1 Introduction

In chapter two Dahman (2018e) we have discussed in details the the available probability distributions. We have seen the measurement of the central tendency (the mean μ in particular) as well as the dispersion (the variance σ^2). We found that in the univariate case the mean μ will follow a normal distribution $\bar{X} \sim N(1, 0)$. On the other hand, the variance σ^2 will follow χ^2 distribution. In chapter three Dahman (2018d) we have discussed the central limit theorem (CLT). We found that $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$. However, in the multivariate case, the formula will be altered. The $\bar{x}$ (in the case of statistic) or μ (in case of population) will be a vector as $\bar{X}_{px1}$ or μ_{px1}. The value of s^2 (in case of statistic) or σ^2 in case of population will be a matrix (covaraince matrix) as S or $\sum$.Thus, the distribution, in the multivariate case, will be $\bar{X} \sim N_p(\mu, \frac{\sum}{n})$. For the variance that will be S ,and it follows a wishart distribution with n-1 degree of freedom.

In the case of the central limited theorem, however this time it's for a multivariate case, the formula will be as following, $\sqrt{n}(\bar{x} - \mu) \sim N_p(0, \sum)$. and the covariance matrix $n(\bar{x} - \mu)^T S^{-1}(\bar{x} - \mu) \sim \chi^2_p$. That's what known as multivariate central theorem (MCLT). Noteworthy, that's apply if $n - p > 40$. What happens if the sample size $n - p < 40$.

2 Hotelling's T^2

We have closed the line in our introduction above with the question if the sample size $n - p < 40$. Remember from chapter four Dahman (2018a) that we have developed a map of the type and conditions of distribution (t distribution in particular). The formula was $t_{n-1} = \frac{\bar{x}-\mu}{s/\sqrt{n}}$. We have learned that this distribution is used when the sample size is less than 40. Here we can transfer the exact same formula into the equivalent multivariate counterpart.

To do so, we know now that μ will be a vector μ_{px1} and σ will be a covariance matrix $\sum_{pxp}$, In addition, $\bar{x}$ will be a vector $\bar{X}_{px1}$. Before we develop the multivariate counterpart, let's rearrange the "t" distribution formula above. We can write the formula as $t = \sqrt{n}(\bar{x} - \mu)(s)^{-1}$. Let's take square root of both terms, $t^2 = n(\bar{x} - \mu)(s^2)^{-1}(\bar{x} - \mu)$. As a result, the Hotelling's T^2 equation will be: : Note, it follows

F distribution:

$$T^2 = n(\bar{x} - \mu)^T (S)^{-1} (\bar{X} - \mu) \sim \frac{(n-1)p}{n-p} F_{p,n-p}. \quad (1)$$

Example: random sample with n=20, were collected . The sample mean vector and covariance matrix are given bellow: (a) obtain the Hotelling's T-Square, (b) what will be the distribution of it.

$$\bar{X} = \begin{pmatrix} 10 \\ 20 \end{pmatrix} \Bigg| \; S = \begin{pmatrix} 40 & -50 \\ -50 & 100 \end{pmatrix}$$

Answer: From equation (1). we can identify all required figures. We have n=20, we have $\bar{X}$ is given, as well as S is given. and μ will be μ_1 and μ_2. If I blog these information in the formula I will get:

$$T^2 = 20 \left[\begin{pmatrix} 10 \\ 20 \end{pmatrix} - \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \right]^T \begin{pmatrix} 40 & -50 \\ -50 & 100 \end{pmatrix}^{-1} \left[\begin{pmatrix} 10 \\ 20 \end{pmatrix} - \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \right]$$

Some linear algebra skills are required to solve the above formula. The final result will be:

$$1.33(10 - \mu_1)^2 + 1.34(10 - \mu_1)(20 - \mu_2) + 0.53(20 - \mu_2)^2$$

For the second part of the question. We know that Hotelling's T^2 follows F distribution as: $T^2 \sim \frac{(n-1)p}{n-p} F_{p,n-p}$. We have all the information that will lead into:

$$\frac{(20-1)2}{20-2} F_{2,(20-2)} = \frac{38}{18} F_{2,18}$$

3 Confidence Region

In chapter four Dahman (2018a) we have discussed the concept of confidence interval. That was in terms of a univariate concept. We have seen how to calculate the confidence interval in terms of any distribution we have. Now, let's have a close look at equation (1). What we see in that equation, having two variables, is an ellipse. In other forms, it can be written as:

$$\mathrm{P} \left[\; \mathrm{n}[(\bar{x} - \mu)^T S^{-1} (\bar{x} - \mu)] <= \frac{(n-1)p}{n-p} F_{p,n-p} \; \right] = 1 - \alpha$$

In chapter seven Dahman (2018c) we have seen different scenarios of ellipse. Please to understand all possible shapes ellipse might take, you may refer to chapter

seven. I just want to stress the point is that, the "size" of the ellipse, not "the shape" is decided by the quantity of alpha distribution. Having we said that, I want you to have a good look at Figure 1 below, we can look at any observation from two perspectives: the first one if from a univariate, and the second is from bio-variate or multi-variate. Now in the figure we see (CI) confidence interval for each variable $\bar{x}_1$ and $\bar{x}_2$. As well as we have an ellipse which represents the confidence region for joint variables $\bar{x}_1, \bar{x}_2$.

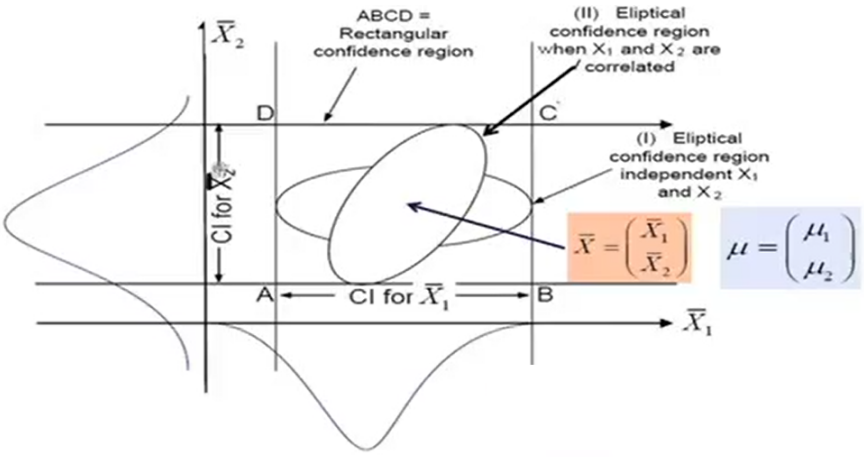

Figure 1: Confidence Region and Interval Views

Now, we can see the same Figure 2 below that X observation from CI point of view it's within control of $\bar{x}_1$ as well as $\bar{x}_2$. However it's out of the CR when both $\bar{x}_1, \bar{x}_2$ are joint.

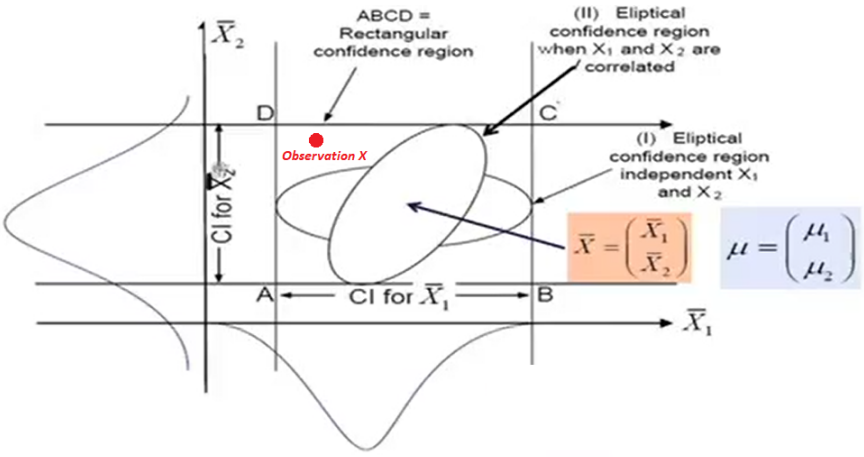

Figure 2: CR and CI with a scenario

To conclude this section, it's very important to understand that first we must develop the CR, and in case we find any point out of control, then we are back to CI to observe and decide which variable is responsible for that observation to be out of

control. That's what's know as simultaneous confidence interval. Our next section will discus this in details. However, before I leave this section let me show you the steps to develop confidence region (CR):

1. define population
2. obtain sample statistics
3. identify appropriate sampling distribution
4. develop CR

Now, in the second step of sampling we might encounter three different scenarios. Please keep them always as a guideline.

1. Scenario one: sampling from multivariate normal population with known $\sum$
2. Scenario two: sampling from multivariate normal population with unknown $\sum$, but large sample,
3. Scenario three: from multivariate normal population with unknown $\sum$, and small sample size.

In scenario one, we follow χ^2 distribution. In scenario two we follow the same distribution, χ^2, however the formula will use S instead of using $\sum$. In scenario three we use F distribution. For all scenarios of course we use equation (1). Note: scenario one instead of using S we use $\sum$ because it's known.

4 Simultaneous Confidence Intervals

The percentage of times that a group of confidence intervals will all include the true population parameters or true differences between factor levels if the study were repeated multiple times. The simultaneous confidence level is based on both the individual confidence level and the number of confidence intervals. For a single comparison, the simultaneous confidence level is equal to the individual confidence level. However, each additional confidence interval causes the simultaneous confidence level to decrease in a cumulative way.

It is important to consider the simultaneous confidence level when you examine multiple confidence intervals because your chances of excluding a parameter or true difference between factor levels for a family of confidence intervals is greater than for any one confidence interval.

Tukey's method, Fisher's least significant difference (LSD), Hsu's multiple comparisons with the best (MCB), and Bonferroni confidence intervals are methods for calculating and controlling the individual and simultaneous confidence levels.

We will consider two methods to discus in this summary paper. The linear combination approach. and Bonferroni approach. Both approaches are available in chapter five Dean W. and Wichern (2007).

4.1 Linear Combination Approach

we suppose that X is normally distributed, $X \sim N_p(\mu, \sum)$. Let's assume we have some constant transpose vector $[a_1, a_2, .., a_p]^T$. So the linear combination can be $a^T X$. Now, if i have $\bar{x}$, i assume it's normally distributed as $\bar{x} \sim N_p(\mu, \sum /n)$. Also here i can obtain a linear combination as $a^T \bar{x}$. Remember from chapter seven that one property of the multivariate normal distribution is that a linear combination of variables will be a univariate normal distribution for these variables. This property can lead me to the fact $a^T \bar{x} \sim N(a^T \mu, a^T \sum \frac{a}{n})$. see this is a univariate level. see the two quantities will be two numbers. As final finding we can drive the final formula as:

$$z = \frac{a^T \bar{X} - a^T \mu}{\sqrt{\frac{a^T \sum a}{n}}} \sim z(0, 1) \tag{2}$$

From equation (2) I can construct the interval as:

$$a^T \bar{X} - z_{\alpha/2} \sqrt{\frac{a^T \sum a}{n}} \leqslant a^T \mu \leqslant a^T \bar{X} + z_{\alpha/2} \sqrt{\frac{a^T \sum a}{n}} \tag{3}$$

from equation (3) let's assume I have $a^T = [0, 0, 0, ..j, ..0, 0]$, then:

$$\bar{x}_j - z_{\alpha/2} \sqrt{\frac{\sigma_{jj}}{n}} \leqslant \mu_j \leqslant \bar{x}_j + z_{\alpha/2} \sqrt{\frac{\sigma_{jj}}{n}}$$

However for a reason the equation in (3) will be modified using maximization lemma. The new transformation of the equation will be:

$$a^T \bar{X} - \sqrt{\chi^2_p(\alpha)} \sqrt{\frac{a^T \sum a}{n}} \leqslant a^T \mu \leqslant a^T \bar{X} + \sqrt{\chi^2_p(\alpha)} \sqrt{\frac{a^T \sum a}{n}} \tag{4}$$

As a result, from equation (4) let's assume I have $a^T = [0, 0, 0, ..j, ..0, 0]$, then:

$$\bar{x}_j - \sqrt{\chi^2_p(\alpha)} \sqrt{\frac{\sigma_{jj}}{n}} \leqslant \mu_j \leqslant \bar{x}_j + \sqrt{\chi^2_p(\alpha)} \sqrt{\frac{\sigma_{jj}}{n}}$$

From the above three different scenarios, we can apply the formula in (4). In scenario one we use the σ, in scenario two we use s_{jj}. In scenario three we use F distribution. Consider the example again given as below, find SCI for the mean of

variables x_1, x_2, $\alpha = 0.05$. Given, n=20, p=2, and $\sum$ is unknown. it's scenario three.

$$\bar{X} = \begin{pmatrix} 10 \\ 20 \end{pmatrix} \Bigg| \; S = \begin{pmatrix} 40 & -50 \\ -50 & 100 \end{pmatrix}$$

apply the equation in (4) with alternating part as:

$$a^T\bar{X} - \sqrt{\tfrac{(n-1)p}{n-p}F^{\alpha}_{p,n-p}}\sqrt{\tfrac{a^T\sum a}{n}} \leqslant a^T\mu \leqslant a^T\bar{X} + \sqrt{\tfrac{(n-1)p}{n-p}F^{\alpha}_{p,n-p}}\sqrt{\tfrac{a^T\sum a}{n}}$$

Let's find first the critical value from $\sqrt{\frac{(n-1)p}{n-p}F^{\alpha}_{p,n-p}} = 7.51$
after you plug into the formula above you get the SCI as:

$$\mu_1 : 6.12 \leqslant \mu_1 \leqslant 13.88$$
$$\mu_2 : 13.88 \leqslant \mu_2 \leqslant 26.12$$

4.2 Bonferroni Approach

The Bonferroni method is a simple method that allows many comparison statements to be made (or confidence intervals to be constructed) while still assuring an overall confidence coefficient is maintained. Let's revisit the three scenarios as:

1. Scenario one: $\bar{x}_j - z_{\alpha_j/2}\sqrt{\frac{\sigma_{jj}}{n}} \leqslant \mu_j \leqslant \bar{x}_j + z_{\alpha_j/2}\sqrt{\frac{\sigma_{jj}}{n}}$
2. Scenario two: $\bar{x}_j - z_{\alpha_j/2}\sqrt{\frac{s_{jj}}{n}} \leqslant \mu_j \leqslant \bar{x}_j + z_{\alpha_j/2}\sqrt{\frac{s_{jj}}{n}}$
3. Scenario three: $\bar{x}_j - t_{n-1}(\alpha_j/2)\sqrt{\frac{\sigma_{jj}}{n}} \leqslant \mu_j \leqslant \bar{x}_j + t_{n-1}(\alpha_j/2)\sqrt{\frac{\sigma_{jj}}{n}}$

5 Hypothesis Testing

5.1 For Single Population

in Chapter five Dahman (2018b) we have discussed in details all what you need to know about univariate scalar of hypothesis teasing. Now we will discus from the multivariate point of view. It's important to keep in mind that Hypothesis testing is the use of statistics to determine the probability that a given hypothesis is true. Again if we consider the three scenarios I should follow the steps:

1. form the hypothesis as $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$. Please note, the μ is a vector of means but not a scalar.

2. test statistic from the Hotelling's T^2 as discussed in equation (1).

3. choose the sampling

4. make the decision

Let's try to see that in each scenario as following:

1. scenario one: form the hypothesis, use the statistic formula with the known $\sum$, distribution is χ^2, make decision: Reject H_0 if equation $\geq \chi^2_p(\alpha)$

2. scenario two: form the hypothesis, use the statistic formula with S, distribution is χ^2, make decision: Reject H_0 if equation $\geq \chi^2_p(\alpha)$

3. scenario three: form the hypothesis, use the statistic formula with S, distribution is F $\frac{(n-1)p}{n-p}F_p, n-p$, make decision: Reject H_0 if equation $\geq \frac{(n-1)p}{n-p}F_p, n-p$

Example let's consider the example to conduct hypothesis tesing for population mean vector $\mu_1 = 9$, $\mu_2 = 18$, $\alpha = 0.05$. n=20, p=2, and $\sum$ is unknown. obviously it falls under scenario three. given:

$$\bar{X} = \begin{pmatrix} 10 \\ 20 \end{pmatrix} \Bigg| \mu_0 = \begin{pmatrix} 9 \\ 18 \end{pmatrix} \Bigg| S = \begin{pmatrix} 40 & -50 \\ -50 & 100 \end{pmatrix}$$

Answer we can follow the steps as we have learned: (a) the hypothesis $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$, (b) apply the formula

$$1.33(10 - \mu_1)^2 + 1.34(10 - \mu_1)(20 - \mu_2) + 0.53(20 - \mu_2)^2$$

$$1.33(10 - 9)^2 + 1.34(10 - 9)(20 - 18) + 0.53(20 - 18)^2 = 6.13$$

(c) distribution will follow F, $\frac{(n-1)p}{n-p}F_p, n - p = 7.51$, (d) we see that 6.13 is $\leqslant$ 7.51, then we fail to reject the null hypothesis,

5.2 For two population

In fact for two population the scenarios and the formulas as well as the steps are ll remain the same. a paper by Dr. J Maiti (Paul and Maiti, 2007) discus the scenario of two populations and the techniques of investigating it using two population scenario.

References

Dahman, M. R. (2018a). AMSM- Estimation (Point and Interval)- Chapter Four. *OSF Preprints. doi.org/10.31219/OSF.IO/FC9ZH.*

Dahman, M. R. (2018b). AMSM- Hypothesis testing- Chapter Five. *OSF Preprints. doi.org/10.31219/OSF.IO/9UP6T.*

Dahman, M. R. (2018c). AMSM- Multivariate Normal Distribution- Chapter Seven. *OSF Preprints. doi.org/10.31219/OSF.IO/UF86G.*

Dahman, M. R. (2018d). AMSM- Sampling Distribution- Chapter Three. *OSF Preprints. doi.org/10.31219/OSF.IO/H5AUC.*

Dahman, M. R. (2018e). AMSM- Univariate Descriptive Statistics-Chapter Two. *OSF Preprints. doi.org/10.31219/OSF.IO/THD3C.*

Dean W., R. A. and Wichern, J. (2007). *Applied Multivariate Statistical Analysis.* Pearson Prentice Hall, 6th edition.

Paul, P. and Maiti, J. (2007). The role of behavioral factors on safety management in underground mines. *Safety Science*, 45(4):449–471.