STATISTICAL MODELS

Course 2024-2025

SEMINAR 1. Multivariate Statistics (Part 1)

1 Basic Problems

PROBLEM 1.- Given the random sample $\mathbf{X}$ below, composed of 3 samples in 2 dimensions with bivariate normal distribution, use Hotelling's T^2 test to evaluate if $\mu_0 = [0, 2]^T$ is different from the population mean at significance level $\alpha = 0.1$.

$$\mathbf{X} = \begin{bmatrix} 2 & 8 & 8 \\ 9 & 6 & 3 \end{bmatrix}$$

The corresponding critical value for the F-distribution is $F_{2,1}(0.9) = 49.5$

PROBLEM 2.- Repeat Problem 1 with the following data:

$$\mathbf{X} = \begin{bmatrix} -7 & 2 & 5 \\ 16 & 4 & -2 \end{bmatrix} \qquad \mu_0 = \begin{bmatrix} -2 \\ 4 \end{bmatrix}$$

PROBLEM 3.- The results of Problem 2 can be explained by realizing that its data corresponds to a transformation of data from Problem 1, as follows:

$$\mathbf{X}' = \mathbf{C}\,\mathbf{X} \qquad \mu_0{}' = \mathbf{C}\,\mu_0 \qquad \mathbf{C} = \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix}$$

Demonstrate algebraically the observed invariance for a generic 2×2 matrix $\mathbf{C}$

PROBLEM 4.- Given the bivariate random sample $\mathbf{X}$ below:

$$\mathbf{X} = \begin{bmatrix} 2 & 4 & 3 & 3 \\ 1 & 3 & 2 & 4 \end{bmatrix}$$

we want to assess whether $\mu_0 = [5, 5]^T$ is a plausible value for the population mean.

a) State the null and alternative hypotheses required to test the question indicated above. Include the mathematical formulation.

b) Indicate what method would you use to test the hypotheses stated in (a). Indicate what assumptions would be needed in order for your method to be valid. For this problem, you do not need to test whether the assumptions hold; we will assume that they do.

c) Test the hypotheses stated in (a) at significance level $\alpha = 0.05$. Hint: use $F_{2,2}(0.95) = 19$

d) Clearly state your conclusion and interpret it to answer the above research question.

PROBLEM 5.- Given a random sample $\mathbf{X}$ below:

$$\mathbf{X} = \begin{bmatrix} 2 & 9 & 6 & 8 \\ 12 & 9 & 9 & 10 \end{bmatrix}$$

a) Evaluate T^2 for testing $H_0 : \mu = [9, 11]^T$

b) Specify what would be the distribution for T^2 in this case.

c) Use your results to test H_0 at $\alpha = 0.05$ significance level. What conclusion do you reach?

d) Use your results to test H_0 at $\alpha = 0.1$ significance level. What conclusion do you reach?

PROBLEM 6.- Indicators from 30 healthy females were analyzed. Three components were measured, resulting in the following (sample) mean and covariance matrix:

$$\overline{\mathbf{x}} = \begin{bmatrix} 7.1 \\ 11.4 \\ 15.1 \end{bmatrix} \qquad \mathbf{S} = \begin{bmatrix} 19.2 & 38.4 & -35.2 \\ 38.4 & 246.4 & 54.4 \\ -35.2 & 54.4 & 160 \end{bmatrix}$$

Test the hypothesis $H_0 : \mu = [6, 11, 16]^T$ against $H_1 : \mu \neq [6, 11, 16]^T$ at significance level $\alpha = 0.05$.

PROBLEM 7.- Given a random sample $\mathbf{X}$ as follows:

$$\mathbf{X} = \begin{bmatrix} 2 & 9 & 6 & 8 \\ 12 & 9 & 9 & 10 \end{bmatrix}$$

a) Calculate and draw the 90% confidence region for μ (see footnote[1])

b) Calculate and draw the 95% confidence region for μ

c) Calculate and draw the 98% confidence region for μ

d) What can you conclude about $H_0 : \mu = [9, 11]^T$

PROBLEM 8.- Using data from Problem 7, we wish to compute simultaneous 95% confidence intervals...

a) ...for each component of the population mean, using Hotelling's T^2.

b) ...for each component of the population mean, but using an approach that produces tighter intervals than Hotelling's T^2

[1]The axes of the ellipsis are in the direction of the eigenvectors of $\mathbf{S}$ and their half-lengths are $c\sqrt{\frac{\lambda_i}{n}}$, where λ_i are the corresponding eigenvalues of $\mathbf{S}$ and c^2 is the critical T^2 value of the confidence ellipse at the specified confidence level.

PROBLEM 9.- A given data sample with $n = 42$ observations has the following mean and covariance:

$$\overline{\mathbf{x}} = \begin{bmatrix} 5.62 \\ 5.89 \end{bmatrix} \qquad \mathbf{S} = \begin{bmatrix} 1.44 & 1.17 \\ 1.17 & 1.46 \end{bmatrix}$$

a) Calculate a 95% confidence ellipse for the mean of the population.

b) Use the confidence ellipse to decide whether $H_0 : \mu = [5.9, 5.7]^T$ can be rejected at $\alpha = 0.05$ significance level.

c) Confirm the above by computing Hotelling's T^2.

PROBLEM 10.- Using the same data from Problem 9:

a) Calculate simultaneous 95% confidence intervals of the mean for each dimension, using T^2.

b) Calculate Bonferroni-corrected individual 95% confidence intervals of the mean for each dimension.

c) Compare the conclusions of the above 2 methods to calculate individual intervals with the conclusion from Problem 9.

PROBLEM 11.- Repeat Problem 10 after transforming both the data sample $\mathbf{X}$ and μ_0 as follows:

$$\mathbf{X} = \mathbf{C}\,\mathbf{X} \qquad \mu_\mathbf{0} = \mathbf{C}\,\mu_\mathbf{0} \qquad \mathbf{C} = \begin{bmatrix} 0.704 & -0.710 \\ 0.710 & 0.704 \end{bmatrix}$$

PROBLEM 12.- A given data sample with $n = 61$ observations has the following mean and covariance:

$$\overline{\mathbf{x}} = \begin{bmatrix} 95.52 \\ 93.39 \end{bmatrix} \qquad \mathbf{S} = \begin{bmatrix} 3266.46 & 1175.50 \\ 1175.50 & 474.98 \end{bmatrix}$$

a) Obtain the large sample 95% simultaneous confidence intervals for the mean of each component.

b) Compare the above calculation to the intervals that would be obtained without assuming a large sample.

c) Obtain the 95% Bonferroni-corrected confidence intervals for the mean of each component.

d) Calculate the large sample 95% confidence ellipse of the mean.

PROBLEM 13.- File `SM22_Seminar_1_13.xlsx` contains data from the marks obtained by students in 2 out of the 5 labs of Statistical Models in the course 2019-2020 (first 2 columns in each row). The last column contains a binary indicator of whether the corresponding student passed the course (1) or not (0). Therefore, we can consider that we have 2 random samples of 2 dimensions each (students who passed the course and those who did not pass). For each of these two random samples, and assuming that they follow a multivariate normal distribution:

a) Calculate and draw 95% confidence regions for the population means. Recall to consider separately the two populations of students (who passed and who did not pass) and therefore you will construct two separate confidence regions.

b) Calculate 95% confidence intervals for the marks of the two groups of students for each lab.

c) Based on your results in (a) and (b), is it possible to extract any conclusion regarding the average marks of these two labs between students who passed or failed the subject?

2 Answers to Selected Problems

PROBLEM 2.- $\hat{\theta}_{T^2} = 105.33 < (4 \times 49.5)$ and the null hypothesis is not rejected.

PROBLEM 3.- Johnson & Whichern (2014), proof of equation (5-9) pages $215 - 216$.

PROBLEM 5.-

a) $\hat{\theta}_{T^2} \simeq 31.37$

b) $\hat{\theta}_{T^2} \sim 3 \times F_{2,2}$

c) $\hat{\theta}_{T^2} = 31.3689 < 3 \times 19$ and the null hypothesis is not rejected. Conclusion: we have not enough evidence to reject that μ_0 could be the population mean at significance level $\alpha = 0.05$

d) $\hat{\theta}_{T^2} = 31.3689 > 3 \times 9$ and the null hypothesis is rejected. Conclusion: we can reject that μ_0 is the population mean at significance level $\alpha = 0.10$

PROBLEM 6.- $\hat{\theta}_{T^2} \simeq 51.68 > (29 \times 3/27) \times 2.9604 = 9.539 \rightarrow$ the null hypothesis is rejected.

PROBLEM 9.-

a) Fig. 1 shows the confidence ellipse.

b) In Fig. 1 we see that μ_0 is outside the confidence region $\rightarrow H_0$ is rejected.

c) $\hat{\theta}_{T^2} = 16.6589 > T^2_{critic} = 6.625 \rightarrow$ reject H_0.

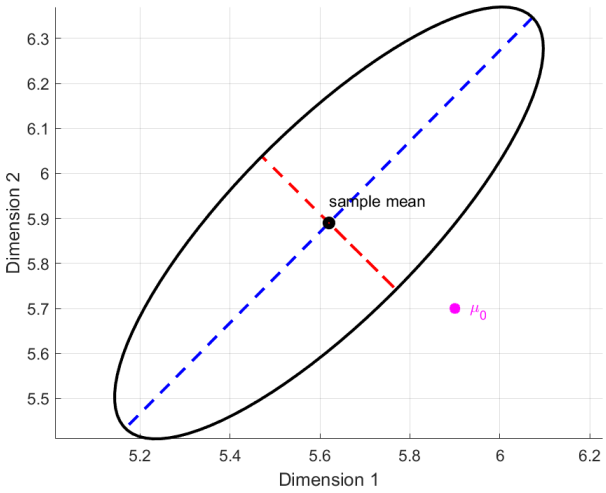

Figure 1: Confidence region for problem 9.

PROBLEM 10.- Using the same data from Problem 9:

a) $(5.1434; \quad 6.0966)$ and $(5.4101; \quad 6.3699)$

b) $(5.1892;\quad 6.0508)$ and $(5.4562;\quad 6.3238)$

c) We cannot reject H_0 in any of the 2 cases.

PROBLEM 11.- The confidence elipse rotates and gets *aligned* with the coordinate axes.

a) $(-0.4356;\quad -0.0152)$ and $(7.494;\quad 8.7995)$

b) $(-0.4154;\quad -0.0354)$ and $(7.5557;\quad 8.7178)$

c) We can reject H_0 with either Bonferroni or Hotelling confidence intervals, as also illustrated in Fig 2.

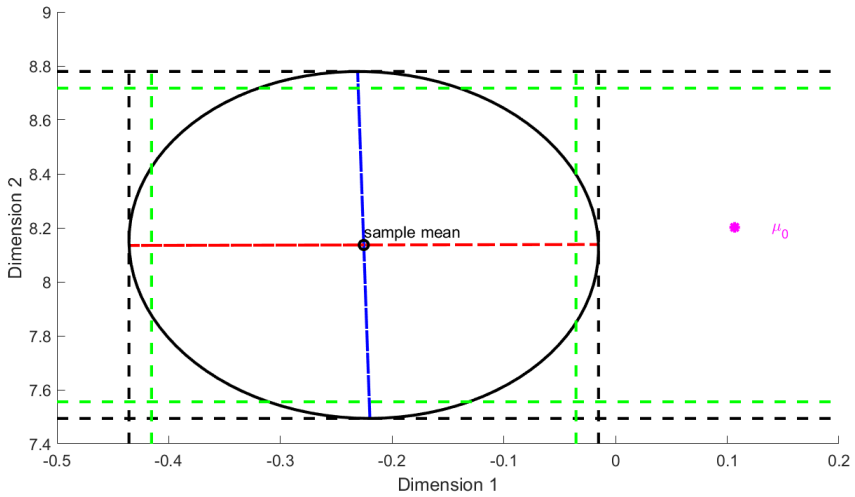

Figure 2: Confidence region and intervals for Problem 11. Intervals based on Hotelling's T^2 in black; Bonferroni-corrected t-intervals in green.

PROBLEM 12.-

a) $(77.61;\quad 113.43)$ and $(86.56;\quad 100.22)$

b) $(76.99;\quad 114.05)$ and $(86.32;\quad 100.46)$

c) $(78.70;\quad 112.34)$ and $(86.97;\quad 99.81)$ without large sample assumption.