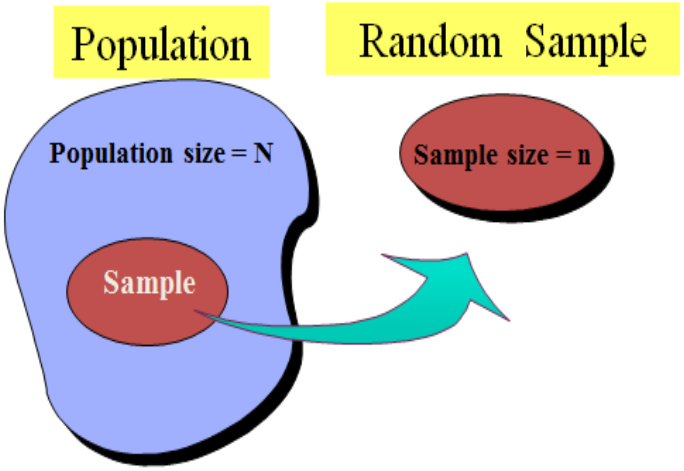
Summary---Topic 4: Sampling Distributions

Terminology



- Sample: a subset of population
- Statistics: characteristics / measures describe random sample
- A statistic is a random variable (A parameter is a constant, not a random variable)
- Sampling distribution: probability distribution of a statistic

Terminology

Population Distribution of Random Variable, X	Sampling Distribution of Sample Means
Population mean, $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$, or= Expected value, $\mathrm{E}[X] = \sum_i x_i \mathrm{P}(x_i)$ (**Sample mean**, $\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$, is an estimator of μ)	Mean of sample means, $\mu_{\bar{X}} = \sum_i \bar{x}_i \mathrm{P}(\bar{x}_i) = \mu$
Population standard deviation,σ: $\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$ or= $\sqrt{\sum_i (x_i - \mathrm{E}[X])^2 \mathrm{P}(x_i)}$ (**Sample standard deviation**, $S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$, is an estimator of σ)	Standard deviation of sample means, $\sigma_{\bar{X}} = \sqrt{\sum_i (\bar{x}_i - \mu_{\bar{X}})^2 \mathrm{P}(\bar{x}_i)} = \sigma/\sqrt{n}$

Central Limit Theorem(CLT)

- When sample size is large enough, the **sample means distribution** is approximately normal regardless of the population distribution.

$$\bar{X} \sim N(\mu_{\bar{X}}, \sigma_{\bar{X}}{}^2) \sim N(\mu, (\sigma/\sqrt{n})^2)$$

- Standardization of sample means distribution:

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

- Rule of thumb: when sample size is larger than or equal to 30 (i.e. n ≥ 30), we can apply the Central Limit Theorem.

Case	Population Distribution	Sample Size	Sampling Distribution of Means	Assumption Requirement
1	Normal	Whatever	Normal	Nil
2	Unknown / not normal	n ≥ 30	By the CLT, the sample means distribution is approximately normal.	Nil
3	Unknown	n < 30	Not normal	Assume population distribution is normal so that the sample means also follow normal distribution.
4	Known, and not normal	n < 30	Not normal	Cannot make assumption as population distribution is known.

Exercises and Solutions

Q1. Given a normal distribution with μ = 100 and σ = 12, if you select a sample of n = 36, what is the probability that $\bar{X}$ is

a) Less than 95?

b) Between 95 and 97.5?

c) Above 102.2?

d) There is a 65% chance that $\bar{X}$ is above what value?

*Mean of sample means, $\mu_{\bar{X}} = \sum_i \bar{x}_i \mathrm{P}(\bar{x}_i) = \mu$

*Standard deviation of sample means, $\sigma_{\bar{X}} = \sqrt{\sum_i (\bar{x}_i - \mu_{\bar{X}})^2 \mathrm{P}(\bar{x}_i)} = \sigma/\sqrt{n}$

Solution:

According to the question, we know that $X \sim N(100, 12^2)$.

Since $\mu_{\bar{X}} = 100$, *and* $\sigma_{\bar{X}} = \frac{12}{\sqrt{36}} = 2$,

and X is normally distributed, we conclude that the sampling distribution of the sample mean $\bar{X} \sim N(100, 2^2)$.

To compute non-standard normal probabilities:1. Do standardization; 2. Check the Standard Normal Table.

a) $\mathrm{P}(\bar{X} < 95) = \mathrm{P}\left(\frac{\bar{X}-100}{2} < \frac{95-100}{2}\right) = \mathrm{P}(Z < -2.5) = 0.0062$

Exercises and Solutions

Q1. Given a normal distribution with μ = 100 and σ = 12, if you select a sample of n = 36, what is the probability that $\bar{X}$ is

a) Less than 95?

b) Between 95 and 97.5?

c) Above 102.2?

d) There is a 65% chance that $\bar{X}$ is above what value?

Solution: $\bar{X} \sim N(100, 2^2)$.

b) $\mathrm{P}(95 < \bar{X} < 97.5) = \mathrm{P}\left(\frac{95-100}{2} < \frac{\bar{X}-100}{2} < \frac{97.5-100}{2}\right)$

$= \mathrm{P}(-2.5 < Z < -1.25) = 0.1056 - 0.0062 = 0.0994$

c) $\mathrm{P}(\bar{X} > 102.2) = \mathrm{P}\left(\frac{\bar{X}-100}{2} > \frac{102.2-100}{2}\right) = \mathrm{P}(Z > 1.1)$

$= 1 - \mathrm{P}(Z < 1.1) = 1 - 0.8643 = 0.1357$

d) We need find the value of "a" such that $\mathrm{P}(\bar{X} > a) = 0.65$ → $\mathrm{P}(\bar{X} < a) = 0.35$

→ $\mathrm{P}\left(\frac{\bar{X}-100}{2} < \frac{a-100}{2}\right) = 0.35$ → $\mathrm{P}\left(Z < \frac{a-100}{2}\right) = 0.35$ → $\frac{a-100}{2} = -0.39$ → $a = 99.22$

Q2. The diameter of a brand of Ping-Pong balls is normally distributed, with a mean of 1.30 inches and a standard deviation of 0.05 inch. If you select a random sample of 25 Ping-Pong balls,

a) What is the sampling distribution of the mean?

b) What is the probability that the sample mean is less than 1.28 inches?

c) What is the probability that the sample mean is between 1.31 and 1.33 inches?

d) The probability is 60% that the sample mean will be between what two values, symmetrically distributed around the population mean?

*Mean of sample means, $\mu_{\bar{X}} = \sum_i \bar{x}_i \mathrm{P}(\bar{x}_i) = \mu$

*Standard deviation of sample means, $\sigma_{\bar{X}} = \sqrt{\sum_i (\bar{x}_i - \mu_{\bar{X}})^2 \mathrm{P}(\bar{x}_i)} = \sigma/\sqrt{n}$

Solution:

a) Let X be the diameter of a brand of Ping-Pong balls, $X \sim N(1.3, 0.05^2)$.

Since $\mu_{\bar{X}} = 1.3$, *and* $\sigma_{\bar{X}} = \frac{0.05}{\sqrt{25}} = 0.01$, and X is normally distributed,

we conclude that the sample means is normally distributed, $\bar{X} \sim N(1.3, 0.01^2)$.

b) $\mathrm{P}(\bar{X} < 1.28) = \mathrm{P}\left(\frac{\bar{X}-1.3}{0.01} < \frac{1.28-1.3}{0.01}\right) = \mathrm{P}(Z < -2) = 0.0228$

Q2. The diameter of a brand of Ping-Pong balls is normally distributed, with a mean of 1.30 inches and a standard deviation of 0.05 inch. If you select a random sample of 25 Ping-Pong balls,

c) What is the probability that the sample mean is between 1.31 and 1.33 inches?

d) The probability is 60% that the sample mean will be between what two values, symmetrically distributed around the population mean?

Solution:

c) $\mathrm{P}(1.31 < \bar{X} < 1.33) = \mathrm{P}\left(\frac{1.31-1.3}{0.01} < \frac{\bar{X}-1.3}{0.01} < \frac{1.33-1.3}{0.01}\right) = \mathrm{P}(1 < Z < 3)$

$= \mathrm{P}(Z < 3) - \mathrm{P}(Z < 1) = 0.99865 - 0.8413 = 0.15735$

d) 60% of the values are between two values → $\mathrm{P}(\mathrm{a} < \bar{X} < \mathrm{b}) = 0.6$

symmetrically distributed around the mean→ $\begin{cases} \mathrm{P}(\bar{X} < a) = 0.2 \\ \mathrm{P}(\bar{X} < b) = 0.8 \end{cases}$

Do the standardization, we have

$$\begin{cases} \mathrm{P}\left(\frac{\bar{X}-1.3}{0.01} < \frac{a-1.3}{0.01}\right) = 0.2 \\ \mathrm{P}\left(\frac{\bar{X}-100}{0.01} < \frac{b-1.3}{0.01}\right) = 0.8 \end{cases} \Rightarrow \begin{cases} \mathrm{P}\left(Z < \frac{a-1.3}{0.01}\right) = 0.2 \\ \mathrm{P}\left(Z < \frac{b-1.3}{0.01}\right) = 0.8 \end{cases} \Rightarrow \begin{cases} \frac{a-1.3}{0.01} = -0.84 \\ \frac{b-1.3}{0.01} = 0.84 \end{cases}$$

$\Rightarrow$ a =1.2916, b = 1.3084

Q3. Time spent using e-mail per session is normally distributed with μ= 8 minutes and σ= 2 minutes. If you select a random sample of 16 sessions,

a) **What is the probability that the sample mean is between 7.8 and 8.2 minutes?**

b) What is the probability that the sample mean is between 7.5 and 8 minutes?

c) If you select a random sample of 100 sessions, what is the probability that the sample means is between 7.8 and 8.2 minutes?

d) Explain the difference in the results of (a) and (c).

*Mean of sample means, $\mu_{\bar{X}} = \sum_i \bar{x}_i \mathrm{P}(\bar{x}_i) = \mu$

*Standard deviation of sample means, $\sigma_{\bar{X}} = \sqrt{\sum_i (\bar{x}_i - \mu_{\bar{X}})^2 \mathrm{P}(\bar{x}_i)} = \sigma/\sqrt{n}$

Solution:

Let X be the time spent using e-mail per session , $X \sim N(8, 2^2)$.

When $\mathrm{n} = 16$, since $\mu_{\bar{X}} = 8$, *and* $\sigma_{\bar{X}} = \frac{2}{\sqrt{16}} = 0.5$, and X is normally distributed,

we conclude that the sample means is normally distributed, $\bar{X} \sim N(8, 0.5^2)$.

a) $\mathrm{P}(7.8 < \bar{X} < 8.2) = \mathrm{P}\left(\frac{7.8-8}{0.5} < \frac{\bar{X}-8}{0.5} < \frac{8.2-8}{0.5}\right) = \mathrm{P}(-0.4 < Z < 0.4)$

$= \mathrm{P}(Z < 0.4) - \mathrm{P}(Z < -0.4) = 0.6554 - 0.3446 = 0.3108$

Q3. Time spent using e-mail per session is normally distributed with μ= 8 minutes and σ = 2 minutes. If you select a random sample of 16 sessions,

b) What is the probability that the sample mean is between 7.5 and 8 minutes?

c) If you select a random sample of 100 sessions, what is the probability that the sample means is between 7.8 and 8.2 minutes?

d) Explain the difference in the results of (a) and (c).

Solution:

We conclude that, $\bar{X} \sim N(8, 0.5^2)$.

b) $$\mathrm{P}(7.5 < \bar{X} < 8) = \mathrm{P}\left(\frac{7.5-8}{0.5} < \frac{\bar{X}-8}{0.5} < \frac{8-8}{0.5}\right) = \mathrm{P}(-1 < Z < 0)$$
$$= \mathrm{P}(Z < 0) - \mathrm{P}(Z < -1) = 0.5 - 0.1587 = 0.3413$$

c) When $\mathrm{n} = 100$, since $\mu_{\bar{X}} = 8$, *and* $\sigma_{\bar{X}} = \frac{2}{\sqrt{100}} = 0.2$, and X is normally distributed,

we conclude that the sample means is normally distributed, $\bar{X} \sim N(8, 0.2^2)$.

$$\mathrm{P}(7.8 < \bar{X} < 8.2) = \mathrm{P}\left(\frac{7.8-8}{0.2} < \frac{\bar{X}-8}{0.2} < \frac{8.2-8}{0.2}\right) = \mathrm{P}(-1 < Z < 1)$$
$$= \mathrm{P}(Z < 1) - \mathrm{P}(Z < -1) = 0.8413 - 0.1587 = 0.6826$$

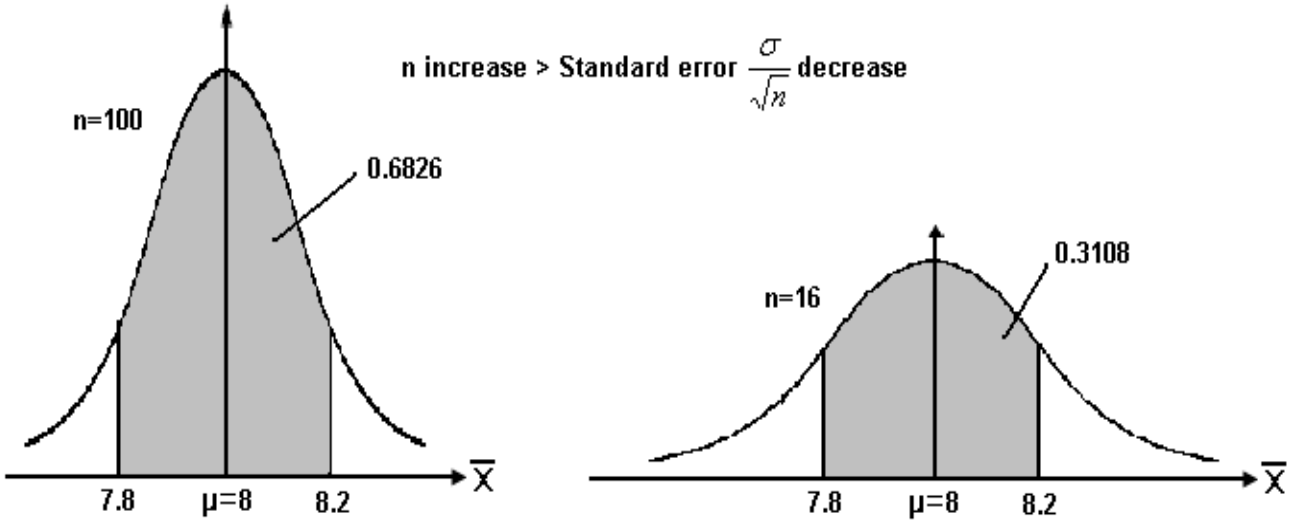
Q3. Time spent using e-mail per session is normally distributed with μ= 8 minutes and σ = 2 minutes. If you select a random sample of 16 sessions,

d) Explain the difference in the results of (a) and (c).

Solution:

c) When $n = 100$,
$\bar{X} \sim N(8, 0.2^2)$;
$P(7.8 < \bar{X} < 8.2) = 0.6826$

a) When $n = 16$, $\bar{X} \sim N(8, 0.5^2)$;
$P(7.8 < \bar{X} < 8.2) = 0.3108$



With the sample size increasing from n = 16 to n = 100, **more sample means will be closer to the population mean**. The standard error of the sample mean of size 100 is much smaller than that of size 16, so the likelihood(probability) that the sample mean will fall within 0.2 minutes of the population mean is much higher for samples of size 100 (probability = 0.6826) than for samples of size 16 (probability = 0.3108).

Q4. In a recent survey concerning the age (to the nearest year) and weight (to the nearest 10 lb) of first-year university students, the following probability distribution was obtained:

Age	Weight				
	100	**110**	**120**	**130**	**140**
19	0.02	0.09	0.09	0.01	0.02
20	0.06	0.15	α	0.05	0.03
21	0.02	0.06	0.11	0.04	0.05

0.2

A sample of 36 first-year students is taken. Find the approximate chance that their total weight is at most 4350 lb.

Solution:

Since $\alpha = 1 - 0.02 - 0.09 - \cdots - 0.05 = 0.2$, we can summarize the distribution of weight, denoted as X:

Weight x_i	**100**	**110**	**120**	**130**	**140**
$\mathbf{P}(x_i)$	0.1	0.3	0.4	0.1	0.1

The population mean: $\mu = \mathrm{E}[X] = \sum_i x_i \mathrm{P}(x_i) = 100 \times 0.1 + \cdots + 140 \times 0.1 = 118$

The population std:

$$\sigma = \sqrt{\sum_i (x_i - \mathrm{E}[X])^2 \mathrm{P}(x_i)} = \sqrt{(100-118)^2 \times 0.1 + \cdots + (140-118)^2 \times 0.1}$$
$$= \sqrt{116} = 10.77$$

Q4. A sample of 36 first-year students is taken. Find the approximate chance that their total weight is at most 4350 lb.

Solution:

Weight x_i	**100**	**110**	**120**	**130**	**140**
$\mathbf{P}(x_i)$	0.1	0.3	0.4	0.1	0.1

The population mean: $\mu = 118$

The population std: $\sigma = 10.77$

When $n = 36$, we obtain that $\mu_{\bar{X}} = 118$, *and* $\sigma_{\bar{X}} = \frac{10.77}{\sqrt{36}} = 1.795$.

Although X is not normally distributed, by Central Limit Theorem, the sampling distribution of mean is normal due to $n > 30$.

Therefore, we conclude that the sample means is normally distributed, $\bar{X} \sim N(118, 1.795^2)$.

$$\begin{aligned} \mathrm{P}(\text{total weight} \le 4350) &= \mathrm{P}\left(\bar{X} \le \frac{4350}{36}\right) = \mathrm{P}\left(\frac{\bar{X}-118}{1.795} \le \frac{4350/36-118}{10.77}\right) \\ &= \mathrm{P}(Z \le 1.58) = 0.9429 \end{aligned}$$

Q5*. At the CityU Computer Service Centre, the loading time for e-Portal page on Internet Explorer is normally distributed with mean 3 seconds.

A random sample of 5 computers is drawn. What is the chance that their total loading time is at least 15 seconds?

*Mean of sample means, $\mu_{\bar{X}} = \sum_i \bar{x}_i \mathrm{P}(\bar{x}_i) = \mu$

*Standard deviation of sample means, $\sigma_{\bar{X}} = \sqrt{\sum_i (\bar{x}_i - \mu_{\bar{X}})^2 \mathrm{P}(\bar{x}_i)} = \sigma/\sqrt{n}$

Solution:

Let X be the loading time, then $X \sim N(3, \sigma^2)$.

Since $\mu_{\bar{X}} = 3$, *and* $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{5}}$, and X is normally distributed,

we get that the average loading time $\bar{X} \sim N(3, (\sigma/\sqrt{5})^2)$.

$$\mathrm{P}(\text{total loading time} \geq 15) = \mathrm{P}\left(\bar{X} \geq \frac{15}{5}\right) = \mathrm{P}(\bar{X} \geq 3)$$
$$= \mathrm{P}\left(\frac{\bar{X}-3}{\sigma/\sqrt{5}} \geq \frac{3-3}{\sigma/\sqrt{5}}\right) = \mathrm{P}(Z \geq 0) = 0.5$$

Q6. Suppose there is a population with population size N = 3. The variable of interest is the salary (X) of individuals. The values of X are 18, 20 and 22 (in thousand dollars).

a) Find the mean (μ) and standard deviation (σ) for the population distribution.

In the process of developing sampling distribution, all possible samples (taken with replacement) of size n = 2 are obtained. The sample mean ($\bar{X}$) is considered as the sample statistic.

b) What are the possible values of this sample mean random variable? Develop the probability distribution of the sample mean.

c) Show that the sample statistic $\bar{X}$ is an unbiased estimator of μ.

d) Denote the standard deviation of $\bar{X}$, verify the following relationship: $\sigma_{\bar{X}} = \sigma/\sqrt{n}$.

e) Does the sampling distribution of $\bar{X}$ follows a Normal Distribution? Explain.

Solution:

a) The population mean: $\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{18+20+22}{3} = 20$

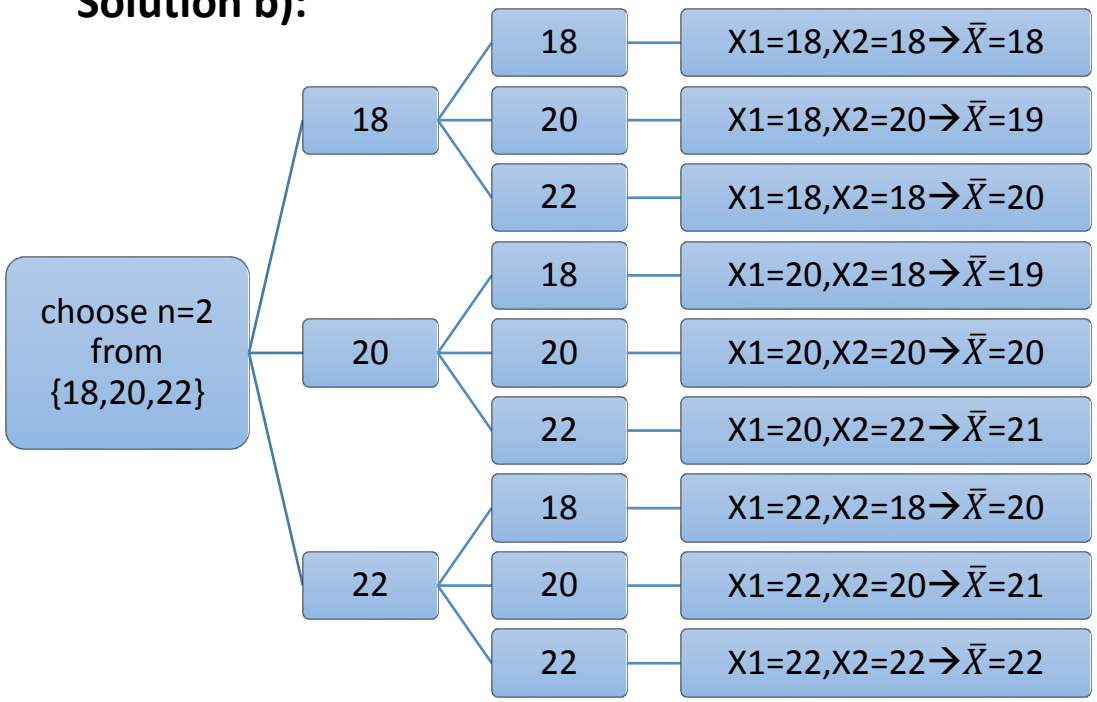
The population std: $\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}} = \sqrt{\frac{(18-20)^2+(20-20)^2+(22-20)^2}{3}} = 1.63$

Q6. Suppose there is a population with population size N = 3. The variable of interest is the salary (X) of individuals. The values of X are 18, 20 and 22 (in thousand dollars).

In the process of developing sampling distribution, all possible samples (taken with replacement) of size n = 2 are obtained. The sample mean ($\bar{X}$) is considered as the sample statistic.

b) What are the possible values of this sample mean random variable? Develop the probability distribution of the sample mean.

Solution b):



$\bar{x}_i$	Frequency	$P(\bar{x}_i)$
18	1	1/9
19	2	2/9
20	3	3/9
21	2	2/9
22	1	1/9

Q6. Suppose there is a population with population size N = 3. The variable of interest is the salary (X) of individuals. The values of X are 18, 20 and 22 (in thousand dollars).

In the process of developing sampling distribution, all possible samples (taken with replacement) of size n = 2 are obtained. The sample mean ($\bar{X}$) is considered as the sample statistic.

c) Show that the sample statistic $\bar{X}$ is an unbiased estimator of μ.

d) Denote the standard deviation of $\bar{X}$, verify the following relationship: $\sigma_{\bar{X}} = \sigma/\sqrt{n}$.

e) Does the sampling distribution of $\bar{X}$ follows a Normal Distribution? Explain.

$\bar{x}_i$	$P(\bar{x}_i)$
18	1/9
19	2/9
20	3/9
21	2/9
22	1/9

Solution: Recall that $\mu = 20, \sigma = 1.63$.

c) $\mu_{\bar{X}} = \sum_i \bar{x}_i P(\bar{x}_i) = 18 \times \frac{1}{9} + \cdots + 22 \times \frac{1}{9} = 20 = \mu.$

d) $\sigma_{\bar{X}} = \sqrt{\sum_i (\bar{x}_i - \mu_{\bar{X}})^2 P(\bar{x}_i)}$

$= \sqrt{(18-20)^2 \times \frac{1}{9} + \cdots + (22-20)^2 \times \frac{1}{9}} = \sqrt{\frac{4}{3}} = 1.15$

$= \sigma/\sqrt{2}.$

e) The sampling distribution of $\bar{X}$ does not follow a Normal Distribution.

- population distribution of X is not normal
- n < 30, cannot apply Central Limit Theorem

Q7. To investigate the length of time working for an employer, researchers at the CityU sampled 344 business students and asked them a question: Over the course of your lifetime, what is the maximum number of years you expect to work for any one employer? The resulting sample had sample mean $\bar{X}$ =19.1 years and sample standard deviation S=6 years. Assume the sample of students was randomly selected from the 5800 undergraduate students in CityU.

a) What are reasonable estimators of population mean and population standard deviation?

b) What is the sampling distribution of $\bar{X}$? Why?

Solution:

a) Sample mean and sample standard deviation are reasonable estimators of population mean and population standard deviation.

b) According to Central Limit Theorem, with large sample size (n = 344), sample mean $\bar{X}$ follows a normal distribution approximately, with mean 19.1 and standard deviation $\frac{6}{\sqrt{344}} = 0.3235.$

Q7. To investigate the length of time working for an employer, researchers at the CityU sampled 344 business students and asked them a question: Over the course of your lifetime, what is the maximum number of years you expect to work for any one employer? The resulting sample had sample mean $\bar{X}$ =19.1 years and sample standard deviation S=6 years. Assume the sample of students was randomly selected from the 5800 undergraduate students in CityU.

c) If the population mean was 18.5 years, what is P($\bar{X}$ ≥19.1 years)?

d) If the population mean was 19.5, what is P($\bar{X}$ =19.1 years)?

e) If P($\bar{X}$ ≥19.1 years) = 0.5, what is the population mean?

Solution:

c) $\bar{X} \sim N(18.5, 0.3235^2)$

$$\mathrm{P}(\bar{X} \geq 19.1\ years) = \mathrm{P}\left(\frac{\bar{X}-18.5}{0.3235} \geq \frac{19.1-18.5}{0.3235}\right) = \mathrm{P}(Z \geq 1.8547)$$
$$= 1 - \mathrm{P}(Z < 1.8547) = 1 - 0.9678 = 0.0322$$

d) $\bar{X} \sim N(19.5, 0.3235^2)$, $\mathrm{P}(\bar{X} = 19.1\ years) = 0$

e) We know that $\mathrm{P}(\bar{X} \geq 19.1) = 0.5$ → $\mathrm{P}(\bar{X} < 19.1) = 0.5$

→ $\mathrm{P}\left(\frac{\bar{X}-\mu}{0.3235} < \frac{19.1-\mu}{0.3235}\right) = 0.5$ → $\frac{19.1-\mu}{0.3235} = 0$ → $\mu = 19.1$ years

Q7. To investigate the length of time working for an employer, researchers at the CityU sampled 344 business students and asked them a question: Over the course of your lifetime, what is the maximum number of years you expect to work for any one employer? The resulting sample had sample mean $\bar{X}$ =19.1 years and sample standard deviation S=6 years. Assume the sample of students was randomly selected from the 5800 undergraduate students in CityU.

f) If P($\bar{X}$ ≥19.1 years) = 0.2, without calculation, can you tell that the population mean is greater or less than 19.1 years? Explain.

Solution:

f) Yes. The population mean will be less than 19.1 years.

Because $\bar{X} \sim N(\mu, 0.3235^2)$ and for a normal distribution mean is equal to median.

However, now we observe $\mathrm{P}(\bar{X} \geq 19.1\ years) = 0.2 < 0.5$→median<19.1

Therefore, the population mean should be less than 19.1 years.