Course: Mathematical statistics

Week 2: Sampling Distribution

Lecturer: Nagulama Moses

Kumi University

March 24, 2025

Outline

1. Sampling distribution for difference between two sample mean
2. Sampling distribution of sample proportion
3. Sampling distribution for the difference between two sample proportion $(\bar{p_1} - \bar{p_2})$

Intended learning outcomes

- Describe the sampling distribution of $\bar{x_1} - \bar{x_2}$ for independent random samples.
- Apply the Central Limit Theorem (CLT) to approximate the distribution of $\bar{x_1} - \bar{x_2}$.
- State the conditions under which the sampling distribution of $\bar{p}$ is approximately normal.
- Explain the sampling distribution of $\bar{p_1} - \bar{p_2}$ for two independent samples.

sampling distribution for difference between two sample mean

- Let us consider the case where two populations are independent.
- Suppose the first population has μ_1 and σ_1^2 and the second population has μ_2 and σ_2^2.
- Suppose the two population are normally distributed since two populations are independent then their linear combinations are also normally distributed.

- Suppose a sample of size n_1, selected randomly from the first population has a sample mean $\bar{x_1}$ and sample of size n_2 selected randomly from second population has a sample mean of $\bar{x_2}$.
- Then the sampling distribution for the difference between two sample means will be normal $(\bar{x_1} = \bar{x_2})$

- we denote $\mu_{\bar{x_1}-\bar{x_2}}$ to stand for the mean for the difference between two sample mean and $\sigma^2_{\bar{x_1}-\bar{x_2}}$ variance between the difference between the two sample means.

$$\mu_{\bar{x_1}-\bar{x_2}} = E(\bar{x_1} - \bar{x_2}) = E(\bar{x_1}) - E(\bar{x_2}) = \mu_1 - \mu_2$$

$$\sigma^2_{\bar{x_1}-\bar{x_2}} = var(\bar{x_1} - \bar{x_2})$$

$$= var(\bar{x_1}) + var(\bar{x_2})$$

variance of independent is additive

$$= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

$$\sigma^2_{\bar{x_1}-\bar{x_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- **Theorem: CLT**

  if we have two independent population with μ_1, μ_2 and variances σ_1^2, σ_2^2 and $\bar{x_1}, \bar{x_2}$ are the sample means of two independent random samples of sizes n_1 and n_2 respectively taken from these populations,

  then the sampling distribution for the difference between two sample means is approximately normally distributed with $N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$

  consequently

$$z = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

  is approximately a standard normal random variable i.e.
  $n_1 \geq 30, n_2 \geq 30$

Example

from two normal and independent populations were the mean of the second population is less than that of the first by 0.5. pairs of samples are drawn such that each pair contains one sample from each population. if the sample from the first population contains 90 variates and from the second population contain 120 variates and if the $\sigma_1 = 10, \sigma_2 = 8$. find the probability that in a pair of samples the difference between the first and second sample means will be

(i) be less than 1.5

(ii) More than 0.8

solution

Given

$$\mu_1, \mu_2, \mu_1 - \mu_2 = 0.5, \sigma_1 = 10, \sigma_2 = 8, n_1 = 90, n_2 = 120$$

independent normal population

sampling distribution of $\bar{x_1} - \bar{x_2}$ is normal $N(0.5, \frac{100}{90} + \frac{64}{120})$

(i)

$$z = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$P(\bar{x_1} - \bar{x_2} < 1.5)$$

$$z = \frac{1.5 - 0.5}{\sqrt{\frac{100}{90} + \frac{64}{120}}} = 0.78$$

$$0.2823 + 0.5 = 0.7823$$

(ii)

$$P(\bar{x_1} - \bar{x_2}) > 0.8$$

$$z = \frac{0.8 - 0.5}{\sqrt{\frac{100}{90} + \frac{64}{120}}} = 0.23$$

$$0.5 - 0.09095 = 0.4090$$

sampling distribution of sample proportion

- Let sample proportion $\hat{p}$. let x be a number of successes in n independent trials with a probability p which is referred to as the population proportion and $\hat{p}$ is the probability of success in each trial

$$\hat{p} = \frac{x}{n}, E(\hat{p}) = E(\frac{x}{n}) = \frac{1}{n}(x) = \frac{1}{n}.np$$

$$\mu_{\hat{p}} = p$$

$$var(\hat{p}) = var(\frac{x}{n}) = \frac{1}{n^2} var(x) = \frac{1}{n^2} npq = \frac{1}{n} pq$$

$$\frac{1}{n} p(1-p)$$

$$\sigma^2_{\hat{p}} = \frac{p(1-p)}{n}$$

if x is the number of successes in n independent trials with a constant population proportion p of success of each trial then the sampling distribution of a sample is approximately normal with $\mu_{\hat{p}} = p$ and $\sigma^2_{\hat{p}} = \frac{p(1-p)}{n}$ provided the sample is large.

consequently the value of

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

is a standard normal random variable

conditions

sampling distribution

sample is large if $np \geq 5$ and $n(1-p) \geq 5$

Example

in the lottery club of Mbale, 30% of its members are 5 years old in the club. a random sample of 100 members are selected from the club . if the sample proportion of members $\hat{p}$ who are 5 years old in the club

(a) find the sample distribution of a sample proportion.

(b) What is the probability that a sample proportion will be between 0.2 and 0.4.

(c) what is the probability that a sample proportion will be within ± 0.05 of the population proportion

solution

$n = 100, p = 0.3, np = 0.3X100 = 30 > 5, n(1-p) = 100(1-0.3) = 70 > 5$

The sample is large. By CLT the sampling distribution of $\hat{p}$ is $\approx$

$$N(\mu_{\hat{p}}, \sigma^2_{\hat{p}}) = N(0.3, \frac{(0.3)(0.7)}{100}$$

(b)

$$p(0.2 < \hat{p} < 0.4)$$

standardise

$$z_1 = \frac{0.2 - 0.3}{\sqrt{\frac{(0.3)(1-0.3)}{100}}} = -2.18$$

$$z_2 = \frac{0.4 - 0.3}{\sqrt{\frac{(0.3)(1-0.3)}{100}}} = 2.18$$

$$p(0.2 < \hat{p} < 0.4) = p(-2.18 < z < 2.18)$$

$$= 0.4854 + 0.4854 = 0.9708$$

(c)

$$p(\hat{p} = p \pm 0.05)$$

$$= p(p - 0.05 < \hat{p} < p + 0.05)$$

$$= p(0.3 - 0.05 < \hat{p} < 0.3 + 0.05)$$

$$= p(0.25 < \hat{p} < 0.35)$$

$$z_1 = \frac{0.25 - 0.3}{\sqrt{\frac{(0.3)(0.7)}{100}}} = -1.09$$

$$z_2 = \frac{0.35 - 0.3}{\sqrt{\frac{(0.3)(0.7)}{100}}} = 1.09$$

$$p(\hat{p} = p \pm 0.05) = p(-1.09 < z < 1.09)$$

$$0.3621 + 0.3621 = 0.7242$$

Example

Medical evidence show that 80% of all patients having a particular disease will fully recover within 3 days after receiving a new drug. a random sample of 20 patients suffering from the disease are treated with a new drug to estimate the sample proportion $\hat{p}$ of patients who recover in 3 days. A data analyst suggests to use a normal probability distribution approximation for the sampling distribution of a sample proportion. Explain whether this approximation is valid.

solution

$$n = 20, p = 0.8$$

w need to first test if the sample is large

$$np = 20 \times 0.8 = 16 \geq 5, n(1-p) = 20(1-0.8) = 4 \leq 5)$$

the approximation is not valid because the sample is not large

sampling distribution for the difference between two sample proportion $(\bar{p}_1 - \bar{p}_2)$

let x_1 be the number of successes in n_1 independent trials taken from a population with proportion p_1 of success and x_2 is the number of successes in n_2 independent trials taken from another population with proportion p_2 of successes

assume that two population are independent

$$\hat{p}_1 = \frac{x_1}{n_1}, \hat{p}_2 = \frac{x_2}{n_2}$$

$$E(\hat{p_1}) = E(\frac{x_1}{n_1}) = \frac{1}{n_1}E(x_1) = \frac{1}{n_1}n_1p_1 = p_1$$

$$E(\hat{p_2}) = E(\frac{x_2}{n_2}) = \frac{1}{n_2}E(x_2) = \frac{1}{n_2}n_2p_2 = p_2$$

$$E(\hat{p_1} - \hat{p_2}) = E(\hat{p_1}) - E(\hat{p_2}) = p_1 - p_2$$

$$var(\hat{p_1}) = \frac{p_1(1-p_1)}{n_1}$$

$$var(\hat{p_2}) = \frac{p_2(1-p_2)}{n_2}$$

$$var(\hat{p_1} - \hat{p_2}) = var(\hat{p_1}) + var(\hat{p_2}) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

if n_1 and n_2 are large samples, then $\hat{p_1}$ and $\hat{p_2}$ are approximately normally distributed and

consequently the $\hat{p_1} - \hat{p_2}$ are also approximately normally distributed with $\mu_{\hat{p_1}-\hat{p_2}} = p_1 - p_2$ and

$$\sigma^2_{\hat{p_1}-\hat{p_2}} = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

$$z = \frac{(\hat{p_1} - \hat{p_2}) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

is the standard random normal variable.

Example

in a certain sub county 70% of all the legible support a particular candidate in election. in another sub county 60% of eligible voters support that particular candidate for a parliamentary election. a random sample of size $n_1 = 200$ is taken from a first sub county and in a random sample of size $n_2 = 400$ is taken from the second sub county. if $\hat{p}_1, \hat{p}_2$ are the sample proportions to support that particular candidate in two sub county

(i) find the sampling distribution for the difference of voters in two sub counties who support that candidate

(ii) find the probability that the difference between two sample proportion in two sub county who support that candidate is atleast 0.02.

(iii) Find the probability between two sample proportion of voters in two sub county that candidate is between 0.3 and 0.12.

solution

$$p_1 = 0.7, p_2 = 0.6, n_1 = 200, n_2 = 400$$

$$n_1 p_1 = 140, n_1(1 - p_1) = 60, n_2 p_2 = 240, n_2(1 - p_2) = 160$$

sampling distribution of $\hat{p_1} - \hat{p_2}$ is approximately normal

$$\mu_{\hat{p_1} - \hat{p_2}} = p_1 - p_2 = 0.7 - 0.6 = 0.1$$

$$\sigma^2_{\hat{p}_1-\hat{p}_2} = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

$$= \frac{(0.7)(0.3)}{200} + \frac{(0.6)(0.4)}{400}$$

$$= 0.00165$$

$$N(\mu_{\hat{p}_1-\hat{p}_2}, \sigma^2_{\hat{p}_1-\hat{p}_2}) = N(0.1, 0.00165)$$

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0.1}{\sqrt{\frac{(0.7)(0.3)}{200} + \frac{(0.6)(0.4)}{400}}}$$

(ii)

$$P(\hat{p_1} - \hat{p_2} \geq 0.02), z = \frac{0.02 - 0.1}{\sqrt{\frac{(0.7)(0.3)}{200} + \frac{(0.6)(0.4)}{400}}}$$

$$z = -1.97$$

$$0.5 + 0.4756 = 0.9756$$

(iii)

$$p(0.03 < \hat{p_1} - \hat{p_2} < 0.12)$$

standardise

$$z_1 = \frac{(0.03) - 0.1}{\sqrt{\frac{(0.7)(0.3)}{200} + \frac{(0.6)(0.4)}{400}}} = -1.72$$

$$z_2 = \frac{(0.12) - 0.1}{\sqrt{\frac{(0.7)(0.3)}{200} + \frac{(0.6)(0.4)}{400}}} = 0.49$$

$$p(-1.72 < z < 0.49)$$

$$p(0 < z < 0.49) + p(0 < z < 1.72) = 0.1879 + 0.4573 = 0.6452$$

References

- Hogg,R;Mckean,J;Craig,A(2012).Introduction to mathematical statistics, 7th edition, pearson Prentice Hall, 2012.

- Hastings K.J,(1997) Probability and statistics, Addison Wesley reading,massachusetts.

Thank You!

Next Lecture: Estimation