3. Probability Theory

3.1 Basic Ideas, Definitions, and Properties

**POPULATION** = Unlimited supply of five types of fruit, in equal proportions.

- \( O_1 = \text{Macintosh apple} \)
- \( O_2 = \text{Golden Delicious apple} \)
- \( O_3 = \text{Granny Smith apple} \)
- \( O_4 = \text{Cavendish banana} \)
- \( O_5 = \text{Plantain banana} \)

**Experiment 1**: Randomly select one fruit from this population, and record its type.

**Sample Space**: The set \( S \) of all possible **elementary outcomes** of an experiment.

\[
S = \{O_1, O_2, O_3, O_4, O_5\} \\
\#(S) = 5
\]

**Event**: Any subset of a sample space \( S \). (“Elementary outcomes” = **simple events**.)

- \( A = \text{“Select an apple.”} = \{O_1, O_2, O_3\} \) \( \#(A) = 3 \)
- \( B = \text{“Select a banana.”} = \{O_4, O_5\} \) \( \#(B) = 2 \)

**Event** \( P(\text{Event}) \)

- \( A \) \( \frac{3}{5} = 0.6 \)
- \( B \) \( \frac{2}{5} = 0.4 \)

As \( \# \text{ trials} \to \infty \) \[ \begin{align*}
    P(A) & = 0.6 & \text{“The probability of randomly selecting an apple is 0.6.”} \\
    P(B) & = 0.4 & \text{“The probability of randomly selecting a banana is 0.4.”}
\end{align*} \]
General formulation may be facilitated with the use of a Venn diagram:

**Experiment \implies Sample Space:** \( S = \{O_1, O_2, \ldots, O_k\} \quad \#(S) = k \)

**Event** \( A = \{O_1, O_2, \ldots, O_m\} \subseteq S \quad \#(A) = m \leq k \)

**Definition:** The **probability of event** \( A \), denoted \( P(A) \), is the long-run relative frequency with which \( A \) is expected to occur, *as the experiment is repeated indefinitely*.

---

**Fundamental Properties of Probability**

For any event \( A = \{O_1, O_2, \ldots, O_m\} \) in a sample space \( S \),

1. \( 0 \leq P(A) \leq 1 \)

2. \( P(A) = \sum_{i=1}^{m} P(O_i) = P(O_1) + P(O_2) + P(O_3) + \ldots + P(O_m) \)

**Special Cases:**

- \( P(\emptyset) = 0 \)
- \( P(S) = \sum_{i=1}^{k} P(O_i) = 1 \) "certainty"

3. If all the elementary outcomes of \( S \) are **equally likely**, i.e.,

\[
P(O_1) = P(O_2) = \ldots = P(O_k) = \frac{1}{k},
\]

then...

\[
P(A) = \frac{\#(A)}{\#(S)} = \frac{m}{k}.
\]

**Example:** \( P(A) = \frac{3}{5} = 0.6, \quad P(B) = \frac{2}{5} = 0.4 \)
Experiment 2: Select a card at random from a standard deck (and replace).

Sample Space: \( S = \{ \text{A }\spadesuit, \ldots, \text{K }\spadesuit \} \)  
\(#(S) = 52\)

Events:  
\( A = \text{“Select a 2.”} = \{ \text{2 }\spadesuit, \text{2 }\clubsuit, \text{2 }\heartsuit, \text{2 }\diamondsuit \} \)  
\(#(A) = 4\)

\( B = \text{“Select a }\clubsuit.\text{”} = \{ \text{A }\clubsuit, \text{2 }\clubsuit, \ldots, \text{K }\clubsuit \} \)  
\(#(B) = 13\)

Probabilities: Since all elementary outcomes are equally likely, it follows that

\[ P(A) = \frac{#(A)}{#(S)} = \frac{4}{52} \quad \text{and} \quad P(B) = \frac{#(B)}{#(S)} = \frac{13}{52}. \]

**New Events from Old Events**

(1) \( A^C = \text{“not } A\text{”} = \{ \text{All outcomes that are in } S, \text{ but not in } A.\} \)

\[ P(A^C) = 1 - P(A) \]

Example: \( A^C = \text{“Select either A, 3, 4, \ldots, or K.”} \quad P(A^C) = 1 - \frac{4}{52} = \frac{48}{52} \).

Example: **Experiment** = Toss a coin once.

Events: \( A = \{ \text{Heads} \} \quad A^C = \{ \text{Tails} \} \)

Probabilities:

*Fair* coin… \( P(A) = 0.5 \quad \Rightarrow \quad P(A^C) = 1 - 0.5 = 0.5 \)

*Biased* coin… \( P(A) = 0.7 \quad \Rightarrow \quad P(A^C) = 1 - 0.7 = 0.3 \)
intersection

(2) \( A \cap B = \{ \text{All outcomes in } S \text{ that } A \text{ and } B \text{ share in common.} \} \)

\[ = \{ \text{All outcomes that result when events } A \text{ and } B \text{ occur simultaneously.} \} \]

**Example:** \( A \cap B = \text{“Select a 2 and a ♣” } = \{2♣\} \Rightarrow P(A \cap B) = \frac{1}{52}. \)

**Definition:** Two events \( A \) and \( B \) are said to be **disjoint**, or **mutually exclusive**, if they cannot occur simultaneously, i.e., \( A \cap B = \emptyset \), hence \( P(A \cap B) = 0. \)

![Venn Diagram](image)

**Example:** \( A = \text{“Select a 2”} \text{ and } C = \text{“Select a 3”} \) are disjoint events.

**Exercise:** Are \( A = \{2^4, 3^4, 4^4, 5^4, \ldots\} \text{ and } B = \{2^6, 3^6, 4^6, 5^6, \ldots\} \) disjoint?

If not, find \( A \cap B. \)

union

(3) \( A \cup B = \{ \text{All outcomes in } S \text{ that are either in } A \text{ or } B, \text{ inclusive.} \} \)

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

\[= 0, \text{ if } A \text{ and } B \text{ are disjoint.} \]

**Example:** \( A \cup B = \text{“Select either a 2 or a ♣”} \) has probability

\[
P(A \cup B) = \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52}.
\]

**Example:** \( A \cup C = \text{“Select either a 2 or a 3”} \) has probability

\[
P(A \cup C) = \frac{4}{52} + \frac{4}{52} - 0 = \frac{8}{52}.
\]
**Note:** Formula (3) extends to \( n \geq 3 \) disjoint events in a straightforward manner:

\[
P(A_1 \cup A_2 \cup \ldots \cup A_n) = P(A_1) + P(A_2) + \ldots + P(A_n).
\]

**Question:** How is this formula modified if the \( n \) events are not necessarily disjoint?

**Example:** Take \( n = 3 \) events…

Then \( P(A \cup B \cup C) = \)

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

**Exercise:** For \( S = \{\text{January,..., December}\} \), verify this formula for the three events \( A = \text{“Has 31 days,”} \) \( B = \text{“Name ends in r,”} \) and \( C = \text{“Name begins with a vowel.”} \)

---

**Exercise:** A single tooth is to be randomly selected for a certain dental procedure. Draw a Venn diagram to illustrate the relationships between the three following events: \( A = \text{“upper jaw,”} \) \( B = \text{“left side,”} \) and \( C = \text{“molar,”} \) and indicate all corresponding probabilities. Calculate the probability that all of these three events, \( A \) and \( B \) and \( C \), occur. Calculate the probability that none of these three events occur. Calculate the probability that exactly one of these three events occurs. Calculate the probability that exactly two of these three events occur. (Think carefully.) Assume equal likelihood in all cases.

The three “set operations” – union, intersection, and complement – can be unified via...

**DeMorgan’s Laws**

\[
(A \cup B)^C = A^C \cap B^C
\]

\[
(A \cap B)^C = A^C \cup B^C
\]

**Exercise:** Using a Venn diagram, convince yourself that these statements are true in general. Then verify them for a specific example, e.g., \( A = \text{“Pick a picture card”} \) and \( B = \text{“Pick a black card.”} \)
**FACT:** Random variables can be used to define events that involve measurement!

**Experiment 3a:** Roll one fair die... **Discrete random variable** $X =$ “value obtained”

**Sample Space:** $S = \{1, 2, 3, 4, 5, 6\}$ \quad $(|S| = 6)$

Because the die is fair, each of the six faces has an equally likely probability of occurring, i.e., 1/6. The **probability distribution** for $X$ (determined by the **probability mass function** (pmf)) $f(x)$ can be organized in a **probability table**, and displayed via a corresponding **probability histogram**, as shown.

<table>
<thead>
<tr>
<th>$x$</th>
<th>$f(x) = P(X = x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1/6</td>
</tr>
<tr>
<td>2</td>
<td>1/6</td>
</tr>
<tr>
<td>3</td>
<td>1/6</td>
</tr>
<tr>
<td>4</td>
<td>1/6</td>
</tr>
<tr>
<td>5</td>
<td>1/6</td>
</tr>
<tr>
<td>6</td>
<td>1/6</td>
</tr>
</tbody>
</table>

**Experiment 3b:** Roll two fair dice. \Rightarrow **Outcome** = (Die 1, Die 2)

**Sample Space:** $S = \{(1, 1), \ldots, (6, 6)\}$ \quad $(|S| = 6^2 = 36)$

**Discrete random variable** $X =$ “Sum of the two dice (2, 3, 4, …, 12).”

**Events:**

- “$X = 2$” = \{(1, 1)\} \quad $(|X = 2| = 1)$
- “$X = 3$” = \{(1, 2), (2, 1)\} \quad $(|X = 3| = 2)$
- “$X = 4$” = \{(1, 3), (2, 2), (3, 1)\} \quad $(|X = 4| = 3)$
- “$X = 5$” = \{(1, 4), (2, 3), (3, 2), (4, 1)\} \quad $(|X = 5| = 4)$
- “$X = 6$” = \{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\} \quad $(|X = 6| = 5)$
- “$X = 7$” = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\} \quad $(|X = 7| = 6)$
- “$X = 8$” = \{(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)\} \quad $(|X = 8| = 5)$
- “$X = 9$” = \{(3, 6), (4, 5), (5, 4), (6, 3)\} \quad $(|X = 9| = 4)$
- “$X = 10$” = \{(4, 6), (5, 5), (6, 4)\} \quad $(|X = 10| = 3)$
- “$X = 11$” = \{(5, 6), (6, 5)\} \quad $(|X = 11| = 2)$
- “$X = 12$” = \{(6, 6)\} \quad $(|X = 12| = 1)$
Again, the probability distribution for $X$ can be organized in a probability table, and displayed via a probability histogram, both of which enable calculations to be done easily:

<table>
<thead>
<tr>
<th>$x$</th>
<th>$f(x) = P(X = x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1/36</td>
</tr>
<tr>
<td>3</td>
<td>2/36</td>
</tr>
<tr>
<td>4</td>
<td>3/36</td>
</tr>
<tr>
<td>5</td>
<td>4/36</td>
</tr>
<tr>
<td>6</td>
<td>5/36</td>
</tr>
<tr>
<td>7</td>
<td>6/36</td>
</tr>
<tr>
<td>8</td>
<td>5/36</td>
</tr>
<tr>
<td>9</td>
<td>4/36</td>
</tr>
<tr>
<td>10</td>
<td>3/36</td>
</tr>
<tr>
<td>11</td>
<td>2/36</td>
</tr>
<tr>
<td>12</td>
<td>1/36</td>
</tr>
</tbody>
</table>

Note that “$X = 7$” and “$X = 11$” are disjoint!

\[
P(X = 7 \text{ or } X = 11) = P(X = 7) + P(X = 11) \quad \text{via Formula (3) above}
\]

\[
= 6/36 + 2/36 = 8/36
\]

\[
P(5 \leq X \leq 8)
\]

\[
= P(X = 5 \text{ or } X = 6 \text{ or } X = 7 \text{ or } X = 8)
\]

\[
= P(X = 5) + P(X = 6) + P(X = 7) + P(X = 8)
\]

\[
= 4/36 + 5/36 + 6/36 + 5/36
\]

\[
= 20/36
\]

\[
P(X < 10) = 1 - P(X \geq 10) \quad \text{via Formula (1) above}
\]

\[
= 1 - [P(X = 10) + P(X = 11) + P(X = 12)]
\]

\[
= 1 - [3/36 + 2/36 + 1/36] = 1 - 6/36 = 30/36
\]

**Exercise:** How could event $E = “\text{Roll doubles}”$ be characterized in terms of a random variable? (Hint: Let $Y = “\text{Difference between the two dice.}”$)
The previous example motivates the important topic of...

**Discrete Probability Distributions**

In general, suppose that all of the distinct population values of a *discrete* random variable $X$ are sorted in increasing order: $x_1 < x_2 < x_3 < ..., \text{ with corresponding probabilities of occurrence } f(x_1), f(x_2), f(x_3), ...$ Formally then, we have the following.

**Definition:** $f(x)$ is a **probability distribution function** for the *discrete* random variable $X$ if, for all $x$,

$$f(x) \geq 0 \quad \text{AND} \quad \sum_{\text{all } x} f(x) = 1.$$

In this case, $f(x) = P(X = x)$, the *probability* that the value $x$ occurs in the population.

The **cumulative distribution function** (cdf) is defined as, for all $x$,

$$F(x) = P(X \leq x) = \sum_{\text{all } x_i \leq x} f(x_i) = f(x_1) + f(x_2) + ... + f(x).$$

Therefore, $F$ is **piecewise constant**, increasing from 0 to 1.

Furthermore, for any two population values $a < b$, it follows that

$$P(a \leq X \leq b) = \sum_{a}^{b} f(x) = F(b) - F(a^-)$$

where $a^-$ is the value just preceding $a$ in the sorted population.

**Exercise:** Sketch the cdf $F(x)$ for Experiments 3a and 3b above.
Population Parameters $\mu$ and $\sigma^2$ (vs. Sample Statistics $\bar{x}$ and $s^2$)

- **population mean** = the “expected value” of the random variable $X$
  
  = the “arithmetic average” of all the population values

If $X$ is a *discrete* numerical random variable, then…

$$
\mu = E[X] = \sum x f(x), \quad \text{where } f(x) = P(X = x), \text{ the probability of } x.
$$

Compare this with the *relative frequency* definition of *sample mean* given in §2.3.

**Properties of Mathematical Expectation**

1. For any constant $c$, it follows that $E[cX] = cE[X]$.
2. For any two random variables $X$ and $Y$, it follows that

   - $E[X + Y] = E[X] + E[Y]$ and, via Property 1,

Any “operator” on variables satisfying 1 and 2 is said to be *linear*.

- **population variance** = the “expected value” of the squared deviation of the random variable $X$ from its mean ($\mu$)

If $X$ is a *discrete* numerical random variable, then…

$$
\sigma^2 = E[(X - \mu)^2] = \sum (x - \mu)^2 f(x).
$$

Equivalently,*

$$
\sigma^2 = E[X^2] - \mu^2 = \sum x^2 f(x) - \mu^2,
$$

where $f(x) = P(X = x)$, the probability of $x$.

Compare the first with the definition of *sample variance* given in §2.3. (The second is the analogue of the *alternate computational formula.*) Of course, the *population standard deviation* $\sigma$ is defined as the square root of the variance.

---

*Exercise: Algebraically expand the expression $(X - \mu)^2$, and use the properties of expectation given above.
Experiment 4: At a party, guests randomly select one pastry from two trays (continually refilled), where the distribution of calories $X_1$ and $X_2$ are indicated below.

**Tray 1**

Probability Table

<table>
<thead>
<tr>
<th>$x$</th>
<th>$f_1(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>90</td>
<td>1/3</td>
</tr>
<tr>
<td>120</td>
<td>1/3</td>
</tr>
<tr>
<td>150</td>
<td>1/3</td>
</tr>
</tbody>
</table>

Equally likely outcomes

- Mean($X_1$) = $\mu_1 = (90)(1/3) + (120)(1/3) + (150)(1/3) = 120$ cals
- Var($X_1$) = $\sigma_1^2 = (-30)^2(1/3) + (0)^2(1/3) + (30)^2(1/3) = 600$ cals$^2$

**Tray 2**

Probability Table

<table>
<thead>
<tr>
<th>$x$</th>
<th>$f_2(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>3/6</td>
</tr>
<tr>
<td>60</td>
<td>2/6</td>
</tr>
<tr>
<td>90</td>
<td>1/6</td>
</tr>
</tbody>
</table>

- Mean($X_2$) = $\mu_2 = (30)(3/6) + (60)(2/6) + (90)(1/6) = 50$ cals
- Var($X_2$) = $\sigma_2^2 = (-20)^2(3/6) + (10)^2(2/6) + (40)^2(1/6) = 500$ cals$^2$
**Summary** (Also refer back to 2.4 - Summary)

**POPULATION**

*Discrete random variable* \( X \)

**Probability Table** → **Probability Histogram**

<table>
<thead>
<tr>
<th>( x )</th>
<th>( f(x) = P(X = x) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 )</td>
<td>( f(x_1) )</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>( f(x_2) )</td>
</tr>
<tr>
<td>. . . . .</td>
<td>. . . . .</td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

\[ \mu = E[X] = \sum x f(x) \]

\[ \sigma^2 = \begin{cases} E[(X - \mu)^2] = \sum (x - \mu)^2 f(x) \\ \text{or} \\ E[X^2] - \mu^2 = \sum x^2 f(x) - \mu^2 \end{cases} \]

**SAMPLE, size** \( n \)

**Relative Frequency Table** → **Density Histogram**

<table>
<thead>
<tr>
<th>( x )</th>
<th>( f(x) = \frac{\text{freq}(x)}{n} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 )</td>
<td>( f(x_1) )</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>( f(x_2) )</td>
</tr>
<tr>
<td>. . . . .</td>
<td>. . . . .</td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

\[ \bar{x} = \sum x f(x) \]

\[ s^2 = \begin{cases} \frac{n}{n-1} \sum (x - \bar{x})^2 f(x) \\ \text{or} \\ \frac{n}{n-1} \left[ \sum x^2 f(x) - \bar{x}^2 \right] \end{cases} \]

\( \bar{x} \) and \( s^2 \) can be shown to be unbiased estimators of \( \mu \) and \( \sigma^2 \), respectively. That is, \( E[\bar{x}] = \mu \), and \( E[S^2] = \sigma^2 \). (In fact, they are MVUE.)
Some Notes on General Parameter Estimation

Suppose that $\theta$ is a fixed population parameter (e.g., $\mu$), and $\hat{\theta}$ is a sample-based estimator (e.g., $\bar{X}$). Consider all the random samples of a given size $n$, and the resulting “sampling distribution” of $\hat{\theta}$ values. Formally define the following:

- **Mean** (of $\hat{\theta}$) = $E[\hat{\theta}]$, the expected value of $\hat{\theta}$.

- **Bias** = $E[\hat{\theta}] - \theta$, the difference between the expected value of $\hat{\theta}$, and the “target” parameter $\theta$.

- **Variance** (of $\hat{\theta}$) = $E\left(\left(\hat{\theta} - E[\hat{\theta}]\right)^2\right)$, the expected value of the squared deviation of $\hat{\theta}$ from its mean $E[\hat{\theta}]$, or equivalently,* $E\left[\hat{\theta}^2\right] - E[\hat{\theta}]^2$.

- **Mean Squared Error (MSE)** = $E\left(\left(\hat{\theta} - \theta\right)^2\right)$, the expected value of the squared difference between estimator $\hat{\theta}$ and the “target” parameter $\theta$.

**Exercise:** Prove* that $\text{MSE} = \text{Variance} + \text{Bias}^2$.

**Comment:** A parameter estimator $\hat{\theta}$ is defined to be **unbiased** if $E[\hat{\theta}] = \theta$, i.e., $\text{Bias} = 0$. In this case, $\text{MSE} = \text{Variance}$, so that if $\hat{\theta}$ minimizes $\text{MSE}$, it then follows that it has the smallest variance of any estimator. Such a highly desirable estimator is called MVUE (Minimum Variance Unbiased Estimator). It can be shown that the estimators $\bar{X}$ and $S^2$ (of $\mu$ and $\sigma^2$, respectively) are MVUE, but finding such an estimator $\hat{\theta}$ for a general parameter $\theta$ can be quite difficult in practice. Often, one must settle for either not having minimum variance or having a small amount of bias.

* using the basic properties of **mathematical expectation** given earlier
Related (but not identical) to this is the idea that, of all linear combinations 
\[c_1 x_1 + c_2 x_2 + \ldots + c_n x_n\] of the data \(\{x_1, x_2, \ldots, x_n\}\) (such as \(\bar{X}\), with \(c_1 = c_2 = \ldots = c_n = 1/n\)) which are also unbiased, the one that minimizes MSE is called BLUE (Best Linear Unbiased Estimator). It can be shown that, in addition to being MVUE (as stated above), \(\bar{X}\) is also BLUE. To summarize,

**MVUE** gives: \(\text{Min Variance among all unbiased estimators} \leq \text{Min Variance among linear unbiased estimators} = \text{Min MSE among linear unbiased estimators} \) (since \(\text{MSE} = \text{Var} + \text{Bias}^2\)), given by **BLUE** (by def).

The Venn diagram below depicts these various relationships.

Comment: If \(\text{MSE} \to 0\) as \(n \to \infty\), then \(\hat{\theta}\) is said to have mean square convergence to \(\theta\). This in turn implies “convergence in probability” (via “Markov's Inequality,” also used in proving Chebyshev’s Inequality), i.e., \(\hat{\theta}\) is a consistent estimator of \(\theta\).