Chapter 2.

1.

a) $y_i =$ the number of dots in unit $i$.

\[
\begin{array}{cccccccc}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\
1 & 1 & 1 & 1 & 2 & 1 & 1 & 0 & 4 & 0 \\
\end{array}
\]

\[
\bar{y} = \frac{11}{10}, \quad \hat{\gamma} = 100 \times \frac{11}{10} = 110
\]

\[
\text{Var}(x) = N(N-n) \cdot \frac{S^2}{n} = 100 \times 90 \times \frac{1}{10} S^2 = 1290
\]

\[
S^2 = \frac{1}{N} \sum (y_i - \bar{y})^2 = 1.43333
\]

b) Using R, getting random sample, "sample \((1:100, 10)\)"

\[
\begin{array}{cccccccc}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 3 & 1 \\
\end{array}
\]

\[
\bar{y} = \frac{12}{10} = 1.2, \quad \hat{\gamma} = 100 \times \frac{12}{10} = 120
\]

\[
\text{Var}(x) = N(N-n) \cdot \frac{S^2}{n} = 100 \times 90 \times \frac{1}{10} S^2 = 960
\]

\[
S^2 = \frac{1}{N} \sum (y_i - \bar{y})^2 = 1.066667
\]

c) $P(\text{the unit in the upper left-hand corner } \in \text{ sample})$

\[
P = \frac{10}{100}
\]

The probability of selecting the sample $I$ obtain in part a)

\[
\frac{10}{100} \times \frac{9}{99} \times \frac{8}{98} \times \frac{7}{97} \times \frac{6}{96} \times \frac{5}{95} \times \frac{4}{94} \times \frac{3}{93} \times \frac{2}{92} \times \frac{1}{91} = 5.741 \times 10^{-11}
\]
#2

a) $\bar{y} = \frac{\sum_{i=1}^{10} y_i}{10} = \frac{31}{10}$
$\hat{z} = 100 \times \frac{31}{10} = 310.$

$\text{var}(\hat{z}) = N(N-n) \frac{S^2}{n} = 100 \times 90 \times \frac{1690}{10} = 1690.$

$b) \bar{y} = \frac{31}{10}$
$\text{var}(\bar{y}) = \frac{N-n}{N} \times \frac{S^2}{n} = \frac{90}{100} \times \frac{1690}{10} = 0.169$

#3

$N_i \ 1 \ 2 \ 3 \ 4 \ 5$

$y_i \ 3 \ 1 \ 0 \ 1 \ 5$

a) $\bar{y} = \frac{\sum_{i=1}^{5} y_i}{5} = \frac{10}{5} = 2$
$\hat{z} = 5 \times 2 = 10$

$S^2 = \frac{1}{4 \times 5} (y_i - \bar{y})^2 = 4.$

The possible samples are, $(N_i \ N_j \ N_k) \ 1 \leq i < j < k \leq 5,$

$(1, 2, 3) \ (1, 2, 4) \ (1, 2, 5) \ (1, 3, 4) \ (1, 3, 5) \ (1, 4, 5) \ (2, 3, 4) \ (2, 3, 5) \ (2, 4, 5) \ (3, 4, 5)$

$\therefore$ the probability that the one sample selected $P(5) = \frac{1}{10}.$
b) 
\[ E(\bar{y}) = \frac{1}{10} \times \left[ \frac{1}{3} (4 + 5 + 9 + 4 + 8 + 9 + 2 + 6 + 7 + 6) \right] \]
\[ = 2. \]

\[ E(\text{median}) = \frac{1}{10} \times (1 + 1 + 3 + 1 + 3 + 3 + 1 + 1 + 1 + 1) = 1.6. \]

The sample mean is unbiased for population mean (same as \( \mu \)), but the sample median = 1.6, the population median = 1. Thus, the sample median is biased.

#4
Suppose \( S = y_1, \ldots, y_n \) is a simple random sample from a population with \( \mu \) and finite variance \( \sigma^2 < \infty \).

\[ \frac{n}{n-1} \left( \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2 \right) = \frac{n}{n-1} \left[ \frac{1}{n} \sum_{i=1}^{n} (y_i - \mu)^2 \right] - \frac{n}{n-1} \left( \frac{1}{n} \sum_{i=1}^{n} (y_i - \mu)(\bar{y} - \mu) \right) + \frac{n}{n-1} \left( \frac{1}{n} (\bar{y} - \mu)^2 \right) \]

Since \( \frac{n}{n-1} \left( \frac{1}{n} \sum_{i=1}^{n} (y_i - \mu)(\bar{y} - \mu) \right) = \frac{n}{n-1} \left( \frac{1}{n} \sum_{i=1}^{n} y_i - \mu \right) \left( \frac{1}{n} \sum_{i=1}^{n} y_i - \mu \right) = \frac{n}{n-1} \left( \frac{1}{n} \sum_{i=1}^{n} y_i - \mu \right) \left( \frac{1}{n} \sum_{i=1}^{n} y_i - \mu \right) = \frac{n}{n-1} (\bar{y} - \mu)^2 \)

\[ \frac{n}{n-1} \left( \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2 \right) = \frac{n}{n-1} \left[ \frac{1}{n} \sum_{i=1}^{n} (y_i - \mu)^2 \right] - \frac{2n}{n-1} (\bar{y} - \mu)^2 + \frac{n}{n-1} (\bar{y} - \mu)^2 \]

\[ = \frac{n}{n-1} \left[ \frac{1}{n} \sum_{i=1}^{n} (y_i - \mu)^2 \right] - n(\bar{y} - \mu)^2. \]

\[ E(S^2) = E \left[ \frac{1}{n-1} \left( \frac{n}{n-1} \left( \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2 \right) \right) \right] = \frac{1}{n-1} \cdot E \left[ \frac{n}{n-1} \left( \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2 \right) \right] \]

\[ = \frac{1}{n-1} \cdot \left[ \frac{n}{n-1} E(\bar{y} - \mu)^2 - n \cdot E(\bar{y} - \mu)^2 \right]. \]
Note that: \( E(\bar{Y} - \mu)^2 = \text{var}(\bar{Y}) = \sigma^2 \).
\[ E(\bar{Y} - \mu)^2 = \text{var}(\bar{X}) = \frac{1}{n} \text{var}(X_i) = \frac{\sigma^2}{n}. \]
\[ E(S^2) = \frac{n \sigma^2 - n \cdot \frac{1}{n} \sigma^2}{n-1} = \sigma^2. \]

#5.

a) Sometimes, it is easy for us to become subjective in selecting the sample. For example, if in question 14 of chapter 2, some people argue that there are two units in the box however others may not agree. Thus, we should define clear definitions in order to avoid ambiguity when we choose samples.

b) Using random number generator program such as R and SAS, we can carry out the sample selection.

c) It is always possible that we make mistakes when we observe the units selected. For example, we may forget record our units or miss our units.

d) Some populations can not be measured. For example several populations have infinite units.

e) In real world, there are some populations with infinite variance.
h) Keep in mind that we should reduce our mistakes. Also, when we design the sample survey, we may consider stratified sampling or cluster sampling depending on our purpose. Finally, before sampling, we should think about the nature and characteristic of population.

Here is my data, using R: "Sample (1:50, 10)". I get random number from 1 to 50. n=10, N=50.

```
1  2  3  4  5  6  7  8  9  10
11  22  23  28  50  33  1  44  6  19
```

Out data, we will handle it chapter 3, question #4.

#6.

In R, using "trees" data.
a) Here is my R code:

trees

y <- trees$Volume

N <- 31

n <- 10

s <- sample(1:N, n)

s

[1] 19  8  21  3  31  11  15  1  2  2

y[s]

[1] 25.7 18.2 34.5 10.2 77.0 24.2 19.1 10.3 31.7 15.6

mean(y[s])

[1] 26.65

mu <- mean(y)

mu

[1] 30.17097

b <- 6

ybar <- numeric(6)

for(k in 1:b){
    s <- sample(1:N, n)
    ybar[k] <- mean(y[s])
}

ybar

b <- 10000

ybar <- numeric(6)

for(k in 1:b){
    s <- sample(1:N, n)
    ybar[k] <- mean(y[s])
}

}
ybar
hist(ybar)
mean(ybar)
[1] 30.15471
var(ybar)
[1] 18.52338
>(1-n/N)*var(y)/N
[1] 5.904536
sd(ybar)
[1] 4.30388
sqrt((1-n/N)*var(y)/n)

Histogram of ybar

[1] 4.278324

> mean((ybar-mu)^2)
My results are the same as those in the text.

b) n = 15
> trees
> y <- trees$Volume
> N <- 31
> n <- 15
> s <- sample(1:N, n)
> s

[1]  4  2  20  25  23  18  7  5  22  31  13  3 16  6  27
> y[s]

[1] 16.4 10.3 24.9 42.6 36.3 27.4 15.6 18.8 31.7 77.0 21.4 10.2 22.2 19.7 55.7
> mean(y[s])
[1] 28.68
> mu <- mean(y)
> mu
[1] 30.17097
> b <- 6
> ybar <- numeric(6)
> for(k in 1:b){
+ s <- sample(1:N, n)
+ ybar[k] <- mean(y[s])
+ }
> ybar

hist(ybar)
Compare the histogram, the shapes are very similar to each other however, we should focus on the mean, variance x-axis and y-axis.

It is clear that when \( n=15 \), there are more frequency around mean so, it has more higher peak and less tail in histogram.

Thus, the mean is similar to each other, \( (n=10 \rightarrow 30.15 \pm 0.15) \) / \( n=15 \rightarrow 30.166 \) 1

when \( n=15 \), we could get smaller variances such as \( \text{var}(\overline{y}) \), \( \text{var}(\overline{y}) \).
Chapter 3.

41. For data part a) of exercise 1 of chapter 2.
\[ \bar{y} = \frac{110}{10}, \quad s^2 = 1.43, \quad n = 10, \quad N = 100 \]

95% confidence interval for \( \mu \) ("z" based)
\[ \frac{110}{10} \pm 1.96 \frac{1.43}{\sqrt{10}} \left( \sqrt{1 - \frac{10}{100}} \right) = \frac{110}{10} \pm 1.96 \times \frac{1.1972}{\sqrt{100}} \left( \sqrt{0.9} \right) \]
\[ = (0.396, 1.804) \]

95% confidence interval for \( \bar{z} \) ("z" based)
\[ \hat{z} = 110, \quad \text{var}(\hat{z}) = 1290 \]
\[ 110 \pm 1.96 \sqrt{\text{var}(\hat{z})} = 110 \pm 1.96 \sqrt{1290} \]
\[ = (39.6035, 180.3965) \]

4b. data part b) of exercise 1 of chapter 2.
\[ \bar{y} = 1.2, \quad s^2 = 1.06, \quad n = 10, \quad N = 100 \]

95% confidence interval for \( \mu \) ("z" based)
\[ 1.2 \pm 1.96 \frac{1.06}{\sqrt{10}} \left( \sqrt{1 - \frac{10}{100}} \right) = 1.2 \pm 1.96 \times \frac{1.0328}{\sqrt{100}} \left( \sqrt{0.9} \right) \]
\[ = (0.5927, 1.8073) \]

95% confidence interval for \( \bar{z} \) ("z" based)
\[ \hat{z} = 120, \quad \text{var}(\hat{z}) = 960 \]
\[ 120 \pm 1.96 \sqrt{\text{var}(\hat{z})} = 120 \pm 1.96 \sqrt{960} \]
We assume that \( \hat{\mu} \) is a normally distributed, unbiased estimator for a population mean and total respectively. In this question, the sample sizes are not enough large so we should consider \( t \) distribution with degree of freedom \((n-1=9)\).

\#2.

a) 90\% confidence interval for \( \hat{\sigma} \).
\[ \hat{\sigma} = 310 \quad \text{Var}(\hat{\sigma}) = 1690 \]
\[ t_{0.05}(9) = 1.8333 \]
\[ 310 \pm 1.8333 \sqrt{1690} = 310 \pm 75.3676. \]
\[ = (234.6324, 385.3676) \]

b) 90\% confidence interval for \( \mu \).
\[ \overline{y} = 3.1 \quad S^2 = 1.87 \quad n = 10 \]
\[ 3.1 \pm 1.8333 \ \frac{S}{\sqrt{n}} \left( \sqrt{1 - \frac{10}{100}} \right) = 3.1 \pm 1.8333 \times \frac{1.3708}{\sqrt{10}} \times \frac{\sqrt{90}}{100} \]
\[ = 3.1 \pm 0.1537 = (2.9463, 3.8537) \]

\#3.

a) Sample \( \overline{y} \)

<table>
<thead>
<tr>
<th>Sample</th>
<th>( \overline{y} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1.2.3)</td>
<td>4/3</td>
</tr>
<tr>
<td>(1.2.4)</td>
<td>5/3</td>
</tr>
<tr>
<td>(1.2.5)</td>
<td>3</td>
</tr>
<tr>
<td>(1.3.4)</td>
<td>4/3</td>
</tr>
<tr>
<td>(1.3.5)</td>
<td>8/3</td>
</tr>
</tbody>
</table>

Sample \( \overline{y} \)

<table>
<thead>
<tr>
<th>Sample</th>
<th>( \overline{y} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1.4.5)</td>
<td>7/3</td>
</tr>
<tr>
<td>(2.3.4)</td>
<td>2/3</td>
</tr>
<tr>
<td>(2.3.5)</td>
<td>2</td>
</tr>
<tr>
<td>(2.4.5)</td>
<td>7/3</td>
</tr>
<tr>
<td>(3.4.5)</td>
<td>6/3</td>
</tr>
</tbody>
</table>

\[ E(\overline{y}) = 2. \]
\[
\text{Var}(\bar{y}) = \frac{1}{10} \left[ (4/3-2)^2 + (5/3-2)^2 + \ldots + (6/3-2)^2 \right] = 0.53
\]
\[
\text{Var}(\text{median}) = \frac{1}{10} \left[ (1-1.6)^2 + (1-1.6)^2 + \ldots + (1-1.6)^2 \right] = 0.84
\]

\( \text{Var}(y_i) \)

\begin{array}{cccccc}
\text{Sample i} & (1, 2, 3) & (1, 2, 4) & (1, 2, 5) & (1, 3, 4) & 4 \\
y_i & 413 & 513 & 3 & 413 & \\
\text{Var}(y_i) & 2.3333 & 1.3333 & 4 & 2.3333 & \\
\text{Var}(\hat{y}_i) & 0.3111 & 0.17777 & 0.5333 & 0.8111 & \\
\end{array}

\begin{array}{cccccc}
(1, 3, 5) & 5 & (1, 4, 5) & 6 & (2, 3, 4) & 7 & (2, 3, 5) & 8 & (2, 4, 5) & 9 \\
8/3 & 3 & 2/3 & 2 & 7/3 & \\
6.3333 & 4 & 0.3333 & 7 & 5.3333 & \\
0.844 & 0.53333 & 0.0444 & 0.93333 & 0.71111 & \\
(3, 4, 5) & 10 & \\
2 & 7 & \\
0.9333 & N_{1} & 1 & 2 & 3 & 4 & 5 & \\
\end{array}

Where:
\[
\bar{y} = \frac{1}{3} (y_1 + y_2 + y_3)
\]
\[
\text{Var}(\bar{y}) = \frac{n-1}{n} \cdot \text{var}(\bar{y}_i) = \frac{2}{5} \times \frac{1}{3} \cdot \text{Var}(y_i)
\]

(c) \( t_{0.025} (2) = 4.303 \)

Confidence Interval for sample i, i=1,2,3,...,10.

1. \( 413 \pm 4.303 \times \sqrt{0.3111} = (-1.0668, 3.9334) \)
2. \( 513 \pm 4.303 \times \sqrt{0.17777} = (-0.1476, 3.1476) \)
3. \( 3 \pm 4.303 \times \sqrt{0.5333} = (-0.1425, 6.1425) \)
4. $413 \pm 4.303 \sqrt{0.3111} = (-1.0668, 3.7334)$
5. $813 \pm 4.303 \sqrt{0.8444} = (-1.2875, 6.6209)$

6. $3 \pm 4.303 \sqrt{0.5333} = (-0.1425, 6.1425)$
7. $213 \pm 4.303 \sqrt{0.0444} = (-0.2404, 1.5739)$
8. $2 \pm 4.303 \sqrt{0.9333} = (-2.1571, 6.1571)$
9. $913 \pm 4.303 \sqrt{0.1111} = (-1.2953, 5.9619)$
10. $2 \pm 4.303 \sqrt{0.9333} = (-2.1571, 6.1571)$

We could find that the only one confidence interval does not contain true value which is 2.

Thus, the actual coverage probability is 0.9.

**#4.**


In R, sample (1:50, 10)

i: 1 2 3 4 5 6 7 8 9 10.

y: 22 23 3 28 50 33 1 44 6 19.

$\bar{y} = 22.9$

$\frac{\bar{y}}{2} = \frac{50 \times 22.9}{2} = 114.5$

$S^2 = E(x^2) - [E(x)]^2 = 1772.9 - (22.9)^2 = 248.49$

$\hat{\text{Var}}(\bar{z}) = \frac{\text{Var}(z)}{n} = \frac{50 \times 40 \times \frac{248.49}{10}}{10} = 496.98$

$\hat{\text{Var}}(\bar{y}) = \frac{N-n}{N} \cdot \frac{S^2}{n} = \frac{40}{50} \cdot \frac{248.49}{10} = 19.8792$. 
95\% confidence interval for μ, \( \pm_{0.025} (\bar{x}) = 2.262 \)

\[
22.9 \pm 2.262 \frac{S}{\sqrt{n}} \sqrt{\left(1-\frac{1}{n}\right)} = 22.9 \pm 2.262 \frac{15.7636}{\sqrt{10}} \times \frac{\sqrt{40}}{\sqrt{150}}
\]

= (12.8146, 32.9854)

95\% confidence interval for \( Z \)

\[
1145 \pm 2.262 \sqrt{\text{Var}(Z)} = 1145 \pm 2.262 \times 222.9305
\]

= (640.7312, 1649.2688)

99\% confidence interval for μ, \( \pm_{0.005} (\bar{x}) = 3.25 \)

\[
22.9 \pm 3.25 \frac{15.7636}{\sqrt{10}} \times \frac{\sqrt{40}}{\sqrt{150}}
\]

= (8.4095, 37.3905)

99\% confidence interval for \( Z \)

\[
1145 \pm 3.25 \sqrt{4969.8}
\]

= (1137.7548, 1869.5241)