CHAPTER 2

Selection from K normal populations based on means

2.1 Introduction

Let $\pi_1,\pi_2,\pi_3,\ldots,\pi_k$ be K populations, the i^{th} population being distributed as normal with mean μ_i and variance σ_i^2 for $i = 1,..,k$. We designate the population as best as the one which has minimum mean. In order to identify or select the population corrosponding to the minimum value of μ_i we take n_i observations from the i^{th} population and find an estimate of μ_i as $\bar{X}_i$ where

$$\bar{X}_i = \frac{\sum_{j=1}^{n_i} X_{ij}}{n_i}$$

Here X_{ij} is the j^{th} observation from i^{th} population.

Now, we order these estimates as

$$\bar{X}_{[1]} \leq \bar{X}_{[2]} \cdot \cdot \cdot \leq \bar{X}_{[k]} \qquad 2.1.1$$

We use the natural selection rule, which says that : select the i^{th} population as the best if and only if

$$\bar{X}_i = \min(\bar{X}_1,\ldots,\bar{X}_k) = \bar{X}_{[1]} \qquad 2.1.2$$

Let the ordering of true means be

$$\mu_{[1]} \leq \mu_{[2]} \cdot \cdot \leq \mu_{[k]}$$

We denote the sample means corrosponding to $\mu_{[1]},\mu_{[2]},\ldots,\mu_{[k]}$, by $\bar{X}_{(1)},\bar{X}_{(2)},\ldots,\bar{X}_{(k)}$.

After finding out the best population by the above procedure we have to ascertain that the probability of correct selection for this selected population is at least P* over all the configurations in the preference zone. That is, the best population has to satisfy the following expression.

$$P_{(CS/\underline{\mu})} \geq P^* \qquad \text{for all} \quad \delta \geq \delta^* \qquad 2.1.3$$

The above expression will automatically be satisfied if the minimum probability of correct selection is equal to P*. From the general selection theory outlined in chapter 1 this minimum probability is attained at least favourable configuration, which in this case is given by vector $(\mu_1, \mu_2, \ldots, \mu_k)$ satisfying the condition

$$\delta^* + \mu_{[1]} = \mu_{[2]} = \cdots \quad \mu_{[k]} \qquad 2.1.4$$

Thus we should obtain the probability of correct selection at least favourable configuration denoted by $P_{(CS/LFC)}$.

[Remark : If the value of $P_{(CS/LFC)}$ is required to be only 1/k then no statistical procedure is necessary. The procedure which selects one of the populations randomly using a simple random sampling procedure achieves the $P_{CS} = K^{-1}$]

Thus we have

$$P_{(CS/LFC)} = P^* \qquad \text{for all } \delta \geq \delta^* \qquad 2.1.5$$

We obtain in sections 2.2 to 2.4 the expressions for $P_{(CS/LFC)}$ in the following cases.

i) Variances are known and equal.

ii) Variances are unknown and equal.

iii) Variances are known and unequal.

Of these, the results obtained in the first case are used in the third chapter for the problem of interest. This expression for probability of correct selection under LFC gives a relationship between the sample size n, the indiffer~~a~~nce constant δ^* and P*, the required probability of correct selection. Given any two the third can be obtained, either exactly or approximately in the above cases. This may be further explained as follows :

Whenever the problem is to carryout a selection procedure for the given set of data we have to find out the appropriate value of P* for given n and a δ^* satisfying $\delta \geq \delta^*$.

Many a times the problem is that of designing a selection procedure. We have to determine the appropriate sample size which satisfies the probability requirement 2.1.5 In that case we solve 2.1.5 and obtain an expression for n.

2.2 Computation of $P_{(CS/LFC)}$ when variances are equal and known.

In this section we suppose that $\sigma_1^2 = \sigma_2^2 = .. \sigma_k^2$. Let σ^2 denote the known common value. The probability of correct selection for any configuration $\underline{\mu}$ is the probability that population selected as best based on sample mean is actually the best with minimum population mean.

Thus

$$P_{(CS/\underline{\mu})} = \text{Prob } [\bar{X}_{(1)} = \bar{X}_{[1]}]$$

$$= \text{Prob } [\bar{X}_{(1)} = \min [\bar{X}_{(1)}, \bar{X}_{(2)}, \ldots, \bar{X}_{(k)}]$$

$$= \text{Prob}\left(\frac{\bar{X}_{(1)}-\bar{\mu}_{[1]}}{\sigma/\sqrt{n}} \leq \frac{\bar{X}_{(i)}-\mu_{[i]}}{\sigma/\sqrt{n}} + \frac{\mu_{[i]}-\mu_{[1]}}{\sigma/\sqrt{n}} \text{ for all } 1=2,..k \right)$$

$$= \text{Prob}\left(Y_1 \leq Y_i + \frac{\mu_{[i]} - \mu_{[1]}}{\sigma/\sqrt{n}} \text{ for all } i = 2,..k \right)$$

where $Y_i = \frac{\bar{X}_{(i)}-\mu_{[i]}}{\sigma/\sqrt{n}}$ is a standard normal variate for all $1 = 2,..k$ and $Y_1, Y_2, \ldots, Y_k$ are independant.

Giving Y_1 a perticular value y and integrating over y, we get

$$P_{(CS/\underline{\mu})} = \int_{-\infty}^{\infty} \pi_{i=2}^{k} \text{ Prob } (Y_{(i)} \geq y - \frac{\mu_{[i]}-\mu_{[1]}}{\sigma/\sqrt{n}}) \quad f(y) \, dy$$

$$= \int_{-\infty}^{\infty} \pi_{i=2}^{k} (1-\phi \, (y - \frac{\mu_{[i]}-\mu_{[1]}}{\sigma/\sqrt{n}}) \, f(y) \, dy \qquad 2.2.1$$

where $f(y)$ is the density function and $\phi(y)$ is the distribution function of standard normal distribution.

Now for least favourable configuration

$\mu_{[i]} - \mu_{[1]} = \delta^*$ for all $i = 2 .. k$

Since from the expression 2.2.1 ($\mu_{[i]}-\mu_{[1]}= \delta_i$) is an increasing function of $P_{(CS/\underline{\mu})}$ and it will be minimum at minimum value of δ_i which is δ^* , in the preference zone. Thus

$$P_{(CS/LFC)} = \int_{-\infty}^{\infty} (1 - \phi\, (y - \frac{\delta^* \sqrt{n}}{\sigma}\,)^{k-1} \; f(y)\; dy$$

$$= \int_{-\infty}^{\infty} \phi\, (- y + \frac{\delta^* \sqrt{n}}{\sigma}))^{k-1} \; f(y)\; dy$$

by putting $u = - y$, we get

$$P_{(CS/LFC)} = \int_{-\infty}^{\infty} \phi\, (u + \frac{\delta^* \sqrt{n}}{\sigma}\,)^{k-1} \; f(u)\; du \qquad 2.2.2$$

We should have $P_{(CS/LFC)} = P^*$. Hence

$$P^* = \int_{-\infty}^{\infty} \phi\, (u + \frac{\delta^* \sqrt{n}}{\sigma}\,)^{k-1} \; f(u) \; du \qquad 2.2.3$$

If we have to find the optimal sample size we should solve 2.2.3 for n.

Let
$$h = \frac{\delta^* \sqrt{n}}{\sigma} \qquad 2.2.4$$

The values of h are tabulated for different values of

K and P* and are given in

i) ‘selecting and ordering populations’ by Gibbons, Sobel and Olkin ; and

ii) ‘ Introduction to Statistics and Probability’ by Dudewicz.

After obtaining h we find n as minimum integer such that

$$n \geq \frac{h^2\sigma^2}{\delta^{*2}} .$$

2.2 b) Sample sizes unequal - In many practical situations we donot have complete control over the sample sizes and they are not all equal. For such a case an exact solution for $P_{(CS/LFC)} = P^*$ cannot be obtained.

An approximate solution for the selection problem with unequal sample sizes can be obtainedd by computing a certain generalized avarage sample.size, denoted by n_o, by squre mean root formula given by

$$n_o = \frac{(\sqrt{n_1} + \sqrt{n_2} \cdot . . . . \sqrt{n_k})^2}{K} \qquad 2.2.4$$

By using this n_o in place of n we can solve the expression 2.2.3. However, we are unable to provide any justification for the use of n_o given in 2.2.4 or we could not find one in the literature.

2.3 Computation of $P_{(CS/LFC)}$ when variances are equal but unknown.

a) Sample sizes equal -

We have K populations normal with means $\mu_1,\mu_2,\ldots,\mu_k$ respectively and variances equal to σ^2, where σ^2 is unknown We first estimate σ^2 by S^2 where

$$S^2 = \frac{\sum_{i=1}^{n}\sum_{i=1}^{p}(X_{ij}-\bar{X}_i)^2}{N-K} \qquad 2.3.1$$

Where N is the total sample size. and and $\mu_i = \bar{X}_i$

Now

$$P_{(CS/\underline{\mu})} = \text{Prob}\,[\ \bar{X}_{[1]} = \bar{X}_{(1)}\ ] \qquad 2.3.2$$

$$= \text{Prob}\,[\ \bar{X}_{(1)} = \min\,(\bar{X}_{(1)},\bar{X}_{(2)},\ldots,\bar{X}_{(k)})$$

$$= \text{Prob}\,[\ \bar{X}_{(1)} \leq \bar{X}_{(i)} \text{ for all } i = 2 \,..\, k\ ]$$

$$= \text{Prob}\,\left[\ \frac{\bar{X}_{(1)}-\mu_{[1]}}{\sigma/\sqrt{n}} \leq \frac{\bar{X}_{(i)}-\mu_{[i]}}{\sigma/\sqrt{n}} + \frac{\mu_{[i]}-\mu_{(i)}}{\sigma/\sqrt{n}} \text{ for all } i = 2 \,..\, k\ \right]$$

Let

$$\frac{\bar{X}_{(i)}-\mu_{[i]}}{\sigma/\sqrt{n}} = Y_i \qquad \text{for all } i = 1,2,\,..\,k$$

We know that $Y_1,Y_2,\ldots,Y_k$ are independant standard normal variates.

Hence

$$P_{(CS/\underline{\mu})} = \text{Prob } (Y_1 \leq Y_i + \frac{\mu_{[i]} - \mu_{[1]}}{\sigma/\sqrt{n}} \text{ for all } i=2 \,..k)$$

$$= \text{Prob } (Y_i \geq Y_1 - \frac{\mu_{[i]} - \mu_{[1]}}{\sigma/\sqrt{n}}) \text{ for all } i = 2 \,..\, k)$$

Letting $Y_1 = y$ and integrating over all possible values of y we get

$$P_{(CS/\underline{\mu})} = \int_{-\infty}^{\infty} \prod_{i=2}^{k} \text{Prob } (Y_i \geq y - \frac{\mu_{[i]} - \mu_{[1]}}{\sigma/\sqrt{n}}) \, f(y) \, dy$$

where f(y) is standard normal density.

$$P_{(CS/\underline{\mu})} = \int_{-\infty}^{\infty} \prod_{i=2}^{k} [1 - \phi(y - \frac{\mu_{[i]} - \mu_{[1]}}{\sigma/\sqrt{n}})] \, f(y) \, dy$$

where ϕ is standard normal distribution.

$$P_{(CS/\underline{\mu})} = \int_{-\infty}^{\infty} \prod_{i=2}^{k} \phi \, (-y + \frac{\mu_{[i]} - \mu_{[1]}}{\sigma/\sqrt{n}}) \, f(y) \, dy$$

Let u = -y, we get

$$P_{(CS/\underline{\mu})} = \int_{-\infty}^{\infty} \prod_{i=2}^{k} (\phi(u + \frac{\mu_{[i]} - \mu_{[1]}}{\sigma/\sqrt{n}}) \, f(u) \, du$$

For LFC $\mu_{[i]} - \mu_{[1]} = \partial^*$ for all i = 2 .. k

Thus

$$P_{(CS/LFC)} = \int_{-\infty}^{\infty} \phi^{k-1} (u + \frac{\partial^* \sqrt{n}}{\sigma}) \, f(u) \, du \qquad 2.3.3$$

The above expression cannot be evaluated because the value of σ is unknown. So we estimate σ by S and proceed as if the variance was known and is equal to s^2. If sample size is large this provides an approximate solution, to the original problem.

2.4 Computation of $P_{(CS/LFC)}$ when the variances are known and unequal.

We estimate the means of the populations as

$$\mu_{[i]} = \bar{X}_{(i)}$$

In this case since variance are unequal the assumption of equality of sample size for different populations is not justified. In fact we should take the sample sizes proportional to the variances.

Let us take

$$\frac{\sigma_1}{\sqrt{n_1}} = \frac{\sigma_2}{\sqrt{n_2}} = \cdot \cdot \cdot \cdot \quad \frac{\sigma_k}{\sqrt{n_k}} = C \qquad 2.4.1$$

where C is a constant.

Our selection rule is to select population as the best one if it has smallest mean and

$$P_{(CS/LFC)} = P^* \quad \text{for all } \sigma \geq \sigma^*$$

The probability of correct selection under $\boldsymbol{\mu}$ is

$$P_{(CS/\mu)} = \text{Prob} \, [\ \bar{X}_{(1)} = \bar{X}_{[1]}\ ]$$

$$= \text{Prob}\ [\ \bar{X}_{(1)} = \min\ (\bar{X}_{(1)},\ldots,\bar{X}_{(k)}]$$

$$= \text{Prob}\ [\ \bar{X}_{(1)} \leq \bar{X}_{(i)} \quad \text{for all } i = 2\ ..\ k]]$$

$$= \text{Prob}\ [\ \bar{X}_{(1)} \leq \bar{X}_{(i)} \quad \text{for all } i = 2\ ..\ k\ ]$$

$$= \text{Prob}[\ \frac{\bar{X}_{(1)} - \mu_{[1]}}{\sigma_1/\sqrt{n_1}} \leq \frac{\bar{X}_{(i)} - \mu_{[1]}}{\sigma_i/\sqrt{n_i}} + \frac{\mu_{[i]} - \mu_{[1]}}{\sigma_i/\sqrt{n_i}} + \quad \text{for all } i = 2\ ..\ k\ ]$$

We know that

$$Y_i = \frac{\bar{X}_{(i)} - \mu_{[i]}}{\sigma_i/\sqrt{n_i}}$$

is a standard normal variate for all i= i,2,...,k and $Y_1,Y_2,\ldots,Y_k$ are independant.

$$P_{(CS/\underline{\mu})} = \text{Prob}\ (\ Y_1 \leq Y_i + \frac{\mu_{[i]} - \mu_{[1]}}{\sigma_i/\sqrt{n_i}} \quad \text{for all } i = 2\ ..\ k\ )$$

$$= \text{Prob}\ (\ Y_i \geq Y_1 - \frac{\mu_{[i]} - \mu_{[1]}}{\sigma_i/\sqrt{n_i}} \quad \text{for all } i = 2\ ..\ k\ )$$

From 2.1.21 we have letting $Y_1 = y$ and integrating over all possible values of y

$$P_{(CS/\underline{\mu})} = \int_{-\infty}^{\infty} \prod_{i=2}^{k} \text{Prob}\ (Y_i \geq y - \frac{\mu_{[i]} - \mu_{[1]}}{C}) \ f(y)\ dy$$

for all i

$$= \int_{-\infty}^{\infty} \prod_{i=2}^{k} 1 - \phi\ (\ y - \frac{\mu_{[i]} - \mu_{[1]}}{C}) \ f(y)\ dy$$

For LFC we have

$$P_{(CS/LFC)} = \int_{-\infty}^{\infty} [\ 1 - \emptyset\ (y - \frac{\partial^*}{C})^{k-1} f(u)\ dy$$

$$= \int_{-\infty}^{\infty} \emptyset\ (\ -y + \frac{\partial^*}{C})^{k-1} \quad f(y)\ dy$$

Let u = -y

$$P_{(CS/LFC)} = \int_{-\infty}^{\infty} \emptyset^{k-1}\ (\ u + \frac{\partial^*}{C}\ ) \quad f(u)\ du \qquad 2.4.3$$

$$= P^* \qquad 2.4.3$$

according to selection rule.

This equation is similar to that in case i, if we define h = ∂^*/C, hence we can refer to the same tables and find out h. Then for different σ_i's we find

$$n_i = \frac{\sigma_i^2 \ h^2}{\partial^{*2}} \qquad \underline{\underline{2.1.}}4 \quad ?$$

Remark : For the'variances unknown and unequal' case one has to follow a stage procedure, since it is not required for the problem of dissertation it is omitted.

2.5 Application of the rules considered earlier for non-normal populations

In this section, we suppose that $\pi_1, \pi_2, \ldots, \pi_k$ are K populations, not necessarily normal, characterized by a parameter Θ, which is equal to Θ_i for the i^{th} population. Suppose that the i^{th} population is best if $\Theta_i = \min(\Theta_1, \ldots, \Theta_k)$ Now, the problem is to identify the population with parametervalue $\Theta_{[1]}$. Suppose that, it is decided to use the stalistic T ; which takes value T_i for the observations from the i^{th} population. Suppose that, the selection rule is to select the population corresponding to the minimum value of $T_1, \ldots, T_k$.

In order to compute the probability of correct selection and obtain the given probability of correct selections, we need the distribution and density functions of T.

If the finite sample distribution of T is difficult to derive, an alternative is to use the asymptotic distribution of T and derive the PCS and other required quantities as we do if we were to select from the K populations which have asymptotic distributions. This is the large sample approximate solution for the given problem.

Now, we shall consider the case when T is a best Asymptotically normal estimator of θ; that is $\sqrt{n}(T-\theta)$ converges in distribution to normal with mean zero and variance $\nu(\theta)$. This fact, sometimes we write as

$$\sqrt{n}\,(T-\theta) \xrightarrow{D} Y$$

when

$$Y \sim N(0, \nu(\theta))$$

Now, in order to obtain the asymptotic solution, we behave as if we have to select the best of K normal populations, i^{th} populations having mean θ_i and variance $\frac{(\theta_i)}{n}$. The existing results given in the previous sections are not useful in solving this problem. However, if the variance is independant of θ, then the results of section 2.2 can be used. The problem can be reduced to this case. if the original populations are ordered according to a strictly monotone function $g(\theta)$ of θ instead of θ, and find a g such that the asymptotic distribution of $g(T)$ is normal with variance independent of θ. If such a function g exists, it is called the variance stabilizing transformation for θ.

Thus the first step in the problem is to find a function g such that

$$\sqrt{n}\,(g(T) - g(\theta)] \xrightarrow{D} Y \qquad 2.5.1$$

When

$Y \sim N(0,C^2)$ and C^2 is a constant independent of Θ.

We know that, if (2.4.1) holds, then (2.4.2) is also true for any function g such that $g'(\Theta) \neq 0$; in fact, for such a g, we have

$$\sqrt{n}\,[g(T)-g(\Theta)\ ] \xrightarrow{D} Y \qquad 2.5.2$$

where

$$Y \sim N(0,\ g'(\Theta)^2 \ \ (\Theta))$$

Now, we want $[g'(\Theta)]^2\ \nu(\Theta)]$ to be equal to C^2 and to be independent of Θ, that is

$$[g'(\Theta)]^2 = \frac{C^2}{\nu(\Theta)}$$

which is equal to

$$g'(\Theta) = \frac{C}{\sqrt{\nu(\Theta)}} \qquad 2.5.3$$

We note that g which satisfies (2.5.3) automatically will be strictly, increasing and so we can order the original population in terms of $g(\Theta)$ instead of Θ. From 2.5.3 we set that

$$g(\Theta) = C \int_0^{\Theta} (\nu(s)\ ]^{-1/2}\ ds \qquad 2.5.4$$

In order that the right hand side integral to exist $g(\Theta)$ to satisfy 2.5.3 we assume that $\nu\ (\Theta)'$ is contineous in Θ.

Let $\mu = g(\Theta)$

and

$$\mu_i = g(\Theta_i).$$

Now, our selection rule reduces to selecting the population corresponding to the minimum value of $g(T_1),..,g(T_k)$ The probability of correct selection and other relevant quantities are calculated from the equation 2.2.3 of section 2.2. Since we have from (2.4.3), for each i,

$$\sqrt{n}\ (g(T_i) - \mu_i) \xrightarrow{D} Y \qquad \text{or}$$

where

$$Y \sim N(0, C^2)$$

$g(T_i)$ is approximately normal with mean μ_i and variance C^2/n , known.

However, there is a difficulty in using this approach. One has to fix the preferance zone in terms of $g(\Theta)$ but not in terms of Θ. Only in few cases, one can find a correspondance betweent the two variances. Stabilizing transformations are given below for three ? distributions and another one as obtained in section 3.3.

i) Suppose X has a bernoulli distribution with parameter Θ. Then $\bar{X}$ is a BAN estimator for Θ based on n observations and

$$\sqrt{n}\ (\bar{X} - \Theta) \xrightarrow{D} Y$$

where

$$Y \sim N(0, \gamma(\theta))$$

where

$$\gamma(\theta) = [\theta(1-\theta)]^{-1}.$$

In this case

$$g(\theta) = \sqrt{2}\ \sin^{-1}(\sqrt{\theta}).$$

ii) Suppose X has a Poisson distribution with mean θ.

Then $\bar{X}$ is a BAN estimator for θ based on n observations and

$$\sqrt{n}(\bar{X} - \theta) \xrightarrow{D} Y$$

where

$$Y \sim N(0, \gamma(\theta))$$

here

$$\gamma(\theta) = \theta$$

so that

$$g(\theta) = \sqrt{\theta}$$

iii) Suppose X has an exponential distributions with dens-de density $f(x, \theta)$ given by

$$f(x, \theta) = \begin{cases} \frac{1}{\theta} \quad \exp \quad -x/\theta , & x > 0 \\ \quad 0 & \text{otherwise.} \end{cases}$$

Again $\bar{X}$ is a BAN estimator with $\gamma(\theta) = \theta^2$ and hence

$g(\Theta)$ in this case, is equal to $\log(\Theta)$.

Only in the third example above, the indifference zone in term $g(\Theta)$ and in term of Θ can be selected as follows:

$$\frac{\Theta_{[2]}}{\Theta_{[1]}} \geq \delta^* \quad \text{iff} \quad g(\Theta_{[2]} - g(\Theta_{[1]}) \geq \log(\delta^*).$$

In the next chapter, we consider the problem of selecting from K populations based on quantal response data and we provide an asymptotic solution based on variance stabilizing transformation.