Definitions:

*element* – the object or event on which measurements are taken.

*variables* – the types of information being measured on the elements comprising the population

*population* – a collection of values for one or more variables that would be measured on elements; it is the collection about which we wish to make an inference (NOTE: this is not exactly the same as the definition given in the book)

*finite population* – composed of a finite number of elements; \( N = \text{number of elements in the population} \)

*infinite population* – composed of an infinite number of elements

*sampling unit* – collection of one or more elements. Sampling units are mutually exclusive (exceptions exist though) and the set of sampling units is exhaustive

*sampling frame* – a complete list of the sampling units comprising the population

*sample* – the measurements of the variables taken on a subset of the sampling units drawn from the sampling frame

*sampling with replacement* – each time a sampling unit is selected to be in the sample, the measurement is taken and the unit is then ‘returned’ to the population before the next sampling unit is selected. This allows units to appear in a sample more than once.
*sampling without replacement* – once a sampling unit has been selected for inclusion in a sample, it is removed from the sampling frame and therefore not returned to the population before the next sampling unit is selected. This ensures that a sampling unit can appear in a sample only once.

When a population is infinite in size, sampling may be without replacement but is usually treated as with replacement.

**Purpose of Sampling:** to obtain accurate and precise estimates of population parameters

*parameters* – information that summarizes the population of interest. Common ones are the mean, median, mode, variance, standard deviation, range, minimum, maximum, mean absolute deviation, median absolute deviation, proportion of a category, etc. Which parameters are of interest depends partly on the type of measurement being taken.

For example, the variables under study are categorical. The parameter of interest here likely is the proportions of one or more of the possible categories.

The variables of interest could be numeric in which case the parameters of interest are often the mean, median, variance, and/or similar.

Although not technically a parameter, another item of interest for a population could be the frequency distribution or density distribution of the measurements that comprise the population.

*Categorical variables:* For categorical variables, it is the frequency distribution, i.e. the proportions of the categories in the population. The mode (the most common category) is often the parameter of interest.
Instead of a graphic, we could write the frequency distribution as $P(\text{category}_i) = P(c_i)$, $i=1,\ldots,C$, where $C$ is the number of distinct categories and $\sum_{i=1}^{C} P(c_i) = 1$.

**Discrete Numerical Variable:** For discrete numerical variables (variables with a countable set of possible values), it is the “probability mass function”
Instead of a graphic, we could write the probability mass function as $P(y), y \in S$ where $S$ is the set of possible values for $y$ and $\sum_{y \in S} P(y) = 1$ (see the example under mathematical expectation).

For continuous numerical variables, it is the “probability density function”

![Probability Density Function](image)

We write the probability density function as an equation $f(y)$ where $y \in I$, $I$ is the interval or set of intervals in which $y$ has positive probability, and $\int_{y \in I} f(y) dy = 1$.

**Accurate and Precise Estimators:**

**Accurate** – the estimator of a population parameter is unbiased for the parameter being estimated

*Unbiasedness* – an estimator is unbiased when the average value of the estimator (from repeated sampling) equals the population parameter value

**Precise** – the estimator of a population parameter has small error
Some Sources of Error:

*Sampling Error* – the uncertainty in the estimator due to sampling, that is due to the variability in its value from sample to sample

*Measurement Error* – the variability in the estimator due to inaccurate measurement of the sampling unit(s)

*Errors of Nonobservation* – errors introduced due to lack of complete coverage of the population. Examples include missing or lost data (e.g. due to sensor malfunction in field work or non-response in a questionnaire survey) and lack of complete coverage (e.g. ignoring some portion of the population)

Mathematical Expectation I:

The following holds in these two cases:

1) Random sampling with replacement of a finite population
2) Random sampling of an infinite population

In both cases every element is equally likely to be sampled at any step of the sampling effort.

Suppose we have a population for which the variable of interest is discrete. Since, the Y-values are discrete, we can use the summarized information about the population values, the probability mass function \( P(y), y \in S \).

Mathematical expectation for a discrete value: let \( g(y) \) be a function of \( y \). Then the expected value of \( g(y) \) is

\[
E[g(y)] = \sum_{y \in S} g(y)P(y).
\]
Example: the probability mass function for the number of children in a family can be given in a table as:

<table>
<thead>
<tr>
<th>Y</th>
<th>P(y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.223</td>
</tr>
<tr>
<td>1</td>
<td>0.335</td>
</tr>
<tr>
<td>2</td>
<td>0.251</td>
</tr>
<tr>
<td>3</td>
<td>0.126</td>
</tr>
<tr>
<td>4</td>
<td>0.047</td>
</tr>
<tr>
<td>5</td>
<td>0.013</td>
</tr>
<tr>
<td>6</td>
<td>0.004</td>
</tr>
<tr>
<td>7</td>
<td>0.001</td>
</tr>
</tbody>
</table>

The mean of the population is calculated using $g(y) = y$:

\[
E[g(y)] = \sum_{y \in S} g(y)P(y) = \sum_{y \in S} yP(y) = 0 \times 0.223 + 1 \times 0.335 + 2 \times 0.251 + ... + 7 \times 0.001 = 1.5 = \mu
\]

The variance of the population is obtained by setting $g(y) = (y - \mu)^2$:

\[
E[g(y)] = \sum_{y \in S} g(y)P(y) = \sum_{y \in S} (y - \mu)^2 P(y) = (0 - 1.5)^2 \times 0.223 + (1 - 1.5)^2 \times 0.335 + ... + (7 - 1.5)^2 \times 0.001 = 1.5 = \sigma^2
\]
Mathematical expectation for a continuous variable: Similar equations exist for the mean and variance of a population composed of continuous values:

\[
E[y] = \mu = \int_{y \in I} y f(y) \, dy
\]

\[
E[(y - \mu)^2] = V(y) = \sigma^2 = \int_{y \in I} (y - \mu)^2 f(y) \, dy
\]

We can use the concept of mathematical expectation to derive properties of estimators, specifically to show whether the estimator is unbiased for the parameter it is being used to estimate and to determine the sampling error (variance) of the estimator.

For example, under simple random sampling with replacement with a sample size of \( n \), the usual estimator of the population mean \( \mu \) is the sample mean \( \bar{y} \). It is straightforward to show that

\[
E[\bar{y}] = \mu
\]

\[
E[(\bar{y} - \mu)^2] = V(\bar{y}) = \frac{\sigma^2}{n}
\]

These equations imply the following:

1) the sample mean \( \bar{y} \) calculated from simple random sampling (with replacement) is unbiased for the population mean \( \mu \), i.e. if we repeatedly took samples of size \( n \) and each time calculated the sample mean, then the average of those sample means equals the population mean

2) the unbiasedness of the sample mean does not depend on the sample size
3) the sample mean has sampling error that depends on the variability of the population of Y-values $\sigma^2$ and on the sample size $n$

4) the sampling variability of the sample mean does depend on the sample size, decreasing as the sample size increases

**Mathematical Expectation II:**

The following holds for random sampling without replacement from a finite population of size $N$.

The reason we used the equations we just saw is that random sampling of infinite populations or of finite populations sampled with replacement has the characteristics:

1) every sampling unit is equally likely to be selected to be in the sample (i.e. has the same selection probability)

2) the selections are independent of each other

Sometimes, selection probabilities differ among units either deliberately (by the sampling design) or due to sampling a finite population without replacement or both.

When the selection probabilities vary among the sampling units, estimators other than the usual ones you are familiar with ($\bar{Y}$, $s^2$, etc) are still unbiased for the quantities they are estimating but they will have large variances, i.e. are not precise. Better estimators exist which exploit the unequal selection probabilities to get estimators with lower variances than $\bar{Y}$, $s^2$, etc.

I’ll introduce these as we go through the semester and cover their expected values as we encounter them.