Statistics & Probability

Chapter 7: SAMPLING DISTRIBUTIONS AND POINT ESTIMATION OF PARAMETERS

FPT University

Department of Mathematics

Quy Nhon, 2023

Table of Contents

Table of Contents

1. **Statistical inference**: statistical methods are used to make decisions and draw conclusions about populations.
2. Statistical inference may be divided into two major areas:
   - Parameter estimation
     $\rightarrow$ Using sample data to compute a number that is in some sense a reasonable value (a good guess), is called **point estimate** (or a reasonable value), of the true population mean.
   - Hypothesis testing
     $\rightarrow$ This is a method used to decide whether the data at hand sufficiently support a particular hypothesis and it allows us to make probabilistic statements about population parameters.

Table of Contents

Suppose we want to obtain a **point estimate** of a population parameter.

1. Before the data are collected, the observations are considered to be random variables, say, $X_1, X_2, \ldots, X_n$.
2. Any function of the observations, or any **statistic**, is also a random variable. $\underline{Example}$: The sample mean $\overline{X}$ and the sample variance S^2 are statistics. Hence, they are also random variables.
3. A statistic is a random variable, so, it has a probability distribution.
   $\rightarrow$ We call the probability distribution of a statistic is a **sampling distribution**.
4. The symbol θ is used to represent the **parameter** such as the mean μ, the variance σ^2 or any parameter of interest.
5. The objective of point estimation is to select a single number, based on sample data, that is the most plausible value for θ. A numerical value of a sample statistic will be used as the point estimate.
6. In general, if X is a random variable with probability distribution $f(x)$, characterized by the unknown parameter θ, and if $X_1, X_2, \ldots, X_n$ is a random sample of size n from X; the statistic $\hat{\Theta} = h(X_1, X_2, \ldots, X_n)$ is called a **point estimator** of θ.

   Note: $\hat{\Theta}$ is a random variable and after the sample has been selected $\hat{\Theta}$ takes on a particular numerical value $\hat{\theta}$ called the **point estimate** of θ.

Point Estimator

A **point estimator** of some population parameter θ is a single numerical value $\hat{\theta}$ of a statistic $\hat{\Theta}$. The statistic $\hat{\Theta}$ is called the **point estimator**.

Example

Suppose the random variable X is normally distributed with an unknown mean μ.

1. The sample mean is a point estimator of the unknown population mean μ, i.e., $\hat{\mu} = \overline{X}$.
2. After the sample has been selected, the numerical value $\overline{x}$ is the point estimate of μ.
3. Therefore, if $x_1 = 25,\ x_2 = 30,\ x_3 = 29$ and $x_4 = 31$, the point estimate of μ is

$$\overline{x} = \frac{x_1 + x_2 + x_3 + x_4}{4} = \frac{25 + 30 + 29 + 31}{4} = 28.75.$$

Quiz

Using the data as in the above example to find the point estimate s^2 of the unknown population variance σ^2.

Estimation problems occur frequently in engineering. We often need to estimate

1. The mean μ of a single population.
2. The variance σ^2 (or standard deviation σ) of a single population.
3. The proportion p of items in a population that belong to a class of interest.
4. The difference in means of two populations $\mu_1 - \mu_2$.
5. The difference in two population proportions $p_1 - p_2$.

Reasonable point estimates of these parameters are as follows:

1. For μ, the estimate is $\hat{\mu} = \overline{x}$, the sample mean.
2. For σ^2, the estimate is $\hat{\sigma}^2 = s^2$, the sample variance.
3. For p, the estimate is $\hat{p} = \dfrac{x}{n}$, the sample proportion, where x is the number of items in a random sample size n that belong to the class of interest.
4. For $\mu_1 - \mu_2$, the estimate is $\hat{\mu}_1 - \hat{\mu}_2 = \overline{x}_1 - \overline{x}_2$, the difference between the sample means of two independent random samples.
5. For $p_1 - p_2$, the estimate is $\hat{p}_1 - \hat{p}_2$, the difference between two sample proportions computed from two independent random samples.

Example

In a sample of 73 products, there are 7 defective products. So a point estimate for proportion p of all defective products is

$$\hat{p} = \frac{x}{n} = \frac{7}{73} \approx 9.6\%.$$

Example

Market researchers use the number of sentences per advertisements. The following represent a random sample of the number of sentences found in 15 advertisements.

$$9,\ 20,\ 18,\ 16,\ 9,\ 9,\ 11,\ 13,\ 22,\ 16,\ 5,\ 18,\ 6,\ 6,\ 5.$$

1. Find a point estimate of the population mean μ.

   Solution. We have $\overline{x} = \dfrac{\sum_{i=1}^{n} x_i}{n} = \dfrac{\sum_{i=1}^{20} x_i}{20} = 12.2.$

2. Find a point estimate of the population standard deviation σ.

   Solution. We have $s = \sqrt{\dfrac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}} = \sqrt{\dfrac{\sum_{i=1}^{n} (x_i - 12.2)^2}{20-1}} \approx 5.77.$

Note

1. We may have some different choices for the point estimator of a parameter.
   $\underline{Example}$: To estimate the mean of a population; we might consider the sample mean, the sample median, or perhaps the average of the smallest and largest observations in the sample as point estimators.
2. To decide whether which point estimator of a particular parameter is the best one to use, we need to examine their statistical properties and develop several criteria for comparing estimators.

Table of Contents

Sampling Distribution

Random Sample

The random variables $X_1, X_2, \ldots, X_n$ are a **random sample** of size n if

1. the X_i's are independent random variables, and
2. every X_i has the same probability distribution.

Example (Random Sample)

The names of 100 students being chosen out of a hat from a university of 1000 students. In this case, the population is all 1000 students, and the sample is random because each student has an equal chance of being chosen.

Statistic

A **statistic** is any function of the observations in a random sample.

Example (Statistic)

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n. Then, the sample mean $\overline{X}$, the sample variance S^2, the sample standard deviation S are statistics.

Sampling Distribution

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

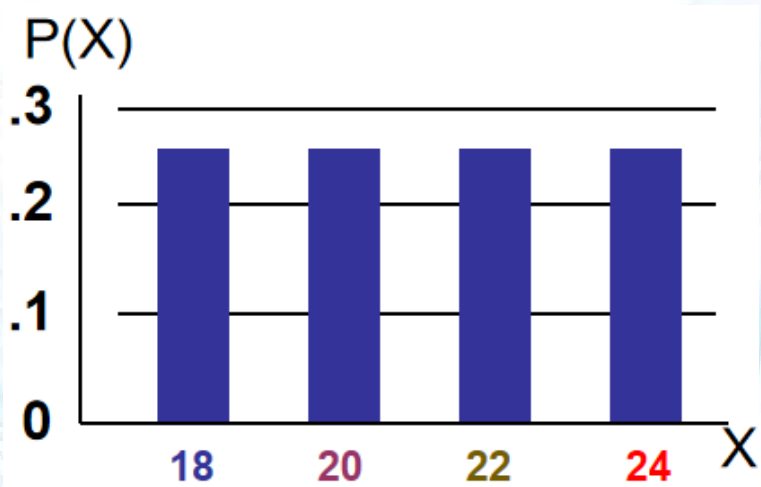
Example (Sampling Distribution)

The probability distribution of $\overline{X}$ is called the sampling distribution of the mean.

Developing a Sampling Distribution

Example. Assume that there is a population of size $N = 4$ and the variable of interest is, X, age of individuals. The values of X are given by $18,\ 20,\ 22,\ 24$ (years).

- $\mu = \dfrac{\sum_{i=1}^{N} X_i}{N} = 21.$
- $\sigma = \sqrt{\dfrac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}} \approx 2.236.$



Uniform distribution.

Consider now all possible samples of size $n = 2$.

1st Obs	2nd Observation			
	18	20	22	24
18	18,18	18,20	18,22	18,24
20	20,18	20,20	20,22	20,24
22	22,18	22,20	22,22	22,24
24	24,18	24,20	24,22	24,24

16 possible samples (sampling with replacement)

16 Sample Means

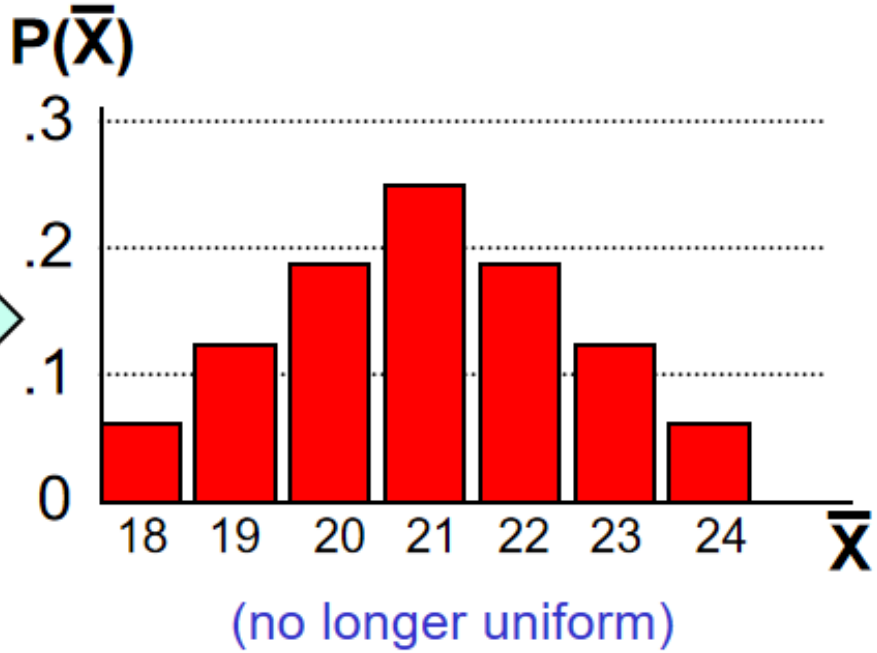
1st Obs	2nd Observation			
	18	20	22	24
18	18	19	20	21
20	19	20	21	22
22	20	21	22	23
24	21	22	23	24

Sampling distribution of all sample means.

16 Sample Means

1st Obs	2nd Observation			
	18	20	22	24
18	18	19	20	21
20	19	20	21	22
22	20	21	22	23
24	21	22	23	24

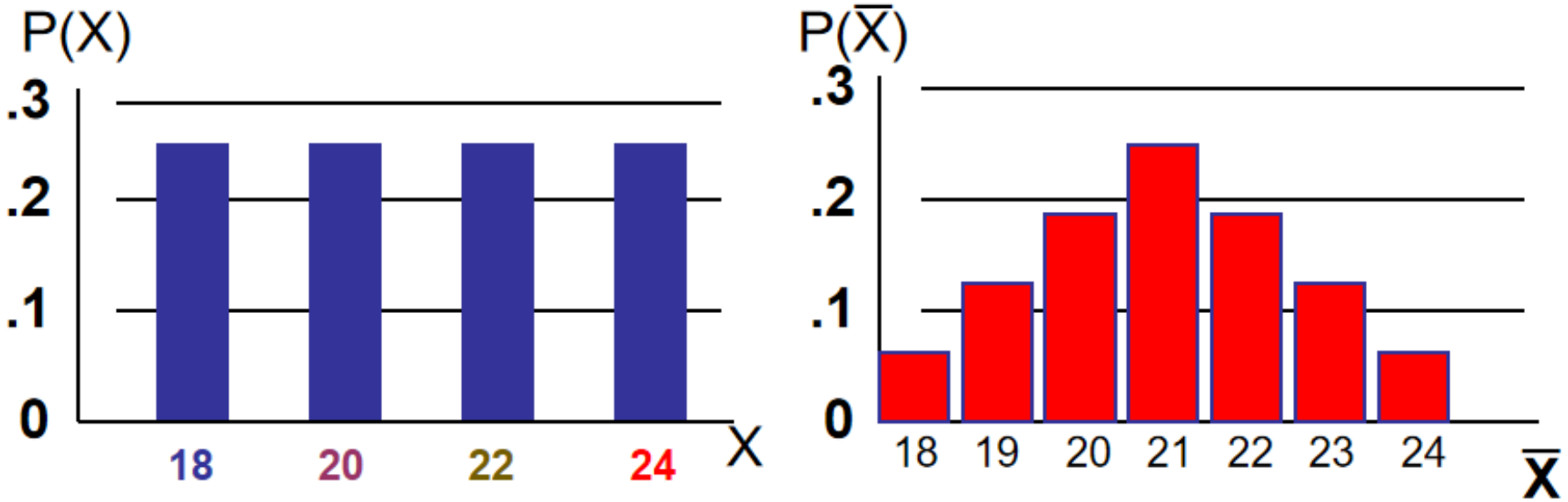
Sample Means Distribution



- $\mu_{\overline{X}} = \dfrac{18 + 19 + \cdots + 24}{16} = 21.$
- $\sigma_{\overline{X}} = \sqrt{\dfrac{(18-21)^2 + (19-21)^2 + \cdots + (24-21)^2}{16}} = 1.58.$

Note. Here, we divide by 16 since there are 16 different samples size of 2.

Comparing the Population Distribution to the Sample Means Distribution



Population $N = 4$

- $\mu = 21.$
- $\sigma \approx 2.236.$

Sample Means Distribution $n = 2$

- $\mu_{\overline{X}} = 21.$
- $\sigma_{\overline{X}} \approx 1.58.$

Problem

Problem: Determine the sampling distribution of the sample mean $\overline{X}$.

1. Suppose that a random sample of size n is taken from a normal population with mean μ and variance σ^2.
2. Each observation in this sample, say, $X_1, X_2, \ldots, X_n$, is a normally and independently distributed random variable with mean μ and variance σ^2.
3. We conclude that the sample mean

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

has a normal distribution with mean

$$\mu_{\overline{X}} = \frac{\mu + \mu + \cdots + \mu}{n} = \mu$$

and variance

$$\sigma^2_{\overline{X}} = \frac{\sigma^2 + \sigma^2 + \cdots + \sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Remark: If the population already has normal distribution, then for any sample size

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Question. Assume we are sampling from a population that has an unknown probability distribution. How to determine the sampling distribution of the sample mean?

Central Limit Theorem (CLT)

If $X_1, X_2, \ldots, X_n$ is a random sample of size n taken from a population (either finite or infinite) with mean μ and finite variance σ^2, and if $\overline{X}$ is the sample mean, the limiting form of the distribution of

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

as $n \to +\infty$, is the standard normal distribution.

Remarks

- The CLT says that the sampling distribution of the mean will always be normally distributed, as long as the sample size is large enough $(n \geq 30)$. Regardless of whether the population has a normal, Poisson, binomial, or any other distribution, the sampling distribution of the mean will be normal.
- $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ where $\overline{X}$ is the sampling distribution of the sample means, μ is the mean of the population, σ is the standard deviation of the population and n is the sample size (note $n \geq 30$).

Note

1. We can apply the Central Limit Theorem for even smaller sample sizes if the population distribution is approximately bell-shaped.
2. If the distribution of the population is fairly symmetrical, the sampling distribution of the mean is approximately normal for samples as small as size 5.
3. In the case in which the distribution of a variable is extremely skewed or has more than one mode, you may need sample sizes larger than 30 to ensure normality in the sampling distribution of the mean.

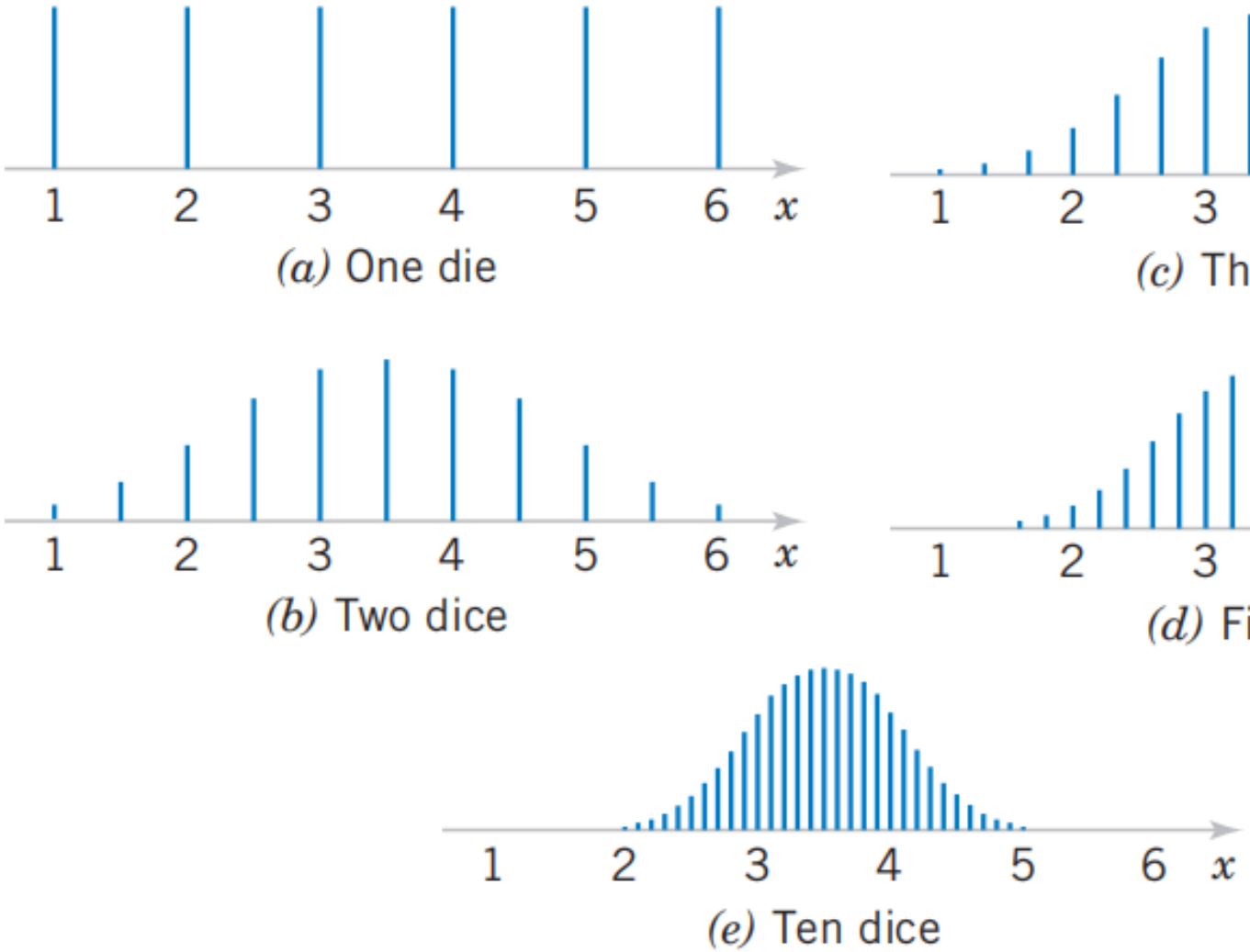


Figure. Distributions of average scores from throwing dice.

Example (Central Limit Theorem)

Suppose that a random variable X has a continuous uniform distribution

$$f(x) = \begin{cases} \frac{1}{2} & \text{if } 4 \le x \le 6 \\ 0 & \text{otherwise.} \end{cases}$$

Find the distribution of the sample mean of a random sample of size $n = 40$.

Solution. The mean and the variance of X are

$$\mu = \frac{b+a}{2} = 5 \quad \text{and} \quad \sigma^2 = \frac{(b-a)^2}{12} = \frac{1}{3}.$$

The CLT indicates that the distribution of $\overline{X}$ is approximately normal with mean and variance are

$$\mu_{\overline{X}} = \mu = 5 \quad \text{and} \quad \sigma^2_{\overline{X}} = \frac{\sigma^2}{n} = \frac{1}{120}.$$

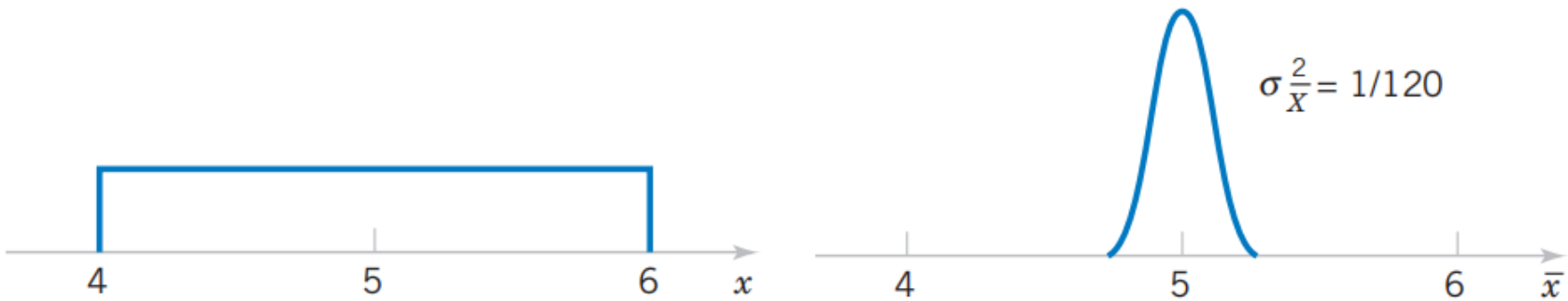


Figure. The distribution of X and $\overline{X}$.

Quiz 1

One year, professional players salaries averaged 1.5 million with a standard deviation of 0.9 million. Suppose a sample of 100 players was taken. Find the approximate probability that the average salary of these 100 players does not exceed 1.4 million.

Quiz 2

An electrical firm manufactures light bulbs that have a length of life that is approximately normally distribution, with mean equal to 800 hours and a standard deviation of 40 hours. Find the probability that a random sample of 16 bulbs will have an average life of less than 775 hours.

Approximate Sampling Distribution of a Difference in Sample Means

If we have two independent population with means μ_1, μ_2 and variance σ_1^2, σ_2^2, and if $\overline{X}_1, \overline{X}_2$ are the sample means of two independent random samples of sizes n_1, n_2 from the populations; then the sampling distribution of

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

is approximately standard normal, if the conditions of the CLT apply. In the case, the two populations are normal, the sampling distribution of Z is exactly standard normal.

Remark

Under the conditions of the CLT,

$$\overline{X}_1 - \overline{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

Example (Aircraft Engine Life)

The effective life of a component used in a jet-turbine aircraft engine is a random variable with mean $\mu_1 = 5000$ hours and standard deviation $\sigma_1 = 40$ hours. The distribution of effective life is fairly close to a normal distribution. The engine manufacturer introduces an improvement into the manufacturing process for this component that increases the mean life to $\mu_2 = 5050$ hours and decreases the standard deviation to $\sigma_2 = 30$ hours. Suppose that a random sample of $n_1 = 16$ components in selected from the old process and a random sample $n_2 = 25$ components is selected from the improved process. What is the probability that the difference in the two sample means $\overline{X}_2 - \overline{X}_1$ is at least 25 hours? Assume that the old and improved processes can be regarded as independent populations.

Solution. We have

$$\overline{X}_1 \sim N\left(\mu_1 = 5000, \frac{\sigma_1^2}{n_1} = 100\right) \quad \text{and} \quad \overline{X}_2 \sim N\left(\mu_2 = 5050, \frac{\sigma_2^2}{n_2} = 36\right)$$

which implies $\overline{X}_2 - \overline{X}_1 \sim N\left(\mu_2 - \mu_1 = 50, \frac{\sigma_2^2}{n_2} + \frac{\sigma_1^2}{n_1} = 136\right) =: N(\mu = 50, \sigma^2 = 136)$.

Therefore,

$$P(\overline{X}_2 - \overline{X}_1 \geq 25) = P\left(\frac{(\overline{X}_2 - \overline{X}_1) - \mu}{\sigma} \geq \frac{25 - \mu}{\sigma}\right) = P(Z \geq -2.14) = 0.9838.$$

Recall

$$\overline{X}_1 - \overline{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

Quiz

The television picture tubes of manufacturer A have a mean lifetime of 6.5 years and a standard deviation of 0.9 year, while those of manufacturer B have mean lifetime of 6.5 years and a standard deviation of 0.8 year. What is the probability that a random sample of 36 tubes from manufacture A will have mean lifetime that is at least 1 year more than the mean lifetime of a sample of 49 tubes from manufacturer B?

Hint.

Population 1	Population 2
$\mu_1 = 6.5$	$\mu_2 = 6.5$
$\sigma_1 = 0.9$	$\sigma_2 = 0.8$
$n_1 = 36$	$n_2 = 49$

Thank you!