4. Classical Probability Distributions

4.1 Discrete Models

**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:** $S = \{1, 2, 3, 4, 5, 6\}$ \#($S$) = 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 $X$ can be defined by a so-called *probability mass function* (pmf) $f(x)$, organized in a *probability table*, and displayed via a corresponding *probability histogram*, as shown.

<table>
<thead>
<tr>
<th>Event $x$</th>
<th>Probability $f(x) = P(X = x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\frac{1}{6}$</td>
</tr>
<tr>
<td>2</td>
<td>$\frac{1}{6}$</td>
</tr>
<tr>
<td>3</td>
<td>$\frac{1}{6}$</td>
</tr>
<tr>
<td>4</td>
<td>$\frac{1}{6}$</td>
</tr>
<tr>
<td>5</td>
<td>$\frac{1}{6}$</td>
</tr>
<tr>
<td>6</td>
<td>$\frac{1}{6}$</td>
</tr>
</tbody>
</table>

Comment on notation:

*Translation:* “The *probability* of rolling 4 is 1/6.”

Likewise for the other probabilities $P(X = 1)$, $P(X = 2)$, ..., $P(X = 6)$ in this example. A mathematically succinct way to write such probabilities is by the notation $P(X = x)$, where $x = 1, 2, 3, 4, 5, 6$. In general therefore, since this *depends* on the value of $x$, we can also express it as a mathematical *function* of $x$ (specifically, the pmf; see above), written $f(x)$. Thus the two notations are synonymous and interchangeable. The previous example could just as well have been written $f(4) = 1/6$. 
Experiment 3b: Roll two distinct, fair dice. \( \Rightarrow \) Outcome = (Die 1, Die 2)

Sample Space: \( S = \{(1, 1), \ldots, (6, 6)\} \) \( \#(S) = 6^2 = 36 \)

<table>
<thead>
<tr>
<th>(1, 1)</th>
<th>(1, 2)</th>
<th>(1, 3)</th>
<th>(1, 4)</th>
<th>(1, 5)</th>
<th>(1, 6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(2, 1)</td>
<td>(2, 2)</td>
<td>(2, 3)</td>
<td>(2, 4)</td>
<td>(2, 5)</td>
<td>(2, 6)</td>
</tr>
<tr>
<td>(3, 1)</td>
<td>(3, 2)</td>
<td>(3, 3)</td>
<td>(3, 4)</td>
<td>(3, 5)</td>
<td>(3, 6)</td>
</tr>
<tr>
<td>(4, 1)</td>
<td>(4, 2)</td>
<td>(4, 3)</td>
<td>(4, 4)</td>
<td>(4, 5)</td>
<td>(4, 6)</td>
</tr>
<tr>
<td>(5, 1)</td>
<td>(5, 2)</td>
<td>(5, 3)</td>
<td>(5, 4)</td>
<td>(5, 5)</td>
<td>(5, 6)</td>
</tr>
<tr>
<td>(6, 1)</td>
<td>(6, 2)</td>
<td>(6, 3)</td>
<td>(6, 4)</td>
<td>(6, 5)</td>
<td>(6, 6)</td>
</tr>
</tbody>
</table>

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

Events:

\( \text{“}X = 2\text{”} = \{(1, 1)\} \) \( \#(X = 2) = 1 \)

\( \text{“}X = 3\text{”} = \{(1, 2), (2, 1)\} \) \( \#(X = 3) = 2 \)

\( \text{“}X = 4\text{”} = \{(1, 3), (2, 2), (3, 1)\} \) \( \#(X = 4) = 3 \)

\( \text{“}X = 5\text{”} = \{(1, 4), (2, 3), (3, 2), (4, 1)\} \) \( \#(X = 5) = 4 \)

\( \text{“}X = 6\text{”} = \{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\} \) \( \#(X = 6) = 5 \)

\( \text{“}X = 7\text{”} = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\} \) \( \#(X = 7) = 6 \)

\( \text{“}X = 8\text{”} = \{(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)\} \) \( \#(X = 8) = 5 \)

\( \text{“}X = 9\text{”} = \{(3, 6), (4, 5), (5, 4), (6, 3)\} \) \( \#(X = 9) = 4 \)

\( \text{“}X = 10\text{”} = \{(4, 6), (5, 5), (6, 4)\} \) \( \#(X = 10) = 3 \)

\( \text{“}X = 11\text{”} = \{(5, 6), (6, 5)\} \) \( \#(X = 11) = 2 \)

\( \text{“}X = 12\text{”} = \{(6, 6)\} \) \( \#(X = 12) = 1 \)

Recall that, by definition, each event “\( X = x \)” (where \( x = 2, 3, 4, \ldots, 12 \)) corresponds to a specific subset of outcomes from the sample space (of ordered pairs, in this case). Because we are still assuming equal likelihood of each die face appearing, the probabilities of these events can be easily calculated by the “shortcut” formula

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

Question for later: What if the dice are “loaded” (i.e., biased)?
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>

$$P(X = 7 \text{ or } X = 11) \quad \text{Note that } "X = 7" \text{ and } "X = 11" \text{ are disjoint!}$$

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

$= \frac{6}{36} + \frac{2}{36} = \frac{8}{36}$

$$P(5 \leq X \leq 8)$$

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

$= P(X = 5) + P(X = 6) + P(X = 7) + P(X = 8)$

$= \frac{4}{36} + \frac{5}{36} + \frac{6}{36} + \frac{5}{36}$

$= \frac{20}{36}$

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

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

$= 1 - \left[ \frac{3}{36} + \frac{2}{36} + \frac{1}{36} \right] = 1 - \frac{6}{36} = \frac{30}{36}$

**Exercise:** How could event $E =$ “Roll doubles” be characterized in terms of a random variable? *(Hint: Let $Y =$ “Difference between the two dice.”)*

\[x\]

\[\begin{array}{c|c}
 x & f(x) = P(X = x) \\
\hline
 2 & 1/36 \\
 3 & 2/36 \\
 4 & 3/36 \\
 5 & 4/36 \\
 6 & 5/36 \\
 7 & 6/36 \\
 8 & 5/36 \\
 9 & 4/36 \\
10 & 3/36 \\
11 & 2/36 \\
12 & 1/36 \\
\end{array}\]

\[\text{Probability Histogram}\]

\[\begin{array}{c}
 x \\
\hline
 2 & \frac{1}{36} \\
 3 & \frac{2}{36} \\
 4 & \frac{3}{36} \\
 5 & \frac{4}{36} \\
 6 & \frac{5}{36} \\
 7 & \frac{6}{36} \\
 8 & \frac{5}{36} \\
 9 & \frac{4}{36} \\
10 & \frac{3}{36} \\
11 & \frac{2}{36} \\
12 & \frac{1}{36} \\
\end{array}\]
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: $x_1 < x_2 < x_3 < \ldots$, with corresponding probabilities of occurrence $f(x_1), f(x_2), f(x_3), \ldots$ Formally then, we have the following.

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

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

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 $x$,

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

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

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 $E[cX] = c E[X]$.
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:** Two populations, where the daily number of calories consumed is designated by $X_1$ and $X_2$, respectively.

**Population 1**

![Probability Table for Population 1](image)

<table>
<thead>
<tr>
<th>$x$</th>
<th>$f_1(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2300</td>
<td>0.1</td>
</tr>
<tr>
<td>2400</td>
<td>0.2</td>
</tr>
<tr>
<td>2500</td>
<td>0.3</td>
</tr>
<tr>
<td>2600</td>
<td>0.4</td>
</tr>
</tbody>
</table>

- **Mean** ($\mu_1$) = (2300)(0.1) + (2400)(0.2) + (2500)(0.3) + (2600)(0.4) = **2500 cals**
- **Var** ($\sigma_1^2$) = (–200)$^2$(0.1) + (–100)$^2$(0.2) + (0)$^2$(0.3) + (+100)$^2$(0.4) = **10000 cals$^2$**

**Population 2**

![Probability Table for Population 2](image)

<table>
<thead>
<tr>
<th>$x$</th>
<th>$f_2(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2200</td>
<td>0.2</td>
</tr>
<tr>
<td>2300</td>
<td>0.3</td>
</tr>
<tr>
<td>2400</td>
<td>0.5</td>
</tr>
</tbody>
</table>

- **Mean** ($\mu_2$) = (2200)(0.2) + (2300)(0.3) + (2400)(0.5) = **2330 cals**
- **Var** ($\sigma_2^2$) = (–130)$^2$(0.2) + (–30)$^2$(0.3) + (70)$^2$(0.5) = **6100 cals$^2$**
**Summary** (Also refer back to 2.4 - Summary)

**POPULATION**

*Discrete* random variable $X$

<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>$\sum x f(x)$</td>
</tr>
</tbody>
</table>

**Parameters**

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

$$\sigma^2 = \left\{ \begin{array}{l} E[(X - \mu)^2] = \sum (x - \mu)^2 f(x) \\ \text{or} \\ E[X^2] - \mu^2 = \sum x^2 f(x) - \mu^2 \end{array} \right.$$  

**SAMPLE, size $n$**

*Relative Frequency Table* → *Density Histogram*

<table>
<thead>
<tr>
<th>$x$</th>
<th>$f(x) = \frac{freq(x)}{n}$</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>$x_k$</td>
<td>$f(x_k)$</td>
</tr>
<tr>
<td>1</td>
<td>$\frac{1}{n} \sum x f(x)$</td>
</tr>
</tbody>
</table>

**Statistics**

$$\bar{x} = \sum x f(x)$$

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

$\bar{X}$ 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 Advanced Notes on General Parameter Estimation ~

Suppose that \( \theta \) is a \textit{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 \textit{“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)** = \( E \left[ (\hat{\theta} - \theta)^2 \right] \), 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 \textit{unbiased} if \( E[\hat{\theta}] = \theta \), i.e., Bias = 0. In this case, MSE = Variance, so that if \( \hat{\theta} \) minimizes 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 \textit{mathematical expectation} given earlier
Related (but not identical) to this is the idea that of all **linear** combinations $c_1x_1 + c_2x_2 + \ldots + c_nx_n$ 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: \[
\text{Min Variance among all unbiased estimators} \\
\leq \text{Min Variance among linear unbiased estimators} \\
= \text{Min MSE among linear unbiased estimators} \quad (\text{since } \text{MSE} = \text{Var} + \text{Bias}^2),
\]

given by **BLUE** \quad (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 $\theta$. 
**Experiment 4 - revisited:** Recall the previous example, where \( X_1 \) and \( X_2 \) represent the daily number of calories consumed in two populations, respectively.

**Population 1**

![Pie chart for Population 1]

**Population 2**

![Pie chart for Population 2]

<table>
<thead>
<tr>
<th>( x )</th>
<th>( f_1(x) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2300</td>
<td>0.1</td>
</tr>
<tr>
<td>2400</td>
<td>0.2</td>
</tr>
<tr>
<td>2500</td>
<td>0.3</td>
</tr>
<tr>
<td>2600</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Mean(\( X_1 \)) = \( \mu_1 = 2500 \) cals;
Var(\( X_1 \)) = \( \sigma_1^2 = 10000 \) cals\(^2\)

<table>
<thead>
<tr>
<th>( x )</th>
<th>( f_2(x) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2200</td>
<td>0.2</td>
</tr>
<tr>
<td>2300</td>
<td>0.3</td>
</tr>
<tr>
<td>2400</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Mean(\( X_2 \)) = \( \mu_2 = 2330 \) cals;
Var(\( X_2 \)) = \( \sigma_2^2 = 6100 \) cals\(^2\)

**Case 1:** First suppose that \( X_1 \) and \( X_2 \) are **statistically independent**, as shown in the **joint probability distribution** given in the table below. That is, each cell probability is equal to the **product** of the corresponding row and column **marginal** probabilities. For example, \( P(X_1 = 2300 \cap X_2 = 2200) = 0.02 \), but this is equal to the product of the column marginal \( P(X_1 = 2300) = 0.1 \) with the row marginal \( P(X_2 = 2200) = 0.2 \). Note that the marginal distributions for \( X_1 \) and \( X_2 \) remain the same as above, as can be seen from the single-underlined values for \( X_1 \), and respectively, the double-underlined values for \( X_2 \).

**\( X_1 \) = # calories for Pop 1**

<table>
<thead>
<tr>
<th>( X_1 ) = # calories</th>
</tr>
</thead>
<tbody>
<tr>
<td>2300</td>
</tr>
<tr>
<td>2200</td>
</tr>
<tr>
<td>2300</td>
</tr>
<tr>
<td>2400</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( X_2 ) = # calories for Pop 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X_2 )</td>
</tr>
<tr>
<td>---------------------</td>
</tr>
<tr>
<td>2200</td>
</tr>
</tbody>
</table>
Now imagine that we wish to compare the two populations, by considering the probability distribution of the calorie difference $D = X_1 - X_2$ between them. (The sum $S = X_1 + X_2$ is similar, and left as an exercise.)

As an example, there are two possible ways that $D = 300$ can occur, i.e., two possible outcomes corresponding to the event $D = 300$: Either $A = \{X_1 = 2500 \text{ and } X_2 = 2200\}$ or $B = \{X_1 = 2600 \text{ and } X_2 = 2300\}$, that is, $A \cup B$. For its probability, recall that $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. However, events $A$ and $B$ are disjoint, for they cannot both occur simultaneously, so that the last term is $P(A \cap B) = 0$. Thus, $P(A \cup B) = P(A) + P(B)$ with $P(A) = .06$ and $P(B) = .12$ from the joint distribution.

<table>
<thead>
<tr>
<th>Events $D = d$</th>
<th>Sample Space: Outcomes in the form of ordered pairs $(X_1, X_2)$</th>
<th>Probabilities from joint distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D = -100$:</td>
<td>(2300, 2400)</td>
<td>.05</td>
</tr>
<tr>
<td>$D = 0$:</td>
<td>(2300, 2300), (2400, 2400)</td>
<td>.13 = .03 + .10</td>
</tr>
<tr>
<td>$D = +100$:</td>
<td>(2300, 2200), (2400, 2300), (2500, 2400)</td>
<td>.23 = .02 + .06 + .15</td>
</tr>
<tr>
<td>$D = +200$:</td>
<td>(2400, 2200), (2500, 2300), (2600, 2400)</td>
<td>.33 = .04 + .09 + .20</td>
</tr>
<tr>
<td>$D = +300$:</td>
<td>(2500, 2200), (2600, 2300)</td>
<td>.18 = .06 + .12</td>
</tr>
<tr>
<td>$D = +400$:</td>
<td>(2600, 2200)</td>
<td>.08</td>
</tr>
</tbody>
</table>

$\text{Mean}(D) = \mu_D =$

$(-100)(.05) + (0)(.13) + (100)(.23) + (200)(.33) + (300)(.18) + (400)(.08) = 170 \text{ cals}$

i.e., $\mu_D = \mu_1 - \mu_2$ (Check this!)

$\text{Var}(D) = \sigma_D^2 =$

$(-270)^2(.05) + (-170)^2(.13) + (-70)^2(.23)$
$+ (30)^2(.33) + (130)^2(.18) + (230)^2(.08)$

$= 16100 \text{ cals}^2$

i.e., $\sigma_D^2 = \sigma_1^2 + \sigma_2^2$ (Check this!)
**Case 2:** Now assume that $X_1$ and $X_2$ are not statistically independent, as given in the joint probability distribution table below.

<table>
<thead>
<tr>
<th>$X_2$ = # calories for Pop 2</th>
<th>2200</th>
<th>2300</th>
<th>2400</th>
<th>2500</th>
<th>2600</th>
</tr>
</thead>
<tbody>
<tr>
<td>2300</td>
<td>.01</td>
<td>.02</td>
<td>.07</td>
<td>.09</td>
<td>.20</td>
</tr>
<tr>
<td>2400</td>
<td>.05</td>
<td>.10</td>
<td>.13</td>
<td>.18</td>
<td>.30</td>
</tr>
<tr>
<td>2400</td>
<td>.07</td>
<td>.12</td>
<td>.13</td>
<td>.18</td>
<td>.50</td>
</tr>
<tr>
<td></td>
<td>.10</td>
<td>.20</td>
<td>.30</td>
<td>.40</td>
<td>1.00</td>
</tr>
</tbody>
</table>

The events “$D = d$” and the corresponding sample space of outcomes remain unchanged, but the last column of probabilities has to be recalculated, as shown. This results in a slightly different probability histogram (Exercise) and parameter values.

Mean($D$) = $\mu_D = (-100)(.07) + (0)(.14) + (100)(.19) + (200)(.33) + (300)(.18) + (400)(.08) = 170$ cals, i.e., $\mu_D = \mu_1 - \mu_2$.

Var($D$) = $\sigma_D^2 = (-270)^2(.07) + (-170)^2(.14) + (-70)^2(.19) + (30)^2(.31) + (130)^2(.20) + (230)^2(.09) = 18517$ cals$^2$

It seems that “the mean of the difference is equal to the difference in the means” still holds, even when the two populations are dependent. But the variance of the difference is no longer necessarily equal the sum of the variances, as with independent populations.
These examples illustrate a general principle that can be rigorously proved with mathematics.

**GENERAL FACT ~**

| Mean(X + Y) = Mean(X) + Mean(Y) | and | Mean(X − Y) = Mean(X) − Mean(Y) |
| Var(X + Y) = Var(X) + Var(Y) | and | Var(X − Y) = Var(X) + Var(Y). |

**Comments:**

- These formulas actually apply to both *discrete* and *continuous* variables (next section).
- The difference relations will play a crucial role in 6.2 - Two Samples inference.
- If X and Y are *dependent*, then the two bottom relations regarding the variance also involve an additional term, Cov(X, Y), the population *covariance* between X and Y. See problems 4.3/29 and 4.3/30 for details.
- The variance relation can be interpreted visually via the Pythagorean Theorem, which illustrates an important *geometric* connection, expanded in the Appendix.

![Pythagorean Theorem Diagram](image)

Certain discrete distributions (or discrete models) occur so frequently in practice, that their properties have been well-studied and applied in many different scenarios. For instance, suppose it is known that a certain population consists of 45% males (and thus 55% females). If a random sample of 250 individuals is to be selected, then what is the probability of obtaining exactly 100 males? *At most* 100 males? *At least* 100 males? What is the “expected” number of males? This is the subject of the next topic:
POPULATION = Women diagnosed with breast cancer in Dane County, 1996-2000

Among other things, this study estimated that the rate of “breast cancer in situ (BCIS),” which is diagnosed almost exclusively via mammogram, is approximately 12-13%. That is, for any individual randomly selected from this population, we have a binary variable

$$
BCIS = \begin{cases} 
1, & \text{with probability } 0.12 \\
0, & \text{with probability } 0.88.
\end{cases}
$$

In a random sample of $n=100$ breast cancer diagnoses, let

$$X = \# \text{ BCIS cases } (0,1,2,\ldots,100).$$

Questions:

- How can we model the probability distribution of $X$, and under what assumptions?
- Probabilities of events, such as $P(X = 0)$, $P(X = 20)$, $P(X \leq 20)$, etc.?
- Mean # BCIS cases = ?
- Standard deviation of # BCIS cases = ?

Full article available online at this link.
**Binomial Distribution**  (Paradigm model = coin tosses)

<table>
<thead>
<tr>
<th>Binary random variable:</th>
<th>Probability:</th>
</tr>
</thead>
</table>
| \( Y = \begin{cases} 
1, \text{ Success (Heads)} & \text{with } P(\text{Success}) = \pi \\
0, \text{ Failure (Tails)} & \text{with } P(\text{Failure}) = 1 - \pi 
\end{cases} \) |

**Experiment:** \( n = 5 \) independent coin tosses

**Sample Space** \( S = \{(H \ H \ H \ H \ H), \ldots, (T \ T \ T \ T \ T)\} \)

\(#(S) = 2^5 = 32\)

**Random Variable:** \( X = \# \text{ Heads in } n = 5 \text{ independent tosses (0, 1, 2, 3, 4, 5)} \)

**Events:**
- \( X = 0 \) = Exercise
- \( #(X = 0) = \binom{5}{0} = 1 \)
- \( X = 1 \) = Exercise
- \( #(X = 1) = \binom{5}{1} = 5 \)
- \( X = 2 \) = Exercise
- \( #(X = 2) = \binom{5}{2} = 10 \)
- \( X = 3 \) = see above
- \( #(X = 3) = \binom{5}{3} = 10 \)
- \( X = 4 \) = Exercise
- \( #(X = 4) = \binom{5}{4} = 5 \)
- \( X = 5 \) = Exercise
- \( #(X = 5) = \binom{5}{5} = 1 \)

**Recall:** For \( x = 0, 1, 2, \ldots, n \), the combinatorial symbol \( \binom{n}{x} \) – read “\( n \)-choose-\( x \)” – is defined as the value \( \frac{n!}{x!(n-x)!} \), and counts the number of ways of rearranging \( x \) objects among \( n \) objects. See Appendix > Basic Reviews > Perms & Combos for details.

**Note:** \( \binom{n}{r} \) is computed via the mathematical function “nCr” on most calculators.
**Probabilities:**

First assume the coin is **fair** \((\pi = 0.5 \Rightarrow 1 - \pi = 0.5)\), i.e., **equally likely** elementary outcomes \(H\) and \(T\) on a single trial. In this case, the probability of any event \(A\) above can thus be easily calculated via \(P(A) = \#(A) / \#(S)\).

<table>
<thead>
<tr>
<th>(x)</th>
<th>(P(X = x) = \frac{1}{2^5} \binom{5}{x})</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>(1/32 = 0.03125)</td>
</tr>
<tr>
<td>1</td>
<td>(5/32 = 0.15625)</td>
</tr>
<tr>
<td>2</td>
<td>(10/32 = 0.31250)</td>
</tr>
<tr>
<td>3</td>
<td>(10/32 = 0.31250)</td>
</tr>
<tr>
<td>4</td>
<td>(5/32 = 0.15625)</td>
</tr>
<tr>
<td>5</td>
<td>(1/32 = 0.03125)</td>
</tr>
</tbody>
</table>

Now consider the case where the coin is **biased** (e.g., \(\pi = 0.7 \Rightarrow 1 - \pi = 0.3\)). Calculating \(P(X = x)\) for \(x = 0, 1, 2, 3, 4, 5\) means summing \(P\)(all its outcomes).

**Example:** \(P(X = 3) = \)

\[
\begin{align*}
P(H H H H T) &= (0.7)(0.7)(0.7)(0.3)(0.3) = (0.7)^3(0.3)^2 \\
+ P(H H H T H) &= (0.7)(0.7)(0.3)(0.7)(0.3) = (0.7)^3(0.3)^2 \\
+ P(H H T T H) &= (0.7)(0.7)(0.3)(0.3)(0.7) = (0.7)^3(0.3)^2 \\
+ P(H T H H T) &= (0.7)(0.3)(0.7)(0.7)(0.3) = (0.7)^3(0.3)^2 \\
+ P(H T H T H) &= (0.7)(0.3)(0.7)(0.3)(0.7) = (0.7)^3(0.3)^2 \\
+ P(H T T H H) &= (0.7)(0.3)(0.3)(0.7)(0.7) = (0.7)^3(0.3)^2 \\
+ P(T H H H T) &= (0.3)(0.7)(0.7)(0.7)(0.3) = (0.7)^3(0.3)^2 \\
+ P(T H H T H) &= (0.3)(0.7)(0.7)(0.3)(0.7) = (0.7)^3(0.3)^2 \\
+ P(T H T H H) &= (0.3)(0.7)(0.3)(0.7)(0.7) = (0.7)^3(0.3)^2 \\
+ P(T T H H H) &= (0.3)(0.3)(0.7)(0.7)(0.7) = (0.7)^3(0.3)^2 \\
\end{align*}
\]

\[
= \binom{5}{3} (0.7)^3 (0.3)^2
\]
Hence, we similarly have...

<table>
<thead>
<tr>
<th>x</th>
<th>$P(X=x)$ = $\binom{5}{x} (0.7)^x (0.3)^{5-x}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$\binom{5}{0}(0.7)^0 (0.3)^5$ = 0.00243</td>
</tr>
<tr>
<td>1</td>
<td>$\binom{5}{1}(0.7)^1 (0.3)^4$ = 0.02835</td>
</tr>
<tr>
<td>2</td>
<td>$\binom{5}{2}(0.7)^2 (0.3)^3$ = 0.13230</td>
</tr>
<tr>
<td>3</td>
<td>$\binom{5}{3}(0.7)^3 (0.3)^2$ = 0.30870</td>
</tr>
<tr>
<td>4</td>
<td>$\binom{5}{4}(0.7)^4 (0.3)^1$ = 0.36015</td>
</tr>
<tr>
<td>5</td>
<td>$\binom{5}{5}(0.7)^5 (0.3)^0$ = 0.16807</td>
</tr>
</tbody>
</table>

Example: Suppose that a certain medical procedure is known to have a 70% successful recovery rate (assuming independence). In a random sample of $n = 5$ patients, the probability that three or fewer patients will recover is:

**Method 1:**

\[
P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)
\]

\[
= 0.00243 + 0.02835 + 0.13230 + 0.30870 = 0.47178
\]

**Method 2:**

\[
P(X \leq 3) = 1 - [ P(X = 4) + P(X = 5) ]
\]

\[
= 1 - [0.36015 + 0.16807] = 1 - 0.52822 = 0.47178
\]

Example: The mean number of patients expected to recover is:

\[
\mu = E[X] = 0 (0.00243) + 1 (0.02835) + 2 (0.13230) + 3 (0.30870) + 4 (0.36015) + 5 (0.16807)
\]

\[
= 3.5 \text{ patients}
\]

This makes perfect sense for $n = 5$ patients with a $\pi = 0.7$ recovery probability, i.e., their product. In the probability histogram above, the “balance point” fulcrum indicates the mean value of 3.5.
General formulation:

The Binomial Distribution

Let the discrete random variable \( X = \text{"# Successes in } n \text{ independent Bernoulli trials (0, 1, 2, …, } n, \text{\"),\} each having constant probability } P(\text{Success}) = \pi, \text{ and hence } P(\text{Failure}) = 1 - \pi. \) Then the probability of obtaining any specified number of successes \( x = 0, 1, 2, \ldots, n, \) is given by:

\[
P(X = x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}.
\]

We say that \( X \) has a Binomial Distribution, denoted \( X \sim \text{Bin}(n, \pi). \) Furthermore, the mean \( \mu = n \pi, \) and the standard deviation \( \sigma = \sqrt{n \pi(1 - \pi)}. \)

Example: Suppose that a certain spontaneous medical condition affects 1% (i.e., \( \pi = 0.01 \)) of the population. Let \( X = \text{"number of affected individuals in a random sample of } n = 300. \) Then \( X \sim \text{Bin}(300, 0.01), \) i.e., the probability of obtaining any specified number \( x = 0, 1, 2, \ldots, 300 \) of affected individuals is:

\[
P(X = x) = \binom{300}{x} (0.01)^x (0.99)^{300-x}.
\]

The mean number of affected individuals is \( \mu = n \pi = (300)(0.01) = 3 \) expected cases, with a standard deviation of \( \sigma = \sqrt{(300)(0.01)(0.99)} = 1.723 \) cases.

Probability Table for Binomial Dist.

<table>
<thead>
<tr>
<th>( x )</th>
<th>( f(x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 0 )</td>
<td>( \binom{n}{0} \pi^0 (1 - \pi)^{n-0} )</td>
</tr>
<tr>
<td>( 1 )</td>
<td>( \binom{n}{1} \pi^1 (1 - \pi)^{n-1} )</td>
</tr>
<tr>
<td>( 2 )</td>
<td>( \binom{n}{2} \pi^2 (1 - \pi)^{n-2} )</td>
</tr>
<tr>
<td>( \text{etc.} )</td>
<td>( \text{etc.} )</td>
</tr>
<tr>
<td>( n )</td>
<td>( \binom{n}{n} \pi^n (1 - \pi)^{n-n} )</td>
</tr>
</tbody>
</table>

**Exercise:** In order to be a valid distribution, the sum of these probabilities must = 1. Prove it.

**Hint:** First recall the Binomial Theorem: How do you expand the algebraic expression \((a+b)^n\) for any \( n = 0, 1, 2, 3, \ldots? \) Then replace \( a \) with \( \pi, \) and \( b \) with \( 1 - \pi. \) Voilà!
Comments:

- The assumption of independence of the trials is absolutely critical! If not satisfied – i.e., if the “success” probability of one trial influences that of another – then the Binomial Distribution model can fail miserably. (Example: \( X = \) “number of children in a particular school infected with the flu”) The investigator must decide whether or not independence is appropriate, which is often problematic. If violated, then the correlation structure between the trials may have to be considered in the model.

- As in the preceding example, if the sample size \( n \) is very large, then the computation of \( \binom{n}{x} \) for \( x = 0, 1, 2, \ldots, n \), can be intensive and impractical. An approximation to the Binomial Distribution exists, when \( n \) is large and \( \pi \) is small, via the Poisson Distribution (coming up…).

- Note that the standard deviation \( \sigma = \sqrt{n \pi (1 - \pi)} \) depends on the value of \( \pi \). (Later…)
How can we estimate the parameter $\pi$ using a sample-based statistic $\hat{\pi}$?

**POPULATION**

Binary random variable

\[
Y = \begin{cases} 
1, & \text{Success with probability } \pi \\
0, & \text{Failure with probability } 1 - \pi 
\end{cases}
\]

**Experiment:** $n$ independent trials

**SAMPLE**

\[
y_1 + y_2 + y_3 + y_4 + y_5 + \ldots + y_n
\]

Let $X = \# \text{Successes in } n \text{ trials} \sim \text{Bin}(n, \pi)$

$(n - X = \# \text{Failures in } n \text{ trials})$.

Therefore, dividing by $n$...

\[
\frac{X}{n} = \text{proportion of Successes in } n \text{ trials}
\]

\[
\hat{\pi} = p \quad (= \bar{y}, \text{ as well})
\]

and hence...

\[
q = 1 - p = \text{proportion of Failures in } n \text{ trials}.
\]

**Example:** If, in a sample of $n = 50$ randomly selected individuals, $X = 36$ are female, then the statistic $\hat{\pi} = \frac{X}{n} = \frac{36}{50} = 0.72$ is an estimate of the true probability $\pi$ that a randomly selected individual from the population is female. The probability of selecting a male is therefore estimated by $1 - \hat{\pi} = 0.28$. 
**Poisson Distribution**  (Models rare events)

**Discrete Random Variable:**

\[ X = \# \text{ occurrences of a (rare) event } E, \text{ in a given interval of time or space, of size } T. \quad (0, 1, 2, 3, \ldots) \]

Assume:

1. All the occurrences of \( E \) are independent in the interval.

2. The mean number \( \mu \) of expected occurrences of \( E \) in the interval is proportional to \( T \), i.e., \( \mu = \alpha T \). This constant of proportionality \( \alpha \) is called the rate of the resulting Poisson process.

Then...

**The Poisson Distribution**

The probability of obtaining any specified number \( x = 0, 1, 2, \ldots \) of occurrences of event \( E \) is given by:

\[
P(X = x) = \frac{e^{-\mu} \mu^x}{x!}
\]

where \( e = 2.71828 \ldots \) (“Euler’s constant”).

We say that \( X \) has a **Poisson Distribution**, denoted \( X \sim \text{Poisson}(\mu) \). Furthermore, the mean is \( \mu = \alpha T \), and the variance is \( \sigma^2 = \alpha T \) also.

Examples:  # bee-sting fatalities per year, # spontaneous cancer remissions per year, # accidental needle-stick HIV cases per year, hemocytometer cell counts
Example (see above): Again suppose that a certain spontaneous medical condition $E$ affects 1% (i.e., $\alpha = 0.01$) of the population. Let $X =$ “number of affected individuals in a random sample of $T = 300$.” As before, the mean number of expected occurrences of $E$ in the sample is $\mu = \alpha T = (0.01)(300) = 3$ cases. Hence $X \sim \text{Poisson}(3)$, and the probability that any number $x = 0, 1, 2, \ldots$ of individuals are affected is given by:

$$P(X=x) = \frac{e^{-\mu} \mu^x}{x!}$$

which is a much easier formula to work with than the previous one. This fact is sometimes referred to as the Poisson approximation to the Binomial Distribution, when $T$ (respectively, $n$) is large, and $\alpha$ (respectively, $\pi$) is small. Note that in this example, the variance is also $\sigma^2 = 3$, so that the standard deviation is $\sigma = \sqrt{3} = 1.732$, very close to the exact Binomial value.

<table>
<thead>
<tr>
<th>$x$</th>
<th>Binomial $P(X=x) = \left(\frac{300}{x}\right)(0.01)^x(0.99)^{300-x}$</th>
<th>Poisson $P(X=x) = \frac{e^{-3}3^x}{x!}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.04904</td>
<td>0.04979</td>
</tr>
<tr>
<td>1</td>
<td>0.14861</td>
<td>0.14936</td>
</tr>
<tr>
<td>2</td>
<td>0.22441</td>
<td>0.22404</td>
</tr>
<tr>
<td>3</td>
<td>0.22517</td>
<td>0.22404</td>
</tr>
<tr>
<td>4</td>
<td>0.16888</td>
<td>0.16803</td>
</tr>
<tr>
<td>5</td>
<td>0.10099</td>
<td>0.10082</td>
</tr>
<tr>
<td>6</td>
<td>0.05015</td>
<td>0.05041</td>
</tr>
<tr>
<td>7</td>
<td>0.02128</td>
<td>0.02160</td>
</tr>
<tr>
<td>8</td>
<td>0.00787</td>
<td>0.00810</td>
</tr>
<tr>
<td>9</td>
<td>0.00258</td>
<td>0.00270</td>
</tr>
<tr>
<td>10</td>
<td>0.00076</td>
<td>0.00081</td>
</tr>
<tr>
<td>etc.</td>
<td>$\rightarrow 0$</td>
<td>$\rightarrow 0$</td>
</tr>
</tbody>
</table>
Why is the Poisson Distribution a good approximation to the Binomial Distribution, for large \(n\) and small \(\pi\)?

**Rule of Thumb:** \(n \geq 20\) and \(\pi \leq 0.05\); excellent if \(n \geq 100\) and \(\pi \leq 0.1\).

Let \(f_{\text{Bin}}(x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}\) and \(f_{\text{Poisson}}(x) = \frac{e^{-\lambda} \lambda^x}{x!}\), where \(\lambda = n\pi\).

We wish to show formally that, for fixed \(\lambda\), and \(x = 0, 1, 2, \ldots\), we have:

\[
\lim_{n \to \infty, \pi \to 0} f_{\text{Bin}}(x) = f_{\text{Poisson}}(x).
\]

**Proof:** By elementary algebra, it follows that…

\[
f_{\text{Bin}}(x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x} \]

\[
= \frac{n!}{x! (n-x)!} \pi^x (1 - \pi)^n (1 - \pi)^{-x} \]

\[
= \frac{1}{x!} n (n-1) (n-2) \ldots (n-x+1) \pi^x \left(1 - \frac{\lambda}{n}\right)^n (1 - \pi)^{-x} \]

\[
= \frac{1}{x!} \frac{n (n-1) (n-2) \ldots (n-x+1)}{n^x} \pi^x \left(1 - \frac{\lambda}{n}\right)^n (1 - \pi)^{-x} \]

\[
= \frac{1}{x!} \frac{n (n-1) (n-2) \ldots (n-x+1)}{n \frac{(n-1)}{n} \frac{(n-2)}{n} \ldots \frac{(n-x+1)}{n}} (n\pi)^x \left(1 - \frac{\lambda}{n}\right)^n (1 - \pi)^{-x} \]

\[
= \frac{1}{x!} \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \ldots \left(1 - \frac{x-1}{n}\right) \lambda^x \left(1 - \frac{\lambda}{n}\right)^n (1 - \pi)^{-x} \]

As \(n \to \infty\), \(\pi \to 0\),

\[
\frac{1}{x!} 1(1)(1) \ldots (1) = 1 \quad \lambda^x \quad e^{-\lambda} \quad 1^{-x} = 1 \]

\[
= \frac{e^{-\lambda} \lambda^x}{x!} = f_{\text{Poisson}}(x). \quad \text{QED}
\]
Classical Discrete Probability Distributions

**Binomial** (probability of finding \(x\) “successes” and \(n-x\) “failures” in \(n\) independent trials)

\[
X = \# \text{ successes} \text{ (each with probability } \pi) \text{ in } n \text{ independent Bernoulli trials}, \quad n = 1, 2, 3, \ldots
\]

\[
f(x) = P(X = x) = \binom{n}{x} \pi^x (1-\pi)^{n-x}, \quad x = 0, 1, 2, \ldots, n
\]

**Negative Binomial** (probability of needing \(x\) independent trials to find \(k\) successes)

\[
X = \# \text{ independent Bernoulli trials for } k \text{ successes} \text{ (each with probability } \pi), \quad k = 1, 2, 3, \ldots
\]

\[
f(x) = P(X = x) = \binom{x-1}{k-1} \pi^k (1-\pi)^{x-k}, \quad x = k, k+1, k+2, \ldots
\]

**Geometric:** \(X = \# \) independent Bernoulli trials for \(k = 1\) success

\[
f(x) = P(X = x) = \pi (1-\pi)^{x-1}, \quad x = 1, 2, 3, \ldots
\]

**Hypergeometric** (modification of Binomial to sampling without replacement from “small” finite populations, relative to \(n\.)

\[
X = \# \text{ successes in } n \text{ random trials taken from a population of size } N \text{ containing } d \text{ successes, } \quad n \geq \frac{N}{10}
\]

\[
f(x) = P(X = x) = \frac{\binom{d}{x} \binom{N-d}{n-x}}{\binom{N}{n}}, \quad x = 0, 1, 2, \ldots, d
\]

**Multinomial** (generalization of Binomial to \(k\) categories, rather than just two)

For \(i = 1, 2, 3, \ldots, k,\)

\[
X_i = \# \text{ outcomes in category } i \text{ (each with probability } \pi_i) \text{ in } n \text{ independent Bernoulli trials}, \quad n = 1, 2, 3, \ldots
\]

\[
\pi_1 + \pi_2 + \pi_3 + \ldots + \pi_k = 1
\]

\[
f(x_1, x_2, \ldots, x_k) = P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \frac{n!}{x_1! x_2! \ldots x_k!} \pi_1^{x_1} \pi_2^{x_2} \ldots \pi_k^{x_k},
\]

\[
x_i = 0, 1, 2, \ldots, n \quad \text{with} \quad x_1 + x_2 + \ldots + x_k = n
\]

**Poisson** (“limiting case” of Binomial, with \(n \to \infty\) and \(\pi \to 0\), such that \(n\pi = \lambda\), fixed)

\[
X = \# \text{ occurrences of a rare event (i.e., } \pi \approx 0) \text{ among many (i.e., } n \text{ large), with fixed mean } \lambda = n\pi
\]

\[
f(x) = P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, 2, \ldots
\]