihood function is

$$L(\theta \mid x_1, x_2, \ldots, x_n) = f_X(\theta \mid x_1) f_X(\theta \mid x_2) \cdots f_X(\theta \mid x_n) = \prod_{i=1}^{n} f_X(\theta \mid x_i).$$

Definition 8.26 (Maximum Likelihood Estimator (MLE))**.** The maximum likelihood estimator of θ is the (feasible) value of θ that maximizes $L(\theta \mid x_1, x_2, \ldots, x_n)$.

Result 8.27. The value of θ that maximizes L also maximizes any continuous monotonic function of L, such as $\ln(L)$.

Remark 53. *The estimators $(\overline{X}, \hat{\sigma}^2)$ are the MLEs for the normal distribution parameters (μ, σ^2), where $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})$.*

Result 8.28. Under mild conditions, when n is large and $\hat{\theta}_{MLE}$ is an MLE of θ,

- $\hat{\theta}$ is an approximately unbiased estimator for θ,
- the variance of $\hat{\theta}$ is nearly as small as the variance that could be obtained with any other estimator,
- $\hat{\theta}$ has an approximately normal distribution.