Solution Tutorial 2

1. (a) To answer this question let's calculate the cumulative distribution function (column 3)

x	P(x)	F(x)
100	0.05	0.05
150	0.20	0.25
200	0.5	0.75
250	0.20	0.95
300	0.05	1.00

From the cumulative probability distribution we see that the probability that the company does not export more than 250 tonnes is 0.95.

(b) $E[X] = \mu = \sum_x xP(x) = 100 \times 0.05 + 150 \times 0.20 + 200 \times 0.5 + 250 \times 0.20 + 300 \times 0.05 = 200$

(c) $\sigma = \sqrt{\sum_x (x-\mu)^2 P(x)}$
$= \sqrt{(100-200)^2 \times 0.05 + (150-200)^2 \times 0.20 + ... + (300-200)^2 \times 0.05}$
$= \sqrt{2000} \approx 44.72$

(d) As we are now dealing with a normal distribution, we can transform our variable of interest Y into the standard normal distribution Z. This will allow us to use standard normal tables in order to find the probability that the company will not export more than 160 tonnes. We transform the value of 160 tonnes to Z units (Z scores)

$$\frac{X-\mu}{\sigma} = \frac{160-120}{30} \approx 1.33$$

$$F(1.33) = P(Z < 1.33) \approx 0.908$$

Note: We looked up $P(Z < -1.33)$ from a standard normal distribution statistical table.

(e) To answer this question we use the fact that the normal distribution is symmetric. We first calculate the probability that the value is not larger than 80 tonnes.

$$\frac{X-\mu}{\sigma} = \frac{80-120}{30} \approx -1.33$$

$$F(-1.33) = P(Z < -1.33) = 1 - P(Z < 1.33) \approx 1 - 0.908 = 0.092$$

We then subtract the probability that the amount of tonnes will not surpass 80 tonnes (i.e., that the amount of tonnes will be equal to or smaller than 80 tonnes) from the probability that the amount of tonnes will not surpass 160 tonnes (i.e., that the amount of tonnes will be equal to or smaller than 160 tonnes).

$$F(1.33) - F(-1.33) = 0.908 - 0.092 = 0.816$$

2. (a) "The sampling distribution of the sample mean" tells us how sample means are distributed across samples - it tells us how likely are different possible values of $\bar{x}$ given the population parameters and the sample size.

   (b) Looking at the lecture slides (unit 3) or textbook (page 258), we know that the sampling distribution of the sample mean is approximately normal when the sample size is large enough. Sample size $n = 47$ is sufficiently large to have an approximately normally distributed sampling distribution of the sample mean. Remember that as long as n is large enough (approximately $n \geq 25$), this is true even if the population distribution of x is not normal!

   We denote: $\overline{X} \sim \mathcal{N}(\mu_{\overline{X}}, \sigma^2_{\overline{X}})$.

   This result follows from the Central Limit Theorem.

   (c) Next we wish to find the values of $\mu_{\overline{X}}$, the mean of the sampling distribution, and $\sigma_{\overline{X}}$, the standard deviation of the sampling distribution (also called the standard error).

   We know (lecture slides, unit 3, or textbook, page 255) that the mean of the sampling distribution for the sample mean will be the same as the population mean, so

   $$\mu_{\overline{X}} = 3$$

   We know (lecture slides, unit 3, or textbook, page 256) that the standard deviation of the sampling distribution for the sample mean (also called standard error of the sample mean!) is given by

   $$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{1.6}{\sqrt{47}} \approx 0.23$$

   (d) We know that the sample mean follows a normal distribution, so we can use our knowledge about the normal distribution to characterize the probability of different values.

   We know that for a random variable which follows a normal distribution, more than 95% of the probability mass is located with

a distance of at most 2 standard deviations from the mean. For example, if a random variable r has a normal distribution with mean a and standard deviation b, we write: $r \sim N(a, b^2)$ and we know that $\mathbb{P}(a - 2b \leq r \leq a + 2b) > 95\%$.

See also: `https://en.wikipedia.org/wiki/68%E2%80%9395%E2%80%9399.7_rule`

In the case of the sample mean, we know that $\overline{X} \sim \mathcal{N}(\mu_{\overline{X}}, \sigma^2_{\overline{X}})$, and therefore it should be the case that

$\mathbb{P}(\mu_{\bar{X}} - 2\sigma_{\bar{X}} \leq \bar{x} \leq \mu_{\bar{X}} + 2\sigma_{\bar{X}}) > 95\%$, or

$\mathbb{P}(3 - 2 \cdot 0.23 \leq \bar{x} \leq 3 + 2 \cdot 0.23) > 95\%$, or

$\mathbb{P}(2.54 \leq \bar{x} \leq 3.46) > 95\%$.

In other words, values of the sample mean that are below 2.54 or above 3.46 happen relatively rarely, if population mean is 3, population standard deviation is 1.6, and sample size is 47. On the other hand, a sample mean like 3.5 is still not impossible!

3. a. A sampling distribution is a probability distribution of the possible values of a statistic for a given size sample selected from a population. Here, the statistic of interest is the sample mean.

Drawing 4 observations from a population of zeros and ones, we have 5 cases (5 different possible samples): we can draw either 0, 1, 2, 3, or 4 times 1. In each of these cases, it is then straightforward to calculate the sample average: For example, if "1" is drawn 3 times and "0" is drawn 1 times, the sample average is equal to 3/4.

Next, we compute the probabilities of each of these outcomes. Recalling the formula for the probability of x successes in n trials,

$$P(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}, \qquad x = 0, 1, 2, 3, 4.$$

Let $p = 9/11$ and $n = 4$, we can compute the probability of each number of successes.

Collecting together different possible values of the sample mean and the probability of each outcome, the sampling distribution of $\bar{X}$ is given by

x	P(x)	$\bar{x}$
0	0.001092822	0
1	0.01967079	1/4
2	0.1327778	1/2
3	0.3983334	3/4
4	0.4481251	1

b. If the data consists of zeros and ones, then there is a direct link between sample means and sample proportions: Sample mean is equal to the number of times "1" occurs, divided by the sample size ($\bar{X} = \hat{p} = X/n$). We can therefore make use of the normal approximation for the binomial distribution, after verifying that n is large enough ($n \cdot p \cdot (1-p) > 5$).

Hence, asymptotically,

$$\bar{X} \sim N\left(p, \frac{p(1-p)}{n}\right).$$

where $p = 9/11$ and $\frac{p(1-p)}{n} = \frac{(9/11)\cdot(2/11)}{50}$

4. (a)

$$\mu_X = \frac{2+4+6+6+7+8}{6} = 5.5$$

and

$$\sigma_X^2 = \frac{(2-\mu_X)^2 + \cdots (8-\mu_x)^2}{6} = 23.5/6 \approx 3.9167.$$

(b) The problem suggests that the sampling is without replacement (the work group should be formed by two persons). There are 15 different possible work groups, with ages

Table 1

	2	4	6	6	7	8
2	$*$	$(2,4)$	$(2,6)$	$(2,6)$	$(2,7)$	$(2,8)$
4	$*$	$*$	$(4,6)$	$(4,6)$	$(4,7)$	$(4,8)$
6	$*$	$*$	$*$	$(6,6)$	$(6,7)$	$(6,8)$
6	$*$	$*$	$*$	$*$	$(6,7)$	$(6,8)$
7	$*$	$*$	$*$	$*$	$*$	$(7,8)$
8	$*$	$*$	$*$	$*$	$*$	$*$

We consider only the element above the main diagonal, because the matrix is symmetric and on the main diagonal we would have work groups made of one person.

Next we compute the sample average for each sample in Table 1:

Table 2

	4	6	6	7	8
2	$(2+4)/2=3$	$(2+6)/2=4$	$(2+6)/2=4$	$(2+7)/2=4.5$	$(2+8)/2=5$
4	$*$	$(4+6)/2=5$	$(4+6)/2=5$	$(4+7)/2=5.5$	$(4+8)/2=6$
6	$*$	$*$	$(6+6)/2=6$	$(6+7)=6.5$	$(6+8)/2=7$
6	$*$	$*$	$*$	$(6+7)/2=6.5$	$(6+8)/2=7$
7	$*$	$*$	$*$	$*$	$(7+8)/2=7.5$
8	$*$	$*$	$*$	$*$	$*$

Next we consider the frequency of each value for the sample averages. The sampling distribution can be summarized as follow

$\bar{x}_2$	3	4	4.5	5	5.5	6	6.5	7	7.5
$p(\bar{x}_2)$	$1/15$	$2/15$	$1/15$	$3/15$	$1/15$	$2/15$	$2/15$	$2/15$	$1/15$

(c) Let us denote by j the different possible sample averages, j ranges from 1 to 9. The expectation is equal to the sum over each possible value times the probability of that value.

$\mathbb{E}(\bar{x}_{2,j}) = \sum_{j=1}^{9} \bar{x}_{2,j} p(\bar{x}_{2,j}) = 5.5.$

The expectation of the sampling distribution of the estimator corresponds to the population value, so the estimator is unbiased.

In other words, even when sampling is done without replacement, the sample mean is an unbiased estimator of the population mean.

(d) The exercise asks us to compute the variance of the *sampling distribution*, not the variance of X. To compute the variance of the sampling distribution, we calculate the squared deviations of each possible sample mean from the mean of the sampling distribution,

$\sum_{j=1}^{9} (\bar{x}_{2,j} - \mu_X)^2 p(\bar{x}_{2,j}) = \frac{23.5}{15} \approx 1.567.$

(e) The left-hand side is the standard deviation of the sampling distribution of the sample mean, $\sigma_{\bar{X}} = \sqrt{Var(\bar{X}_2)} \approx \sqrt{1.567} \approx 1.25$

On the right-hand side, we have $\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \approx \frac{3.9167}{\sqrt{2}}\sqrt{\frac{6-2}{6-1}} \approx 1.25$

Indeed, the claim in the exercise is true: when sampling is done without replacement and the sample size is large relative to the population, then

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

The reason is that when n is large relative to N, individual sample members are not distributed independently of one another. The term $\frac{N-n}{N-1}$ is often called a finite population correction factor.

Recall also that when we had sampling with replacement, or $n << N$, we had that

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

(f) Proceeding along the line of Table 2, and applying the formula we find for example that the sample variance for the sample $(2, 4)$ is

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{2-1}(2-3)^2 + (4-3)^2 = 2$$

Similarly, for the sample $(2, 8)$ we will get

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{2-1}(2-5)^2 + (8-5)^2 = 18$$

Repeating the same calculation for each sample we find that the distribution of s^2 is

s_2^2	0.0	0.5	2.0	4.5	8	12.5	18
$p(s_2^2)$	$1/15$	$3/15$	$5/15$	$1/15$	$3/15$	$1/15$	$1/15$

(g) $\mathbb{E}(s_2^2) = \sum_{j=1}^{7} s_{2,j}^2 p(s_{2,j}^2) = 4.7$. The expectation of the estimator for sample variance is different from the population variance (3.9167). This means that the estimator for sample variance is biased, when sampling is done without replacement.

(Remember that the sample variance is an unbiased estimator of the population variance in the case of a simple random sampling with replacement).

(h) Solving for σ^2 we find $\sigma^2 = 5\,\mathbb{E}(s^2)/6 \approx 3.9167$ which indeed corresponds to the population variance.