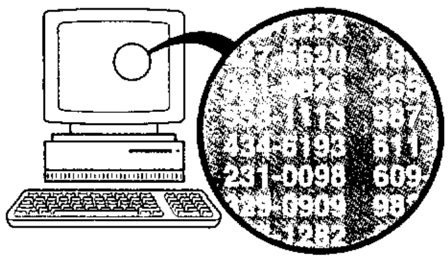
2 SIMPLE PROBABILITY SAMPLES

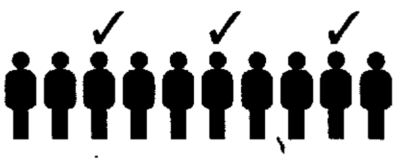
2.1 Types of Probability Samples

- A **simple random sample (SRS)** of size n is a sample of n units selected from a population in such a way that every sample of size n has an equal chance of being selected.
  - The Board of Regents at a university is deciding whether or not every one of its students must have two semesters of statistics coursework in order to fulfill their degree requirements. The student senate selected 200 students at random from the entire student body and asked their opinion on this issue. It was found that 1% of the students sampled supported a requirement of two semesters of statistics coursework.
  - A sample of twenty MSU instructors is selected based on the following scheme: An alphabetical list of all the instructors is prepared, then a unique number is assigned in sequence $(1, 2, 3, \ldots)$ to each instructor, and finally, using a random number generator, twenty instructors are randomly chosen.
- A **stratified random sample** is a sample selected by first dividing the population into non-overlapping groups called **strata** and then taking a simple random sample within each stratum. Dividing the population into strata should be based on some criterion so that units are *similar within* a stratum but are *different between* strata.
  - The Career Services staff wants to know if they are adequately meeting the needs of the students they serve. They design a survey which addresses this question, and they send the survey to 50 students randomly chosen from each of the university's colleges (50 agriculture students, 50 arts and science students, 50 engineering students, etc.).
  - A biologist is interested in estimating a deer population total in a small geographic region. The region contains two habitat types which are known to influence deer abundance. From each habitat type, 10 plots are randomly selected to be surveyed.
  - Note: Stratum sample sizes do not have to be equal.
- A **systematic sample** is a sample in which units are selected in a "systematic" pattern in the population of interest. To take a systematic sample you will need to divide the sampling frame into groups of units, randomly choose a set of starting points in the first group, and then sample from every group using the same positions of the starting points.
  - You are a quality engineer at Intel and are testing the quality of newly-produced computer chips. You need to take a sample of chips and test their quality. As the chips roll off the production line, you decided to test every 50^{th} chip produced starting with the third chip (i.e., sample chips 3, 53, 103, 153, ...).
- Suppose the observation units in a population are grouped into non-overlapping sets called *clusters*. A **cluster sample** is a SRS of clusters.
  - You work for the Department of Agriculture and wish to determine the percentage of farmers in the United States who use organic farming techniques. It would be difficult and costly to collect a SRS or a systematic sample because both of these sampling designs would require visiting many individual farms that are located far from each other. A convenience sample of farmers from a single county would be

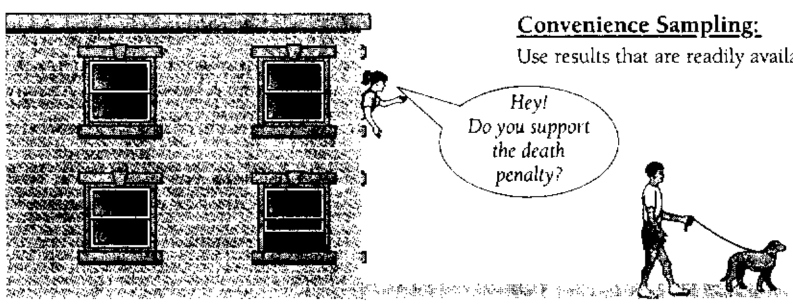
biased because farming practices vary from region to region. You decide to select several dozen counties from across the United States and then survey *every* farmer in each of these selected counties. Each county contains a cluster of farmers and data is collected on every farm within the randomly selected counties (clusters).



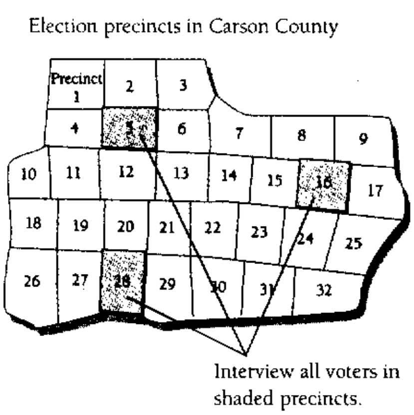
Simple Random Sampling:
Every sample of the same size has an equal chance of being selected. Computers are often used to generate random telephone numbers.



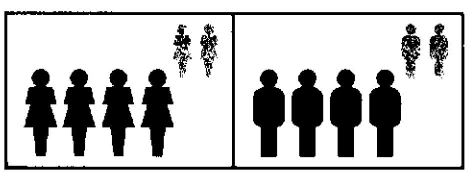
Systematic Sampling:
Select every *k*th member.



Convenience Sampling:
Use results that are readily available.



Cluster Sampling:
Divide the population area into sections, randomly select a few of those sections, and then choose all members in them.



Stratified Sampling:
Partition the population into at least two strata, then draw a sample from each.

- A **multistage sample** is a sample acquired by successively selecting smaller groups within the population in stages. The selection process at any stage may employ any sampling design (such as a SRS or a stratified sample).

  – A city official is investigating rumors about the landlord of a large apartment building complex. To get an idea of the tenants' opinions about their landlord, the official takes a SRS of buildings in the complex followed by a SRS of apartments from each selected building. From each chosen apartment a resident is interviewed.

– A U.S. national opinion poll was taken as follows: First, the U.S. is stratified into 4 regions. Then a random sample of counties was selected from each region followed by a random sample of townships within each of these counties. Finally, a random sample of households (clusters) within each township is taken.

2.2 Probability Sampling Designs

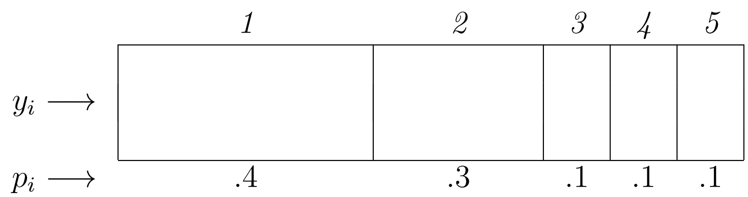
- Suppose N is the population size. That is, there are N units in the **universe** or **finite population** of interest.
- The N units in the universe are denoted by an **index set** of labels:

$$\mathcal{U} = \{ 1, 2, 3, \ldots N \}$$

  Note: Some texts will denote $\mathcal{U} = \{u_1, u_2, u_3, \ldots u_N\}$.

- From this universe (or population) a sample of n units is to be taken. Let $\mathcal{S}$ represent a sample of n units from $\mathcal{U}$.
- Associated with each of the N units is a measurable value related to the population characteristic of interest. Let y_i be the value associated with unit i, and the population of y-values is $\{y_1, y_2, \ldots, y_N\}$.
- A point of clarification is needed here. On page 28, Lohr states that $\mathcal{S}$ is "a subset consisting of n of the units in $\mathcal{U}$.
  - $\mathcal{S}$ will only be a 'subset' in a strict mathematical sense if sampling is done *without replacement*. That is, a unit can only appear in the sample at most once.
  - However, $\mathcal{S}$ will not be a 'subset' if sampling is done *with replacement*. That is, a unit can appear in the sample multiple times.
- Sampling designs that are based on planned randomness are called **probability samples**, and a probability $P(\mathcal{S})$ is assigned to every possible sample $\mathcal{S}$.
- The probability that unit i will be included in a sample is denoted π_i and is called the **inclusion probability** for unit i.
- In many sampling procedures, different units in the population have different probabilities of being included in a sample (i.e., different inclusion probabilities) that depend on either (i) the type of sampling procedure or (ii) the probabilities may be imposed by the researcher to obtain better estimates by including "more important" units with higher probability.
- **Example**: Suppose unit selections are drawn with probability proportional to unit size and that sampling is done with replacement of units.
  - The population total = 16. There are $N = 5$ sampling units. The figure shows the units, labeled 1 to 5, and the five y_i values.

- Sampling plan: You are to select two units. The first unit u_i is selected with probability p_i proportional to its size and the y_i is recorded. The unit is then put back. A second unit u_j is selected using the same method for selecting the first unit, and its y_j value is recorded.
- Note that the same unit can be sampled twice. This is an example of *sampling with replacement*. The following table describes the population.



- The table below shows all 15 possible pairs of sampled y-values (but not ordered). E.g., (1,2) means you selected unit 1 then unit 2 **or** you selected unit 2 then unit 1. In either case, you end up with the same sample of size 2. Note $\sum_{i=1}^{15} P(\mathcal{S}_i) = 1$.

Sample	Units	y-values	$P(\mathcal{S}_j)$	Calculation
$\mathcal{S}_1$	$\{1,2\}$	7 , 4	0.24	$(.4)(.3) + (.3)(.4)$
$\mathcal{S}_2$	$\{1,3\}$	7 , 0	0.08	$(.4)(.1) + (.1)(.4)$
$\mathcal{S}_3$	$\{1,4\}$	7 , 2	0.08	$(.4)(.1) + (.1)(.4)$
$\mathcal{S}_4$	$\{1,5\}$	7 , 3	0.08	$(.4)(.1) + (.1)(.4)$
$\mathcal{S}_5$	$\{2,3\}$	4 , 0	0.06	$(.3)(.1) + (.1)(.3)$
$\mathcal{S}_6$	$\{2,4\}$	4 , 2	0.06	$(.3)(.1) + (.1)(.3)$
$\mathcal{S}_7$	$\{2,5\}$	4 , 3	0.06	$(.3)(.1) + (.1)(.3)$
$\mathcal{S}_8$	$\{3,4\}$	0 , 2	0.02	$(.1)(.1) + (.1)(.1)$
$\mathcal{S}_9$	$\{3,5\}$	0 , 3	0.02	$(.1)(.1) + (.1)(.1)$
$\mathcal{S}_{10}$	$\{4,5\}$	2 , 3	0.02	$(.1)(.1) + (.1)(.1)$
$\mathcal{S}_{11}$	$\{1,1\}$	7 , 7	0.16	$(.4)(.4)$
$\mathcal{S}_{12}$	$\{2,2\}$	4 , 4	0.09	$(.3)(.3)$
$\mathcal{S}_{13}$	$\{3,3\}$	0 , 0	0.01	$(.1)(.1)$
$\mathcal{S}_{14}$	$\{4,4\}$	2 , 2	0.01	$(.1)(.1)$
$\mathcal{S}_{15}$	$\{5,5\}$	3 , 3	0.01	$(.1)(.1)$

The inclusion probability π_i is found by summing $P(\mathcal{S}_i)$ over all samples containing unit i. Thus, the inclusion probabilities when sampling with replacement are

π_1	$=$	$= .24 + .08 + .08 + .08 + .16$
π_2	$=$	$= .24 + .06 + .06 + .06 + .09$
π_3	$=$	$= .08 + .06 + .02 + .02 + .01$
π_4	$=$	$= .08 + .06 + .02 + .02 + .01$
π_5	$=$	$= .08 + .06 + .02 + .02 + .01$

- **Example** Suppose unit selections are drawn with probability proportional to unit size, and that sampling is done without replacement of units.

  - The population is the same as the previous example.

	1	*2*	*3*	*4*	*5*
$y_i \longrightarrow$	7	4	0	2	3
$p_i \longrightarrow$	.4	.3	.1	.1	.1

- Sampling plan: You are to select two units. The first unit u_i is selected with probability p_i proportional to its size and the y_i is recorded. The unit is not put back. A second unit u_j is selected using sampling proportional to size for the remaining 4 units, and its y_j value is recorded.
- Note that the same unit cannot be sampled twice. This is an example of *sampling without replacement*. The following table shows all 10 possible pairs of sampled y-values (but not ordered). Note: $\displaystyle\sum_{i=1}^{10} P(\mathcal{S}_i) = 315/315 = 1$.
- Once that once the first unit is selected, the probabilities for selecting the second unit are no longer p_i. The probabilities are proportional to *sizes of the remaining 4 units*.
- For example, to find the probability of selecting units 1 and 2, we need to calculate the probabilities of (i) selecting unit 1 first, then unit 2 and (ii) selecting unit 2 first, then unit 1.
  * The probability of selecting unit 1 first is .4. Now only units 2, 3, 4, and 5 remain with unit 2 accounting for 3/6 (=.3/(.3+.1+.1+.1) = .3/.6) of the remaining sizes. Thus, the probability of selecting unit 1 then unit 2 = (.4)(.3/.6).
  * The probability of selecting unit 2 first is .3. Now only units 1, 3, 4, and 5 remain with unit 1 accounting for 4/7 (=.4/(.4+.1+.1+.1) = .4/.7) of the remaining sizes. Thus, the probability of selecting unit 2 then unit 1 = (.3)(.4/.7).
  * Thus, the probability of sampling units 1 and 2 is the sum of these two probabilities: (.4)(.3/.6) + (.3)(.4/.7).

Sample	Units	y-values	$P(\mathcal{S}_j)$	Calculation
$\mathcal{S}_1$	$\{1,2\}$	7 , 4	13/35 = 117/315 ≈ .371	(.4)(.3/.6) + (.3)(.4/.7)
$\mathcal{S}_2$	$\{1,3\}$	7 , 0	1/9 = 35/315 ≈ .111	(.4)(.1/.6) + (.1)(.4/.9)
$\mathcal{S}_3$	$\{1,4\}$	7 , 2	1/9 = 35/315 ≈ .111	(.4)(.1/.6) + (.1)(.4/.9)
$\mathcal{S}_4$	$\{1,5\}$	7 , 3	1/9 = 35/315 ≈ .111	(.4)(.1/.6) + (.1)(.4/.9)
$\mathcal{S}_5$	$\{2,3\}$	4 , 0	8/105 = 24/315 ≈ .076	(.3)(.1/.7) + (.1)(.3/.9)
$\mathcal{S}_6$	$\{2,4\}$	4 , 2	8/105 = 24/315 ≈ .076	(.3)(.1/.7) + (.1)(.3/.9)
$\mathcal{S}_7$	$\{2,5\}$	4 , 3	8/105 = 24/315 ≈ .076	(.3)(.1/.7) + (.1)(.3/.9)
$\mathcal{S}_8$	$\{3,4\}$	0 , 2	1/45 = 7/315 ≈ .022	(.1)(.1/.9) + (.1)(.1/.9)
$\mathcal{S}_9$	$\{3,5\}$	0 , 3	1/45 = 7/315 ≈ .022	(.1)(.1/.9) + (.1)(.1/.9)
$\mathcal{S}_{10}$	$\{4,5\}$	2 , 3	1/45 = 7/315 ≈ .022	(.1)(.1/.9) + (.1)(.1/.9)

Thus, the inclusion probabilities when sampling without replacement are

$$
\begin{array}{lclcll}
\pi_1 & = & 74/105 & = & (117+35+35+35)/315 & \approx \\
\pi_2 & = & 63/105 & = & (117+24+24+24)/315 & \approx \\
\pi_3 & = & 73/315 & = & (35+24+7+7)/315 & \approx \\
\pi_4 & = & 73/315 & = & (35+24+7+7)/315 & \approx \\
\pi_5 & = & 73/315 & = & (35+24+7+7)/315 & \approx
\end{array}
$$

2.2.1 Parameters, Statistics, Expectations, and Estimation Bias

- One goal of sampling is to draw conclusions about a population of interest based on the data collected. This process of drawing conclusions is called **statistical inference**.
- A **parameter** is a value which describes some characteristic of a population (or possibly describes the entire population).
- A **statistic** is a value that can be computed from the observed (sample) data without making use of any unknown parameters.
- Unless the statistic and parameter are explicitly stated, we will use $\widehat{\theta}$ and θ to represent an unspecified statistic and parameter of interest, respectively.
- In general, the value of a population parameter is unknown. Statistics computed from sampling data can provide information about the unknown parameter.
- Common statistics of interest: Let $y_1, y_2, \ldots, y_n$ be a sample of y-values.
  - The **sample mean** is $\overline{y} \;=\; \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} y_i$.
  - The **sample variance** is $s^2 \;=\; \dfrac{1}{n-1}\left[(y_1-\overline{y})^2 + (y_2-\overline{y})^2 + \cdots + (y_n-\overline{y})^2\right]$

$$= \frac{1}{n-1}\sum_{i-1}^{n}(y_i-\overline{y})^2 \;=\; \frac{1}{n-1}\left[\sum y_i^2 - \frac{1}{n}\left(\sum y_i\right)^2\right]$$

  - The **sample standard deviation** s is $\sqrt{s^2}$.
- Common parameters of interest:
  - Notation: Let parameter t be the **population total** and parameter $\overline{y}_U$ be the **population mean** from a finite population of size N. Thus,

$$t \;=\; \sum_{i=1}^{N} y_i \qquad\qquad \overline{y}_U \;=\; \frac{1}{N}\sum_{i=1}^{N} y_i \;= \tag{1}$$

  - The **population variance** parameter S^2 is defined as:

$$\begin{aligned} S^2 \;&=\; \frac{1}{N-1}\sum_{i=1}^{N}(y_i-\overline{y}_U)^2 \qquad (2)\\ &=\; \left(\frac{1}{N-1}\right)\left(\sum_{i=1}^{N} y_i^2 - \frac{t^2}{N}\right) \;=\; \left(\frac{1}{N-1}\right)\left(\sum_{i=1}^{N} y_i^2 - N\overline{y}_U^2\right)\end{aligned}$$

  - The **population standard deviation** parameter S is defined as $S = \sqrt{S^2}$.
  - In other texts, τ, μ, σ^2 and σ are used to represent the population total t, mean $\overline{y}_U$, variance S^2, and standard deviation S.

- Because only a part of the population is sampled in any sampling plan, the value of a statistic $\widehat{\theta}$ will vary in repeated random sampling. The inherent variability of $\widehat{\theta}$ associated with sampling is called **sampling variability**.

- The **sampling distribution** of a statistic $\widehat{\theta}$ is the probability distribution of the values that can be observed for the statistic over *all* possible samples for a given sampling scheme.

- The **expected value** of $\widehat{\theta}$, denoted $E[\widehat{\theta}]$ is the mean of the sampling distribution of $\widehat{\theta}$:

$$E[\widehat{\theta}] = \sum_{\mathcal{S}} \widehat{\theta}_{\mathcal{S}} P(\mathcal{S}) \quad (3)$$
$$= \sum_{k} kP(\widehat{\theta} = k) \quad (4)$$

  - In (3): $\widehat{\theta}_{\mathcal{S}}$ is the value of $\widehat{\theta}$ calculated for sample $\mathcal{S}$ and the summation is taken over all possible samples ($\mathcal{S}$). Thus, $E[\widehat{\theta}]$ is the weighted average of $\widehat{\theta}$ calculated over all possible samples with weights $P(\mathcal{S})$.
  - In (4): The summation is taken over k = the set of possible values that can be observed for $\widehat{\theta}$. Thus, $E[\widehat{\theta}]$ is the weighted average of the possible values of $\widehat{\theta}$ with weights $P(\widehat{\theta} = k)$ = the probability of observing $\widehat{\theta} = k$.
  - These represent two approaches for calculating $E[\widehat{\theta}]$.

- The **(estimation) bias** of the estimator $\widehat{\theta}$ for estimating a parameter θ is the numerical difference between $E[\widehat{\theta}]$ and the parameter value θ. That is, $\text{Bias}[\widehat{\theta}] = E[\widehat{\theta}] - \theta$.

- An estimator $\widehat{\theta}$ is **unbiased** to estimate a parameter θ if $\text{Bias}[\widehat{\theta}] = 0$.

- **Estimation example**: Consider the small population of $N = 4$ values: $0, 3, 6, 12$. For this population, the population mean $\overline{y}_U = 21/4 =$ and the population variance

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{y}_U)^2 = \frac{1}{3}\sum_{i=1}^{5}(y_i - 5.25)^2$$
$$= \frac{1}{3}\left[(0-5.25)^2 + (3-5.25)^2 + (6-5.25)^2 + (12-5.25)^2\right]$$
$$= \frac{1}{3}\left[(-5.25)^2 + (-2.25)^2 + (0.75)^2 + (6.75)^2\right]$$
$$= \frac{1}{3}(27.5625 + 5.0625 + 0.5625 + 45.5625) = \frac{78.75}{3} = \quad .$$

  Thus, $S = \sqrt{\mathbf{26.25}} \approx$.

  Now, consider the following two sampling schemes, and assume the probabilities for selecting a sampling unit are all equal within each stage.

  Scheme I: Take a sample of size $n = 2$ with replacement. Because there are $4 \times 4 = 16$ ordered sampling sequences, each one has probability $= 1/16$. See Table 2.1A.

  Scheme II: Take a sample of size $n = 2$ without replacement. Because there are $4 \times 3 = 12$ ordered sampling sequences, each one has probability $= 1/12$. See Table 2.1B.

Table 2.1A: Sample means, variances, and standard deviations for all possible samples selected with replacement and $n = 2$

$\mathcal{S}$	y-values	$P(\mathcal{S})$	$\overline{y}_{\mathcal{S}}$	$s^2_{\mathcal{S}}$	$s_{\mathcal{S}}$	$\overline{y}_{\mathcal{S}}P(\mathcal{S})$	$s^2_{\mathcal{S}}P(\mathcal{S})$	$s_{\mathcal{S}}P(\mathcal{S})$
$\mathcal{S}_1$	0 , 3	$P(\mathcal{S}_1) = 2/16$	1.5	4.5	2.1213	3/16	9/16	0.2652
$\mathcal{S}_2$	0 , 6	$P(\mathcal{S}_2) = 2/16$	3.0	18.0	4.2426	6/16	36/16	0.5303
$\mathcal{S}_3$	0 , 12	$P(\mathcal{S}_3) = 2/16$	6.0	72.0	8.4853	12/16	144/16	1.0606
$\mathcal{S}_4$	3 , 6	$P(\mathcal{S}_4) = 2/16$	4.5	4.5	2.1213	9/16	9/16	0.2652
$\mathcal{S}_5$	3 , 12	$P(\mathcal{S}_5) = 2/16$	7.5	40.5	6.3640	15/16	81/16	0.7955
$\mathcal{S}_6$	6 , 12	$P(\mathcal{S}_6) = 2/16$	9.0	18.0	4.2426	18/16	36/16	0.5303
$\mathcal{S}_7$	0 , 0	$P(\mathcal{S}_7) = 1/16$	0.0	0.0	0	0/16	0/16	0
$\mathcal{S}_8$	3 , 3	$P(\mathcal{S}_8) = 1/16$	3.0	0.0	0	5/16	0/16	0
$\mathcal{S}_9$	6 , 6	$P(\mathcal{S}_9) = 1/16$	6.0	0.0	0	6/16	0/16	0
$\mathcal{S}_{10}$	12 , 12	$P(\mathcal{S}_{10}) = 1/16$	12.0	0.0	0	12/16	0/16	0
						84/16	315/16	≈ 3.4471

Table 2.1B: Sample means, variances, and standard deviations for all possible samples selected without replacement and $n = 2$

$\mathcal{S}$	y-values	$P(\mathcal{S})$	$\overline{y}_{\mathcal{S}}$	$s^2_{\mathcal{S}}$	$s_{\mathcal{S}}$	$\overline{y}_{\mathcal{S}}P(\mathcal{S})$	$s^2_{\mathcal{S}}P(\mathcal{S})$	$s_{\mathcal{S}}P(\mathcal{S})$
$\mathcal{S}_1$	0 , 3	$P(\mathcal{S}_1) = 2/12$	1.5	4.5	2.1213	3/12	9/12	0.3536
$\mathcal{S}_2$	0 , 6	$P(\mathcal{S}_2) = 2/12$	3.0	18.0	4.2426	6/12	36/12	0.7071
$\mathcal{S}_3$	0 , 12	$P(\mathcal{S}_3) = 2/12$	6.0	72.0	8.4853	12/12	144/12	1.4142
$\mathcal{S}_4$	3 , 6	$P(\mathcal{S}_4) = 2/12$	4.5	4.5	2.1213	9/12	9/12	0.3536
$\mathcal{S}_5$	3 , 12	$P(\mathcal{S}_5) = 2/12$	7.5	40.5	6.3640	15/12	81/12	1.0607
$\mathcal{S}_6$	6 , 12	$P(\mathcal{S}_6) = 2/12$	9.0	18.0	4.2426	18/12	36/12	0.7071
						63/12	315/12	≈ 4.9497

<u>Estimation using equation (3)</u>:

- Recall: the parameter values are $\overline{y}_U = 5.25$, $S^2 = 26.25$, and $S \approx 5.1235$.
- Using equation (3) and Table 2.1A, we get

$$\begin{aligned} E[\overline{y}] &= \sum_{\mathcal{S}} \overline{y}_{\mathcal{S}} P(\mathcal{S}) \;=\; 84/16 \;=\; 5.25 && \longrightarrow \text{Bias}[\overline{y}] = 5.25 - 5.25 = \\ E[s^2] &= \sum_{\mathcal{S}} s^2_{\mathcal{S}} P(\mathcal{S}) \;=\; 315/16 \;=\; 19.6875 && \longrightarrow \text{Bias}[s^2] = 19.6875 - 26.25 = \\ E[s] &= \sum_{\mathcal{S}} s_{\mathcal{S}} P(\mathcal{S}) \;\approx\; 3.4471 && \longrightarrow \text{Bias}[s] \approx 3.4471 - 5.1235 = \end{aligned} \tag{5}$$

  Therefore, $\overline{y}$ is an <u>unbiased</u> estimator for $\overline{y}_U$, but s^2 and s are <u>biased</u> estimators for S^2 for S for Scheme I.

- Using equation (3) and Table 2.1B, we get

$$\begin{aligned} E[\overline{y}] &= \sum_{\mathcal{S}} \overline{y}_{\mathcal{S}} P(\mathcal{S}) \;=\; 63/12 \;=\; 5.25 && \longrightarrow \text{Bias}[\overline{y}] = 5.25 - 5.25 = \\ E[s^2] &= \sum_{\mathcal{S}} S^2_{\mathcal{S}} P(\mathcal{S}) \;=\; 315/12 \;=\; 26.25 && \longrightarrow \text{Bias}[s^2] = 315/12 - 26.25 = \\ E[s] &= \sum_{\mathcal{S}} s_{\mathcal{S}} P(\mathcal{S}) \;\approx\; 4.9497 && \longrightarrow \text{Bias}[s] \approx 4.9497 - 5.1235 = \end{aligned} \tag{6}$$

  Therefore, $\overline{y}$ and s^2 are an <u>unbiased</u> estimators for $\overline{y}_U$ and S^2, respectively, but s is a <u>biased</u> estimator for S for Scheme II.

Estimation using equation (4):

- To find the sampling distributions of $\overline{y}$, s^2, and s, you determine all possible values that can be observed for each statistic and the associated probabilities for observing each of these values.
- Let k represent a possible value of a statistic.
- Tables 2.2A and 2.2B show the sampling distributions of $\overline{y}$, s^2, and s for Scheme I (sampling with replacement) and II (sampling without replacement), respectively. A third (product) column is included for each statistic to calculate expectations.
- Note that we will get the same results concerning expected values and biases as those summarized in (5) and (6).

Table 2.2A: Sampling distribution of $\overline{y}$, s^2, and s for Scheme I and expected value calculations for $E(\overline{y})$, $E(s^2)$, and $E(s)$

k	$P(\overline{y}=k)$	$kP(\overline{y}=k)$	k	$P(s^2=k)$	$kP(s^2=k)$	k	$P(s=k)$	$kP(s=k)$
0.0	1/16	0/16	0.0	4/16	0/16	$\sqrt{0.0}$	4/16	0
1.5	2/16	3/16	4.5	4/16	18/16	$\sqrt{4.5}$	4/16	≈ 0.5303
3.0	3/16	9/16	18.0	4/16	72/16	$\sqrt{18.0}$	4/16	≈ 1.0607
4.5	2/16	9/16	40.5	2/16	81/16	$\sqrt{40.5}$	2/16	≈ 0.7955
6.0	3/16	18/16	72.0	2/16	144/16	$\sqrt{72.0}$	2/16	≈ 1.0607
7.5	2/16	15/16						
9.0	2/16	18/16						
12.0	1/16	12/16						
		$E[\overline{y}] = 84/16$			$E[S^2] = 315/16$			$E[S] \approx$

Table 2.2B: Sampling distribution of $\overline{y}$, s^2, and s for Scheme II and expected value calculations for $E(\overline{y})$, $E(s^2)$, and $E(s)$

k	$P(\overline{y}=k)$	$kP(\overline{y}=k)$	k	$P(s^2=k)$	$kP(s^2=k)$	k	$P(s=k)$	$kP(s=k)$
1.5	2/12	3/12	4.5	4/12	18/12	$\sqrt{4.5}$	4/12	≈ 0.7071
3.0	2/12	6/12	18.0	4/12	72/12	$\sqrt{18.0}$	4/12	≈ 1.4142
4.5	2/12	9/12	40.5	2/12	81/12	$\sqrt{40.5}$	2/12	≈ 1.0607
6.0	2/12	12/12	72.0	2/12	144/12	$\sqrt{72.0}$	2/12	≈ 1.4142
7.5	2/12	15/12						
9.0	2/12	18/12						
		$E[\overline{y}] = 63/12$			$E[S^2] = 315/12$			$E[S] \approx$

- **Important**: You cannot make general statements about bias (such as "statistic $\widehat{\theta}$ is an unbiased (biased) estimator of parameter θ". You must also know the probability sampling plan to determine whether or not a statistic is biased or unbiased.
- For example, it will not always be true that $\overline{y}$ will be an unbiased estimator of $\overline{y}_U$. That is, there will be sampling plans for which $E(\overline{y}) \neq \overline{y}_U$.

2.2.2 The Variance and Mean Squared Error of a Statistic

- So far the focus has been on the expected value of a statistic to check for bias. Another natural concern is the variability (spread) of the statistic. It is certainly possible for an unbiased statistic to have large variability.
- We will consider two measures of variability: the variance and the mean squared error.
- The **variance** of the sampling distribution of $\widehat{\theta}$ (or simply, $V(\widehat{\theta})$) is defined to be

$$V(\widehat{\theta}) \;=\; E\left[(\widehat{\theta}_{\mathcal{S}} - E(\widehat{\theta}))^2\right] \;=\; \sum_{\mathcal{S}} P(\mathcal{S}) \left[(\widehat{\theta}_{\mathcal{S}} - E(\widehat{\theta})\right]^2 \tag{7}$$

  where $\widehat{\theta}_{\mathcal{S}}$ is the value of $\widehat{\theta}$ calculated for sample $\mathcal{S}$.

- The **mean squared error** is defined to be

$$\text{MSE}[\widehat{\theta}] \;=\; E\left[(\widehat{\theta} - \theta)^2\right] \;=\; \sum_{\mathcal{S}} P(\mathcal{S}) \left[(\widehat{\theta}_{\mathcal{S}} - \theta)\right]^2 \tag{8}$$

- The MSE, however, can be rewritten:

$$\begin{aligned}
\text{MSE}[\widehat{\theta}] \;&=\; E\left[(\widehat{\theta} \;-\; \theta)^2\right] \\
&=\; E\left[\Big(\,(\widehat{\theta} \;-\; E[\widehat{\theta}]) \;+\; (E[\widehat{\theta}] \;-\; \theta)\,\Big)^2\right] \\
&=\; E\left[(\widehat{\theta} \;-\; E[\widehat{\theta}])^2 \;+\; (E[\widehat{\theta}] \;-\; \theta)^2 \;+\; 2(\widehat{\theta} \;-\; E[\widehat{\theta}])(E[\widehat{\theta}] \;-\; \theta)\right] \\
&=\; E\left[(\widehat{\theta} \;-\; E[\widehat{\theta}])^2\right] \;+\; E\left[(E[\widehat{\theta}] \;-\; \theta)^2\right] \;+\; 2E\left[(\widehat{\theta} \;-\; E[\widehat{\theta}])\right](E[\widehat{\theta}] \;-\; \theta) \\
&=\; E\left[(\widehat{\theta} \;-\; E[\widehat{\theta}])^2\right] \;+\; E\left[(E[\widehat{\theta}] \;-\; \theta)^2\right] \;+\; 2\,(0)\,(E[\widehat{\theta}] \;-\; \theta) \\
&=
\end{aligned} \tag{9}$$

  We used $(A+B)^2 = A^2 + B^2 + 2AB$ with $A = \widehat{\theta} \;-\; E[\widehat{\theta}]$ and $B = E[\widehat{\theta}] \;-\; \theta$.

- The relationship between variance and mean squared error:
  - The variance is the average of the squared deviations of the statistic values about the mean (expected value) of the statistic.
  - The mean squared error is the average of the squared deviations of the statistic values about the parameter.
  - Thus, for an unbiased statistic, the variance and the mean squared error are identical (i.e. $V(\widehat{\theta}) = \text{MSE}(\widehat{\theta})$).