Chapter 3
Fundamental Sampling Distributions

Department of Statistics and Operations Research

Edited by: Reem Alghamdi

February 2020

Plan

Plan

Definitions

1. A population is a set of all individuals, objects or events which are of some interest to make inferences about a specific problem or experiment.
2. A sample is a subset of a population.
3. Any function of the random variables constituting a random sample is called a statistic.

- Sample mean: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.
- Sample median: $\widetilde{X} = \begin{cases} x_{\frac{n+1}{2}}, \text{ if } n \text{ is odd}, \\ \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}), \text{ if } n \text{ is even}. \end{cases}$
- Sample variance: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \overline{X})^2$.

The computed value of S^2 for a given sample is denoted by s^2.

Theorem

If S^2 is the variance of a random sample of size n, we may write

$$S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right]$$

- Sample standard deviation: $S = \sqrt{S^2}$

Plan

Definition

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

Theorem

If $X_1, X_2, ..., X_n$ are independent random variables having normal distributions with means $\mu_1, \mu_2, ..., \mu_n$ and variances $\sigma_1^2, \sigma_2^2, ..., \sigma_n^2$, respectively, then the random variable $Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n$ has a normal distribution with mean

$$\mu_Y = a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n$$

and variance

$$\sigma_Y^2 = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2$$

Suppose that a random sample of n observations is taken from a normal population with mean μ and variance σ^2. Each observation X_i, $i = 1, 2, ..., n$, of the random sample will then have the same normal distribution. Hence, we conclude that

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

has a normal distribution with mean

$$\mu_{\overline{X}} = \frac{1}{n}\{\mu + \mu + ... + \mu\} = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu$$

and variance

$$\sigma^2_{\overline{X}} = \frac{1}{n^2}\left\{\sigma^2 + \sigma^2 + ... + \sigma^2\right\} = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n}.$$

Case I

If $X_1, X_2, ..., X_n$ are independent random variables having normal distributions with means μ and variances σ^2, then the sample mean $\overline{X}$ is normally distributed with mean equal to μ and standard deviation equal to $\sigma/\sqrt{n}$. Consequently the random variable

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

is a standard normal distribution.

Case II

If $X_1, X_2, ..., X_n$ are independent random variables having normal distributions with means μ and unknown variances, then the sample mean $\overline{X}$ is normally distributed with mean equal to μ and standard deviation equal to $s/\sqrt{n}$. Consequently the random variable

$$T = \frac{(\overline{X} - \mu)}{s/\sqrt{n}} \sim t_{(n-1)}$$

is a student t- distribution with $(n-1)$ degrees of freedom (df).
Note: Usually if n is large ($n \geq 30$) the t-distribution is approximated by a standard normal.

Case III (Central limit theorem):

If $\overline{X}$ is the mean of a random sample of size n taken from any non-normal population with mean μ and finite variance σ^2, then the limiting form of the distribution of

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

is approximately standard normal distribution as $n \to \infty$ (generally if $n \geq 30$).

Example

An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed, 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.

Solution

Here $\mu = 800$, $\sigma = 40$ and $n = 16$. The random variable $\overline{X}$ is normally distributed with mean $\mu_{\overline{X}} = \mu = 800$ and standard deviation $\sigma_{\overline{X}} = \sigma_X/\sqrt{n} = 10$.
Then $Z = (\overline{X} - 800)/10 \sim N(0, 1)$. Hence,

$$P(\overline{X} < 775) = P\left(\frac{\overline{X} - 800}{10} < \frac{775 - 800}{10}\right) = P(Z < -2.5) = 0.0062.$$

Example

Traveling between two campuses of a university in a city via shuttle bus takes, on average, 28 minutes with a standard deviation of 5 minutes. In a given week, a bus transported passengers 40 times. What is the probability that the average transport time was more than 30 minutes?

Solution

In this case, $\mu = 28$ and $\sigma = 5$. We need to calculate the probability $P(\overline{X} > 30)$ with $n = 40$. Hence,

$$\begin{aligned} P(\overline{X} > 30) &= P\left(\frac{\overline{X} - 28}{5/\sqrt{40}} > \frac{30 - 28}{5/\sqrt{40}}\right) \\ &= P\left(Z > 2.53\right) = 1 - P\left(Z \leq 2.53\right) \\ &= 1 - 0.9943 = 0.0057. \end{aligned}$$

There is only a slight chance that the average time of one bus trip will exceed 30 minutes.

Plan

Case I theorem

If two independent samples of size n_1 and n_2 are drawn at random from two normal populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively, then the sampling distribution of the differences of means, $\overline{X}_1 - \overline{X}_2$, is normally distributed with mean and variance given by

$$\mu_{\overline{X}_1 - \overline{X}_2} = \mu_1 - \mu_2 \text{ and } \sigma^2_{\overline{X}_1 - \overline{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Hence,

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

is a standard normal distribution.

Case II theorem

If two independent samples of size n_1 and n_2 are drawn at random from two normal populations with means μ_1 and μ_2 and the variances σ_1^2 and σ_2^2 are unknown but equal, then the sampling distribution of the differences of means, $\overline{X}_1 - \overline{X}_2$, is normally distributed with mean and variance given by

$$\mu_{\overline{X}_1-\overline{X}_2} = \mu_1 - \mu_2 \text{ and } \sigma^2_{\overline{X}_1-\overline{X}_2} = \frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}$$

Hence,

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2/n_1 + s_p^2/n_2}} \sim t(n_1 + n_2 - 2)$$

where S_p^2 is the pooled variance as,

$$S_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}$$

Case III theorem

If two independent samples of size n_1 and n_2 are drawn at random from any non-normal populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively, and n_1 and n_2 are greater than or equal to 30, then the sampling distribution of the differences of means, $\overline{X}_1 - \overline{X}_2$, is approximately normally distributed with mean and variance given by

$$\mu_{\overline{X}_1 - \overline{X}_2} = \mu_1 - \mu_2 \text{ and } \sigma^2_{\overline{X}_1 - \overline{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Hence,

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

is approximately standard normal distribution.

Example

Eighteen specimens are painted using type A, and the drying time, in hours, is recorded for each. The same is done with type B. The population standard deviations are both known to be 1.0. Assuming that the mean drying time is equal for the two types of paint, find $P(\overline{X}_A - \overline{X}_B > 1.0)$, where $\overline{X}_A$ and $\overline{X}_B$ are average drying times for samples of size $n_A = n_B = 18$.

Solution

From the sampling distribution of $\overline{X}_A - \overline{X}_B$, we know that the distribution is approximately normal with mean $\mu_{\overline{X}_A - \overline{X}_B} = \mu_A - \mu_B = 0$ and variance $\sigma^2_{\overline{X}_A - \overline{X}_B} = \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B} = 1/9$.

$$
\begin{aligned}
P(\overline{X}_A - \overline{X}_B > 1) &= P\left(\frac{\overline{X}_A - \overline{X}_B - (\mu_A - \mu_B)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} > \frac{1-0}{\sqrt{1/9}}\right) \\
&= P(Z > 3) \\
&= 1 - P(Z \leq 3.0) = 1 - 0.9987 = 0.0013.
\end{aligned}
$$

Example

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 a mean lifetime of 6.0 years and a standard deviation of 0.8 year. What is the probability that a random sample of 36 tubes from manufacturer A will have a mean lifetime that is at least 1 year more than the mean lifetime of a sample of 49 tubes from manufacturer B?

Solution

We are given the following information:

Population 1	Population 2
$\mu_1 = 6.5$	$\mu_2 = 6.0$
$\sigma_1 = 0.9$	$\sigma_2 = 0.8$
$n_1 = 36$	$n_2 = 49$

If we use, the sampling distribution of $\overline{X}_1 - \overline{X}_2$ will be approximately normal and will have a mean and standard deviation

$$\mu_{\overline{X}_1 - \overline{X}_2} = 6.5 - 6.0 = 0.5 \text{ and } \sigma_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{0.81}{36} + \frac{0.64}{49}} = 0.189$$

Hence,

$$\begin{aligned} P(\overline{X}_1 - \overline{X}_2 &\geq 1.0) \simeq P(Z \geq 2.65) = 1 - P(Z < 2.65) \\ &= 1 - 0.9960 = 0.0040. \end{aligned}$$

Plan

Theorem 21

If S^2 is the variance of a random sample of size n taken from a normal population having the variance σ^2, then the statistic

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2} \sim \chi^2_{n-1}$$

has a chi-squared distribution with $\nu = n - 1$ degrees of freedom.

Example

We take the sample variance s^2 of a random sample of sizes $n = 10$ from $N(\mu, 30)$.

1. Find c such that $P(s^2 \leq c) = 0.90$
2. $P(s^2 > 27.8)$

Solution

1. Since $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$, then

$$P(s^2 \leq c) = P(s^2 \times \frac{n-1}{\sigma^2} \leq c \times \frac{n-1}{\sigma^2}) = P(\chi^2 \leq c \times \frac{9}{30}) = 0.90.$$

Or equivalently,

$$P(\chi^2 > c \times \frac{9}{30}) = 0.1.$$

From χ^2- table, $\chi^2(9; 0.1) = 14.684$.
Since $c \times \frac{9}{30} = 14.684$, then $c = 48.947$.

Solution

2.

$$\begin{aligned} P(s^2 > 27.8) &= P(s^2 \times \frac{n-1}{\sigma^2} > 27.8 \times \frac{n-1}{\sigma^2}) \\ &= P(\chi^2 > 8.34) = 0.5. \end{aligned}$$

Plan

Suppose that random samples of size n_1 and n_2 are selected from two normal populations with variances σ_1^2 and σ_2^2, respectively. From Theorem 21, we know that

$$\chi_1^2 = \frac{(n_1 - 1)S_1^2}{\sigma_1^2} \text{ and } \chi_2^2 = \frac{(n_2 - 1)S_2^2}{\sigma_2^2}$$

are random variables having chi-squared distributions with $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$ degrees of freedom. Furthermore, since the samples are selected at random, we are dealing with independent random variables. Then, using F- distribution theorem in chapter 2 with $\chi_1^2 = U$ and $\chi_2^2 = V$, we obtain the following result.

Theorem

If S_1^2 and S_2^2 are the variances of independent random samples of size n_1 and n_2 taken from normal populations with variances σ_1^2 and σ_2^2, respectively, then

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

has an F-distribution with $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$ degrees of freedom.

Example

Let s_1^2 and s_1^2 be the variances of two independent random samples of sizes $n_1 = 10$ and $n_2 = 8$ from $N(\mu_1, 25)$ and $N(\mu_2, 36)$. Find c such that $P(\frac{s_1^2}{s_2^2} < c) = 0.95$

Solution

$$P(\frac{s_1^2}{s_2^2} < c) = P(\frac{s_1^2}{s_2^2} \times \frac{\sigma_2^2}{\sigma_1^2} < c \times \frac{\sigma_2^2}{\sigma_1^2}) = P(F < c \times \frac{36}{25}) = 0.95.$$

Or equivalently,

$$P(F > c \times \frac{36}{25}) = 0.05.$$

From F- table, $f(9, 7; 0.05) = 3.68$.
Then, $c \times \frac{36}{25} = 3.68$, $c = 2.56$.

Plan

In many situations the use of the sample proportion is easier and more reliable because, unlike the mean, the proportion does not depend on the population variance, which is usually an unknown quantity. We will represent the sample proportion by $\widehat{p}$ and the population proportion by p. Construction of the sampling distribution of the sample proportion is done in a manner similar to that of the mean. One has $\widehat{p} = X/n$ where X is a number of success for a sample of size n. It is clear that X is a binomial distribution $B(n, p)$. Its mean $\mu_X = np$ and its variance $\sigma_X^2 = np(1-p)$.

Theorem

The mean $\mu_{\widehat{p}}$ of the sample distribution $\widehat{p}$ is equal to the true population proportion p, and its variance $\sigma^2_{\widehat{p}}$ is equal to $p(1-p)/n$.

Theorem

If $np \geq 5$ and $n(1-p) \geq 5$, then the random variable $\widehat{p}$ is approximation a normal distribution with mean $\mu_{\widehat{p}} = p$ and standard deviation (or standard error) $\sigma_{\widehat{p}} = \sqrt{p(1-p)/n}$. Hence

$$Z = \frac{\widehat{p} - p}{\sqrt{p(1-p)/n}} \simeq N(0,1)$$

is approximately a standard normal distribution.

Example

If the insurance company of Tawuniya says that the proportion of cars in Riyadh subscribe to this company is 60%. In a sample of 500 cars, find the probability that the sample proportion $\widehat{p}$ of cars subscribe to this company is more than 0.55 (or more than 275 cars in the sample)?

Solution

Here $n = 500$, $p = 0.60$. Since $np \geq 5$ and $n(1 - p) \geq 5$, hence $P(\widehat{p} > 0.55) = P(\frac{\widehat{p}-p}{\sqrt{p(1-p)/n}} > \frac{0.55-0.60}{\sqrt{0.6(0.4)/500}}) \simeq P(Z > -2.28) = 0.9887$.

Plan

Theorem

The mean $\mu_{\widehat{p}_1-\widehat{p}_2}$ of the sample distribution of the difference between two sample proportions $\widehat{p}_1 - \widehat{p}_2$ is equal to the difference $p_1 - p_2$ between the true population proportions, and its variance $\sigma^2_{\widehat{p}_1-\widehat{p}_2}$ will be equal to $p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2$.

Theorem

If $n_1 p_1 \geq 5$, $n_1(1-p_1) \geq 5$, $n_2 p_2 \geq 5$, $n_2(1-p_2) \geq 5$, then the random variable $\widehat{p}_1 - \widehat{p}_2$ is approximation a normal distribution with mean $\mu_{\widehat{p}_1 - \widehat{p}_2} = p_1 - p_2$ and standard deviation (or standard error) $\sigma_{\widehat{p}} = \sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}$. Hence

$$Z = \frac{(\widehat{p}_1 - \widehat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

is approximately a standard normal distribution.

Example

Suppose that 25% of the male students and 20% of the female students in some certain university smoke cigarettes. A random sample of 50 male students and another random sample of 100 female students are independently taken from this university. Let $\widehat{p}_1$ and $\widehat{p}_2$ be the proportions of smokers in the two samples, respectively. Then, compute:

1. $\mu_{\widehat{p}_1 - \widehat{p}_2}$ and $\sigma^2_{\widehat{p}_1 - \widehat{p}_2}$.
2. The probability that $\widehat{p}_1$ is greater than $\widehat{p}_2$ by at least 0.06.

Solution

Here $p_1 = 0.25$, $p_2 = 0.20$, $n_1 = 50$ and $n_2 = 100$. It is clear that $n_1 p_1 \geq 5$, $n_1(1 - p_1) \geq 5$, $n_2 p_2 \geq 5$, $n_2(1 - p_2) \geq 5$.

1.
$$\mu_{\widehat{p}_1 - \widehat{p}_2} = p_1 - p_2 = 0.05$$
and
$$\sigma^2_{\widehat{p}_1 - \widehat{p}_2} = p_1(1 - p_1)/n_1 + p_2(1 - p_2)/n_2 = 0.00535.$$

2.
$$\begin{aligned} P(\widehat{p}_1 - \widehat{p}_2 \geq 0.06) &= P\left(\frac{\widehat{p}_1 - \widehat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \geq \frac{0.06 - 0.05}{\sqrt{0.00535}}\right) \\ &\simeq P(Z \geq 0.14) = 0.4443. \end{aligned}$$