Now we study a topic that is very important for statistical inference.

- A **random sample** is a sequence of random variables \( X_1, X_2, \ldots, X_n \) that are *independent* and *identically* distributed.
  - This property is often abbreviated as **i.i.d.**
  - The number \( n \) is called the **sample size**.
- A **statistic** is a function of the random variables in a random sample.
  - Each statistic is itself a random variable and therefore has its own probability distribution, describing how it would vary under repeated random sampling.
  - The probability distribution of a statistic is called a **sampling distribution**.

**Example 12-1: Grade Point Average**

Suppose that a student named Marcus has a .3 probability of getting an A, a .5 probability of getting a B, and a .2 probability of getting a C in a class. Suppose further that this probability distribution holds *independently* for each of two classes that he is taking this term. Let \( X_1 \) denote the number of grade points (A = 4 points, B = 3 points, C = 2 points) that he receives for course 1 and similarly define \( X_2 \) for course 2.

a) What type of random variable is \( X_i \): discrete or continuous?

b) Draw a sketch of the probability mass function (pmf) of \( X_i \).

You could calculate that \( E(X_i) = 3.1 \), \( \text{Var}(X_i) = 0.49 \), and \( \text{SD}(X_i) = 0.7 \).

c) Interpret the expected value in this context.

Consider these two statistics:

- \( \bar{X} = \) average (mean) grade points in the two courses
- \( M = \) maximum number of grade points in the two courses

The table below lists all nine possible pairs of grades that Marcus could earn in these courses.
d) Determine the probabilities of these pairs of grades, and record them in the table. Also report the values of these two statistics for each possible pair of grades:

<table>
<thead>
<tr>
<th>Course 1</th>
<th>Course 2</th>
<th>Probability</th>
<th>Sample mean</th>
<th>Sample max</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>C</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

e) Report the probability (sampling) distribution of the sample mean grade points (denoted by $\bar{X}$) by listing its possible values and the probability of each:

<table>
<thead>
<tr>
<th>$\bar{X}$</th>
<th>$p(\bar{X})$</th>
</tr>
</thead>
</table>

f) Sketch a graph of this pmf of $\bar{X}$. How does this pmf compare to the pmf of the grade points in a single course $X$?

You could calculate that $E(\bar{X}) = 3.1$, $\text{Var}(\bar{X}) = 0.245$, and $\text{SD}(\bar{X}) = 0.495$.

g) How does these compare to their counterparts for the grade points in a single course?

h) Report the probability (sampling) distribution of the sample maximum $M$ by listing its possible values and the probability of each. Also sketch this pmf.

<table>
<thead>
<tr>
<th>$m$</th>
<th>$p(m)$</th>
</tr>
</thead>
</table>

Winter 2017
Example 15-2: Grade Point Average (cont)
Now suppose that you want to investigate Marcus’s academic performance over a year in which he takes 10 courses.

a) If you were to list all possible outcomes (grade permutations) for those 10 courses, how many would there be?

It’s no longer feasible to enumerate all possible outcomes, but we can rely on simulation to approximate the sampling distributions of these statistics. The following code, in the programming language called R, performs such a simulation:

```r
# start with N = number of repetitions, n = number of courses
# also start with pa = Pr(A), pb = Pr(B), pc = Pr(C)
#
# grpts = rep(NA, times = n)
# GPA = rep(NA, times = N)
# GPmax = rep(NA, times = N)
for (i in 1:N) {
  rand = runif(n,0,1)
  for (j in 1:n) {
    if (rand[j] < pa) {grpts[j] = 4}
    if ((rand[j] >= pa) & (rand[j] < pa+pb)) {grpts[j] = 3}
    if (rand[j] >= pa+pb) {grpts[j] = 2}
  }
  GPA[i] = mean(grpts)
  GPmax[i] = max(grpts)
}
hist(GPmax); table(GPmax)
mean(GPmax); sd(GPmax)
hist(GPA); table(GPA)
mean(GPA); sd(GPA)
```

b) Explain the difference between the (i in 1:N) and the (j in 1:n) loops.

c) Run this code for 100,000 simulated years of 10 courses per year. Report the approximate sampling distribution, mean, and SD of the sample maximum.

d) Explain why it makes sense that the sample maximum displays so little variability in this situation.
e) Comment on its shape, mean, and SD of the (approximate) sampling distribution of the sample mean GPA? How do these compare to their counterparts with a sample size of $n = 2$?

f) Now think about increasing the sample size (number of courses) to 40, representing an entire college career. Before you run the simulation, predict what you will see with regard to the distribution of sample maximum and sample mean.

g) Run a simulation with 100,000 simulated college careers of 40 courses. Comment on what the simulation reveals about the sampling distributions of the sample maximum and sample mean.

The most important theoretical result in probability and statistics is the **Central Limit Theorem (CLT)**:

- Let $X_1, X_2, \ldots, X_n$ be i.i.d. with $\mu = \text{E}(X_i)$ and $\sigma^2 = \text{Var}(X_i)$. Also let $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ denote the sample mean. Then the sampling distribution of $\bar{X}$ has:
  - $E(\bar{X}) = \mu$
    - Be careful in reading this statement, which speaks of 3 different means:
      - The sample mean, $\bar{X}$
      - The population mean, $\mu$
      - The mean of the sample means, $E(\bar{X})$
  - $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$, so $\text{SD}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$
    - Averages vary less than individual values.
    - SD decreases proportionally to the square root of sample size.
  - An approximately normal distribution for large values of $n$
    - Regardless of the distribution of the $X_i$’s
    - Exactly normal for any $n$ if the $X_i$’s are normally distributed
    - Becomes closer and closer to normal as the sample size $n$ increases
    - Also closer to normal for $X_i$’s that are closer to normal
Example 12-3: Manufacturing Potato Chips
Suppose that the weights of bags of potato chips coming off an assembly line are normally distributed with mean μ = 12 ounces and standard deviation σ = 0.4 ounces.

a) Determine the probability that one randomly selected bag weighs less than 11.9 ounces. [Hint: Start with a sketch. Also calculate a z-score.]

b) If you take a random sample of 10 bags, would you expect the probability of their sample mean weight being less than 11.9 ounces to be greater or less than the probability found in (a)? Explain, without performing the calculation.

c) Calculate the probability asked about in the previous question. [Hint: Draw and label a sketch of the sampling distribution and shade the region whose area corresponds to this probability.] Does this probability indicate that a sample mean as small as 11.9 ounces would be surprising if the population mean were really 12 ounces?
d) Repeat this analysis, for a random sample of 50 randomly selected bags.

e) Determine the smallest sample size for which the probability of the sample mean being less than 11.9 ounces is less than .01.  [Hints: Find the first percentile of the standard normal distribution as the value \( z \) such that \( P(Z \leq z) < .01 \). Set this percentile equal to the \( z \)-score from standardizing 11.9 and solve for \( n \).]

f) If you were told that a consumer group had weighed randomly selected bags and found a sample mean weight of 11.9 ounces, would you strongly doubt the claim that the true mean weight of all of the potato chip bags is 12 ounces?  On what unspecified information does your answer depend?  Explain.

g) Which of your above answers to would be affected if the distribution of the weights of the bags was not normal but was rather skewed?

h) Find a value \( k \) such that the probability of the sample mean weight of 100 randomly selected bags being between \( 12-k \) and \( 12+k \) is roughly 0.95.  In other words, between what two \( \bar{x} \) values do the middle 95% of the \( \bar{x} \) values fall?