Sampling Distribution

Definition : Let $X_1, \ldots, X_n$ be a sequence of random variables. We say that $X_1, \ldots, X_n$ are identically distributed if they have the same marginal probability distribution.

Definition: Let $X_1, \ldots, X_n$ be random variables. We say that $X_1, \ldots, X_n$ is a random sampling from a certain **population** if $X_1, \ldots, X_n$ are *independent* and *identically distributed* (i.i.d.).

Remarks:

- We will model an experiment with n trials as an i.i.d. random sampling.
- The probability distribution shared by $X_1, \ldots, X_n$ is called the **population**.
- The mean μ that is shared by $X_1, \ldots, X_n$ is called the mean of the population. In other words, $E[X_i] = \mu$ for all $i = 1, 2, \ldots, n$.
- The variance σ^2 that is shared by $X_1, \ldots, X_n$ is called the variance of the population. In other words, $\mathsf{Var}[X_i] = \sigma^2$ for all $i = 1, 2, \ldots, n$.

Definition: A function of a random sampling is called a *statistic*. A statistic is itself a random variable and its probability distribution is called its **sampling distribution.**

Remark: In these notes we are interested in the sampling distribution of the sample mean $\overline{X}$ of a sample of size n from a population with mean μ and variance σ^2.

Linear Combinations of Independent Random Variables

Let $X_1, X_2, \ldots, X_n$ be random variables. A linear combination Y of $X_1, X_2, \ldots, X_n$ is a random variable of the form

$$Y = a_1 X_1 + a_2 X_2 + \ldots + a_n X_n = \sum_{i=1}^{n} a_i X_i,$$

where $a_1, a_2, \ldots, a_n$ are real constants. Here are some facts about linear combinations:

1. The mean of $Y = a_1 X_1 + a_2 X_2 + \ldots + a_n X_n$ is

$$E[Y] = a_1 E[X_1] + a_2 E[X_2] + \ldots + a_n E[X_n] = \sum_{i=1}^{n} a_i E[X_i].$$

2. If $X_1, X_2, \ldots, X_n$ are **independent**, then

$$V[Y] = a_1^2 V[X_1] + a_2^2 V[X_2] + \ldots + a_n^2 V[X_n] = \sum_{i=1}^{n} a_i^2 V[X_i].$$

3. If $X_1, X_2, \ldots, X_n$ are **independent** and **normal**, then

$$Y \sim N(E[Y]; V[Y]).$$

   In other words, Y is also normal.

Remark: The sample mean $\overline{X} = \sum_{i=1}^{n} X_i/n$ is a linear combination with $a_i = 1/n$. So, if the population is $N(\mu, \sigma^2)$, then

$$\overline{X} \sim N\left(\mu_{\overline{X}} = \mu, \sigma^2_{\overline{X}} = \frac{\sigma^2}{n}\right) \quad \text{and} \quad Z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

Exemple 42: Suppose that the lifetime of a battery is normally distributed with a mean of 150 hours and a standard deviation of 25 hours. We take a random sampling of $n = 15$ batteries. What is the probability the sample mean lifetime of these 15 batteries will be greater than 150 hours?
solution: $\mu = 150$, $\sigma = 25$, $n = 15$ since $X_1, \ldots, X_n$ are normal, $\overline{X}$ is a linear combination of normals, so

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

Now,

$$\mu_{\overline{X}} = \mu = 15$$

and

$$\sigma^2_{\overline{X}} = \frac{\sigma^2}{n} = \frac{25^2}{15} = 41.67$$

so

$$\overline{X} \sim N\left(\mu_{\overline{X}} = 150, \sigma^2_{\overline{X}} = 41.67\right).$$

Therefore,

$$\begin{aligned} P(\overline{X} > 150) &= P\left(\frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} > \frac{150 - \mu_{\overline{X}}}{\sigma_{\overline{X}}}\right) \\ &= 1 - P\left(\frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} \leq \frac{150 - \mu_{\overline{X}}}{\sigma_{\overline{X}}}\right) \\ &= 1 - P\left(\frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} \leq \frac{150 - 150}{41.67}\right) \\ &= 1 - P\left(\frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} \leq 0\right) \\ &= 1 - \Phi(0) = 1 - \frac{1}{2} = \frac{1}{2} \end{aligned}$$

Question : If the population is not normal, what is the sampling distribution of $\overline{X}$? The following theorem gives an approximate answer to this question.

Central limit theorem: Suppose that $X_1, X_2, \ldots, X_n$ is a random sampling of a population with mean μ and variance σ^2. Let

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

Then

$$\lim_{n\to\infty} F_{Z_n}(z) = \Phi(z).$$

Remarks:

- So, if the sample size n is large,

$$\overline{X} \sim N\left(\mu; \frac{\sigma^2}{n}\right) \quad \text{approximately.}$$

- What should the sample size be for this to be a good approximation? The answer to this question depends on the population. If the population is close to normal, then the approximation will be good regardless of the size of the sample. However, if the population is far from being normal, then we need n to be larger. In practice we use $n \geq 30$ as a rule of thumb for when this is a good approximation.

Example 43: The pressure tolerance of concrete has mean of 2500 pounds per square inch and standard deviation of 50 pounds per square inch. Please approximate the probability that the sample mean of the pressure tolerance of a random sampling of 40 concrete slabs will be less than 2490 pounds per square inch.

solution:

$$E[\overline{X}] = \mu = 2500$$

$$V[\overline{X}] = \frac{1}{n}\sigma^2 = \frac{1}{40} \times 50^2 = 62.5$$

so by the central limit theorem

$$\overline{X} \sim N\left(\mu = 62.5, \frac{1}{n}\sigma^2 = 62.5\right), \quad \text{approximately.}$$

So

$$P(\overline{X} < 2490) = P\left(\frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} < \frac{2490 - \mu_{\overline{X}}}{\sigma_{\overline{X}}}\right)$$

$$= \Phi\left(\frac{2490 - 2500}{\sqrt{62.5}}\right) = \Phi(-1.26) = 0.1038$$

Note: We did not do the following part of these notes in class ("Comparing the sample means of independent populations"), so the the following section will not be covered on the exam.

Comparing the sample means of independent populations:

Let $\overline{X}$ and $\overline{Y}$ be the sample means of random samples from two populations $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$. Let n_X and n_Y be the sample sizes of these two samplings.

We know that

$$\overline{X} \text{ has distribution } N(\mu_X, \sigma_X^2/n_X)$$

and

$$\overline{Y} \text{ has distribution } N(\mu_Y, \sigma_Y^2/n_Y).$$

Since $X - Y$ is the linear combination of independent normal random variables X and Y, the random variable $X - Y$ is also normal. In other words, we have

$$E[\overline{X}_1 - \overline{X}_2] = E[\overline{X}_1] - E[\overline{X}_2] = \mu_1 - \mu_2$$

and

$$V[\overline{X}_1 - \overline{X}_2] = 1^2 V[\overline{X}_1] + (-1)^2 E[\overline{X}_2] = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

So,

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim N(0,1).$$

Remark:
By the central limit theorem, if $n_X \geq 30$ and $n_Y \geq 30$, then Z is approximately $N(0,1)$, even if the populations themselves are not normal.

Example 44: The lifetime of a component in a jet airplane is normally distributed with a mean of 5050 hours and a standard deviation of 40 hours. Suppose that when a different manufacturing process is used, the lifetime is is normally distributed with a mean of 5000 hours and a standard deviation

of 30 hours. Let $\overline{X}$ be the sample mean lifetime of a random sampling of $n_X = 20$ components from first manufacturing process, and let $\overline{Y}$ be the sample mean lifetime of a random sampling of $n_Y = 20$ components from the second manufacturing process. Determine the following probability:

$$P\left(\overline{X} - \overline{Y} > 10\right).$$