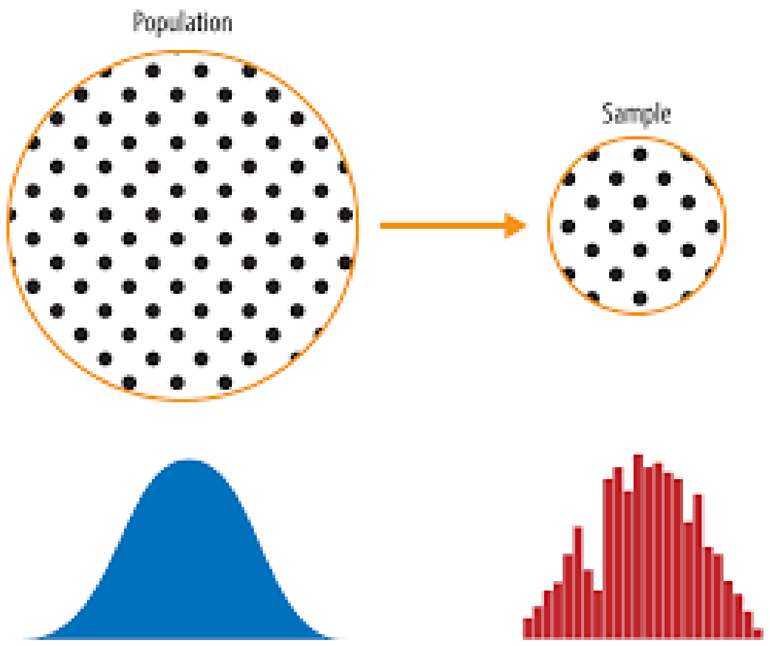
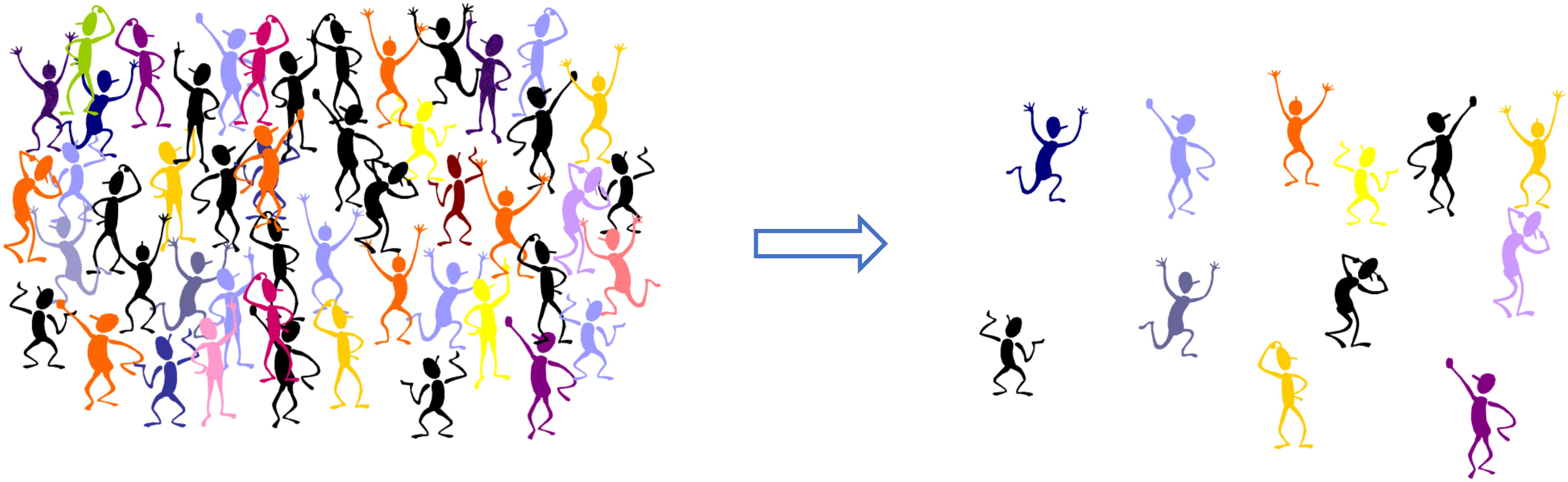
Sampling and sampling distributions



- **Sampling**
- Sampling distribution of a statistic
- Distribution of the sample mean and the central limit theorem

Sampling (1)

- **Population** – A group that *includes all* the cases (individuals, objects, or groups) in which the researcher is interested.
- **Sample** – A relatively *small subset* and *representative* of a population.



Sampling (2)

- Its interesting to know some numerical feature of the population, such as, the mean and standard deviation of the population, or some other numerical measure of center or variability.

> A numerical feature of a population is called a **parameter**.

- Whereas a parameter refers to some numerical characteristic of the population, a sample-based quantity is called a **statistic.**

> A **statistic** is a numerical valued function of the sample observations.

- *Note* that every statistic is a random variable. *Three* points are crucial:
  a) Because a sample is only a part of the population, the numerical value of a statistic *cannot* be expected to give us the *exact value* of the parameter.
  b) The *observed value* of a statistic depends on the particular sample that happens to be selected.
  c) There will be some *variability in the values* of a statistic *over different occasions* of sampling.

- **Sampling**
- **Sampling distribution of a statistic**
- Distribution of the sample mean and the central limit theorem

Sampling distribution of a statistic (1)

- The sample mean and any other statistic varies from sample to sample, it is a random variable and has its own probability distribution.

- The variability of the statistic, in repeated sampling, is described by this probability distribution.

> The probability distribution of a statistic is called its **sampling distribution**.

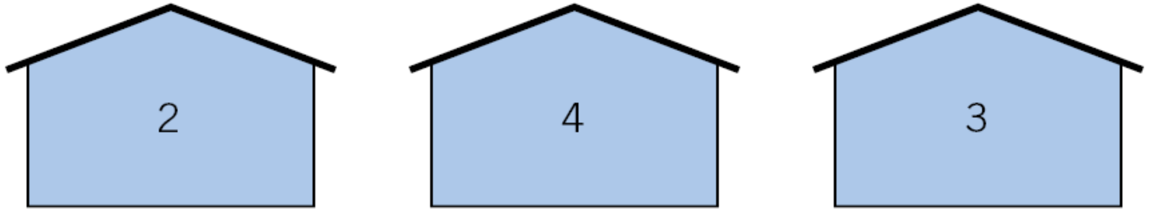
- The sampling distribution of a statistic is *determined from* the distribution $f(x)$ that governs the population, and it also depends on the *sample size* n.

Sampling distribution of a statistic (2)

Example:

- A population consists of three housing units, where the value of X, the number of rooms for rent in each unit, is shown in the illustration.

  Consider drawing a random sample of size 2 with replacement.



- Denote by $X_1$and X_2 the observation of X obtained in the first and second drawing, respectively. Find the sampling distribution of $\bar{X} = (X_1 + X_2)/2$.

Answer:

The population distribution, each of the X values 2, 3, and 4 occurs in $\frac{1}{3}$ of the population of the housing units. Because each unit is equally likely to be selected.

Sampling distribution of a statistic (3)

Answer:…cont'ed

The possible samples (x_1, x_2) of size 2 and the corresponding values of $\bar{X}$ are

The Population Distribution

x	$f(x)$
2	$\frac{1}{3}$
3	$\frac{1}{3}$
4	$\frac{1}{3}$

(x_1, x_2)	(2, 2)	(2, 3)	(2, 4)	(3, 2)	(3, 3)	(3, 4)	(4, 2)	(4, 3)	(4, 4)
$\bar{x} = \frac{x_1 + x_2}{2}$	2	2.5	3	2.5	3	3.5	3	3.5	4

The nine possible samples are equally likely so, for instance, P[$\bar{X}$=2.5]=$\frac{2}{9}$

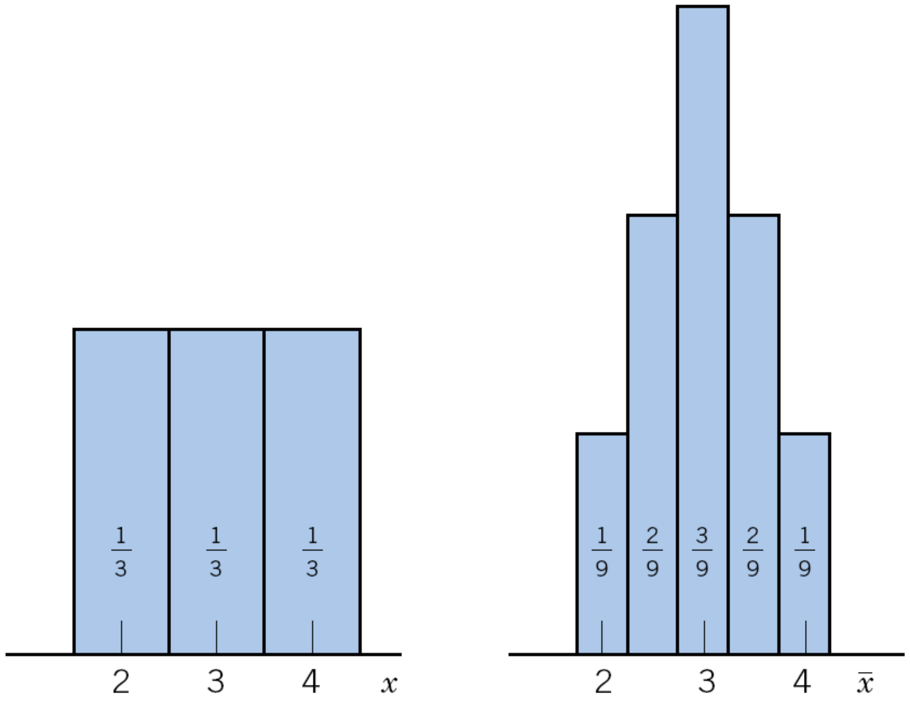
The Probability Distribution of $\bar{X} = (X_1 + X_2)/2$

Value of $\bar{X}$	Probability
2	$\frac{1}{9}$
2.5	$\frac{2}{9}$
3	$\frac{3}{9}$
3.5	$\frac{2}{9}$
4	$\frac{1}{9}$

Sampling distribution of a statistic (4)

***Answer*:**…cont'ed

The probability histograms of the distributions would be:-



(*a*) Population distribution.

(*b*) Sampling distribution of $\overline{X} = (X_1 + X_2)/2$.

Sampling distribution of a statistic (5)

- Before its value becomes available, each observation is modeled as random variable.

The observations $X_1, X_2, \ldots, X_n$ are a **random sample of size *n* from the population distribution** if they result from independent selections and each observation has the same distribution as the population.

- Because of variation in the population, the random sample will vary and so will $\bar{X}$, the sample median, or any other statistic.

Example:

Given a characteristic X, a large population is described by

the probability distribution

Population distribution

x	$f(x)$
0	.2
3	.3
12	.5

Sampling distribution of a statistic (6)

Example:…cont'ed

Let $X_1, X_2\ , X_3$ be a random sample of size 3 from this distribution.

a) List all the possible samples and determine their probabilities.

b) Determine the sampling distribution of the sample mean.

c) Determine the sampling distribution of the sample median.

Answers:

(a)

- Because we have a random sample, each of the three observations $X_1, X_2\ , X_3$ has the same distribution as the population and they are independent.
- The possible sample size become $3x3x3 = \mathbf{27}$

Sampling distribution of a statistic (7)

Answers:…cont'ed

Population distribution

x	$f(x)$
0	.2
3	.3
12	.5

Population mean:

$$E(X) = 0(0.2) + 3(0.3) + 12(0.5) = \mathbf{6.9}$$

Population variance:

$$Var(x) = 0^2(0.2) + 3^2(0.3) + 12^2(0.5) - 6.9^2 = \mathbf{27.09} = \sigma^2$$

	Possible Samples x_1	x_2	x_3	Sample Mean $\bar{x}$	Sample Median m	Probability
1	0	0	0	0	0	(.2)(.2)(.2) = .008
2	0	0	3	1	0	(.2)(.2)(.3) = .012
3	0	0	12	4	0	(.2)(.2)(.5) = .020
4	0	3	0	1	0	(.2)(.3)(.2) = .012
5	0	3	3	2	3	(.2)(.3)(.3) = .018
6	0	3	12	5	3	(.2)(.3)(.5) = .030
7	0	12	0	4	0	(.2)(.5)(.2) = .020
8	0	12	3	5	3	(.2)(.5)(.3) = .030
9	0	12	12	8	12	(.2)(.5)(.5) = .050
10	3	0	0	1	0	(.3)(.2)(.2) = .012
11	3	0	3	2	3	(.3)(.2)(.3) = .018
12	3	0	12	5	3	(.3)(.2)(.5) = .030
13	3	3	0	2	3	(.3)(.3)(.2) = .018
14	3	3	3	3	3	(.3)(.3)(.3) = .027
15	3	3	12	6	3	(.3)(.3)(.5) = .045
16	3	12	0	5	3	(.3)(.5)(.2) = .030
17	3	12	3	6	3	(.3)(.5)(.3) = .045
18	3	12	12	9	12	(.3)(.5)(.5) = .075
19	12	0	0	4	0	(.5)(.2)(.2) = .020
20	12	0	3	5	3	(.5)(.2)(.3) = .030
21	12	0	12	8	12	(.5)(.2)(.5) = .050
22	12	3	0	5	3	(.5)(.3)(.2) = .030
23	12	3	3	6	3	(.5)(.3)(.3) = .045
24	12	3	12	9	12	(.5)(.3)(.5) = .075
25	12	12	0	8	12	(.5)(.5)(.2) = .050
26	12	12	3	9	12	(.5)(.5)(.3) = .075
27	12	12	12	12	12	(.5)(.5)(.5) = .125
						Total = 1.000

Sampling distribution of a statistic (8)

Answers:…cont'ed

(b)

Sampling Distribution of $\overline{X}$

$\overline{x}$	$f(\overline{x})$
0	.008
1	.036 = .012 + .012 + .012
2	.054 = .018 + .018 + .018
3	.027
4	.060 = .020 + .020 + .020
5	.180 = .030 + .030 + .030 + .030 + .030 + .030
6	.135 = .045 + .045 + .045
8	.150 = .050 + .050 + .050
9	.225 = .075 + .075 + .075
12	.125 = .125

$$E(\overline{X}) = \sum \overline{x} f(\overline{x}) = 0(.008) + 1(.036) + 2(.054) + 3(.027) + 4(.060) + 5(.180) + 6(.135) + 8(.150) + 9(.225) + 12(.125)$$
$$= 6.9 \text{ same as } E(X), \text{ pop. mean}$$

$$\mathrm{Var}(\overline{X}) = \sum \overline{x}^2 f(\overline{x}) - \mu^2 = 0^2(.008) + 1^2(.036) + 2^2(.054) + 3^2(.027) + 4^2(.060) + 5^2(.180) + 6^2(.135) + 8^2(.150) + 9^2(.225) + 12^2(.125) - (6.9)^2$$
$$= 9.03 = \frac{27.09}{3} = \frac{\sigma^2}{3}$$

$\mathrm{Var}(\overline{X})$ is one-third of the population variance.

(c)

Sampling Distribution of the Median m

m	$f(m)$
0	.104 = .008 + .012 + .020 + .012 + .020 + .012 + .020
3	.396 = .018 + .030 + .030 + .018 + .030 + .018 + .027 + .045 + .030 + .045 + .030 + .030 + .045
12	.500 = .050 + .075 + .050 + .075 + .050 + .075 + .125

Mean of the distribution of sample median
$= 0(.104) + 3(.396) + 12(.500) = 7.188 \neq 6.9 = \mu$
Different from the mean of the population distribution

Variance of the distribution of sample median
$= 0^2(.104) + 3^2(.396) + 12^2(.500) - (7.188)^2 = 23.897$
[<u>not</u> one-third of the population variance 27.09]

- **Sampling**
- **Sampling distribution of a statistic**
- **Distribution of the sample mean and the central limit theorem**

Distribution of the sample mean (1)

- The sampling distribution of $\bar{X}$ also has a mean $E(\bar{X})$ and a standard deviation $sd(\bar{X})$. These can be expressed in terms of the population mean μ and standard deviation σ.

- The distribution of $\bar{X}$ is centered at the population mean μ in the sense that expectation serves as a measure of center of a distribution.

Mean and Standard Deviation of $\bar{X}$

The distribution of the sample mean, based on a random sample of size n, has

$$E(\bar{X}) = \mu \qquad (= \text{Population mean})$$

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n} \qquad \left(= \frac{\text{Population variance}}{\text{Sample size}} \right)$$

$$\mathrm{sd}(\bar{X}) = \frac{\sigma}{\sqrt{n}} \qquad \left(= \frac{\text{Population standard deviation}}{\sqrt{\text{Sample size}}} \right)$$

- The standard deviation of $\bar{X}$ equals the population standard deviation divided by the square root of the sample size.
  - The variability of the sample mean is governed by the **two** factors: the **population variability $\boldsymbol{\sigma}$** and the **sample size $\boldsymbol{n}$**.

Distribution of the sample mean (2)

Example:

- Calculate the mean and standard deviation for the population distribution and the distribution of $\bar{X}$ given below. Verify the relations $E(\bar{X}) = \mu$ and $sd(\bar{X}) = \sigma/\sqrt{n}$.

The Population Distribution

x	$f(x)$
2	$\frac{1}{3}$
3	$\frac{1}{3}$
4	$\frac{1}{3}$

The Probability Distribution of $\bar{X} = (X_1 + X_2)/2$

Value of $\bar{X}$	Probability
2	$\frac{1}{9}$
2.5	$\frac{2}{9}$
3	$\frac{3}{9}$
3.5	$\frac{2}{9}$
4	$\frac{1}{9}$

Distribution of the sample mean (3)

Answer: …cont'ed

Mean and Variance of $\overline{X} = (X_1 + X_2)/2$

Population Distribution				Distribution of $\overline{X} = (X_1 + X_2)/2$			
x	$f(x)$	$xf(x)$	$x^2f(x)$	$\bar{x}$	$f(\bar{x})$	$\bar{x}f(\bar{x})$	$\bar{x}^2f(\bar{x})$
2	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{4}{3}$	2	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{4}{9}$
3	$\frac{1}{3}$	$\frac{3}{3}$	$\frac{9}{3}$	2.5	$\frac{2}{9}$	$\frac{5}{9}$	$\frac{12.5}{9}$
4	$\frac{1}{3}$	$\frac{4}{3}$	$\frac{16}{3}$	3	$\frac{3}{9}$	$\frac{9}{9}$	$\frac{27}{9}$
Total	1	3	$\frac{29}{3}$	3.5	$\frac{2}{9}$	$\frac{7}{9}$	$\frac{24.5}{9}$
				4	$\frac{1}{9}$	$\frac{4}{9}$	$\frac{16}{9}$
				Total	1	3	$\frac{84}{9}$

$\mu = 3$

$$\sigma^2 = \frac{29}{3} - (3)^2 = \frac{2}{3}$$

$$E(\overline{X}) = 3 = \mu$$

$$\text{Var}(\overline{X}) = \frac{84}{9} - (3)^2 = \frac{1}{3}$$

By direct calculation, $sd(\bar{X}) = 1/\sqrt{3}$.

This is confirmed by the relation

$$sd(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{2}{3}}\Big/\sqrt{2} = \frac{1}{\sqrt{3}}$$

Distribution of the sample mean (4)

$\overline{X}$ Is Normal When Sampling from a Normal Population

In random sampling from a **normal** population with mean μ and standard deviation σ, the sample mean $\overline{X}$ has the normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$.

Example:

Suppose the weights of the contents of cans of mixed nuts have a normal distribution with mean 32.4 ounces and standard deviation 0.4 ounce.

a) If every can is labeled 32 ounces, what proportion of the cans have contents that weigh less than the labeled amount?
b) If two packages are randomly selected, specify the mean, standard deviation, and distribution of the average weight of the contents.
c) If two packages are randomly selected, what is the probability that the average weight is less than 32 ounces?

Distribution of the sample mean (5)

Answer:

Denote X = weight of a package. We are given that X is normal with mean 32.4 and standard deviation 0.4.

a) We convert to the standard normal to obtain

$$P[X < 32] = P\left[\frac{X-32.4}{0.4} < \frac{32-32.4}{0.4}\right] = P[Z < -1] = \mathbf{0.1587}$$

Hence, about 16% of the packages weigh less than the labeled amount.

b) Let X_1 and X_2 denote the weight of two randomly chosen packages. Observe that:

$$E(\bar{X}) = \mathbf{32.4} \qquad sd(\bar{X}) = \frac{0.4}{\sqrt{2}} = \mathbf{0.2828}$$

Hence, $\bar{X} = \frac{X_1+X_2}{2}$ is normal with mean 32.4 and standard deviation 0.2828.

c) Again, we convert to the standard normal (using part (b)) to obtain

$$P[\bar{X} < 32] = P\left[\frac{\bar{X}-32.4}{0.2828} < \frac{32-32.4}{0.2828}\right] = P[Z < -1.414] = \mathbf{0.0786}$$

Hence, there is about an 8% chance that the average weight of two packages will be less than the labeled amount of 32 ounces.

Central limit theorem (1)

- When the sample size n is large, the distribution $\bar{X}$ is approximately normal, regardless of the shape of the population distribution. In practice, the normal approximation is usually adequate when $\boldsymbol{n}$ **is greater than 30**.

Central Limit Theorem

Whatever the population, the distribution of $\overline{X}$ is approximately normal when n is large.

In random sampling from an arbitrary population with mean μ and standard deviation σ, when n is large, the distribution of $\overline{X}$ is approximately normal with mean μ and standard deviation $\sigma/\sqrt{n}$. Consequently,

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \quad \text{is approximately } N(0, 1)$$

Central limit theorem (2)

Example:

The result of a recent survey suggests that one plausible population distribution, for X = number of persons with whom an adult discusses important matters, can be modeled as a population having mean $\mu = 2$ and standard deviation $\sigma = 2$. A random sample of size 100 will be obtained.

a) What can you say about the probability distribution of the sample mean $\bar{X}$?

b) Find the probability that $\bar{X}$ exceeds 2.3.

Answer:

a) We have $E(\bar{X}) = \mu = 2.0$ and $sd(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{100}} = 0.2$. Since $n = 100$ is large, the central limit theorem ensures that the distribution of $\bar{X}$ is *approximately normal* with mean and standard deviation as calculated above.

b) The standardized variable is $Z = \frac{\bar{X}-2}{0.2}$. As such, we have

$$P[\bar{X} > 2.3] = P[Z > \frac{2.3-2}{0.2}] = P[Z > 1.5] = \mathbf{0.0668}$$